When Politically Congruent Opinions Pass for Facts: Self-Worth and Motivated Classification in Media Environments

Natalia Lamberova[*], Alexei Zakharov[†], Sydney Brown, Minh Dao[‡]

Word count: 9,771
Main text word count (Introduction through Conclusion; excluding references and appendices): 7,262

"Ah, it is easy to deceive me; I myself am glad to be deceived."
— Alexander Pushkin, Confession *(adapted translation)*

Abstract

People constantly encounter political claims whose status is ambiguous: they sound factual, but they also carry interpretation, evaluation, or partisan judgment. How do citizens decide whether such claims deserve evidentiary status? We argue that motivated reasoning can begin before belief updating. Citizens first decide whether a claim should function as evidence at all, and that classification judgment can itself be politically motivated. We formalize this earlier decision in a motivated-beliefs model in which stronger self-worth reduces the marginal psychological payoff from protective relabeling relative to the accuracy cost of misclassification. We test that implication in a survey experiment. Respondents complete a brief task designed to bolster or leave unchanged their sense of self-worth, then classify 24 short statements as facts or opinions. In the control group, classification errors are systematically related to statement content rather than random noise. A positive self-worth treatment reduces the treatment-effect difference for politically congruent opinions, relative to intermediate (neither congruent nor opposing) opinion-statement pairs, by 2.79 to 2.95 percentage points across four specifications, with the largest raw p-value equal to 0.023. Non-partisan-statement placebo tests are close to zero, while fact-side effects are less uniform. The results show that self-protective motives can shape a politically central judgment: which claims are allowed to carry factual authority in the first place. The mechanism may be especially consequential in environments where social standing feels precarious, including periods of economic insecurity.

Keywords: motivated reasoning; self-worth; survey experiment; fact–opinion classification; media; political communication; belief formation.

[*]Assistant Professor, Political and Policy Sciences at University of Texas at Dallas, Natalia.Lamberova@utdallas.edu, corresponding coauthor
[†]Assosiate Instructional Professor, University of Chicago, zakharov1@uchicago.edu
[‡]University of Texas at Dallas, Minh.Dao@utdallas.edu

1. Introduction

Before anyone decides whether a claim should change what they believe, they must decide what kind of claim it is: evidence about the world or interpretation of it. Motivated reasoning can therefore begin earlier than belief updating. It can enter at the moment when citizens decide whether a political claim deserves evidentiary status at all. That boundary is increasingly difficult to maintain. Opinion-rich content has expanded across contemporary media environments, including social media, newspapers, newsletters, podcasts, television panels, and campaign communication (Kavanagh and Rich, 2018; Bursztyn et al., 2023). When opinions are treated as facts, rhetoric enters belief formation with the authority of evidence.

The ability and willingness to distinguish opinions from factual statements is highly consequential. Disagreement can arise not only from conflicting values or selective exposure, but from divergent judgments about what counts as factual information (Alesina and Stantcheva, 2020; Mettler and Mondak, 2024; Zumofen, Stadelmann-Steffen and Bühlmann, 2024). A policy argument framed as empirical necessity can sound more legitimate than the same argument presented openly as interpretation. Recent work in *Political Behavior* shows how false headlines, corrective labels, and citizens' confidence in factual answers shape perceived accuracy and political knowledge (Graham, 2020; Clayton et al., 2020). The consequences are not merely semantic: Bursztyn et al. (2023) show that when anti-COVID opinions are received as facts, the result can be higher mortality. Repetition and familiarity make such claims harder to correct (Lewandowsky et al., 2012; Lewandowsky, Ecker and Cook, 2017; Swire, Ecker and Lewandowsky, 2017), and political misperceptions may persist even after correction (Nyhan and Reifler, 2010). Even online search meant to verify a claim can increase belief in both true and false news by surfacing apparent corroboration (Aslett et al., 2021).

We argue that the fact–opinion boundary is politically motivated. Much of the motivated-reasoning literature asks how people update after information has been accepted as evidence. We move one step earlier in the information-processing sequence and ask whether citizens first classify claims in ways that protect the self. Factual messages that cut against one's side may be dismissed as opinion, while congruent opinions may be elevated to the status of fact. Self-

worth is central to this process. Citizens encounter political claims not as detached observers, but as people with identities, loyalties, and a need to maintain a positive view of themselves and their side (Cichocka, Marchlewska and Cislak, 2024*a*). A congruent opinion can therefore do more than persuade; if treated as fact, it can affirm that one's side—and one's own political judgment—was right. Congenial interpretations provide ego value because they let citizens see the world as vindicating their prior beliefs, their political side, and their own competence as political judges. If self-deception partly protects this self-regarding value, then self-worth should reduce the need to grant factual status to congruent opinions (Steele, 1988; Sherman and Cohen, 2006; Cohen and Sherman, 2014; Bénabou and Tirole, 2002, 2016; Bracha and Brown, 2012).

This mechanism is also impacted by economic concerns. Economic insecurity, unemployment, inequality, and status anxiety can make citizens' social standing and competence feel precarious. These conditions may increase the value of self-protective interpretation: a claim that vindicates one's side can also vindicate one's judgment and place in the social order (Krauss and Orth, 2022; Gedikli et al., 2023; Melita, Willis and Rodriguez-Bailon, 2021).

We test our predictions with a survey experiment. Respondents are first randomly assigned to a brief task designed to bolster or leave unchanged their sense of self-worth, and then classify 24 short statements as facts or opinions. The experiment provides evidence about a broader conceptual claim: political reasoning is shaped not only by how citizens process evidence, but also by which claims they allow to become evidence. The manipulation draws on evidence that self-affirmation, reflecting on personal strengths, and gratitude exercises can reinforce self-integrity and well-being (Steele, 1988; Sherman and Cohen, 2006; Cohen and Sherman, 2014; Rash, Matsuba and Prkachin, 2011). A social-media-like setting is a natural place to study this boundary because social media are central venues for political information and opinion exchange, and because reporting and commentary often travel there in the same short format (Westerman, Spence and Van Der Heide, 2014; Tucker et al., 2018). Statements vary along four dimensions: source, type, topic, and partisan slant. In the empirical analysis, source proxies the perceived credibility of the signal: statements attributed to newspapers should appear more

credible than otherwise similar statements attributed to social media. Because source cues can condition responses to political messages (Baum and Groeling, 2009), we treat source as a credibility cue rather than as a merely cosmetic attribute of the statement. Partisan slant applies to facts as well as opinions. A fact can be politically congruent or opposing when it reports verifiable information that would be more welcome to one ideological side than the other.

Our contribution is threefold. First, we extend motivated-reasoning research by moving earlier in the information-processing sequence. Recent work in *Political Behavior* shows that partisan cues, affect, prior attitudes, and information-selection choices shape how citizens respond to political information once it is encountered (Zumofen, Stadelmann-Steffen and Bühlmann, 2024; Fuller, de la Cerda and Rametta, 2026; Barbieri, Petitpas and Sciarini, 2026). The standard question in this literature is whether individuals interpret or update on information in self-serving directions. Our question is whether self-protection also shapes the prior judgment of whether a claim is information at all. To clarify that mechanism, we develop an extension of the motivated-belief framework in Little (2025). In our model, an individual who observes a statement cannot choose an arbitrary posterior directly; instead, the individual chooses among three mental models, corresponding to the statement being a fact or a biased opinion. Second, we provide causal evidence using randomized variation in a short-lived self-worth task rather than observational proxies for personality or ideology. Third, we contribute to research on political information environments by showing that opinion-rich communication may matter not only because citizens encounter congenial content, but because some citizens are willing to grant that content the authority of fact.

Three empirical findings organize the paper. First, even in the control group, classification errors are systematic rather than random. Respondents are much more accurate on opinions than on facts, and classification varies systematically with topic, source, and respondent-statement congruence. Second, assignment to the positive self-worth treatment specifically reduces the promotion of congruent opinions into facts. Across four model specifications, the positive self-worth treatment reduces the treatment-effect difference for congruent opinion-statement pairs by between 2.79 and 2.95 percentage points relative to Intermediate opinion-

statement pairs, with the largest raw p-value equal to 0.023. Third, this pattern is politically specific rather than a general response style: non-partisan-statement placebo tests are close to zero, and the effect is not accompanied by longer time on task. Fact-side effects are less uniform, so the evidence is sharpest where the theory is most directly about political persuasion: the positive self-worth treatment makes respondents less willing to let Congruent opinion borrow the authority of fact.

We do not claim that campaigns or media firms directly target citizens on the basis of self-esteem, nor do we argue that all fact–opinion confusion is driven by self-worth. The point is more fundamental: self-protective motives can shape what citizens are willing to treat as evidence in the first place. This mechanism helps explain why political communication so often blurs reporting, commentary, and persuasion—and why such blurring can be effective even when the underlying claims are plainly evaluative.

The rest of the paper proceeds as follows. Section 2 develops the intuition and empirical hypotheses. Section 3 formalizes the mechanism. Section 4 describes the design and sample. Section 5 presents the main findings. Section 6 discusses the evidence and threats to validity. Section 7 concludes.

2. Theory and hypotheses

Our argument combines a motivated-reasoning mechanism with a simple observation about contemporary political information environments. The first part concerns the motive: self-worth can change the psychological value of interpreting claims in self-serving ways. The second part concerns the sequence: before citizens update on a claim, they decide whether it belongs in the evidentiary set at all. The third part concerns the message: some statements are easier than others to receive as evidence rather than evaluation. In our setting, type, source, topic, and partisan slant jointly determine the scope for motivated classification.

2.1. *The fact–opinion boundary*

We define facts as claims that are, in principle, verifiable against external evidence, and opinions as evaluative or interpretive claims whose truth cannot be settled by observation alone (Mettler and Mondak, 2024). Factual claims can be wrong or later revised; what makes them factual claims is that they describe something about the world that can, at least in principle, be checked against evidence. Opinions add evaluation, interpretation, or prescription, so evidence alone may not settle them. The distinction is not always obvious in political communication, where evaluative claims often use a factual register and factual claims can be rhetorically framed. That ambiguity is precisely why the boundary is theoretically important: before citizens update beliefs, they must decide whether a statement deserves evidentiary status at all. Our outcome therefore captures classification judgments, not simple agreement or disagreement with a claim. The line between fact and opinion is not perfectly sharp, but it is still useful.

2.2. *Motivated classification: Congruence*

In the motivated-reasoning literature, people may depart from purely accuracy-oriented belief formation because some beliefs are psychologically useful. Distorted interpretations can protect self-worth, reduce the sting of unfavorable information, or help sustain valued identities (Kunda, 1990; Taber and Lodge, 2006; Bolsen, Druckman and Cook, 2014; Bénabou and Tirole, 2002, 2016; Bracha and Brown, 2012). Self-affirmation research makes the same logic especially relevant here: when people can sustain a broader sense of self-integrity, they often become less defensive in processing threatening information (Steele, 1988; Sherman and Cohen, 2006; Cohen and Sherman, 2014). Political behavior research on cueing, selective exposure, and misinformation shows that citizens' prior commitments shape which information they seek, how they judge source credibility, and how they update after exposure (Clayton et al., 2020; Zumofen, Stadelmann-Steffen and Bühlmann, 2024; Fuller, de la Cerda and Rametta, 2026; Barbieri, Petitpas and Sciarini, 2026). We focus on a still earlier judgment: whether the encountered claim enters the evidentiary stream at all. If self-protective reasoning is partly useful because it protects the self, then temporarily bolstering self-worth should reduce the marginal payoff

from granting factual status to congenial interpretation.

Our setting lets us study that idea at an especially consequential margin. Citizens often confront short, decontextualized claims whose factual status is not self-evident. Before deciding how much evidentiary weight to place on such a claim, they must decide whether it is factual reporting or interpretation. That decision is precisely where motivated classification can enter: it determines which claims receive factual authority and which are treated as mere opinion.

Our hypotheses focus on directional classification errors. Proposition 1 in our formal model predicts that respondents will be tempted to move claims in a self-protective direction: congruent opinions can be elevated into facts, while opposing facts can be discounted as opinions:

H1. Among opinion statements, politically congruent opinions are more likely than intermediate or opposing opinions to be classified as facts.

H2. Among fact statements, politically opposing facts are more likely than intermediate or congruent facts to be classified as opinions.

2.3. Motivated classification: Self-esteem

The positive self-worth treatment in our experiment asks respondents to write short posts about recent accomplishments and sources of gratitude, a task designed to bolster self-worth through self-affirmation and positive reflection, while the control group completes neutral writing prompts. The goal is to create a positive self-reflection task in a familiar online register and to test whether assignment to that task moves the fact–opinion boundary. When self-worth is under pressure, Congruent opinions may be easier to accept as objective, while Opposing facts may be easier to dismiss as opinion. These two forms of motivated classification need not be psychologically symmetric. Elevating a congruent opinion into a fact may be more discretionary because the claim already contains interpretation. Discounting a recognizably verifiable fact as opinion may carry different informational or cognitive costs. If self-worth is bolstered instead, the need for protective relabeling should weaken. Hence, our expectation, formalized in Proposition 2, is that the positive self-worth treatment should reduce directional classification bias in the presence of political slant:

H1a. The positive self-worth treatment reduces the tendency to classify politically congruent opinions as facts, relative to Intermediate respondent–statement pairs.

H2a. The positive self-worth treatment reduces the tendency to classify politically opposing facts as opinions, relative to Intermediate respondent–statement pairs.

2.4. Motivated classification: Other factors

Our final two predictions are concerned with the source and domain of statements. Source matters because audiences use media brands as shorthand for credibility (Baum and Groeling, 2009; Westerman, Spence and Van Der Heide, 2014). We therefore interpret newspaper versus social-media attribution as a signal-credibility manipulation. Domain matters because in statements related to politics we expect identity concerns (and the accompanying motivated reasoning) to be the strongest. In our design, health is the least identity-laden domain, while economics plausibly lies in between. This leads us to formulate the following hypotheses:

H3. Statements appearing on social media should be more likely to be treated as opinions than ones appearing in newspapers.

H4. Misattribution should vary by statement domain, most likely for politics-related statements and least likely for health-related.

3. Formalizing motivated classification

This section formalizes the earlier classification decision. Existing motivated-belief models are well suited to studying how citizens update after receiving evidence. Our setting differs because citizens must first decide whether a claim should receive evidentiary status. The model provides a conceptual bridge between those two stages. It links self-worth, respondent ideology, statement slant, and fact–opinion classification in a setting in which a claim can be treated either as evidence about the world or as biased interpretation.

The model is meant to discipline a simple behavioral claim: citizens may be motivated not only to interpret evidence in congenial ways, but also to decide whether a claim becomes evidence at all. We use an extension of a motivated-belief model in which belief formation reflects

two competing concerns: accuracy about the world and the psychological value of holding self-protective beliefs (Kunda, 1990; Taber and Lodge, 2006; Little, 2025; Acharya, Park and Zaidman, 2025). The key choice in our setting comes before updating is complete. After observing a statement, an individual must decide whether to treat it as a fact—an unbiased, though noisy, signal about the state of the world—or as a biased opinion. That interpretive choice then shapes the individual's *ex post* belief about the world.

The trade-off is straightforward. Reclassifying a non-congruent statement as "a mere opinion" can be psychologically attractive, while granting factual status to a congruent claim allows it to carry more evidentiary weight. But there is also a cost to departing from the statement's objective status. In that sense, the model captures a tension between self-worth and accuracy in a way that maps directly onto the survey task. Our formal approach is closest to the motivated-belief framework in Little (2025), but the choice problem is different: rather than selecting an arbitrary posterior distribution, the individual chooses among three mental models, corresponding to the statement being a fact, a liberal-biased opinion, or a conservative-biased opinion.

The formal analysis yields two comparative statics that guide the empirical analysis. First, when self-worth is higher, the utility gain from motivated interpretation is smaller, so respondents should be less willing to relabel congruent opinions as facts or non-congruent facts as mere opinions. Second, motivated misclassification is more attractive when a statement creates a stronger identity-relevant temptation to reinterpret its status. These results motivate our focus on Congruent and Opposing respondent–statement pairs, and on whether treatment effects are strongest where ideological alignment gives respondents a reason to police or relax the fact–opinion boundary. The model therefore connects the paper's conceptual claim to its estimand: treatment effects on the evidentiary status citizens assign to political claims.

We now formalize these ideas. We define factual statements as those directly stating or estimating the true state with minimal bias, though possibly with some measurement error or noise. Opinion statements, by contrast, reflect subjective beliefs, preferences, or biases. They systematically differ from the true state because of individual priors, partisan positions, or

cognitive biases.

Consider the state of the world $x \sim \mathcal{N}(0,1)$, and suppose that an individual observes a *statement* (s,b), where

$$s = x + b + \epsilon$$

is the content of the statement, $b \in \{-1,0,1\}$ is the bias of the statement, and $\epsilon \sim \mathcal{N}(0,\sigma^2)$ is the noise with which the statement is observed. We call any statement with $b = 0$ a *fact* and any statement with $b \neq 0$ an *opinion*. We assume that, while the individual observes the bias b accurately, he may be motivated to hold a potentially different belief $\tilde{b} \in \{-1,0,1\}$ about b. If the individual is indifferent between holding the accurate belief $\tilde{b} = b$ and holding some other belief, he chooses $\tilde{b} = b$.[1]

The individual's choice of $\tilde{b}$ (for example, when a factual statement is treated as an opinion) affects his posterior belief about x and his payoff in two ways. First, choosing $\tilde{b} \neq b$ is costly because it entails holding an inaccurate belief about the state of the world. The second term captures the ego value of a congruent opinion. In political settings, believing that the world vindicates one's side can also vindicate one's own judgment, competence, and identity. Treating an aligned opinion as fact therefore does more than change an estimate of the state of the world: it allows the individual to feel that "my side was right" and that one's prior view was justified. We therefore write utility as the sum of an accuracy cost and ego utility from a congenial world-state:

$$U(s,b,\tilde{b}) = -[\text{Accuracy cost}] + \gamma \cdot [\text{Ego utility from a congenial world-state}] \tag{1}$$

The first term depends on both b and $\tilde{b}$ and represents the individual's disutility due to her beliefs potentially being different from objective beliefs. This disutility increases in the absolute difference between b and $\tilde{b}$, with zero disutility for $b = \tilde{b}$.

The second term is one's expected belief-based utility. There are two components that contribute to it. First, the individual enjoys believing that the world is in a way that is aligned with

[1] Our model can be readily extended to the case where the choice of $\tilde{b}$ is not constrained.

the individual's political alignment y, with $y = -1$ denoting a liberal alignment and $y = 1$ a conservative alignment. Correspondingly, this payoff is increasing in x when $y = 1$ and decreasing when $y = -1$. Second, the individual has *intrinsic self-esteem* a that is independent of y and x. We define the quantity $z = a + yx$ as the individual's total belief-based utility when the state of the world is x, assuming that intrinsic self-esteem and the state of the world x are perfect substitutes in the psychological sense when $y = 1$ and perfect complements when $y = -1$. The parameter $\gamma > 0$ represents the importance of holding the right beliefs. The specific functional forms for both components of (1) are defined following Little (2025) and are given in Appendix A1.

We proceed to derive the following result:

Proposition 1. If $y = 1$, then $\tilde{b} \leq b$. If $y = -1$, then $\tilde{b} \geq b$.

This means that an individual will be motivated to misperceive the bias of statements in the direction that opposes the individual's alignment. For example, a conservative will never choose to believe that a liberally biased opinion is a fact or is conservatively biased, while a fact will never be perceived as being conservatively biased. The following result describes how the individual's motivated belief depends on intrinsic self-esteem a.

Proposition 2. There exist $\underline{a}_1 < a_0 < \bar{a}_1$ such that:

1. If $b = -y$, then $\tilde{b} = b$;
2. If $b = 0$, then $\tilde{b} = -y$ if $a < a_0$ and $\tilde{b} = 0$ if $a \geq a_0$;
3. If $b = y$, then $\tilde{b} = -y$ if $a < \underline{a}_1$; $\tilde{b} = 0$ if $a \in [\underline{a}_1, \bar{a}_1)$; and $\tilde{b} = b$ if $a \geq \bar{a}_1$.

Our result shows that for high levels of intrinsic self-esteem, the individual will always adopt the objective belief about the bias of a statement. For lower levels of intrinsic self-esteem, the individual will choose to believe that an opinion with an aligned bias is a fact. If a is lower still, the individual will start treating facts as opinions with bias that is opposite of the individual's alignment. For very low levels of intrinsic self-esteem, any statement, even an opinion with an aligned bias, will be suspected of being an opinion with an opposing bias. Intuitively,

as one's intrinsic self-worth falls, the psychological benefit of defensive reinterpretation increasingly outweighs the accuracy cost of moving away from the statement's objective status.

Finally, we look at the effects of γ and σ^2 on motivated beliefs:

Proposition 3. The values $\underline{a}_1, a_0, \bar{a}_1$ increase with γ, tend to $-\infty$ as $\gamma \to 0$ or $\sigma^2 \to 0$, and tend to $+\infty$ if $\sigma^2 \to \infty$.

As the value of holding beliefs decreases, beliefs differ from objective beliefs under a narrower set of other parameter values. The effect of statement precision is ambiguous. For very precise signals, the individual adopts objective beliefs. As signals become very uninformative, the individual always chooses to believe that the statement has the bias opposite to their political alignment.

These comparative statics motivate our experimental design. Self-esteem is not itself a clean causal object in observational data: it may proxy for attention, effort, confidence, or more general task performance, all of which may also shape fact–opinion classification. A raw cross-sectional relationship between self-esteem and classification judgments is therefore difficult to interpret. In our data, that pre-treatment association in fact runs in the opposite direction of the treatment effect we estimate below. Random assignment to a self-worth task allows us to test whether a temporary psychological state changes classification judgments.

4. Research Design and Data

4.1. Sample

We fielded an online survey experiment on Qualtrics in December 2023 and recruited respondents through CloudResearch, an online panel provider known to perform well relative to other common online recruitment platforms on several standard data-quality dimensions (Douglas, Ewell and Brauer, 2023). The study was registered before fielding.[2]

A total of 1,512 respondents completed the informed consent and demographics sections and were assigned by the Qualtrics randomizer to one of the two primary groups: the positive

[2]The registration is available at https://aspredicted.org/147358.

self-worth treatment arm (755 respondents) or the control arm (757 respondents). A total of 1,392 completed the survey (697 in the treatment group and 695 in the control), and 1,390 completed the survey and passed the attention check (695 in each arm). Attrition does not differ significantly between the positive and control arms ($p = 0.937$, Fisher's exact test), indicating that survey attrition does not drive the estimated treatment effects.

The experiment also included a third treatment arm where the participants were exposed to a task designed to reduce self-worth. Since this task was structurally different from both the positive self-worth treatment and control, the main text focuses on the positive treatment-versus-control comparison, while the results for the negative-treatment arm are reported in Appendix A3.1. There were 755 participants assigned to the negative treatment arm and a total of 2,337 who started the survey.

4.2. The self-worth treatment

Respondents were randomly assigned to treatment or control group, block-stratified by gender. In the positive self-worth arm ($N = 755$ assigned; $N = 695$ in the main analytic sample), respondents were presented with one of two mood induction tasks, writing short posts about gratitude and recent accomplishments, a task designed to bolster self-worth through self-affirmation and positive reflection (Steele, 1988; Sherman and Cohen, 2006; Cohen and Sherman, 2014; Rash, Matsuba and Prkachin, 2011). Here is an example of a task:[3]

> Imagine you're crafting a post about a recent personal success or a valuable lesson you've learned. It could be something big or small, but it made you feel good or helped you grow. Share your story within the 280-character limit.

In the control arm ($N = 757$ assigned; $N = 695$ in the main analytic sample), the respondents were asked to reflect on a neutral event, such as what they ate for breakfast. The study also included an auxiliary negative social-comparison arm ($N = 755$ assigned; $N = 740$ completed the survey and passed the attention check) involving an upward social comparison in a social-media environment; such comparisons have been reported to threaten self-esteem (Vogel et al.,

[3]See Appendix A4.1 for the full list of tasks.

2014). The results for this analysis are reported in Appendix A3.1. Because that task differs in format from both the positive and control prompts and increases time on the classification block, it is not a clean inverse manipulation of self-worth. The main causal comparison therefore remains assignment to the positive-writing task versus the control group.

4.3. *Classifying statements as facts or opinions*

After the treatment, each respondent was presented with 24 statements and had to classify each statement as either a fact or an opinion. The statements differed along several characteristics. First, they varied along topic (politics, economics, and health) and source (*The Wall Street Journal*, *The New York Times*, Reddit, and Twitter/X). Second, they differed in their partisan slant: non-partisan, pro-Democratic, or pro-Republican.

The statements were selected to populate this design rather than sampled at random from the news environment. We identified candidate statements from newspapers and social-media sources, including *The New York Times*, *The Wall Street Journal*, Twitter/X, and Reddit. The search was guided by the experimental cells: candidates had to fit the assigned topic, source type, partisan slant, and fact-versus-opinion classification. We then retained a balanced set of short statements that could be presented in a common survey format while preserving the intended variation across topic, source, slant, and type.

Finally, some statements presented facts, while others presented opinions. For example, a factual statement read: "Malaria cases in Texas and Florida are the first US spread since 2003, CDC says", while an opinion can state: "Walking more is one of the most underrated health hacks out there." The full list of statements appears in Appendix A4.2.

The primary object of interest in this study is the respondent–statement pair. Respondents are grouped into three ideological categories using a seven-point liberal–conservative self-placement item: liberal respondents (values 1–3), moderate respondents (value 4), and conservative respondents (values 5–7). Each statement is coded by its intended partisan slant as pro-Democratic, pro-Republican, or non-partisan. We then code the respondent–statement pair as Congruent, Opposing, or Intermediate. Matching ideological and partisan directions

are Congruent, while mismatching ideological and partisan directions are Opposing. Non-partisan statements are Congruent for moderate respondents and Intermediate for respondents on either ideological side; partisan statements are Intermediate for moderate respondents. In the regression models, Intermediate respondent–statement pairs serve as the baseline category.

Table 1: Mapping of respondent ideology and statement slant to congruence categories

Respondent ideology	**Pro-Democratic statement**	**Pro-Republican statement**	**Non-partisan statement**
Liberal	Congruent	Opposing	Intermediate
Conservative	Opposing	Congruent	Intermediate
Moderate	Intermediate	Intermediate	Congruent

Note: Congruent means the statement's slant aligns with the respondent's ideological side, or that a non-partisan statement is shown to a moderate respondent. Opposing means the statement's slant favors the opposite ideological side. Intermediate covers cases that are neither Congruent nor Opposing: non-partisan statements for liberal and conservative respondents, and pro-Democratic or pro-Republican statements for moderate respondents. Intermediate respondent–statement pairs are the baseline category in the regressions.

4.4. Controls

Pre-treatment controls include gender, education, race (White, Black, Hispanic, Asian, or other), full-time employment, seven-point self-identified ideology, pre-treatment self-esteem, and three news-source rankings: television, newspapers, and social media. Our measure of pre-treatment self-esteem was based on an eight-item subset of the Rosenberg Self-Esteem Scale, a widely used measure of global self-worth based on positively and negatively worded statements about the self (Rosenberg, 1965). Respondents indicated whether they strongly agreed, agreed, disagreed, or strongly disagreed with four positively worded items and four negatively worded items (the items had high internal consistency, Cronbach α=0.93); the standardized first principal component was used in our analysis.[4] Finally, we asked respondents to rank the importance of TV, newspapers, social media, and word of mouth as news sources. Word-of-mouth

[4]We did not include a post-treatment measure of state self-worth. This was a design choice rather than an oversight. Asking about self-worth before the classification task could prime the construct and contaminate the outcome. Asking after the task would make the measure partly endogenous to performance on the classification task itself. Finally, conditioning on whose self-worth changed would create post-treatment selection, because realized psychological response to the prompt is not randomized. We therefore interpret the estimates as intention-to-treat effects of assignment to the positive self-worth task. Response time is used only as auxiliary evidence about task engagement, not as a manipulation check for self-worth.

rank is reported in the descriptive and balance tables, but is omitted from the regression control vector because the four ranking variables are mechanically constrained.

Summary statistics are reported in Appendix Table A-1. Appendix Table A-2 reports balance between the positive self-worth and control arms using treatment assignment from the Qualtrics randomizer fields. The groups are similar on gender, education, ideology, employment, news-consumption rankings, and ethnicity, while the positive arm has a slightly higher pre-treatment self-esteem than the control arm. A joint balance test across the nonredundant regression-control covariates does not reject equality across assignment groups ($p = 0.690$); controls are added to correct for any remaining imbalances.

5. Results

Most people were not able to perfectly differentiate opinions from facts (Figure 1). On average, a respondent misclassified opinion as fact in 14% (sd=35%) of cases, and fact as opinion in 18% (sd=38%) of cases.

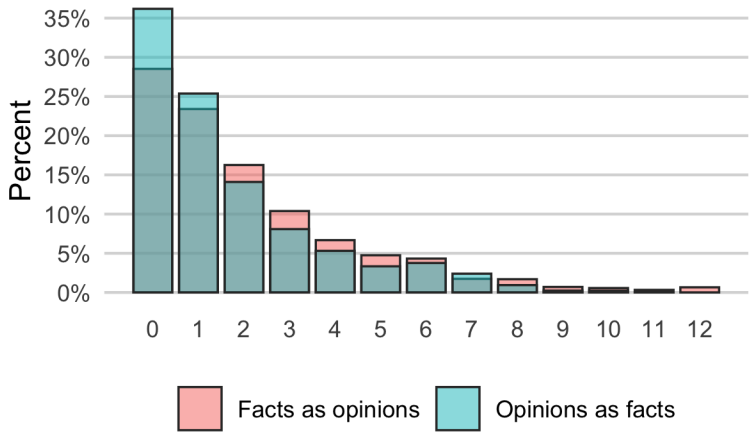

Figure 1: Number of attribution errors

Three substantive findings organize the results. First, classification errors are systematic rather than random: respondents do not apply the same evidentiary standard to all claims. Second, assignment to the positive self-worth task specifically reduces the promotion of Congruent opinions into facts. Third, that effect is politically specific rather than a general tendency

to click "opinion" or spend more time on the task. The hypotheses structure the tests, but the broader pattern is about evidentiary status: which claims citizens allow to function as facts.

5.1. Baseline patterns in classification errors

Our mode of empirical analysis is OLS regression where the dependent variable is the binary variable indicating classification error. In Figure 2, we estimate how statement characteristics predict misclassification. Figure 2 reports adjusted differences in classification-error probabilities within the control group. The coefficients compare statement or alignment categories with the omitted category; they do not estimate effects of the positive self-worth task. The results are reported for three sets of observations: the entire dataset (left panel), facts (central panel), and opinions (right panel). Two specifications are considered: one with respondent controls and one with respondent fixed effects.

The first finding is that classification error is not purely random. Respondents are partly selective about which claims receive the status of a fact. Across the two specifications, opinions produce 3.4 to 3.5 percentage points fewer errors than facts, with the largest raw p-value below 0.001. This indicates that respondents are not simply confused by the task; errors are asymmetric, with factual statements more likely than opinion statements to lose their factual status.

The baseline pattern is partly consistent with H1. Congruent opinions are 0.7 to 1.1 percentage points more likely than Intermediate opinions to be misclassified as facts; this difference is small and imprecise, with the largest raw p-value equal to 0.183. Opposing opinions are 4.7 to 5.5 percentage points less likely than Intermediate opinions to be misclassified as facts, with the largest raw p-value below 0.001. Respondents are therefore least willing to grant factual status to opinions that cut against their side.

The fact-side pattern is not consistent with H2. Congruent facts are 2.2 to 2.6 percentage points less likely than Intermediate facts to be misclassified as opinions, with the largest raw p-value below 0.001. Opposing facts are also 2.6 to 2.8 percentage points less likely to be misclassified, with the largest raw p-value below 0.001.

The difference between congruent and opposing factual statements is small and not statistically distinguishable from zero in the reported tests. Overall, opposing statements are 3.5 to 3.7 percentage points less likely to be misclassified, with the largest raw p-value below 0.001. Congruent and Intermediate statements do not differ statistically. This means the strongest baseline pattern is not a simple symmetric tendency to protect one's side on both margins. The overall ordering is driven primarily by respondents' reluctance to treat Opposing opinions as facts. Fact items show a different pattern: congruent facts in either direction are classified more accurately than Intermediate facts.

The source pattern is consistent with H3: social-media opinions are 6.4 to 6.5 percentage points less likely than newspaper opinions to be misclassified as facts, with the largest raw p-value below 0.001, while social-media facts are 6.8 to 6.9 percentage points more likely than newspaper facts to be misclassified as opinions, with the largest raw p-value below 0.001. The difference for all observations is near zero because, for facts and opinions, the effects of news source on classification error are in opposite directions.

The domain pattern is more nuanced than H4. Political statements are 2.2 to 2.3 percentage points less likely than economics statements to be misclassified overall, with the largest raw p-value below 0.001. But this all-statements estimate combines opposite directional patterns: political facts are 1.5 percentage points more likely than economics facts to be misclassified as opinions, with the largest raw p-value equal to 0.016, while political opinions are 5.8 to 5.9 percentage points less likely than economics opinions to be misclassified as facts, with the largest raw p-value below 0.001. Political content therefore sharpens classification for opinions but makes factual claims more vulnerable to being treated as opinion.

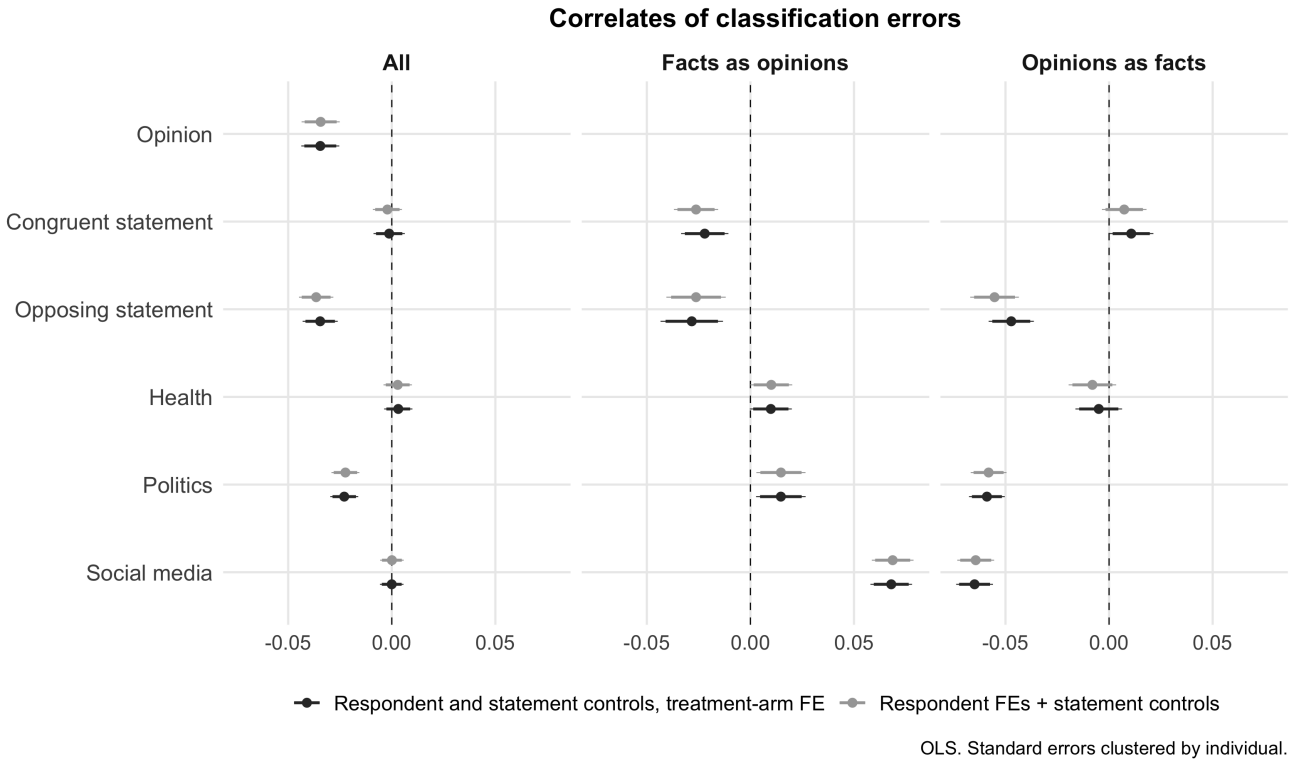

Figure 2: Correlates of classification errors. The figure plots adjusted level differences from linear probability models for classification errors in the control-group descriptive analysis, not randomized treatment effects. The left panel pools all errors; the middle panel restricts the outcome to facts classified as opinions; the right panel restricts the outcome to opinions classified as facts. Positive coefficients indicate a higher probability of error relative to the omitted category. Thin intervals show 95 percent confidence intervals; thicker intervals show 90 percent confidence intervals. Standard errors are clustered by respondent.

5.2. Positive treatment and motivated classification

The second finding concerns the causal effect: assignment to the positive self-worth task changes directional misclassification by reducing the promotion of Congruent opinions into facts.

In Figure 3 we regress misclassification on treatment, assuming that the treatment effect can differ depending on whether the statement is congruent, opposing, or intermediate. We carry out the analysis separately for opinions and facts.[5] As before, intermediate respondent-statement pairs are the omitted congruence category and standard errors are clustered by respondent.

[5] Appendix Figure A-1 reports the corresponding treatment and congruence coefficients.

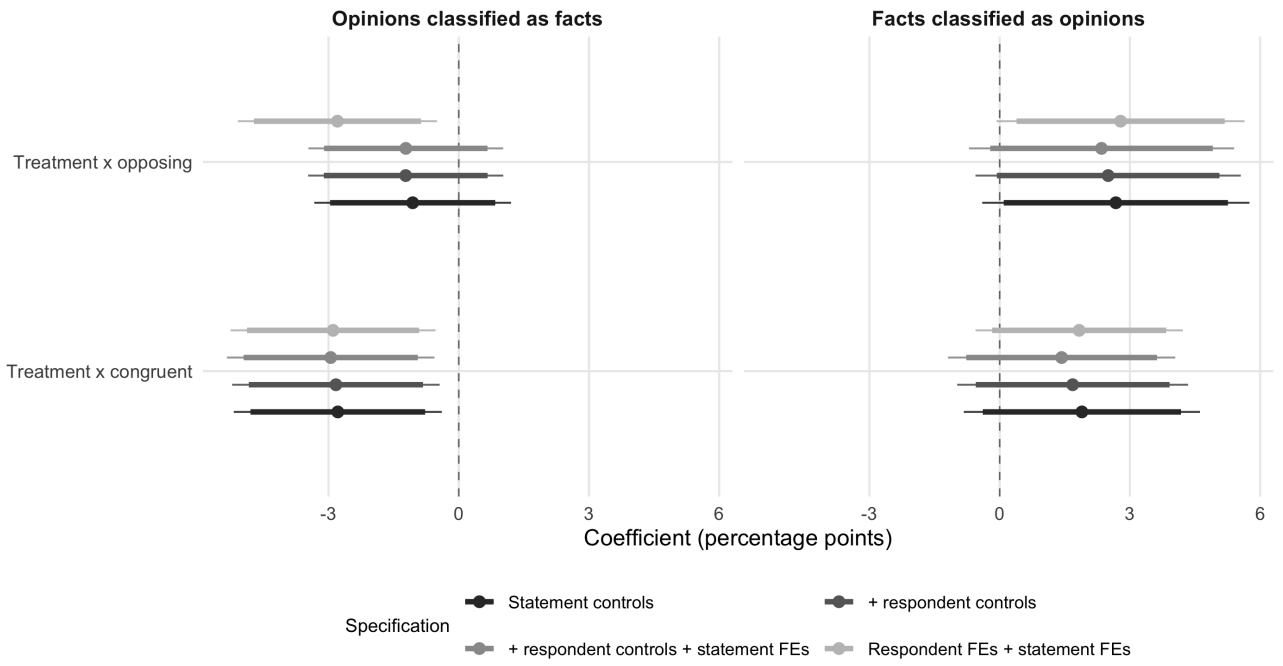

Figure 3: Positive self-worth treatment and directional misclassification across specifications. Points show how much more or less the positive self-worth task changes errors for Congruent or Opposing respondent-statement pairs than for Intermediate pairs. For example, Treatment × Congruent equals the positive-task effect among Congruent pairs minus the positive-task effect among Intermediate pairs; it is not the treatment effect for Congruent statements by itself. Thin intervals show 95 percent confidence intervals; thicker intervals show 90 percent confidence intervals. Estimates are in percentage points and standard errors are clustered by respondent. "Statement controls" includes topic and source. "+ respondent controls" adds the full pre-treatment respondent-control vector described in Section 4.4. "+ respondent controls + statement FEs" replaces topic and source controls with statement fixed effects. "Respondent FEs + statement FEs" includes respondent and statement fixed effects without time-invariant respondent controls. Appendix Figure A-1 reports the corresponding treatment and congruence main effects.

The positive self-worth treatment has a differential effect on Congruent and Intermediate opinions. Across the four specifications in Figure 3, the Congruent-opinion estimate ranges from -2.95 to -2.79 percentage points, and the largest p-value is 0.023. This estimate compares how much the positive task changes classifications for Congruent opinion-statement pairs with how much it changes classifications for Intermediate opinion-statement pairs. It is consistent with H1a. The Opposing-opinion estimate is less stable: it ranges from -2.8 to -1.0 percentage points across specifications, with the largest raw p-value equal to 0.369.

We treat H1a and H2a as the two theory-guided treatment tests. Appendix Section A3.2 reports Holm and Benjamini–Hochberg adjustments for this two-test family and also reports a

conservative adjustment across all four treatment differences shown in Figure 3. The multiple-testing check leaves the Congruent-opinion result below 0.05 for the two theory-guided tests and below 0.10 when all four displayed quantities are adjusted together; the Opposing-fact result remains unsupported. The same substantive pattern also holds when the models add statement fixed effects, which absorb persistent differences in the difficulty of individual statements. Hence, the Congruent-opinion result does not arise simply because Congruent and Intermediate categories contain easier or harder items. The finding suggests that the intervention did not simply improve performance across the entire task or induce a general reluctance to select "fact." Rather, it specifically reduced respondents' willingness to grant factual status to politically congruent opinions.

The fact-as-opinion column is less supportive of a broader claim. The Congruent-fact estimate is positive in every specification, ranging from 1.6 to 1.9 percentage points, with the largest raw p-value equal to 0.248. The estimate for opposing facts is also positive, ranging from 2.35 to 2.79 percentage points across specifications, with the largest raw p-value equal to 0.132. This means H2a is not supported. If anything, the point estimates move in the opposite direction from the prediction, but they are not precise enough to establish a reliable reverse effect. Together with the baseline fact-side pattern, this result suggests that fact classification operates differently from opinion elevation rather than showing simple symmetric partisan defensiveness. We therefore interpret the evidence as supporting the Congruent-opinion implication of the model, but not the stronger claim that bolstering self-worth reduces motivated misclassification on every margin. In substantive terms, the treatment makes respondents less likely to promote congruent opinions into facts; it does not make them uniformly less defensive or uniformly more accurate.

Appendix Figure A-2 shows the underlying treatment effects by statement type and alignment for the model with respondent controls and statement fixed effects. On the opinion side, the Congruent-versus-Intermediate difference comes from a -2.69 percentage-point treatment effect for Congruent pairs and a 0.26 percentage-point treatment effect for Intermediate pairs. Figure 3 shows that this difference is similar across all four displayed specifications.

The aggregate accuracy outcome is less theoretically precise than the directional decomposition. Appendix Figure A-3 provides a descriptive view of the control-group alignment rates that motivate the directional analysis. It shows the same asymmetry as the regression evidence: among opinions, respondents are least likely to promote Opposing claims into facts, while fact-side differences do not show the same simple pattern. The figure is not used as a formal treatment-effect test.

5.3. Additional analysis

Appendix Figure A-4 compares the same respondent-characteristic associations in the full sample and in the control group. The full-sample series includes treatment-arm fixed effects; the control-group-only series is estimated only among respondents assigned to the control writing task. The estimates are descriptive associations and should not be interpreted causally. Conservative ideology is associated with more misclassification, especially facts classified as opinions, while a college degree is associated with fewer opinion-as-fact errors and fewer errors overall. The media-ranking variables should be read as importance ranks: lower values mean the respondent ranked that source as more important. Thus, the negative coefficient for television news implies that respondents who rank television news as more important make more classification errors. These associations are descriptive rather than causal, but they are consistent with prior work showing that fact–opinion classification varies with political predispositions and media environments (Mettler and Mondak, 2024; Mitchell et al., 2018; Bursztyn et al., 2023).

Higher pre-treatment self-esteem is also associated with slightly more errors. Across the full-sample and control-group-only specifications, a one-standard-deviation increase in the self-esteem measure corresponds to a 0.73 to 0.78 percentage-point increase in the pooled error rate, a 0.50 to 0.64 percentage-point increase for opinions classified as facts, and a 0.81 to 1.07 percentage-point increase for facts classified as opinions. The largest raw p-values are 0.178, 0.425, and 0.168, respectively. This does not contradict the experimental result because baseline self-esteem is not randomly assigned and likely captures stable personality, confidence, or response-style differences, whereas the treatment temporarily bolsters self-worth. One inter-

pretation is that stable high self-esteem may be associated with greater confidence in one's own classifications, while the randomized treatment reduces self-protective pressure at the moment of judgment. This distinction is consistent with work separating dispositional self-evaluation from motivated reasoning and identity-protective cognition (Cichocka, Marchlewska and Cislak, 2024*b*; Pantazi, Hale and Klein, 2021; Pennycook and Rand, 2019).

5.4. Robustness and heterogeneity

The third finding is that the focal pattern is politically specific rather than a general response-style shift. The focal opinion-side result is stable against several checks. Leaving out each opinion statement in turn produces estimates with the same sign and similar magnitude: the Congruent-opinion estimate ranges from -3.7 to -2.1 percentage points across the 12 leave-one-out specifications (Appendix Figure A-5), with the effects significant at $p < 0.06$ in all cases. This indicates that no single opinion item drives the result.

The non-partisan-statement placebo weighs against a broad response-style explanation. When the stimulus set is restricted to statements with no partisan slant, these estimates are close to zero, suggesting that the treatment does not simply make respondents uniformly more cautious, more skeptical, or less willing to classify statements as facts (see Online Appendix A3.3). Appendix Section A3.4 reports the numerical baseline estimates underlying Figure 2 and the media-by-type rates used to evaluate whether social-media items are simply more opinion-like.

Treatment assignment also does not produce a detectable increase in classification time. Across the two time-on-task specifications, the positive self-worth treatment changes total classification time by 0.58 to 0.97 percent, with the smallest raw p-value equal to 0.749 (Appendix Figure A-6). The response-time result weighs against an explanation based on longer task engagement, although response time cannot capture every possible change in attention or effort.[6]

Appendix Figure A-7 examines whether the focal treatment effect differs between respon-

[6]The negative social-comparison arm, by contrast, uses a task that is different from both the positive self-worth treatment and the control, and takes 6.45 to 6.70 percent longer than control across the two time-on-task specifications, with the largest raw p-value equal to 0.042. This is why we treat that arm as auxiliary rather than as a clean mirror image of the positive self-worth manipulation.

dents who answered pre-treatment survey pages relatively quickly or slowly. Because the grouping variable is measured before the intervention, this analysis avoids conditioning on post-treatment classification behavior. In the respondent-plus-statement-fixed-effects specification, the focal coefficient is -2.3 percentage points among faster pre-treatment respondents and -3.5 percentage points among slower pre-treatment respondents; a direct test does not show evidence that the two estimates differ ($p = 0.590$).

Similarly, Appendix Figure A-8 calculates the treatment effects separately for liberal and conservative respondents, as the way that political information is processed can vary with ideological alignment (Baron and Jost, 2019). Among liberal and conservative respondents, the corresponding estimates are -2.9 and -3.0 percentage points, respectively, and the direct ideological-group difference test is near zero ($p = 0.975$; Appendix Figure A-8). These subgroup estimates are used to assess whether the focal effect is a one-subgroup artifact, not to claim exact equality. The evidence does not suggest that the result is driven by either liberals or conservatives alone. Appendix Section A3.5 further shows that pre-treatment self-esteem does not identify a clear gradient in responsiveness to the positive-writing task.

6. Discussion

Two conclusions emerge from the evidence. First, respondents do not draw the fact–opinion boundary symmetrically. The clearest baseline pattern concerns opinions: people are especially reluctant to grant factual status to claims that oppose their political side and comparatively more permissive toward congenial interpretations. Factual statements display a less straightforward pattern, suggesting that motivated classification is not simply a mirror-image process on the two sides of the boundary.

Second, temporarily bolstering self-worth narrows this permissiveness toward congruent opinions. The intervention does not produce a general improvement in classification accuracy, nor does it consistently alter the treatment of factual statements. Its clearest effect appears where political interpretation can be promoted into evidence. This asymmetry is theoretically informative: self-protective motives may matter most when citizens decide whether congenial

rhetoric deserves factual authority, rather than whenever they encounter politically uncomfortable information.

These findings shift attention to a classificatory stage that precedes conventional belief updating. Political communication can influence citizens not only by changing the beliefs they hold, but also by changing which claims they permit to function as evidence. That mechanism is particularly relevant in media environments where reporting, commentary, elite cues, and campaign rhetoric regularly appear in similar formats. It also complements recent work showing that partisan cues and misinformation interventions affect how citizens evaluate political information after exposure (Clayton et al., 2020; Fuller, de la Cerda and Rametta, 2026; Barbieri, Petitpas and Sciarini, 2026). A fact-like presentation can matter even when a claim is not literally false: it can invite citizens to treat interpretation as evidence.

This matters for democratic accountability. Accountability depends on citizens being able to distinguish claims that can be checked against external evidence from claims that mainly express evaluation, blame, or persuasion. If citizens assign evidentiary status in politically motivated ways, then campaigns, partisan media, and political elites need not persuade only by changing beliefs after evidence is received. They can also benefit from framing congenial interpretations as the sort of claims that deserve evidentiary weight. Our experiment does not identify long-run media effects or strategic elite behavior. It identifies one mechanism that can make those environments consequential: citizens may differ in which political claims they allow to enter the evidentiary record in the first place.

The scope conditions are also important. The design uses a fixed statement set, a short online classification task, and a brief self-worth intervention, so the results identify a specific psychological mechanism rather than the full ecology of political media consumption.

7. Conclusion

Citizens do not enter public debate armed only with attitudes; they also carry a working map of what counts as evidence. In an information environment crowded with commentary that mimics reportage, that map becomes politically important. A claim need not be false to mislead.

It can also mislead by being received as fact when it is better understood as opinion.

This paper studies one micro-foundation of that process. It extends motivated-reasoning research by moving earlier in the sequence of political information processing. The question is not only how citizens update after receiving evidence, but also whether they classify a claim as evidence before updating begins. In a survey experiment, we show first that fact–opinion classification is systematically structured by respondent-statement congruence, source, and topic even in the control group. Respondents are not simply bad at the task. They are selectively permissive toward some claims and selectively skeptical toward others. Second, assignment to a brief positive self-worth treatment lowers the probability that congruent opinions are classified as facts. Non-partisan-statement placebo tests are close to zero, and the effect is not accompanied by more time on task. Fact-side effects are less uniform, so the clearest evidence locates the mechanism at the point where congenial interpretation is promoted into evidence.

That pattern fits the motivated-reasoning interpretation developed in the introduction. Much of that literature studies whether people shade beliefs in self-serving directions when self-worth or identity is at stake. Our contribution is to test whether the same logic operates earlier, when citizens decide whether a claim should be treated as fact or opinion. The evidence shows that assignment to a task designed to bolster self-worth can affect this prior classificatory step. The positive task makes respondents less willing to let congenial interpretation borrow the authority of fact, revealing one mechanism through which opinion-saturated political communication can become persuasive.

The broader implication for political behavior is that persuasion may work partly by changing the evidentiary status citizens assign to claims. A politician, campaign, or media entrepreneur who phrases policy judgments in fact-like language, and delivers them in rhetoric that resonates with identity, may gain leverage not simply by changing what citizens think but by shaping what citizens treat as information. The mechanism may be especially consequential in environments where social standing feels precarious, including periods of economic insecurity, because such environments may create demand for self-protective interpretations. We do not observe such strategic targeting or test these macro conditions directly, and we do not claim that

self-worth is the only channel through which rhetorical authority is produced. But the results identify one plausible psychological mechanism through which such conditions could matter: the ego utility of treating congruent claims as facts.

The design isolates a specific mechanism. The treatment is brief, the stimuli are short statements rather than full articles, and the model abstracts from the strategic supply side. Those limits are also virtues: they allow us to identify a process that is hard to see in naturally occurring media data. The broader lesson is simple. The politics of information begins before citizens decide what is true. It begins when they decide what sort of claim they are looking at, and whether that claim deserves to count as evidence.

Statements and Declarations

Ethics approval. The study was approved by the Institutional Review Board at the University of Texas at Dallas (IRB-23-392) on October 10, 2023.

Funding. This research was supported by the Institute for Humane Studies at George Mason University through grant IHS017555 to Natalia Lamberova.

Use of generative AI. During the preparation of this manuscript, the authors used generative AI tools to assist with manuscript editing, improving clarity and organization, checking internal consistency, identifying relevant literature, and reviewing code and document structure. The AI tools were not used to generate empirical results, conduct statistical analyses, or make scientific decisions. All analyses, interpretations, and conclusions were developed, verified, and approved by the authors, who take full responsibility for the manuscript.

REFERENCES

Acharya, Avidit, Kyungtae Park and Tomer Zaidman. 2025. "Motivated Reasoning and Information Aggregation." *arXiv preprint arXiv:2512.10125* . 3

Alesina, Alberto and Stefanie Stantcheva. 2020. "The polarization of reality.". **URL:** *http://www.nber.org/papers/w26675* 1

Aslett, Kevin, Zeve Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler, Richard Bonneau and Joshua A Tucker. 2021. "Testing the Effect of Information on Discerning the Veracity of News in Real Time." *Journal of Experimental Political Science* pp. 1–15. 1

Barbieri, Andrea, Adrien Petitpas and Pascal Sciarini. 2026. "Betting on Partisanship: Biased Information Processing and Opinion Change on a Citizen Initiative." *Political Behavior* 48:875–893. 1, 2.2, 6

Baron, Jonathan and John T Jost. 2019. "False equivalence: Are liberals and conservatives in the United States equally biased?" *Perspectives on Psychological Science* 14(2):292–303. 5.4

Baum, Matthew A. and Tim Groeling. 2009. "Shot by the Messenger: Partisan Cues and Public Opinion Regarding National Security and War." *Political Behavior* 31(2):157–186. 1, 2.4

Bénabou, Roland and Jean Tirole. 2002. "Self-confidence and personal motivation." *Quarterly Journal of Economics* pp. 19–57. 1, 2.2

Bénabou, Roland and Jean Tirole. 2016. "Mindful economics: The production, consumption, and value of beliefs." *Journal of Economic Perspectives* 30(3):141–164. 1, 2.2

Bolsen, Toby, James N. Druckman and Fay Lomax Cook. 2014. "The Influence of Partisan Motivated Reasoning on Public Opinion." *Political Behavior* 36(2):235–262. 2.2

Bracha, Anat and Donald J Brown. 2012. "Affective decision making: A theory of optimism bias." *Games and Economic Behavior* 75(1):67–80. 1, 2.2

Bursztyn, Leonardo, Aakaash Rao, Christopher Roth and David Yanagizawa-Drott. 2023. "Opinions as facts." *The Review of Economic Studies* 90(4):1832–1864. 1, 5.3

Cichocka, Aleksandra, Marta Marchlewska and Aleksandra Cislak. 2024*a*. "Self-Worth and Politics: The Distinctive Roles of Self-Esteem and Narcissism." *Political Psychology* 45(S1):43–85. 1

Cichocka, Aleksandra, Marta Marchlewska and Aleksandra Cislak. 2024*b*. "Self-worth and politics: the distinctive roles of self-esteem and narcissism." *Political Psychology* 45:43–85. 5.3

Clayton, Katherine, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, Morgan Sandhu, Rachel Sang, Rachel Scholz-Bright, Austin T. Welch, Andrew G. Wolff, Amanda Zhou and Brendan Nyhan. 2020. "Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media." *Political Behavior* 42:1073–1095. 1, 2.2, 6

Cohen, Geoffrey L and David K Sherman. 2014. "The psychology of change: Self-affirmation and social psychological intervention." *Annual review of psychology* 65(1):333–371. 1, 2.2, 4.2

Douglas, Benjamin D, Patrick J Ewell and Markus Brauer. 2023. "Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA." *Plos one* 18(3):e0279720. 4.1

Fuller, Sam, Nicolás de la Cerda and Jack T. Rametta. 2026. "Affect, Not Ideology: The Heterogeneous Effects of Partisan Cues on Policy Support." *Political Behavior* 48:273–297. 1, 2.2, 6

Gedikli, Ceren, Marianna Miraglia, Sara Connolly, Mark Bryan and David Watson. 2023. "The Relationship Between Unemployment and Wellbeing: An Updated Meta-Analysis of Longitudinal Evidence." *European Journal of Work and Organizational Psychology* 32(1):128–144. 1

Graham, Matthew H. 2020. "Self-Awareness of Political Knowledge." *Political Behavior* 42:305–326. 1

Jiang, Shaohai and Annabel Ngien. 2020. "The effects of Instagram use, social comparison, and self-esteem on social anxiety: A survey study in Singapore." *Social Media+ Society* 6(2):2056305120912488. A3.1

Kavanagh, Jennifer and Michael D. Rich. 2018. *Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life*. Santa Monica, CA: RAND Corporation. 1

Krauss, Samantha and Ulrich Orth. 2022. "Work Experiences and Self-Esteem Development: A Meta-Analysis of Longitudinal Studies." *European Journal of Personality* . 1

Kunda, Ziva. 1990. "The case for motivated reasoning." *Psychological bulletin* 108(3):480. 2.2, 3

Lewandowsky, Stephan, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz and John Cook. 2012. "Misinformation and Its Correction: Continued Influence and Successful Debiasing." *Psychological Science in the Public Interest* 13(3):106–131. PMID: 26173286.
URL: *https://doi.org/10.1177/1529100612451018* 1

Lewandowsky, Stephan, Ullrich K.H. Ecker and John Cook. 2017. "Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era." *Journal of Applied Research in Memory and Cognition* 6(4):353–369.
URL: *https://www.sciencedirect.com/science/article/pii/S2211368117300700* 1

Little, Andrew T. 2025. "How to Distinguish Motivated Reasoning from Bayesian Updating." *Political Behavior* pp. 1–25. 1, 3, 3, A1

Melita, Davide, Guillermo B. Willis and Rosa Rodriguez-Bailon. 2021. "Economic Inequality Increases Status Anxiety Through Perceived Contextual Competitiveness." *Frontiers in Psychology* 12. 1

Mettler, Matthew and Jeffery J Mondak. 2024. "Fact-opinion differentiation." *Harvard Kennedy School Misinformation Review* . 1, 2.1, 5.3

Mitchell, Amy, Jeffrey Gottfried, Michael Barthel and Nami Sumida. 2018. Distinguishing Between Factual and Opinion Statements in the News. Technical report Pew Research Center.

URL: *https://www.pewresearch.org/journalism/2018/06/18/distinguishing-between-factual-and-opinion-statements-in-the-news/* 5.3

Nyhan, Brendan and Jason Reifler. 2010. "When Corrections Fail: The Persistence of Political Misperceptions." *Political Behavior* 32(2):303–330. 1

Pantazi, Myrto, Scott Hale and Olivier Klein. 2021. "Social and Cognitive Aspects of the Vulnerability to Political Misinformation." *Political Psychology* 42(S1):267–304. 5.3

Pennycook, Gordon and David G. Rand. 2019. "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning." *Cognition* 188:39–50. 5.3

Rash, Joshua A, M Kyle Matsuba and Kenneth M Prkachin. 2011. "Gratitude and well-being: Who benefits the most from a gratitude intervention?" *Applied Psychology: Health and Well-Being* 3(3):350–369. 1, 4.2

Rosenberg, Morris. 1965. *Society and the Adolescent Self-Image*. Princeton, NJ: Princeton University Press. 4.4

Sherman, David K and Geoffrey L Cohen. 2006. "The psychology of self-defense: Self-affirmation theory." *Advances in experimental social psychology* 38:183–242. 1, 2.2, 4.2

Steele, Claude M. 1988. The Psychology of Self-Affirmation: Sustaining the Integrity of the Self. In *Advances in Experimental Social Psychology*, ed. Leonard Berkowitz. Vol. 21 of *Advances in Experimental Social Psychology* Academic Press pp. 261–302.

URL: *https://www.sciencedirect.com/science/article/pii/S0065260108602294* 1, 2.2, 4.2

Swire, Briony, Ullrich K. H. Ecker and Stephan Lewandowsky. 2017. "The role of familiarity in correcting inaccurate information." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 43(12):1948–1961. 1

Taber, Charles S and Milton Lodge. 2006. "Motivated skepticism in the evaluation of political beliefs." *American journal of political science* 50(3):755–769. 2.2, 3

Tucker, Joshua A, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal and Brendan Nyhan. 2018. "Social media, political polarization, and political disinformation: A review of the scientific literature." *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)* . 1

Vogel, Erin A, Jason P Rose, Lindsay R Roberts and Katheryn Eckles. 2014. "Social comparison, social media, and self-esteem." *Psychology of popular media culture* 3(4):206. 4.2, A3.1

Westerman, David, Patric R Spence and Brandon Van Der Heide. 2014. "Social media as information source: Recency of updates and credibility of information." *Journal of computer-mediated communication* 19(2):171–183. 1, 2.4

Zumofen, Guillaume, Isabelle Stadelmann-Steffen and Marc Bühlmann. 2024. "No, It Is Not All About Selective Exposure: Information Selection Strategies in Referendums." *Political Behavior* 46:1771–1790. 1, 2.2

Online appendix for

When Politically Congruent Opinions Pass for Facts: Self-Worth and Motivated Classification in Media Environments

Table of Contents

A1. Definition of payoffs and proofs of statements

Formally, we assume that the individual's payoff takes the following form:

$$U = -D_{KL}(f(\cdot|s,\tilde{b})||f(\cdot|s,b)) + \gamma \int u(a+yx)dF(x|s,\tilde{b}). \tag{A1}$$

Following Little (2025), we take the first summand of (A1) to be the Kullback-Leibler divergence between the objective posterior distribution of x

$$D_{KL}(f(\cdot|s,\tilde{b})||f(\cdot|s,b)) = \int \log\left(\frac{f(\cdot|s,\tilde{b})}{f(\cdot|s,b)}\right) f(\cdot|s,\tilde{b})df.$$

It is a way to measure how two probability distributions are different from one another. For $\tilde{b} = b$, the subjective and objective ex post distributions of x are the same, and the divergence $D_{KL}(f(\cdot|s,\tilde{b})||f(\cdot|s,b))$ is zero; otherwise, it is positive.

The second summand of (A1) is one's expected belief-based utility. We assume that the individual is risk-averse with respect to the level of self-esteem, with constant absolute risk aversion: $u(z) = -e^{-z}$.

Proof of Proposition 1. We start by noting that

$$x|s,b \sim \mathcal{N}\left(\frac{s-b}{1+\sigma^2}, \frac{\sigma^2}{1+\sigma^2}\right), \qquad x|s,\tilde{b} \sim \mathcal{N}\left(\frac{s-\tilde{b}}{1+\sigma^2}, \frac{\sigma^2}{1+\sigma^2}\right).$$

This gives us

$$D_{KL}(f(\cdot|s,\tilde{b})||f(\cdot|s,b)) = \frac{(b-\tilde{b})^2}{2\sigma^2(1+\sigma^2)}.$$

The second summand of (A1) becomes

$$-\gamma\sqrt{\frac{1+\sigma^2}{2\pi\sigma^2}} \int \exp\left(-a+xy - \frac{1+\sigma^2}{2\sigma^2}\left(x - \frac{s-\tilde{b}}{1+\sigma^2}\right)^2\right) dx =$$

$$= -\gamma\sqrt{\frac{1+\sigma^2}{2\pi\sigma^2}} \exp\left(-a + \frac{y^2\sigma^2 - 2(s-\tilde{b})y}{2(1+\sigma^2)}\right) \int \exp\left(-\frac{1+\sigma^2}{2\sigma^2}\left(x - \frac{2y\sigma^2 - s + \tilde{b}}{1+\sigma^2}\right)^2\right) dx =$$

$$= -\gamma \exp\left(-a + \frac{y^2\sigma^2 - 2(s-\tilde{b})y}{2(1+\sigma^2)}\right).$$

The individual's motivated belief problem can be rewritten as

$$\max_{\tilde{b}\in\{-1,0,1\}} U(s,b,\tilde{b}) = -\frac{(b-\tilde{b})^2}{2\sigma^2(1+\sigma^2)} - \gamma \exp\left(-a + \frac{y^2\sigma^2 - 2(s-\tilde{b})y}{2(1+\sigma^2)}\right). \tag{A2}$$

Q.E.D.

Proof of Proposition 2. Denote

$$X = \frac{1}{2\gamma\sigma^2(1+\sigma^2)}, \qquad V = \exp\left(-a + \frac{\sigma^2 - 2s}{2(1+\sigma^2)}\right), \qquad W = \exp\left(\frac{1}{1+\sigma^2}\right).$$

We have $U(s,0,-1) \geq U(s,0,0)$ if and only if $V \geq \frac{XW}{W-1}$ or

$$a \leq a_0 = \frac{\sigma^2 - 2s}{2(1+\sigma^2)} + \log(2\gamma\sigma^2(1+\sigma^2)) + \log\left(1 - \exp\left(-\frac{1}{1+\sigma^2}\right)\right).$$

We have $U(s,1,-1) \geq U(s,1,0)$ if and only if $V \geq \frac{3XW}{W-1}$ or

$$a \leq \underline{a}_1 = \frac{\sigma^2 - 2s}{2(1+\sigma^2)} + \log\left(\frac{2}{3}\gamma\sigma^2(1+\sigma^2)\right) + \log\left(1 - \exp\left(-\frac{1}{1+\sigma^2}\right)\right).$$

Finally, we have $U(s,1,0) \geq U(s,1,1)$ if and only if $V \geq \frac{X}{W-1}$ or

$$a \leq \bar{a}_1 = \frac{\sigma^2 - 2s}{2(1+\sigma^2)} + \log\left(2\gamma\sigma^2(1+\sigma^2)\right) + \log\left(\exp\left(\frac{1}{1+\sigma^2}\right) - 1\right).$$

As $\frac{X}{W-1} < \frac{XW}{W-1} < \frac{3XW}{W-1}$, we have $\underline{a}_1 < a_0 < \bar{a}_1$. **Q.E.D.**

Proof of Proposition 3. The comparative statics with respect to γ is trivial. The values tend to $-\infty$ as either $\gamma \to 0$ or $\sigma^2 \to 0$ because of the $\log(2\gamma\sigma^2(1+\sigma^2))$ term. For the $\sigma^2 \to \infty$ case,

we need to evaluate the limit of $\sigma^2(1+\sigma^2)\left(\exp\left(\frac{1}{1+\sigma^2}\right)-1\right)$. Asymptotic expansion gives us

$$\sigma^2(1+\sigma^2)\left(\exp\left(\frac{1}{1+\sigma^2}\right)-1\right)=\sigma^2(1+\sigma^2)\left(\frac{1}{\sigma^2}-\frac{1}{2\sigma^4}+\frac{1}{6\sigma^6}+\ldots\right),$$

with the highest term σ^2, so it tends to ∞ as $\sigma^2\to\infty$. **Q.E.D.**

A2. SUPPLEMENTAL TABLES AND FIGURES

Table A-1: Summary statistics

	N	Mean	Median	SD	Min	Max
Metric Variables						
Pre-treatment self-esteem (legacy sum)	2172	19.95	20.00	5.47	4.00	28.00
Correctly classified facts	2173	9.87	11.00	2.36	0.00	12.00
Correctly classified opinions	2173	10.33	11.00	2.00	1.00	12.00
Total correct classifications	2173	20.20	21.00	3.54	6.00	24.00
	N	Freq.	%			
Categorical Variables						
Gender						
Male	2183	990	0.454			
Female	2183	1183	0.542			
Prefer not to answer	2183	6	0.003			
Non-binary	2183	4	0.002			
Education						
High School	2183	245	0.112			
Some college	2183	436	0.200			
Associates	2183	258	0.118			
Bachelor's	2183	858	0.393			
Advanced	2183	386	0.177			
Political Views						
Extremely Liberal	2183	281	0.129			
Liberal	2183	574	0.263			
Slightly Liberal	2183	290	0.133			
Moderate	2183	415	0.190			
Slightly Conservative	2183	204	0.093			
Conservative	2183	277	0.127			
Extremely Conservative	2183	103	0.047			
Haven't thought much about this	2183	39	0.018			

Table A-2: Balance by treatment assignment

Covariate	Neutral control	Positive self-worth	Negative social comparison	Omnibus p	Pos.-control p
Female	0.54 (0.50)	0.54 (0.50)	0.54 (0.50)	1.000	0.996
Education scale	3.24 (1.28)	3.30 (1.32)	3.32 (1.27)	0.404	0.356
Ideology scale	3.52 (1.88)	3.55 (1.79)	3.47 (1.87)	0.682	0.757
Pre-treatment self-esteem (PCA z-score)	-0.07 (1.01)	0.04 (1.00)	0.03 (0.99)	0.065	0.032

Notes: Cells report means with standard deviations in parentheses by treatment assignment. Treatment assignment is recorded in the analysis file through the assigned task fields. Self-esteem is the first principal component of the eight pre-treatment self-esteem items, standardized to mean zero and standard deviation one. Omnibus p-values come from one-way ANOVA; positive-control p-values come from two-sided t-tests. Because treatment is randomized, these tests are descriptive diagnostics rather than identifying assumptions.

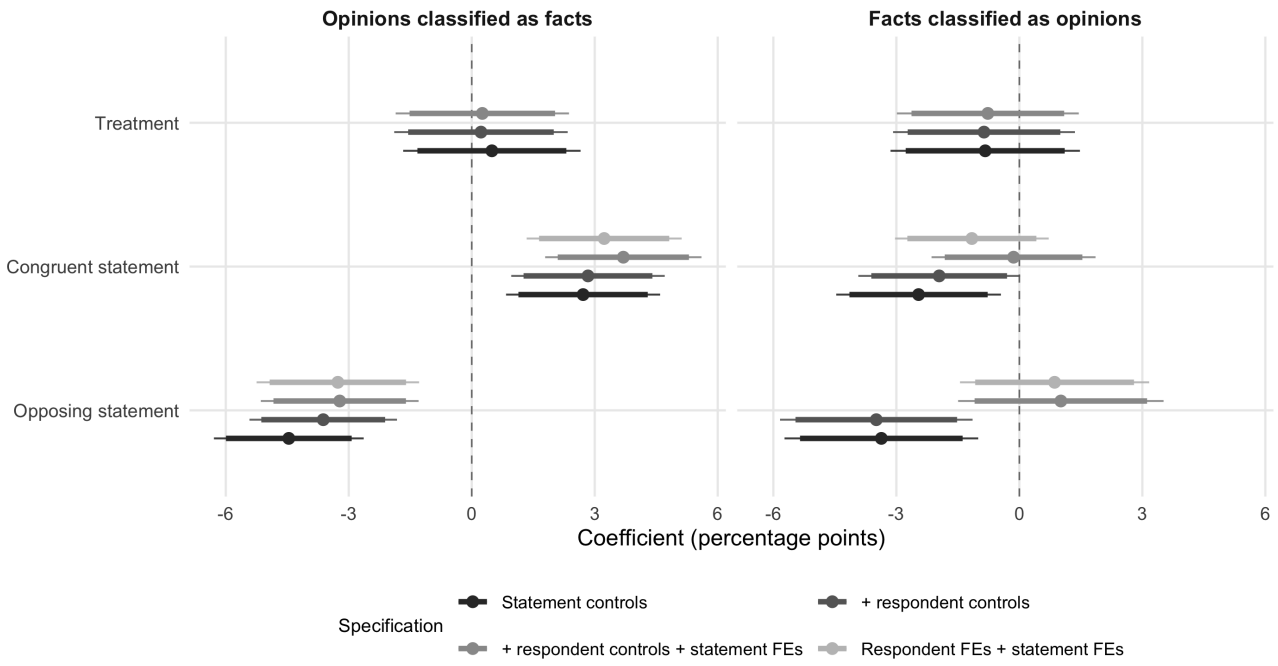

Figure A-1: Main effects from the Figure 3 models. The positive-treatment coefficient is the treatment effect for Intermediate respondent-statement pairs, the omitted alignment category. The Congruent and Opposing coefficients are baseline alignment differences in the control group. These quantities are components of the interaction model rather than the focal treatment-by-alignment tests. Estimates are in percentage points; thin intervals show 95 percent confidence intervals and thicker intervals show 90 percent confidence intervals. Standard errors are clustered by respondent.

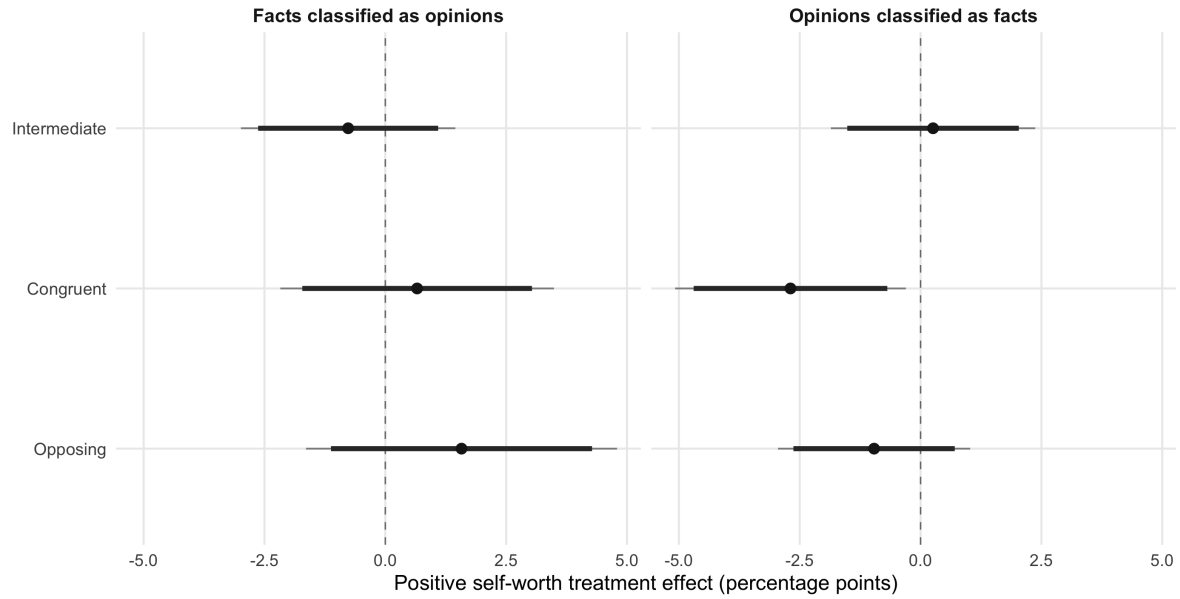

Figure A-2: Cell-specific positive self-worth treatment effects by statement type and respondent–statement alignment. Appendix Figure A-2 breaks out the model from Figure 3 that includes respondent controls and statement fixed effects into the six alignment-specific treatment effects. The Rmd fits two linear probability models–one for opinions classified as facts and one for facts classified as opinions–and computes the treatment effect within each alignment category. The figure therefore does not estimate six independent regressions. Formal tests of differences across alignment categories come from Figure 3; they should not be inferred from overlap or non-overlap of confidence intervals in this figure. Models include respondent controls and statement fixed effects; standard errors are clustered by respondent. Estimates are in percentage points, thin intervals show 95 percent confidence intervals, and thick intervals show 90 percent confidence intervals. Positive estimates indicate more directional misclassification.

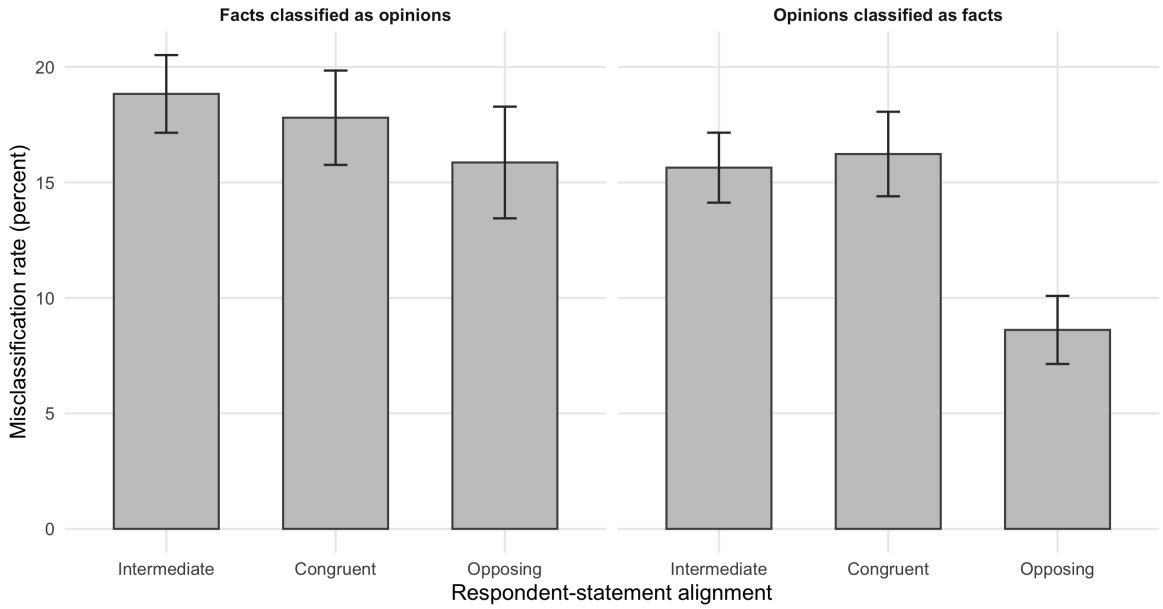

Figure A-3: Control-group directional misclassification rates by statement type and respondent–statement alignment. Bars show mean error rates; intervals are 95 percent confidence intervals based on respondent-clustered standard errors.

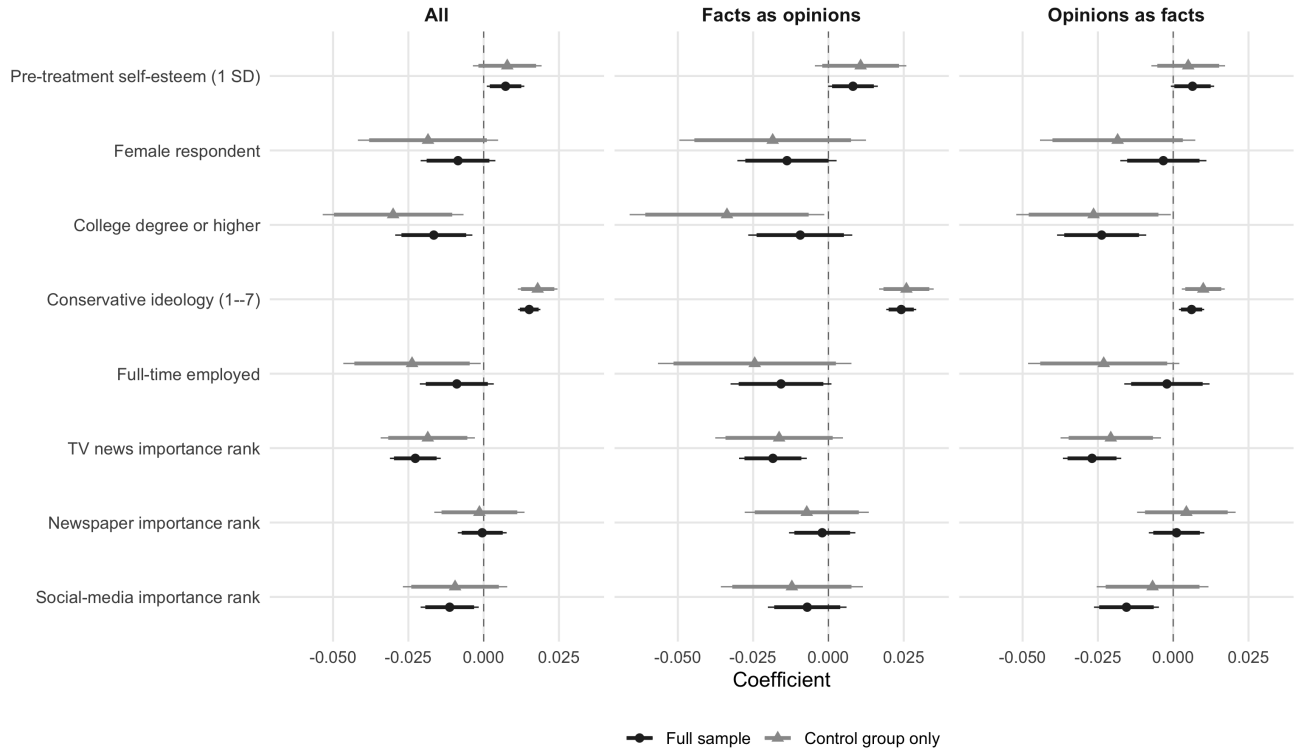

Figure A-4: Associations between pre-treatment respondent characteristics and directional classification errors. The full-sample series includes treatment-arm fixed effects; the control-group-only series is estimated only among respondents assigned to the control writing task. These coefficients are descriptive associations and should not be interpreted causally. Media variables are importance ranks, so lower values indicate greater reported importance of the news source. Standard errors are clustered by respondent. Thin intervals show 95 percent confidence intervals and thick intervals show 90 percent confidence intervals.

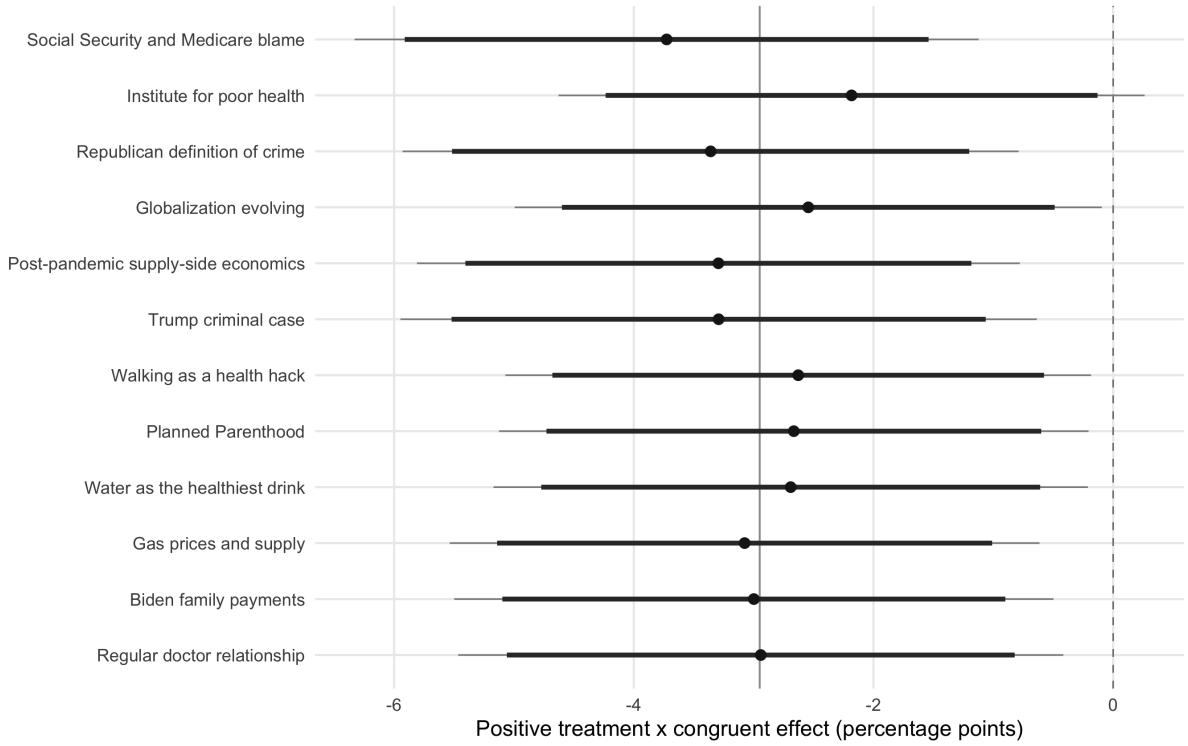

Figure A-5: Leave-one-statement-out robustness for the focal congruent-opinion effect. For each of the 12 opinion statements, the figure reports the Congruent-opinion estimate from the Figure 3 opinion-as-fact model with respondent controls and statement fixed effects after excluding that statement. Statements are ordered by the absolute change in the coefficient relative to the full-sample estimate. The light solid line marks the full-sample coefficient and the dashed line marks zero. This exercise assesses whether one statement drives the focal estimate; it does not estimate statement-specific treatment effects or establish generalizability to statements outside the experiment. Models include respondent controls and statement fixed effects; standard errors are clustered by respondent. Thin intervals show 95 percent confidence intervals and thick intervals show 90 percent confidence intervals.

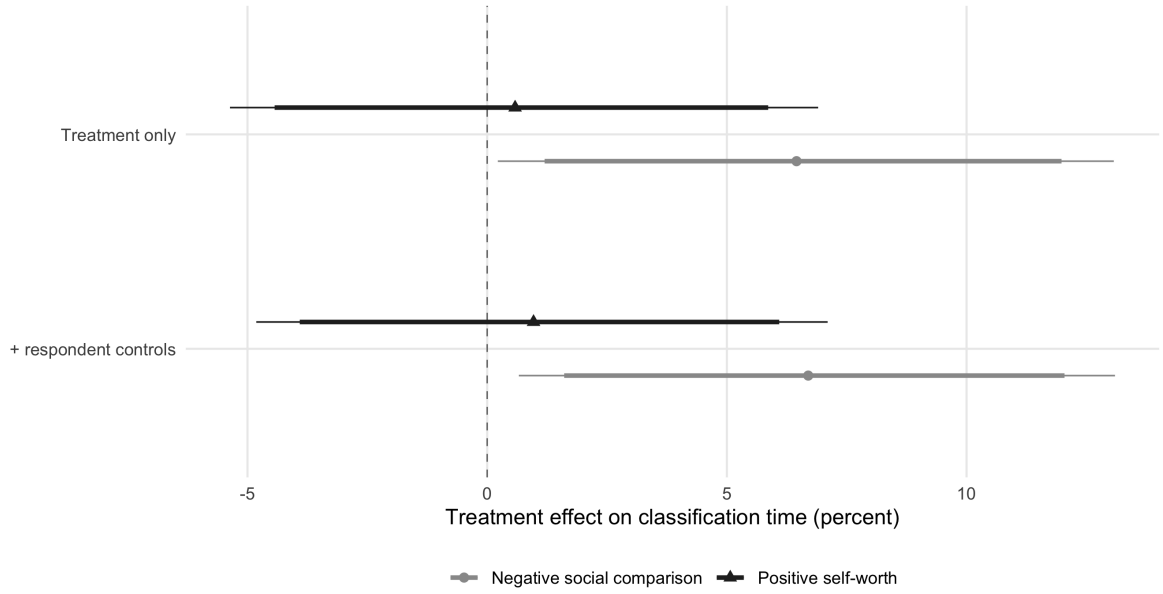

Figure A-6: Treatment effects on total classification time. The dependent variable is the logarithm of the total time each respondent spent across the four classification pages; plotted coefficients are transformed to percentage changes using $100 \times [\exp(\hat{\beta}) - 1]$. Because there is one time outcome per respondent, the specifications include treatment assignment and, where indicated, respondent controls only. Statement controls or statement fixed effects cannot be estimated in this respondent-level model. The figure is auxiliary evidence about effort or engagement, not a manipulation check for self-worth. Thin intervals show 95 percent confidence intervals and thick intervals show 90 percent confidence intervals.

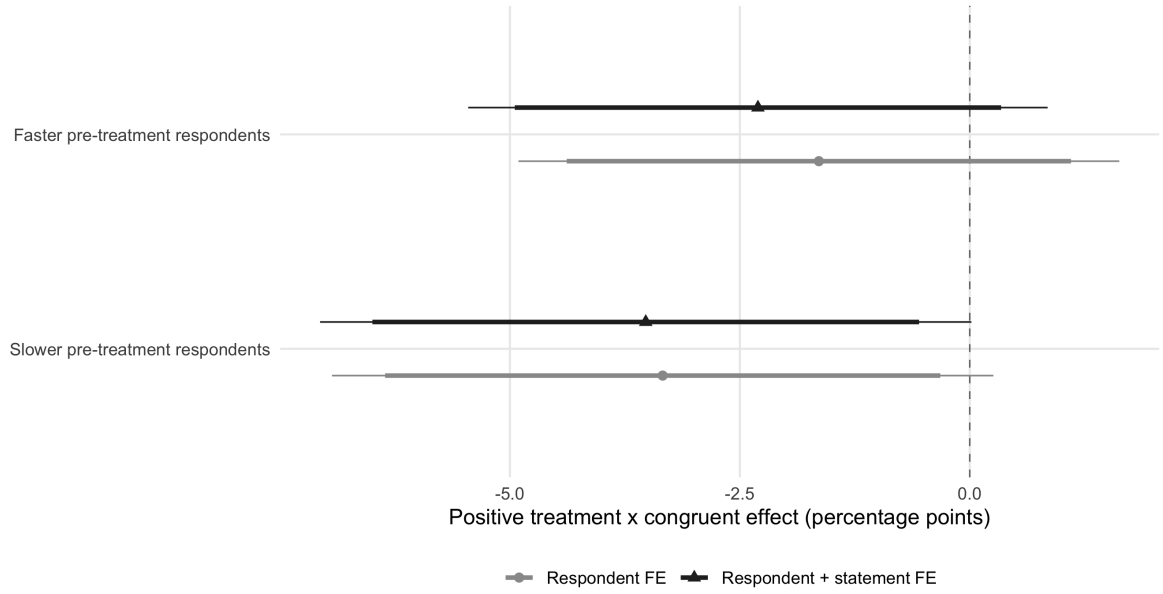

Figure A-7: Heterogeneity by pre-treatment response speed. Points report subgroup-specific Treatment $\times$ Congruent effects for the opinion-as-fact outcome. Speed is constructed exclusively from pre-treatment pages, so the split does not condition on a post-treatment outcome. The pooled triple-interaction model provides the formal test of whether the effects differ across groups; differences in whether individual subgroup confidence intervals cross zero should not be interpreted as evidence of a statistically significant between-group difference. Thin intervals show 95 percent confidence intervals and thick intervals show 90 percent confidence intervals. Standard errors are clustered by respondent.

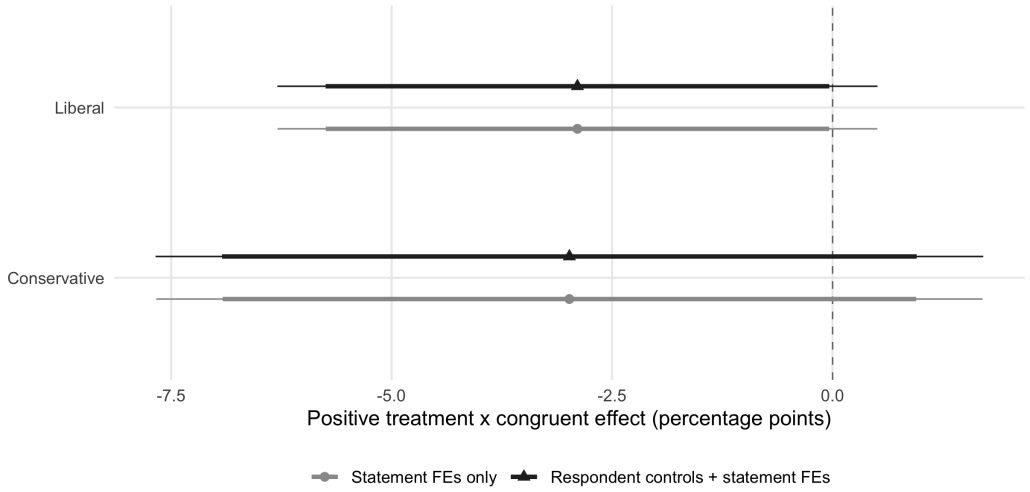

Figure A-8: Ideological-group heterogeneity in the focal congruent-opinion effect. Points report subgroup-specific Treatment $\times$ Congruent effects for the opinion-as-fact outcome, estimated separately among liberal and conservative respondents. The figure shows a statement-fixed-effects specification and a specification that also adds respondent controls. The pooled triple-interaction model provides the formal test of whether the effects differ across groups; differences in whether individual subgroup confidence intervals cross zero should not be interpreted as evidence of a statistically significant between-group difference. Standard errors are clustered by respondent. Thin intervals show 95 percent confidence intervals and thick intervals show 90 percent confidence intervals.

A3. Supplemental analysis

A3.1. Negative treatment

The negative intervention involved exposing participants to a series of idealized, Instagram-like posts from individuals of the same gender, boasting about highly unlikely achievements. This exposure to upward comparison stimuli has been demonstrated to lower self-esteem (Vogel et al., 2014). Jiang and Ngien (2020) show that, while Instagram use per se does not impact self-esteem directly, it does induce social comparisons that have detrimental effect on self-esteem. In our study, participants were asked to assess the success of unrealistic Instagram bloggers to their own, thereby encouraging comparisons to these unrealistic benchmarks.

The auxiliary negative social-comparison task does not generate the mirror-image pattern implied by a simple monotonic self-worth account. Figure A-9 reports how much more or less the negative social-comparison task changes errors for Congruent or Opposing respondent–statement pairs than for Intermediate pairs. The figure does not compare the positive and negative active arms directly.

The negative arm's treatment-by-Congruent estimate for opinion statements is negative, like the corresponding positive-task estimate, rather than positive. Across the four displayed specifications, this estimate ranges from -2.42 to -2.00 percentage points, with the largest raw p-value equal to 0.099. The opposing-opinion estimate also remains negative, ranging from -2.81 to -1.84 percentage points, with the largest raw p-value equal to 0.108. At the same time, the negative task differs substantially in format from the positive and control prompts and increases time spent on the classification block. One possible interpretation is that the two active tasks reach a similar classification outcome through different routes: positive reflection may reduce defensive motivation, whereas upward comparison may increase vigilance, scrutiny, or effort. We therefore report the negative arm as auxiliary evidence that active psychological tasks can move the fact–opinion boundary, but not as a clean inverse manipulation or as a manipulation check for the proposed self-worth mechanism.

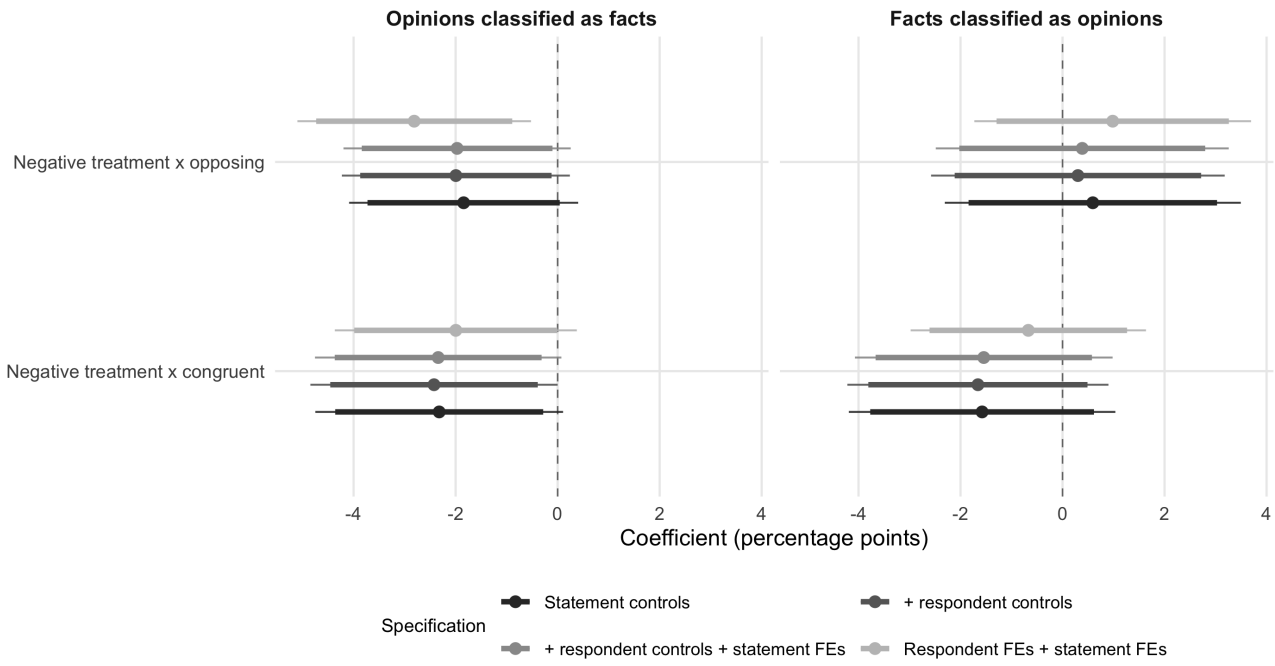

Figure A-9: Negative social-comparison treatment-by-alignment effects. Points show how much more or less the negative social-comparison task changes errors for Congruent or Opposing respondent-statement pairs than for Intermediate pairs. The left panel shows opinions classified as facts and the right panel shows facts classified as opinions. Models progress from statement controls to respondent controls, respondent controls plus statement fixed effects, and respondent plus statement fixed effects. Standard errors are clustered by respondent. Thick intervals show 90 percent confidence intervals and thin intervals show 95 percent confidence intervals. The figure repeats the Figure 3 comparison for negative social comparison versus control; it is not a direct positive-versus-negative comparison.

A3.2. Multiple-testing adjustment

Table A-3 reports Holm and Benjamini–Hochberg adjusted p-values for the two theory-guided treatment tests and, conservatively, for all four treatment differences displayed in Figure 3. The Congruent-opinion result remains below 0.05 when the two theory-guided tests are adjusted together and remains below 0.10 when all four displayed quantities are adjusted together. The Opposing-fact result remains unsupported. H1 and H2 describe baseline control-group patterns and belong to a separate descriptive analysis, so they are not included in this treatment-effect adjustment.

Table A-3: Multiple-testing adjustments for directional treatment effects

Test	Family	Raw p	Holm p: focal two	BH p: focal two	Holm p: all four	BH p: all four
H1a: positive treatment x congruent opinions	Theory-guided treatment effect	0.016	0.031	0.031	0.063	0.063
Secondary: positive treatment x opposing opinions	Secondary diagnostic	0.287			0.571	0.287
Secondary: positive treatment x congruent facts	Secondary diagnostic	0.285			0.571	0.287
H2a: positive treatment x opposing facts	Theory-guided treatment effect	0.132	0.132	0.132	0.397	0.265

Notes: Raw p-values come from the model in Figure 3 that includes respondent controls and statement fixed effects. The theory-guided family contains two randomized treatment tests: H1a and H2a. H1 and H2 describe baseline control-group patterns and belong to a separate descriptive analysis, so they are not included in this treatment-effect correction family. The other two treatment-by-alignment quantities are secondary diagnostics. The final two columns provide a conservative adjustment across all four displayed quantities.

A3.3. Non-partisan-statement placebo tests

Figure A-10 restricts the stimulus set to statements with no partisan slant and reports the average positive-task-versus-control effect on each directional error. Within this restricted sample, non-partisan statements are coded Congruent for moderate respondents and Intermediate for liberal and conservative respondents, so the plotted coefficient averages over both alignment categories. The figure is designed to test a content-independent response-style explanation–for example, a general tendency to click "fact" less often–not to test whether moderate and ideologically sided respondents react differently.

Across the three specifications, the opinion-as-fact estimate ranges from 0.07 to 0.43 percentage points, with the smallest p-value equal to 0.721. The fact-as-opinion estimate ranges from -0.28 to -0.19 percentage points, with the smallest p-value equal to 0.798. These near-zero estimates support the interpretation that the main result is not a general shift in willingness to classify statements as facts or opinions.

Because the restricted sample pools Congruent moderate pairs and Intermediate liberal/conservative pairs, the figure should not be interpreted as a direct comparison of those respondent groups.

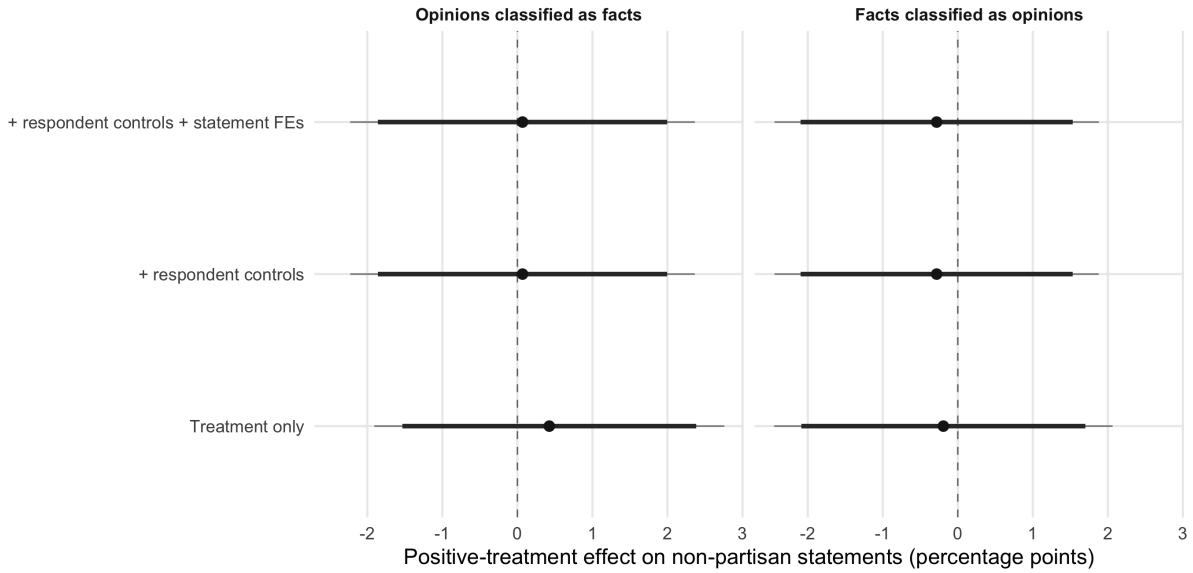

Figure A-10: Positive-treatment effects on non-partisan statements. Points report the coefficient on assignment to the positive self-worth task from models restricted to statements with no partisan slant. The left panel shows opinions classified as facts and the right panel shows facts classified as opinions. Models progress from treatment only to respondent controls and respondent controls plus statement fixed effects; standard errors are clustered by respondent. Thick intervals show 90 percent confidence intervals and thin intervals show 95 percent confidence intervals. Because non-partisan statements are coded as Congruent for moderate respondents and Intermediate for liberal and conservative respondents, the coefficients average over both alignment categories. The figure is a placebo test for a content-independent response shift, not a comparison between moderate and ideologically sided respondents.

A3.4. Baseline classification-error estimates

Appendix Table A-4 is a numerical key to Figure 2. Its entries are adjusted differences in misclassification probability relative to the omitted category, not raw cell means. Rows labeled "all statements" pool fact and opinion items; rows labeled "fact statements" or "opinion statements" use the corresponding subsample. Because the table combines estimates from several related models, the rows should not be read as mutually exclusive pieces of a single regression. Appendix Table A-5 reports control-group misclassification rates by media and statement type. Together, these tables show that the baseline error pattern is structured by statement content and alignment rather than by a simple story in which social-media items are always more opinion-like.

Table A-4: Key estimates for baseline classification-error patterns

Quantity	**Estimate (pp)**	**Std. error (pp)**	p**-value**
Panel A. Statement type and respondent–statement alignment			
Opinion statement vs fact statement, all statements	-3.5	0.5	< 0.001
Congruent vs intermediate alignment, opinion statements	1.1	0.5	0.050
Opposing vs intermediate alignment, opinion statements	-4.7	0.6	< 0.001
Congruent vs intermediate alignment, fact statements	-2.2	0.6	< 0.001
Opposing vs intermediate alignment, fact statements	-2.8	0.8	< 0.001
Congruent vs opposing alignment, fact statements	0.6	0.8	0.436
Congruent vs intermediate alignment, all statements	-0.1	0.4	0.743
Opposing vs intermediate alignment, all statements	-3.5	0.4	< 0.001
Panel B. Source attribution			
Social media vs newspaper source, opinion statements	-6.5	0.5	< 0.001
Social media vs newspaper source, fact statements	6.8	0.5	< 0.001
Social media vs newspaper source, all statements	-0.0	0.3	0.992
Panel C. Statement topic			
Politics vs economics topic, all statements	-2.3	0.3	< 0.001
Politics vs economics topic, fact statements	1.5	0.6	0.016
Politics vs economics topic, opinion statements	-5.9	0.4	< 0.001

Note: Each observation is a respondent–statement classification. Statement type, source, and topic are statement attributes; Congruent, Opposing, and Intermediate are respondent–statement alignment categories. Each row reports an adjusted coefficient from the respondent-controls specification corresponding to the sample named in the row; all statements pools fact and opinion items, while fact statements and opinion statements use the corresponding subsample. Positive values indicate a higher probability of misclassification relative to the stated omitted category; negative values indicate a lower probability. All estimates and standard errors are in percentage points, and standard errors are clustered by respondent. The table is descriptive and is not a treatment-effect analysis.

Table A-5: Baseline misclassification rates by media and statement type

Statement type	Error type	Media	Item-level N	Mean error	Rate (pp)
Fact	Fact classified as opinion	Newspaper	4170	0.157	15.7
Fact	Fact classified as opinion	Social media	4170	0.205	20.5
Opinion	Opinion classified as fact	Newspaper	4170	0.174	17.4
Opinion	Opinion classified as fact	Social media	4170	0.108	10.8

Notes: The sample is restricted to the neutral-control group. Opinion-as-fact errors are calculated among opinion statements; fact-as-opinion errors are calculated among fact statements.

A3.5. Heterogeneity by pre-treatment self-esteem

Appendix Figure A-11 shows no monotonic relationship between pre-treatment self-esteem and the focal treatment effect; the joint test of equality across quartiles yields $p = 0.820$. We therefore do not interpret baseline self-esteem as identifying the respondents most responsive to the positive-writing task. This is consistent with the distinction between a stable self-evaluation measure and a temporary writing task: the experiment tests whether a short positive self-worth intervention changes classification at the moment of judgment, not whether people with lower baseline self-esteem are always more vulnerable to motivated classification.

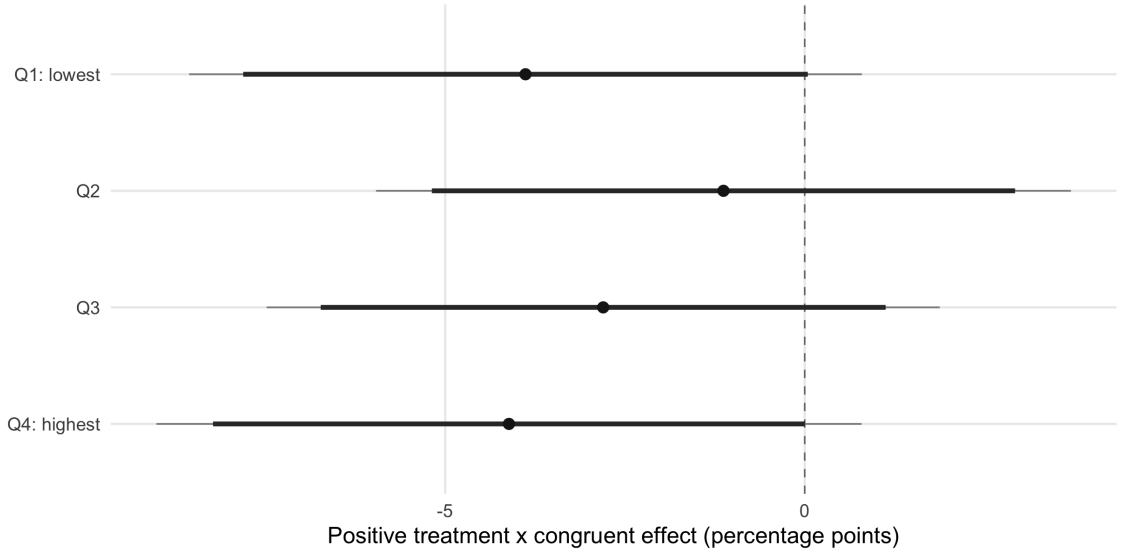

Figure A-11: Exploratory heterogeneity by pre-treatment self-esteem. Points report the Treatment × Congruent effect on opinion-as-fact errors within quartiles of the pre-treatment self-esteem index, recovered as linear combinations from one pooled model with respondent controls other than the continuous self-esteem score and statement fixed effects. The joint Wald test evaluates whether the four effects differ. Q1 is the lowest and Q4 the highest self-esteem quartile. The estimates do not display a monotonic pattern, so baseline self-esteem does not identify a clear gradient in responsiveness to the temporary positive-writing task. Thick intervals show 90 percent confidence intervals and thin intervals show 95 percent confidence intervals.

A4. Survey Instrument and Statement Materials

A4.1. Treatment materials

1. **Control Group**

   - Imagine you're writing a post about a recent book you've read, movie you've watched, or place you've visited. Describe it and share your thoughts in a concise way. Please do not exceed 280 characters.
   - What do you typically eat for breakfast? Describe your usual morning meal using as much detail as you'd like, but don't exceed 280 character limit.

2. **Positive Treatment (Male/Female/Non-Binary)**

   - Imagine you're crafting a post about a recent personal success or a valuable lesson you've learned. It could be something big or small, but it made you feel good or grow in some way. Share your story keeping within the 280 character limit.
   - Imagine you're using Twitter to express gratitude for something positive in your life. It can be a small joy, a major win, or a personal strength you appreciate. How would you word that tweet, given the 280 character limit?

3. **Negative Treatment (Male/Female/Non-Binary)**

   - Instructions: In this activity, you will review 3 Instagram posts. Alongside each post, you will see an average 'success rating' that each individual has received from other participants. Success, for the context of this exercise, is defined by personal achievement, goal fulfilment, and overall life satisfaction. It considers the accomplishment of meaningful aims or purposes, the attainment of popularity or profit, or the sense of fulfilment and satisfaction derived from these accomplishments. Please take your time to carefully read each post.
   - Your Task: After considering the posts and their ratings, you'll have the opportunity to either agree or disagree with each person's success rating. Then, you'll be asked to

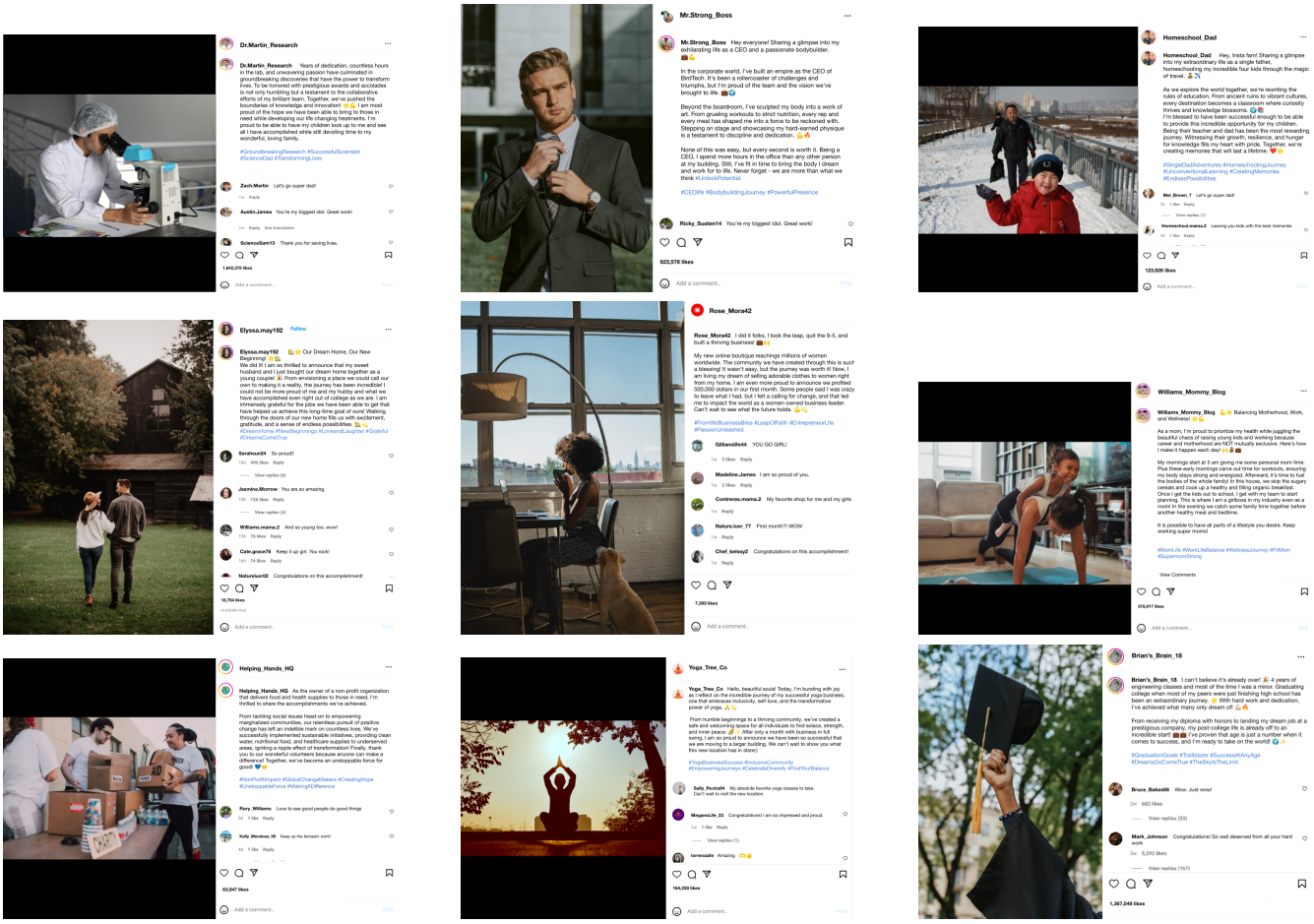

Figure A-12: Instagram Posts

rank the individuals behind these posts and yourself in order of perceived success. Remember, there's no right or wrong answer, we're interested in your unique perspective

- The 3 Instagram posts for each gender the participants identify themselves with are given in Figure A-12. First row is the negative treatment for Male group, second row for Female, and third row for Non-Binary.
- Finally, participants are asked to rank themselves with the three hypothetically successful figures: "Take a moment to reflect on your life and achievements to date. How would you rate yourself? Now, consider the individuals whose posts you've just reviewed. Where would you rank yourself in comparison to them?"

Table A-6: Statements used in the classification task, with type, topic, source, and partisan slant

No.	Type	Topic	Source	Slant	Statement
1	Fact	Politics	WSJ	Pro-R	North Carolina selects its Supreme Court justices by partisan elections, and in November Republicans took two of the seven seats from the Democrats, giving the court a 5-2 GOP tilt. The North Carolina justices agreed to reconsider the 2022 decision, and by their new 5-2 split announced in April that courts had no power to review congressional maps drawn by the legislature.
2	Opinion	Politics	WSJ	Pro-R	But for his president father [Joe Biden], running for re-election, having his son come across like Vito Genovese taking the Fifth dozens of times before questions about shell companies and payments to himself and his dad isn't a great look.
3	Fact	Economics	WSJ	NP	When someone claims a payment is a loan for tax purposes, the IRS looks for three things: You've "got to have a promissory note, you got to have defined interest, and you got to have repayments."
4	Opinion	Economics	WSJ	Pro-R	The only way to get gas prices sustainably as low as they can go is to bring more supply into the market.

No.	Type	Topic	Source	Slant	Statement
5	Fact	Health	WSJ	NP	Medical groups have lowered the recommended starting age for colorectal cancer screenings to 45 from 50, in response to rising colon cancer rates among younger people. But only about 20% of 45- to 49-year-olds are screened for one of the deadliest cancers, according to American Cancer Society data.
6	Opinion	Health	WSJ	NP	The most important thing is to have a relationship with a doctor that you're seeing with some regularity.
7	Fact	Politics	NYT	Pro-D	Mr. Weiss disclosed an agreement on Tuesday under which Mr. [Hunter] Biden would plead guilty to two misdemeanor tax charges and avoid prosecution on an unrelated gun charge by agreeing to a pretrial diversion program.
8	Opinion	Politics	NYT	Pro-D	Since the rise of Donald Trump, the Republican definition of a crime has veered sharply from the law books and become extremely selective.
9	Fact	Economics	NYT	NP	Published salary ranges are getting wider for occupations in pharmacy, medical information, scientific research and development, and other industries. Pay estimates for jobs in food preparation and child care are growing more precise.

No.	Type	Topic	Source	Slant	Statement
10	Opinion	Economics	NYT	NP	Globalization, seen in recent decades as unstoppable a force as gravity, is clearly evolving in unpredictable ways. (The move away from an integrated world economy is accelerating. And the best way to respond is a subject of fierce debate.)
11	Fact	Health	NYT	NP	Those who lived in areas with high levels of transportation noise were more likely to have highly activated amygdalas, arterial inflammation and – within five years – major cardiac events.
12	Opinion	Health	NYT	NP	Our country needs an institute devoted to research on the top cause of poor health.
13	Fact	Politics	Twitter/X	Pro-D	75% of Floridians across the political spectrum do not support Florida's extreme abortion ban.
14	Opinion	Politics	Twitter/X	Pro-R	Planned Parenthood isn't about planning or preparation.
15	Fact	Economics	Twitter/X	NP	Central bankers have lifted their official policy rate to about 5 percent over the past 15 months, which has translated into higher borrowing costs across the economy.
16	Opinion	Economics	Twitter/X	Pro-R	The policy that was needed following the pandemic was supply side economics!
17	Fact	Health	Twitter/X	NP	The amount of exercise you should do each week depends on several factors, including your age, overall health, fitness goals, and any underlying medical conditions.

No.	Type	Topic	Source	Slant	Statement
18	Opinion	Health	Twitter/X	NP	Walking more is one of the most underrated health hacks out there.
19	Fact	Politics	Reddit	Pro-D	On Thursday, former US president and current frontrunner for the 2024 Republican presidential nomination Donald Trump posted to his social media platform that he had been informed by federal prosecutors that he is the target of an ongoing investigation. The probe stems from potential mishandling of classified documents allegedly taken from the White House. Trump has denied all wrongdoing.
20	Opinion	Politics	Reddit	Pro-D	Donald Trump's New Criminal Case Looks Devastating.
21	Fact	Economics	Reddit	Pro-D	United States Crude Oil Production nears record high, and has already surpassed the level it was at in mid-2019.
22	Opinion	Economics	Reddit	Pro-R	Too many readers who are ignorant or lack knowledge blame Social Security and Medicare for draining the federal budget and taking a large amount out of it.
23	Fact	Health	Reddit	NP	Malaria cases in Texas and Florida are the first US spread since 2003, CDC says.
24	Opinion	Health	Reddit	NP	You don't need to drink anything that has calories and water is the choice of healthy people worldwide.

No.	Type	Topic	Source	Slant	Statement

Note: WSJ = *The Wall Street Journal*; NYT = *The New York Times*. Slant refers to the intended partisan congeniality of the statement: **Pro-D** = pro-Democratic, **Pro-R** = pro-Republican, and **NP** = non-partisan (no partisan slant). Slant is a statement-level property and does not depend on the respondent's ideology. In the analysis, we classify respondent-statement pairs as *Congruent*, *Opposing*, or *Intermediate* (neither congruent nor opposing); the Intermediate category applies at the pair level and is separate from the statement slant shown here.

A5. PREREGISTRAGION

This study was pre-registered on AsPredicted (https://aspredicted.org/147358). The design, primary treatment (positive and negative self-worth tasks and neutral control) and the binary fact/opinion classification outcome follow the pre-registration. The total sample size ($N = 2,337$) is slightly lower than $N = 2400$ in the pre-registration.

Our primary analysis uses OLS/linear-probability regressions organized around the respondentâstatement congruence framework (Congruent, Opposing, Intermediate) and the formal model, which is a richer estimation strategy than the marginal-means comparisons named in the pre-registration. The pre-registration named topic, media, political slant, and self-esteem as the factors of interest. Hypotheses H1, H2, H1a, and H2a operationalize the pre-registration's prediction that political slant, particularly in interaction with self-worth, drives directional classification error, though the respondentâstatement congruence framework (Congruent/ Opposing/ Intermediate) and the interaction-regression specifications that structure these tests were developed after registration. H3 states the pre-registered media prediction in directional form; H4 elaborates the pre-registered topic prediction, generalizing it to overall misattribution.