

Understanding Beliefs in Misinformation: Repetition, Partisan Signals and Bayesian Processing

Supporting Materials (SM)

Contents

1	Samples Baseline Characteristics	1
2	Survey Instrument: Headlines	7
3	Measuring Prior Familiarity	11
4	Partisan Heterogeneity Across Four Studies	13
5	Formalizing the Bayesian Framework	15
5.1	The Model	15
5.2	Partisan Motivated Reasoning in the Model	16
5.3	Illusory Truth Effects in the Model	17
5.4	Simulating the Interaction between PMR and ITE	17
5.5	Closed-Form Results for n Repeated Exposures	21
5.6	Allowing Repetition to Increase Signal Credibility	23
6	Regression Tables	25
7	Meta Analytical Results	30
8	Robustness Checks: Four Point Continuous Outcome	32
9	Deviations from the Pre-Analysis Plan	34
10	Research Ethics	35

1 Samples Baseline Characteristics

This section presents demographic characteristics of our samples. Table 1 presents baseline demographics for all participants across the four studies. We see some notable differences across the samples. Mainly, comparing the samples from studies 1 and 2 (recruited via Qualtrics) and studies 3 and 4 (recruited via Connect-CloudResearch), the former are more balanced on age and education, while the latter have a higher share of more educated and younger respondents. Furthermore, the CloudResearch-Connect samples also exhibit a higher presence of Democrats compared to the Qualtrics samples. While these differences are considerable, we do not consider them strong evidence for explaining the results across the four studies, particularly the weak ITE effects observed in studies 3 and 4. First, we do not find that our ITE results differ by education, age, or partisan identification. Second, previous studies on ITE do not theorize or provide empirical evidence for these variables as necessary conditions for the manifestation of ITE Dechêne et al. (2010).

In addition, we provide tables comparing demographics between the two treatment arms and the control group in studies 1 and 2, and between the two treatment arms in study 4. Given that Study 3 consists of a within-subject participant, with no between-participant randomization, we do not provide a similar comparison table for these studies and present only the demographic composition of the samples. We show that treatment and control groups are largely balanced across demographic covariates. Omnibus f-tests comparing treatment and control cannot reject the null of joint orthogonality between these samples.

Lastly, we provide an attrition bias test across the treatment groups, measuring threats of differential attrition between the waves of study 1 and study 2 developed in (Ghanem, Hirshleifer and Ortiz-Becerra, 2023). Table 5 presents the results. We find no evidence of differential attrition in the sample. Columns 3-6 present the mean baseline outcome for treatment respondents (TR), control respondents (CR), treatment attriters (TA), and control attriters (CA), respectively. Column 7 reports the p-value of the hypothesis test with two equality restrictions (*selective attrition*). We cannot reject the joint null hypothesis of no differences in baseline values between treatment and control respondents and between treatment and control attriters for most variables; gender is the only exception, failing the internal validity test.

Table 1: Baseline Demographics Across the Four Studies,

		Study 1 (N=1971)		Study 2 (N=1289)		Study 3 (N=1004)		Study 4 (N=1636)	
		N	Pct.	N	Pct.	N	Pct.	N	Pct.
Age	18 - 24	145	7.4	88	6.8	85	8.5	150	9.2
	25 - 34	406	20.6	247	19.2	254	25.3	408	24.9
	35 - 44	390	19.8	273	21.2	170	16.9	403	24.6
	45 - 54	264	13.4	182	14.1	199	19.8	233	14.2
	55 - 64	544	27.6	354	27.5	179	17.8	233	14.2
	65+	222	11.3	145	11.2	117	11.7	209	12.8
Gender	Female	1227	62.3	777	60.3	502	50.0	989	60.5
	Male	730	37.0	507	39.3	499	49.7	640	39.1
	Other	14	0.7	5	0.4	3	0.3	7	0.4
Race	White	1340	68.0	873	67.7	752	74.9	1156	70.7
	Non-White	631	32.0	416	32.3	252	25.1	480	29.3
Education	Less than high school	68	3.5	42	3.3	7	0.7	7	0.4
	High school graduate	553	28.1	369	28.6	101	10.1	207	12.7
	Some college	459	23.3	303	23.5	177	17.6	355	21.7
	2 year degree	260	13.2	162	12.6	120	12.0	196	12.0
	4 year degree	428	21.7	285	22.1	450	44.8	613	37.5
	Professional degree	181	9.2	115	8.9	125	12.5	211	12.9
	Doctorate	22	1.1	13	1.0	24	2.4	47	2.9
Region	Midwest	433	22.0	288	22.3	192	19.1	315	19.3
	Northeast	350	17.8	227	17.6	208	20.7	313	19.1
	South	767	38.9	490	38.0	400	39.8	654	40.0
	West	421	21.4	284	22.0	204	20.3	354	21.6
Party Id (+ Leaners)	Republican	900	45.7	589	45.7	331	33.0	615	37.6
	Democrat	838	42.5	562	43.6	579	57.7	861	52.6

Study 1: Qualtrics Sample; Study 2: Qualtrics Sample; Study 3: CloudResearch Sample; Study 4: CloudResearch Sample

Table 2: Study 1 - Balance Table for Treatment Arms and Control

Variables		Control (N=639)		T1 - Prior Exposure (N=673)		T2 - Warning Labels (N=659)	
		N	Pct.	N	Pct.	N	Pct.
Age	18 - 24	39	6.1	47	7.0	59	9.0
	25 - 34	133	20.8	135	20.1	138	20.9
	35 - 44	123	19.2	135	20.1	132	20.0
	45 - 54	89	13.9	92	13.7	83	12.6
	55 - 64	179	28.0	195	29.0	170	25.8
	65+	76	11.9	69	10.3	77	11.7
Gender	Female	421	65.9	389	57.8	417	63.3
	Male	216	33.8	278	41.3	236	35.8
	Other	2	0.3	6	0.9	6	0.9
Race	Non-White	202	31.6	201	29.9	228	34.6
	White	437	68.4	472	70.1	431	65.4
Education	Less than high school	22	3.4	22	3.3	24	3.6
	High school graduate	181	28.3	185	27.5	187	28.4
	Some college	152	23.8	164	24.4	143	21.7
	2 year degree	92	14.4	91	13.5	77	11.7
	4 year degree	131	20.5	139	20.7	158	24.0
	Professional degree	52	8.1	65	9.7	64	9.7
	Doctorate	9	1.4	7	1.0	6	0.9
Region	Midwest	150	23.5	129	19.2	154	23.4
	Northeast	97	15.2	135	20.1	118	17.9
	South	246	38.5	268	39.8	253	38.4
	West	146	22.8	141	21.0	134	20.3
Partisanship	Republican	302	47.3	314	46.7	284	43.1
	Democrat	258	40.4	289	42.9	291	44.2
Omnibus Test: F-Statistics = 1.31		p-value	= 0.17				

Table 3: Study 2 - Balance Table for Treatment Arms and Control

Baseline Variables		Control		T1 - Prior Exposure		T2 - Warning Labels	
		N	Pct.	N	Pct.	N	Pct.
Age	18 - 24	25	5.9	29	6.5	34	8.1
	25 - 34	84	19.8	82	18.5	81	19.2
	35 - 44	88	20.8	96	21.6	89	21.1
	45 - 54	59	13.9	66	14.9	57	13.5
	55 - 64	119	28.1	131	29.5	104	24.7
Gender	65+	49	11.6	40	9.0	56	13.3
	Female	276	65.1	251	56.5	250	59.4
	Male	147	34.7	191	43.0	169	40.1
Race	Other	1	0.2	2	0.5	2	0.5
	Non-White	143	33.7	131	29.5	142	33.7
	White	281	66.3	313	70.5	279	66.3
Education	Less than high school	13	3.1	14	3.2	15	3.6
	High school graduate	118	27.8	130	29.3	121	28.7
	Some college	108	25.5	102	23.0	93	22.1
	2 year degree	58	13.7	59	13.3	45	10.7
	4 year degree	93	21.9	89	20.0	103	24.5
	Professional degree	29	6.8	46	10.4	40	9.5
	Doctorate	5	1.2	4	0.9	4	1.0
Region	Midwest	96	22.6	91	20.5	101	24.0
	Northeast	64	15.1	88	19.8	75	17.8
	South	167	39.4	169	38.1	154	36.6
	West	97	22.9	96	21.6	91	21.6
Partisanship	Republican	198	46.7	206	46.4	185	43.9
	Democrat	180	42.5	196	44.1	186	44.2
Omnibus Test: F-Statistics = 0.765		p-value	= 0.763				

Table 4: Study 3 - Balance Table for Treatment Arms

Baseline Variables		T1 - False Headlines(N=815)		T2- Fake Headlines	
		N	Pct.	N	Pct.
Age	18 - 24	72	8.8	78	9.5
	25 - 34	212	26.0	196	23.9
	35 - 44	207	25.4	196	23.9
	45 - 54	103	12.6	130	15.8
	55 - 64	116	14.2	117	14.3
	65+	105	12.9	104	12.7
Gender	Female	496	60.9	493	60.0
	Male	314	38.5	326	39.7
	other	5	0.6	2	0.2
Race	Non-White	235	28.8	245	29.8
	White	580	71.2	576	70.2
Education	Less than high school	3	0.4	4	0.5
	High school graduate	110	13.5	97	11.8
	Some college	180	22.1	175	21.3
	2 year degree	98	12.0	98	11.9
	4 year degree	300	36.8	313	38.1
	Professional degree	110	13.5	101	12.3
	Doctorate	14	1.7	33	4.0
Region	Midwest	148	18.2	167	20.3
	Northeast	166	20.4	147	17.9
	South	335	41.1	319	38.9
	West	166	20.4	188	22.9
Partisanship	Republican	295	36.2	320	39.0
	Democrat	435	53.4	426	51.9
Omnibus Test:	F-Statistics = 1.12, p-values = 0.32				

Table 5: Studies 1 and 2: Sample Internal Validity Test for Differential Attrition

Outcome	Attrition Bias		Mean Baseline Outcome by Group				Test of Internal Validity p-value
	C	Differential	TR	CR	TA	CA	
Age	0.664	0.014 (p-value = 0.537)	3.68	3.73	3.57	3.72	0.44
Gender	0.664	0.014 (p-value = 0.537)	1.43	1.35	1.36	1.33	0.03
Race	0.664	0.014 (p-value = 0.537)	1.68	1.66	1.67	1.73	0.22
Education	0.664	0.014 (p-value = 0.537)	3.54	3.49	3.57	3.53	0.77
Region	0.664	0.014 (p-value = 0.537)	2.58	2.63	2.61	2.57	0.75
Partisanship	0.664	0.014 (p-value = 0.537)	1.49	1.48	1.49	1.43	0.34

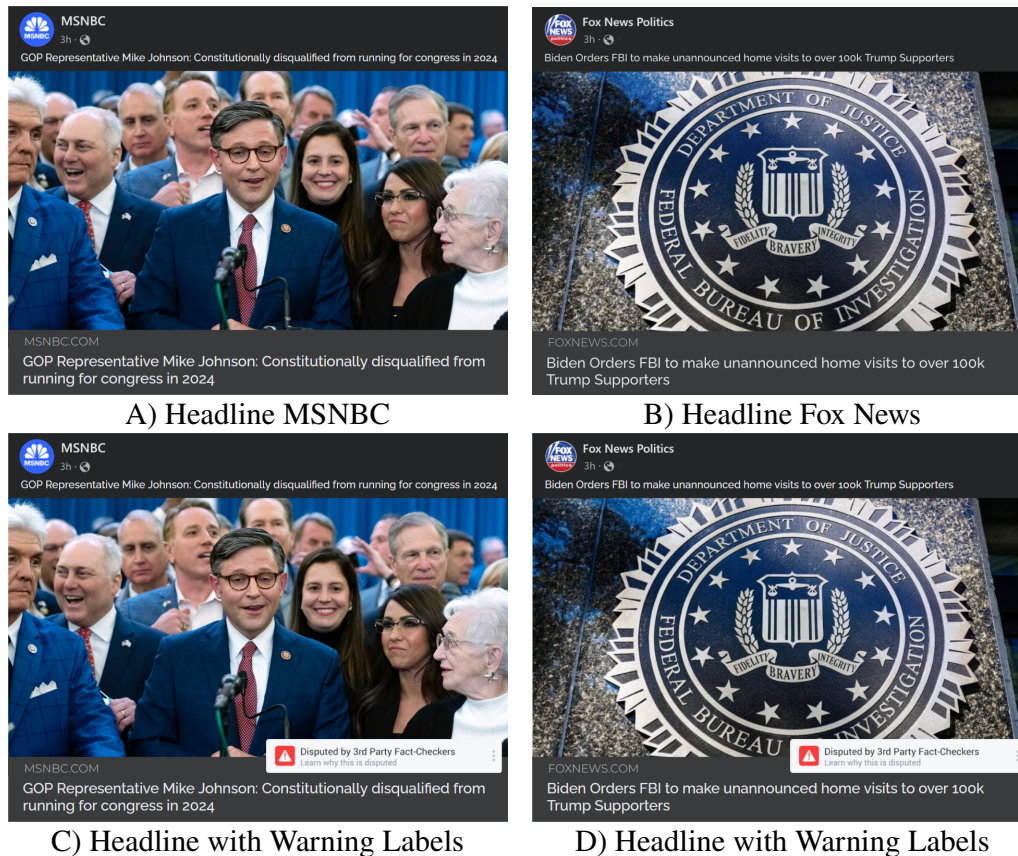
Note:

Column 1 reports the attrition rate for control, and Column 2 reports the differential attrition rate between treatment and control, with the corresponding p-value testing for difference in attrition between the groups (*differential attrition*). Columns 3-6 present the mean baseline outcome for treatment respondents (TR), control respondents (CR), treatment attriters (TA), and control attriters (CA), respectively. Column 7 reports the p-value of the hypothesis test with two equality restrictions (*selective attrition*).

2 Survey Instrument: Headlines

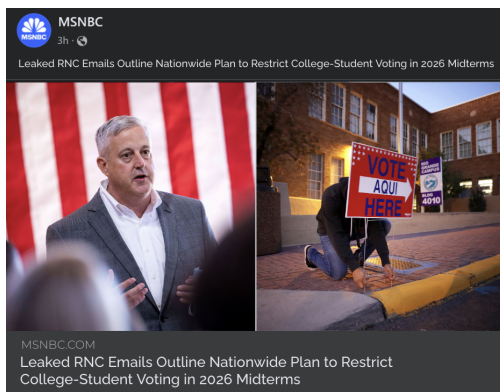
In this section, we present visual examples of the headlines presented to participants during the familiarity and accuracy stages. Figure 2 presents two visual examples of the wholly fabricated headlines (Fake Treatment) included in Study 4. Additionally, we present the full wording of all the headlines used in the three studies in table ??, separating them according to their veracity and political alignment.

Figure 1: Headlines Visual Presentation

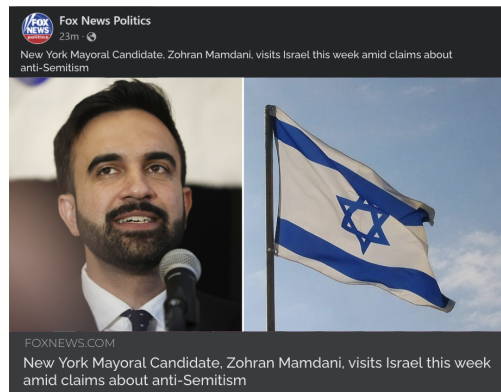


Note: The upper row presents examples of the headlines from different sources shown to participants. The bottom row presents examples of the headlines with and without warning labels.

Figure 2: Study 4: Examples of the wholly fabricated headlines



A) Fake Headline MSNBC



B) Fake Headline Fox News

Table 6: Headlines Across the Four Studies

Studies 1 & 2		
Direction	True Headlines	False Headlines
Pro-Democrats	Biden, aiming at Trump, says he won't use military as a "prop"	GOP Representative Mike Johnson: Constitutionally disqualified from running for congress in 2024
	Biden Admin pledges to phase out coal power plants at COP28	Trump's Older Sister Worries About Him: 'Donnie's Acting Like A Nutjob'
	Liz Cheney is working to make sure Trump never returns to the White House	Mitch McConnell's doctor declares him "unfit" for office
	Team Trump is "utterly clueless" in NY Appeals Court	Unsealed court documents show that Trump owned stock in a company the CDC used for Covid-19 testing
	Meta removes Trump ads displaying symbol used by Nazis	Michigan GOP tries to pass bill marking immigrants' licenses with yellow stars
	The Pentagon has 3 years to strip Confederate names from bases	Biden: "Republicans will cut Social Security benefits by 13% on average for Americans"
Pro-Republicans	AZ border experiencing "increased levels" of illegal migrant entries	Pro Lifers warn blue states are waging a "two pronged" attack to silence pregnancy resource centers
	Hunter Biden paid Joe Biden for Account for biz that received payments from China	
	Two Year Study of Gun Owners Exposes Dem's Gun Control Narrative as Nonsense	Biden Admin considers a federal ban on gas stoves over climate change concerns
	Pro Lifers warn blue states are waging a "two pronged" attacked to silence pregnancy resource centers	Starbucks employee union supports Hamas
	Democrat running for Congress indicted in \$803,000 fraud, embezzlement case	The Obama Foundation is being investigated for racketeering
	Deadline Looms: Dem-controlled Senate passes only 3 of 12 spending bills	10 NFL Players were disqualified in a single week for kneeling during the National Anthem
Study 3		
Pro-Democrats	"Historic" levels of cash pours into the Harris campaign	Trump needs help leaving rally stage, seen holding son's hand as he goes down steps
	JD Vance defends rumors about Haitian immigrants eating pets	Trump has publicly announced his support for Project 2025
Pro-Republicans	Proposed GOP legislation would strip immigrant protesters of visas if they participate in university demonstrations	Study finds that 27% of noncitizens in the United States are registered to vote
	Harris previously said police could pay surprise visits to legal gun owners' homes for safe storage checks	Taylor Swift has lost \$125M in brand deals since endorsing Harris
Non-Political	"Shogun" shatters Emmy record with 18 wins in one season	Schools don't teach cursive anymore because AI technology can't read it
	National Women's Soccer League reaches historic collective bargaining agreement	Kate Middleton won't return to royal duties for many years, kept out of the public eye

Study 4			
Direction	True Headlines	False Headlines	Wholly Fake Headlines
Pro-Democrats	USAID staff saw 'no rhyme or reason' in which programs were saved	FIFA plans to move World Cup outside the United States due to concerns over Trumps immigration policies.	Leaked RNC Emails Outline Nationwide Plan to Restrict College-Student Voting in 2026 Midterms
	Sen. Mike Lee wants to sell 2 million acres of public land to benefit the wealthy	73 world leaders have cancelled US trips over 'safety and dignity' concerns	Republican officials are implicated in participating in Sean Combs sex trafficking parties
	FEMA says it's giving control to the states. They aren't sure what that means.	Trump's ICE is deputizing bounty hunters to arrest undocumented immigrants	Major energy failure plunges New York City into blackout during summer heatwave last week as climate emergency intensifies
	Trump stumbles as he walks up the steps to Air Force One	DoorDash shares driver information with Trump administration immigration officials	Top Trump officials seen hunting near-extinct amur leopards according to leaked images
Pro-Republicans	Top Trump health official slams Democrats for 'misleading' claims about Medicaid reform	New evidence asserts that coronavirus originated as a lab leak is the 'confirmable truth' in line with Trump's claims	DNC-Commissioned Study Labels Traditional Two-Parent Families "Detrimental" to Child Development
	NATO leader praises Trumps 'decisive action on Iran	Gas prices have fallen to \$1.98 a gallon in multiple states in Trump's first six months in office	Leaked Memo: Democratic Leadership Pushing 70% Wealth Tax Retroactive to 2010
	Rep Nancy Mace says Nike should face "consequences" for allegedly funding trans youth athlete study	Biden administration 'shoveled' \$93 billion in loans out of Energy Department in final months of 2024	New York Mayoral Candidate, Zohran Mamdani, visits Israel this week amid claims about anti-semitism
	Protestors in LA spit on and burn American flag	While in office, Obama received over \$22 million in donations from Venezuela for his presidential library	Trump's FTC revokes all regulations on big tech, saying "let the market flourish"

3 Measuring Prior Familiarity

In this section, we provide descriptive evidence about the levels of prior familiarity with the headlines used across studies 1, 3, and 4. We explore prior familiarity with a given headline, measured at the familiarization stage across the studies. Study 2 is not included in the analysis because this follow-up does not have a familiarization stage, as participants were exposed to headlines the day before in the study 1.

To remind the reader, hypotheses 3 of the main manuscript relates to the role of the prior in belief formation. Our core argument states that loose priors (lower familiarity) increases the effects of ITE and PMR. We expect this to manifest in three settings: True vs. False headlines, Political vs Non-Political headlines, and Wholly Fabricated vs False headlines. For these comparisons, we expect effects to be weaker for True, Non-Political, and False headlines because familiarity would be higher in these categories.

Figure 3 presents the results. Overall, the assumption of higher familiarity for True, Non-Political, and False Headlines (vs. Wholly Fabricated) holds. For example, in study 1, participants reported being familiar with 35.6% of the true headlines, compared to 25.3% of the false. Considering only the political vs non-political distinction, our assumption of more diffuse priors for non-political also holds for both true and false cases. Lastly, we show the lowest levels of familiarity, as expected, with the headlines we fully fabricated in study 4.

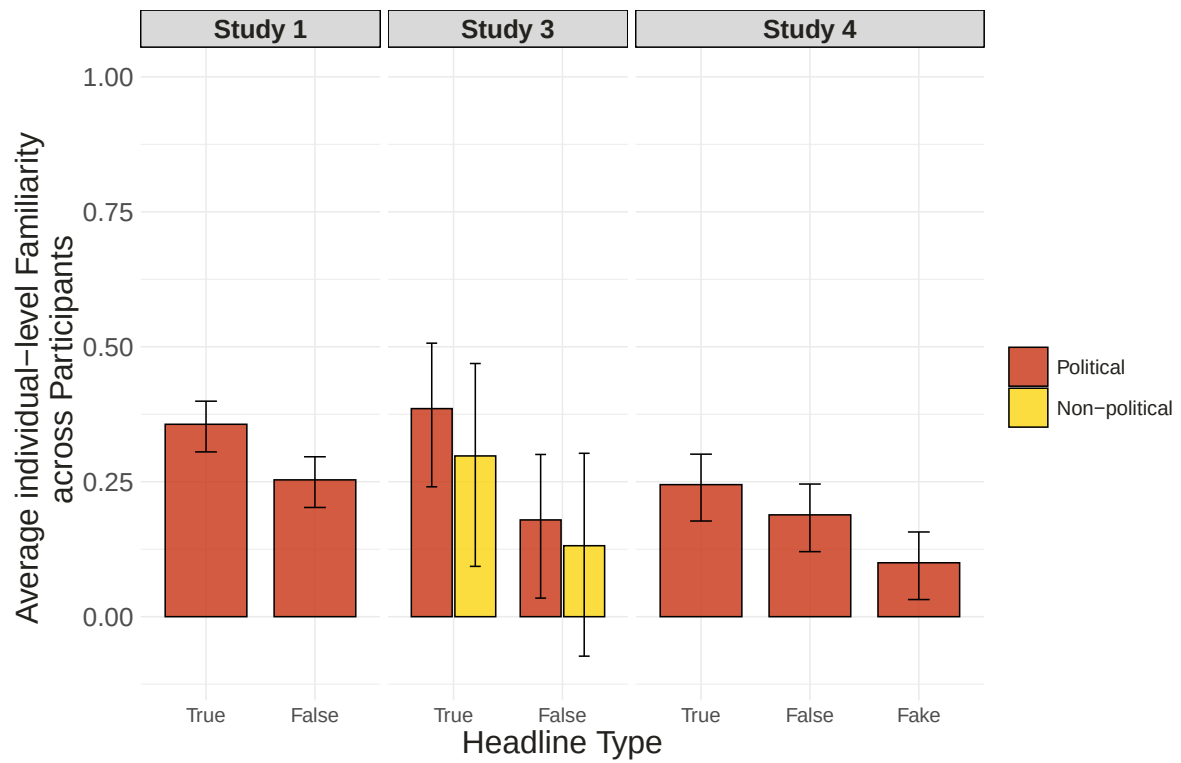
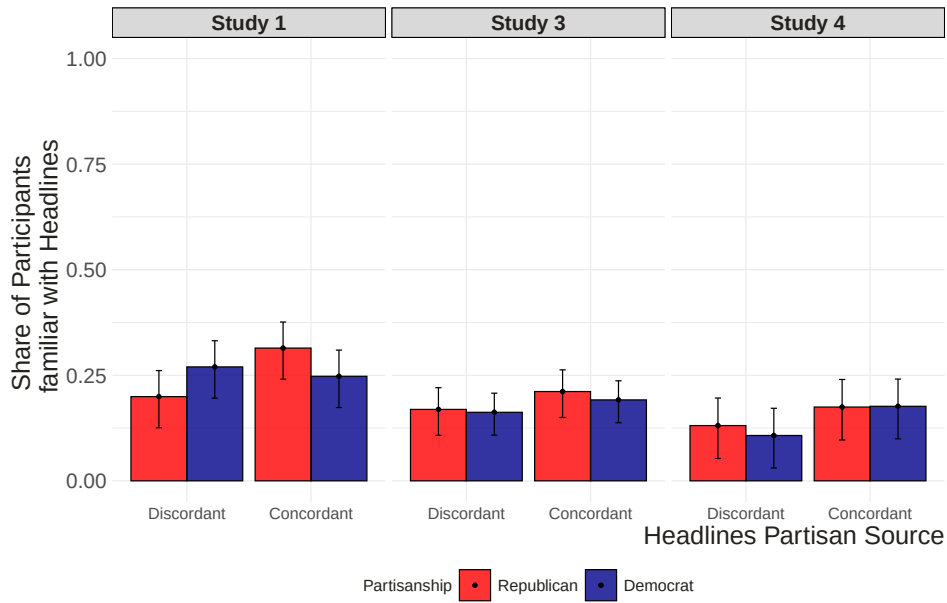


Figure 3: Marginal Means indicating prior familiarity with headlines indicated in the first stage of the experiment across studies 1, 3 and 4. Study 2 does not have a pure familiarity stage.

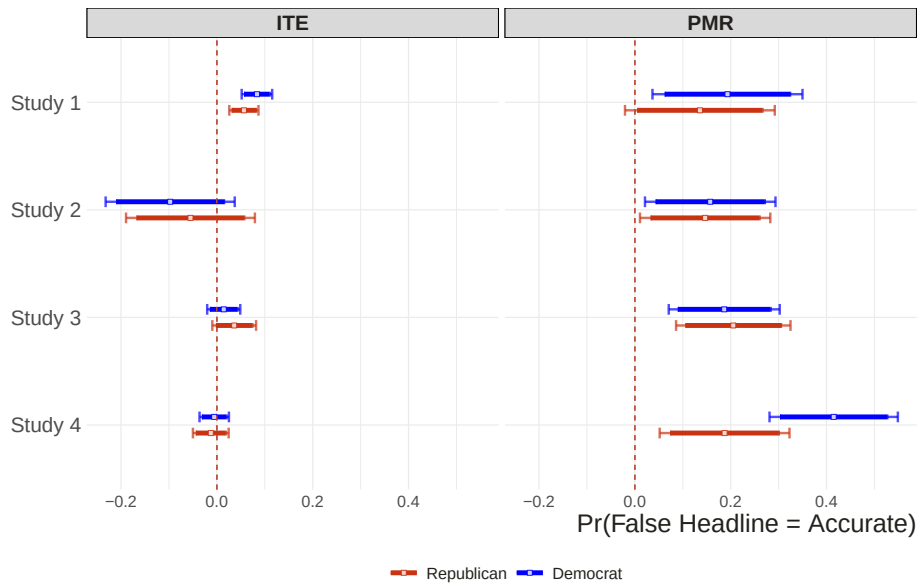
4 Partisan Heterogeneity Across Four Studies

Our experiment was conducted in the United States, where mis- and disinformation is unequally distributed across the parties (Guess, Nagler and Tucker, 2019; Grinberg et al., 2019). The Republican party's de facto leader at the time of our study was Donald Trump, whose proclivity for falsehoods is well-documented. In this section, we therefore explore partisan heterogeneity in the main findings reported in the paper.

While an exhaustive summary is beyond the scope of this paper, we present three revealing patterns in our data in this appendix, shown in 4. First, descriptively, on average Republican respondents are more familiar than Democrats with concordant false headlines, consistent with there being more right-leaning misinformation in this environment. The gap is pronounced in Study 1 (fielded early in 2024) and attenuates in Studies 3 and 4, where the partisan difference is small and the confidence intervals overlap. Second, consistent with theories of selective exposure (Stroud, 2011), partisans from both groups are consistently more familiar with concordant headlines compared to discordant ones. Lastly, we do not find that either ITE or PMR is stronger among Republicans than among Democrats across the four studies presented in the main manuscript. This finding goes in line with studies on partisan asymmetries on information processing and persuasion (Guay and Johnston, 2022; Coppock, 2023)



(A) Familiarity



(B) Accuracy Beliefs

Figure 4: Conditional Effects of Familiarity and Accuracy Beliefs Conditional on Partisanship

5 Formalizing the Bayesian Framework

This section develops in full the Bayesian model of belief formation summarized in the main text. We first present the complete model and show how partisan-motivated reasoning (PMR) and the illusory truth effect (ITE) map onto its parameters. We then present simulated evidence for the dynamics underlying our pre-registered hypotheses, derive closed-form solutions for the effect of n repeated exposures to the same signal, and examine the robustness of these predictions to alternative functional forms for how repetition affects signal credibility.

5.1 The Model

Following the summary of the framework found in Druckman and McGrath (2019), we start from Bayes' Rule in which an individual's posterior belief (denoted $\pi_i(\mu|x)$) about the state of the world μ is a function of a prior (denoted $\pi_i(\mu)$) and a signal (denoted x). For the sake of simplicity, we assume that these beliefs and signals are all distributed according to a normal distribution with a mean denoted generically with μ and standard deviation denoted generically with σ^2 .

$$\textbf{Prior: } \pi(\mu) \sim \mathcal{N}(\hat{\mu}_{i,0}, \hat{\sigma}_{i,0}^2)$$

$$\textbf{Signal: } x \sim \mathcal{N}(\mu_x, \hat{\sigma}_{i,x}^2)$$

The hat symbol $\hat{\cdot}$ denotes beliefs, i indicates an individual, and 0 and x represent the prior and signal, respectively. Note that the subscripts are substantively meaningful here: $\hat{\mu}_{i,0}$ means that individual i 's prior belief is not necessarily the same as individual j 's prior belief – i.e., $\hat{\mu}_{i,0} \neq \hat{\mu}_{j,0}$ – and μ_x means that the signal needn't be centered on the true state of the world μ – i.e., $\mu_x \neq \mu$.¹ But most important for our substantive interest is the subscript on the precision parameter of the signal $\hat{\sigma}_{i,x}^2$. This precision parameter is the centerpiece of our intuition, and can be substantively interpreted as an inverted “credibility”: larger values indicate less credibility or, alternatively, more uncertainty about the source. Importantly, the $\hat{\cdot}$ notation makes explicit that this credibility is *subjective*. While there is a true state of the world μ that produces signals via a true data generating process σ^2 , the heart of cognitive bias rests on the assumption that an individual's *perception* of this distribution is what matters. This setup permits two different individuals to assign different credibility to the same signal, allowing – for example – a Republican and a Democrat to read

¹In our setting, we assume that $\mu_x = \mu$.

the same headline and adjust their beliefs differently.

How does the updating process work? Since both the signal and prior are assumed to be distributed normally, solving Bayes' Rule shows that the posterior belief can be expressed as:

$$\pi_i(\mu|x) \sim \mathcal{N}\left(\hat{\mu}_{i,0} + (\mu_x - \hat{\mu}_{i,0}) \left(\frac{\hat{\sigma}_{i,0}^2}{\hat{\sigma}_{i,0}^2 + \hat{\sigma}_{i,x}^2} \right), \frac{\hat{\sigma}_{i,0}^2 \hat{\sigma}_{i,x}^2}{\hat{\sigma}_{i,0}^2 + \hat{\sigma}_{i,x}^2}\right)$$

Substantively, the updated belief is centered on the prior belief $\hat{\mu}_{i,0}$ adjusted by the difference between the signal μ_x and the prior, weighted by the ratio of the (inverse) precision assigned to the prior $\hat{\sigma}_{i,0}^2$ relative to the net precision of the signal and the prior. The larger the $\hat{\sigma}_{i,0}^2$ term (i.e., the less confidently held is the prior belief) relative to the signal term, the more the signal influences the posterior belief.

5.2 Partisan Motivated Reasoning in the Model

In the partisan motivated reasoning framework, consider the same signal $x \sim \mathcal{N}(\mu_x, \hat{\sigma}_{i,x}^2)$ produced by a partisan source (i.e., a headline written by Fox News). A Democrat and a Republican will read this headline, but update differently based on their subjective precision parameter $\hat{\sigma}_{i,x}^2$. For the Democrat d , they assign very little credibility to headlines from Fox News compared to a Republican r , meaning $\hat{\sigma}_{d,x}^2 \gg \hat{\sigma}_{r,x}^2$. As such, assuming that both individuals' prior beliefs were held with the same certainty ($\hat{\sigma}_{d,0}^2 = \hat{\sigma}_{r,0}^2$), and centered on the same prior belief ($\hat{\mu}_{d,0} = \hat{\mu}_{r,0}$), the degree to which the Democrat's posterior belief is influenced by the headline is substantially smaller than the degree to which the Republican's posterior is.

As discussed in the main text, the explanations for why Democrats and Republicans assign different credibility to the same signal are varied, either explicitly couched in terms of a Bayesian model (Druckman and McGrath, 2019), or implicitly consistent with the Bayesian intuition (Holyoak and Morrison, 2012; Lodge, 2013; Kunda, 1990). The benefit of the Bayesian model for the purposes of this investigation is that any of these explanations can be mapped into the precision parameter $\hat{\sigma}_{i,x}^2$, which should decline along with the political discordance between the source of the signal and its recipient: in other words, more politically concordant sources are assigned smaller values of $\hat{\sigma}_{i,x}^2$, reflecting greater credibility and thus influence over the posterior belief (hypothesis 1a in the main text).²

²Politics and motivated reasoning might also manifest in other parts of a Bayesian model, such as the location of countervailing signals if a partisan is motivated to seek out disconfirming arguments, or gin them up in their own head, when confronted with discordant signals. We leave this uncommon albeit debated "backfire" dynamic to other research, and argue that – at least in our experimental setting – participants are unlikely to exert this level of cognitive or practical effort.

5.3 Illusory Truth Effects in the Model

Repeated exposure to a signal operates through two parts of the Bayesian framework. As with partisan motivated reasoning, seeing the same information multiple times increases the credibility of the signal, meaning that the precision parameter associated with novel information ($\hat{\sigma}_{n,x}$) is larger than that associated with repeated information ($\hat{\sigma}_{r,x}$).

In addition, in a Bayesian framework, the illusory truth effect should also operate through the prior belief itself. By definition, in a dynamic setting where the same individual is exposed to a signal multiple times, the posterior generated from the first exposure becomes the prior for the second exposure, and so on for each subsequent exposure. This will produce two implications from the Bayesian model. On the one hand, each exposure to the same signal should move the posterior closer to the signal than the prior. On the other hand, each posterior's precision is greater than its associated prior by construction, meaning that each subsequent prior should be harder to move further.

These dynamics are illustrated in Figure 5. The top panel displays how, assuming a constant signal, the individual's prior belief becomes more confidently held and moves towards the signal at each repeated exposure. The bottom panel exhibits the same pattern, although the additional influence through the credibility parameter means that the posteriors update more strongly at each subsequent exposure. These two mechanisms – greater credibility for familiar signals and increasingly tight priors – form the core of the comparative expectations derived in the main text, in particular for how the two cognitive biases interact.

5.4 Simulating the Interaction between PMR and ITE

As discussed in the main text, situating both theories in the common Bayesian model puts structure on the less-studied question of whether and how they might interact. Here, two offsetting expectations matter. First, as demonstrated in Figure 5, each subsequent prior belief distribution should grow increasingly narrow, which falls out naturally from the assumption that the prior for each subsequent exposure is the posterior from the previous exposure.³ Second, and conversely, is the process by which the ITE also operates by increasing the credibility of a familiar signal. For simplicity, assume that ITE operates by increasing the credibility of any signal by a fixed amount. The effect on the posterior for a politically concordant signal

³The variance of the posterior must always be less than the variance of the prior by $\frac{\hat{\sigma}_0^2 \hat{\sigma}_x^2}{\hat{\sigma}_0^2 + \hat{\sigma}_x^2}$.

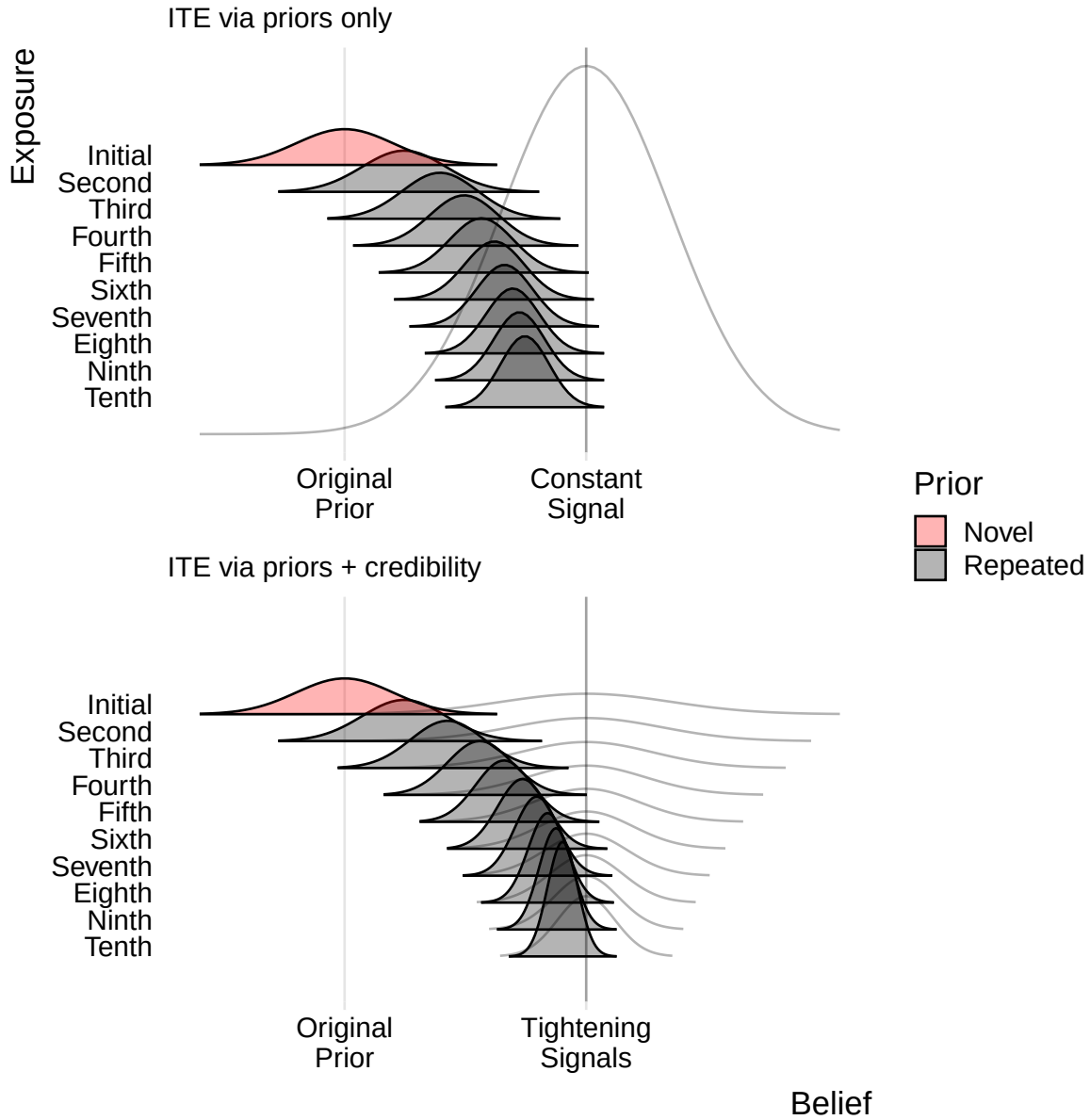


Figure 5: Simulated evidence of illusory truth effects over ten exposures (y-axis) to the same signal (vertical line). Each shaded distribution represents the posterior belief from the previous period’s exposure, which becomes the prior belief for the subsequent exposure to the same signal. The top panel displays the scenario in which the illusory truth effect only operates via dynamic priors. The bottom panel displays the scenario in which ITE has an additional effect on the subjective credibility of a repeated signal $\hat{\sigma}_x^2$.

will be *larger* than the effect on the posterior associated with a politically discordant signal.⁴ This is due to the non-linearity in how reductions in the $\hat{\sigma}_{i,x}^2$ term translate into changes in the posterior.

⁴We explore alternative characterizations by which repeated exposure reduces the variance of the signal in Section 5.6 below, none of which change the substantive expectations laid out here.

To give an example (visualized in Figure 6), assume that the standard deviation of the signal decreases by 0.1 as a result of prior exposure, consistent with the literature on prior and repeated exposure (Fazio, Rand and Pennycook, 2019; Fazio et al., 2015; Pennycook, Cannon and Rand, 2018). Furthermore, assume that the variance of the signal from a politically discordant source is 0.5, while the variance of the signal from a politically concordant source is 0.3, again consistent with the literature on partisan motivated reasoning (Lodge, 2013; Kunda, 1990; Flynn, Nyhan and Reifler, 2017). Focusing on the first row of the left column of Figure 6 (“Initial”), the difference in the updated beliefs upon seeing this signal from a concordant versus discordant source is 0.17.

Now consider the second exposure to the same signal (second row of the left column). Two things happen: first, the posterior belief from the first exposure becomes the prior belief in the second exposure, and second, the variance of *both* signals reduces to 0.4 and 0.2 for the politically discordant and concordant sources, respectively. The net result is an increase in the magnitude of partisan motivated reasoning from 0.17 in the initial exposure to 0.30 in the second exposure. This implies that the interaction term between ITE and PMR should be positive. In parallel, the right column of Figure 6 illustrates this from the perspective of the illusory truth effect: among discordant headlines, repeated exposure amounts to a difference of 0.15, while among concordant headlines this almost doubles to 0.28. These simulations motivate the reinforcing half of the interaction expectation (hypothesis 2 in the main text).

However, the rule of diminishing returns discussed above in Figure 5 starts to obtain here as well. All else equal, increasingly tighter priors should reduce the strength of partisan motivated reasoning since the second exposure’s prior should be more confidently held and harder to move, generating a decay in the effects of partisan signals over time (or any signal). This dynamic can be described as a negative interaction term over repeated exposure. Importantly, this does not imply that PMR itself would ever become negative, only that the difference between a politically concordant and discordant signal should be smaller when an individual has seen the signal multiple times already. As illustrated in the third row of the left column of Figure 6, a third exposure increases the difference in beliefs from a concordant versus discordant signal even further to 0.38, but there is also evidence of the diminishing marginal returns as the increase from 0.30 to 0.38 is less than the increase from 0.17 to 0.30. Eventually the interaction term must become negative by the same logic discussed above in Figure 5: as each posterior becomes the next period’s increasingly tight prior, any new signals have diminishing influence on posterior attitudes. The bottom row of Figure 6 visualizes this

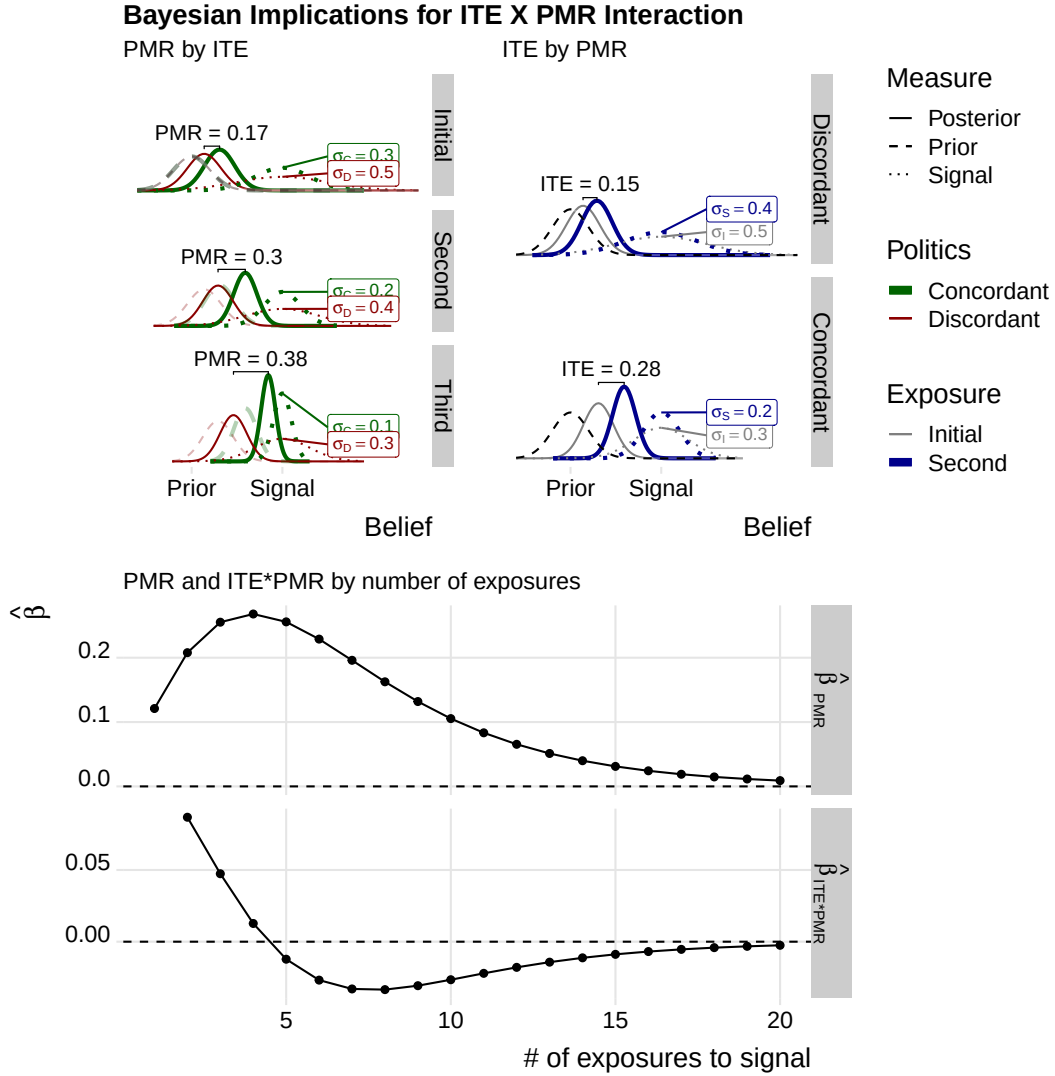


Figure 6: Simulated evidence of the interaction effect between ITE and motivated reasoning. *Left column:* The same increase in the credibility associated with a politically concordant signal (equivalent to a 0.2 decrease in the standard deviation of the signal’s distribution) is much more influential upon the second exposure compared to the first, and modestly more influential upon the third exposure compared to the second. *Right column:* In parallel, the same increase in the credibility of a signal associated with prior exposure (equivalent to a 0.1 decrease in the standard deviation of the signal’s distribution) produces a much larger increase in the posterior belief when the signal is from a politically concordant source compared to a politically discordant source. *Bottom row:* Over repeated exposures, the increases flatten out and start diminishing at the margin, turning the interaction term negative after the fourth exposure.

dynamic for 20 repetitions of the same signal, using the closed-form solution derived in Section 5.5 below: the additional boost to PMR flattens out and starts diminishing after the 4th exposure.⁵

⁵The precise dynamics displayed are a function of the initial prior’s variance, as well as the simulated effect of PMR and of repeated signals on the credibility parameter. The point after which the interaction term becomes negative is a function of both the

These offsetting dynamics motivate the scope conditions pre-registered in the main text: the interaction between PMR and ITE should be positive when a headline is sufficiently novel, and should weaken – or turn negative – as prior beliefs tighten with accumulating exposures (hypothesis 3a in the main text). The same logic extends to the other sources of variation in the strength of priors that our designs exploit: non-political headlines (hypothesis 3b), wholly fabricated headlines (hypothesis 3c), and true headlines (hypothesis 3d).

5.5 Closed-Form Results for n Repeated Exposures

The main text discusses the interaction between ITE and PMR in the context of relatively novel signals, and notes that many repeated exposures should eventually tilt the interaction effect to become negative – i.e., after enough exposures to the same signal, the marginal effect of partisan motivated reasoning should become negative. (NB: this is not to say that individuals update more on signals from discordant sources than concordant. Only that the positive PMR coefficient should be smaller after the 10th exposure compared to the 9th.) We flesh this idea out in more detail here, starting from the model in Section 5.1. For simplicity, we will drop the i subscript and denote each period with an integer subscript instead. I.e., $\pi_2(\mu|x)$ represents the second exposure to the signal.

Assume that each subsequent exposure’s prior is the same as the posterior from the previous period’s exposure. Thus

$$\pi_2(\mu|x) \sim \mathcal{N}\left(\hat{\mu}_1 + (\mu_x - \hat{\mu}_1) \left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_1^2 + \hat{\sigma}_x^2}\right), \frac{\hat{\sigma}_1^2 \hat{\sigma}_x^2}{\hat{\sigma}_1^2 + \hat{\sigma}_x^2}\right) \quad (1)$$

where $\hat{\mu}_1 = \hat{\mu}_0 + (\mu_x - \hat{\mu}_0) \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_0^2 + \hat{\sigma}_x^2}\right)$ and $\hat{\sigma}_1^2 = \frac{\hat{\sigma}_0^2 \hat{\sigma}_x^2}{\hat{\sigma}_0^2 + \hat{\sigma}_x^2}$. For any n , the closed form solution for the variance is thus

$$\hat{\sigma}_n^2 = \frac{\hat{\sigma}_0^2}{1 + n \frac{\hat{\sigma}_0^2}{\hat{\sigma}_x^2}} \quad (2)$$

and the closed form solution for the posterior mean is thus

$$\hat{\mu}_n = \frac{\hat{\sigma}_x^2 \hat{\mu}_0 + n \hat{\sigma}_0^2 \mu_x}{\hat{\sigma}_x^2 + n \hat{\sigma}_0^2} \quad (3)$$

relative variance of the initial prior and the initial signal, as well as the magnitude of the credibility gap between concordant and discordant sources, as we show in Section 5.5.

With these in hand, assume that differences in source credibility are constant across periods, such that the credibility associated with a concordant source is always δ smaller than the credibility associated with a discordant source, such that $\hat{\sigma}_C^2 \leq \hat{\sigma}_D^2$ where $\hat{\sigma}_D^2 = \hat{\sigma}_C^2 + \delta$. We can express the posterior belief for each source at a given moment in time as

$$\hat{\mu}_{C,n} = \frac{\hat{\sigma}_C^2 \hat{\mu}_0 + n \hat{\sigma}_0^2 \mu_x}{\hat{\sigma}_C^2 + n \hat{\sigma}_0^2} \quad (4)$$

$$\hat{\mu}_{D,n} = \frac{\hat{\sigma}_D^2 \hat{\mu}_0 + n \hat{\sigma}_0^2 \mu_x}{\hat{\sigma}_D^2 + n \hat{\sigma}_0^2} \quad (5)$$

Partisan motivated reasoning is simply the difference between these posteriors at a given period:

$$PMR_n = \hat{\mu}_{C,n} - \hat{\mu}_{D,n} \quad (6)$$

$$= \frac{\hat{\sigma}_C^2 \hat{\mu}_0 + n \hat{\sigma}_0^2 \mu_x}{\hat{\sigma}_C^2 + n \hat{\sigma}_0^2} - \frac{\hat{\sigma}_D^2 \hat{\mu}_0 + n \hat{\sigma}_0^2 \mu_x}{\hat{\sigma}_D^2 + n \hat{\sigma}_0^2} \quad (7)$$

$$= \frac{\hat{\sigma}_C^2 \hat{\mu}_0 + n \hat{\sigma}_0^2 \mu_x}{\hat{\sigma}_C^2 + n \hat{\sigma}_0^2} - \frac{(\hat{\sigma}_C^2 + \delta) \hat{\mu}_0 + n \hat{\sigma}_0^2 \mu_x}{(\hat{\sigma}_C^2 + \delta) + n \hat{\sigma}_0^2} \quad (8)$$

In this set-up, we have only incorporated repeated exposure through its influence on the prior at each period. In other words, we do not conceptualize the illusory truth effect (ITE) as biased cognition, but rather a simple mechanical byproduct of sequential Bayesian updating. As such, to understand the interaction between PMR and ITE, we only need to calculate the partial derivative of the above expression with respect to n to formalize our expectations regarding the interaction term under these assumptions.

$$\frac{\partial PMR_n}{\partial n} = \frac{\partial \hat{\mu}_{C,n} - \hat{\mu}_{D,n}}{\partial n} \quad (9)$$

$$= \frac{\partial \hat{\mu}_{C,n}}{\partial n} - \frac{\partial \hat{\mu}_{D,n}}{\partial n} \quad (10)$$

$$= \frac{\partial \frac{\hat{\sigma}_C^2 \hat{\mu}_0 + n \hat{\sigma}_0^2 \mu_x}{\hat{\sigma}_C^2 + n \hat{\sigma}_0^2}}{\partial n} - \frac{\partial \frac{(\hat{\sigma}_C^2 + \delta) \hat{\mu}_0 + n \hat{\sigma}_0^2 \mu_x}{(\hat{\sigma}_C^2 + \delta) + n \hat{\sigma}_0^2}}{\partial n} \quad (11)$$

Applying the quotient rule yields:

$$\frac{\partial \hat{\mu}_{C,n}}{\partial n} = \frac{\hat{\sigma}_0^2 \mu_x (\hat{\sigma}_C^2 + n \hat{\sigma}_0^2) - \hat{\sigma}_0^2 (\hat{\sigma}_C^2 \hat{\mu}_0 + n \hat{\sigma}_0^2 \mu_x)}{(\hat{\sigma}_C^2 + n \hat{\sigma}_0^2)^2} \quad (12)$$

$$= \frac{\hat{\sigma}_0^2 (\mu_x \hat{\sigma}_C^2 - \hat{\sigma}_C^2 \hat{\mu}_0)}{(\hat{\sigma}_C^2 + n \hat{\sigma}_0^2)^2} \quad (13)$$

$$\frac{\partial \hat{\mu}_{D,n}}{\partial n} = \frac{\hat{\sigma}_0^2 (\mu_x (\hat{\sigma}_C^2 + \delta) - (\hat{\sigma}_C^2 + \delta) \hat{\mu}_0)}{((\hat{\sigma}_C^2 + \delta) + n \hat{\sigma}_0^2)^2} \quad (14)$$

Thus the partial derivative of PMR with respect to n can be written and simplified to:

$$\frac{\partial PMR_n}{\partial n} = \frac{\hat{\sigma}_0^2 (\mu_x \hat{\sigma}_C^2 - \hat{\sigma}_C^2 \hat{\mu}_0)}{(\hat{\sigma}_C^2 + n \hat{\sigma}_0^2)^2} - \frac{\hat{\sigma}_0^2 (\mu_x (\hat{\sigma}_C^2 + \delta) - (\hat{\sigma}_C^2 + \delta) \hat{\mu}_0)}{((\hat{\sigma}_C^2 + \delta) + n \hat{\sigma}_0^2)^2} \quad (15)$$

$$= \hat{\sigma}_0^2 (\mu_x - \hat{\mu}_0) \left(\frac{\hat{\sigma}_C^2}{(\hat{\sigma}_C^2 + n \hat{\sigma}_0^2)^2} - \frac{(\hat{\sigma}_C^2 + \delta)}{((\hat{\sigma}_C^2 + \delta) + n \hat{\sigma}_0^2)^2} \right) \quad (16)$$

This expression highlights several dimensions of substantive interest. First, the partial derivative with respect to n contains n in it, highlighting that there are non-linearities in the relationship between repeated exposure and partisan signals. Closer examination and simulation reveals that this is a function of the relative magnitude of $\hat{\sigma}_0^2$ and $\hat{\sigma}_x^2$. Specifically, if the variance of the signal is larger than that of the prior, the PMR gap will grow over initial periods as congruent signals are more influential than incongruent signals. The precise inflection point is a function of the magnitude of the difference between $\hat{\sigma}_x^2$ and $\hat{\sigma}_0^2$. We simulate different values of these parameters in Figure 7, where the slope of the lines capture the interaction term's sign on $\beta_{PMR} * \beta_{ITE}$.

In sum then, assuming that the variance of the signal is not substantially higher than the variance of the prior, the interaction term between ITE and PMR should be negative if we only allow repeated exposure to operate via the sequence linking the previous period's posterior with the current period's prior.

5.6 Allowing Repetition to Increase Signal Credibility

However, the literature on the illusory truth effect emphasizes other pathways by which individuals believe repeated information more than novel information. These include familiarity, more easily accessible details, and less cognitive work. Although none of this work articulates ITE explicitly in the setting of a Bayesian model of attitude change, it is reasonable to interpret these explanations in terms of the credibility parameter $\hat{\sigma}_x^2$. Modeling this idea is complicated by the decision over the specific functional form by which $\hat{\sigma}_x^2$ is

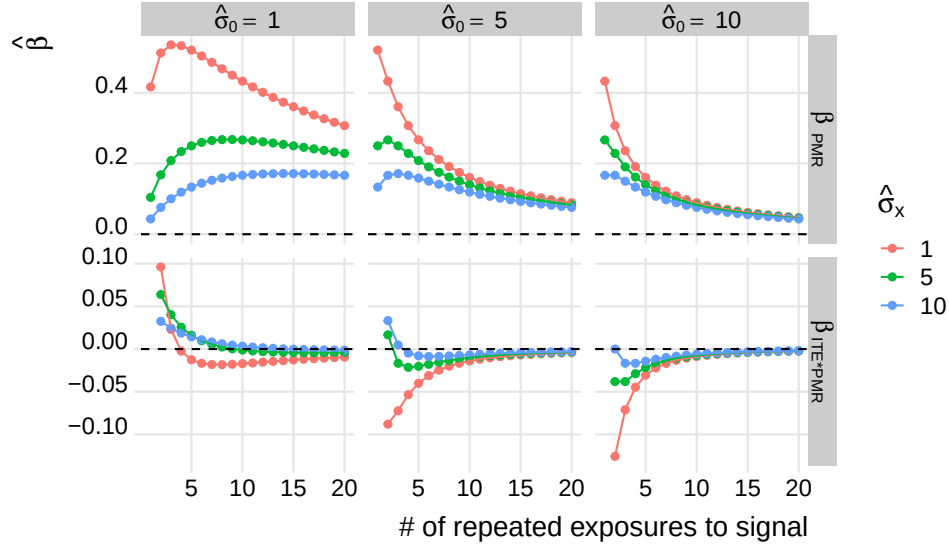


Figure 7: Difference in posteriors between politically concordant and discordant partisan sources (β_{PMR} , top row) and the interaction between repeated exposure and PMR ($\beta_{ITE*PMR}$, bottom row) by different numbers of repeated exposures (x-axes), across different values of $\hat{\sigma}_x^2$ (colors) and $\hat{\sigma}_0^2$ (columns).

influenced by n . All else equal, we would expect that the increased credibility should be diminishing at the margin, capturing the fact that the illusory truth effect should be stronger going from zero previous exposures to one, than it should going from 100 previous exposures to 101.

We formalize this intuition with:

$$\hat{\sigma}_x^2(n) = \hat{\sigma}_x^2(0)e^{-\lambda n} \quad (17)$$

where λ governs how quickly the reduction in repeated exposures obtains. We set $\hat{\sigma}_0^2 = 1$, $\hat{\sigma}_x^2 = 10$ and allow λ to range from a linear decay ($\lambda = 0$) to a rapid decay ($\lambda = 0.5$). Figure 8 visualizes the results, highlighting that the substantive expectations about an interaction term that decays with repeated exposures from the previous analysis hold, with λ simply governing how quickly the positive interaction term disappears. Given the patterns documented in Study 2 of the main text – where the estimated influence of partisan source cues increases after the first exposure while the incremental gains diminish with repeated encounters – it would seem that λ is relatively small in our applied setting.

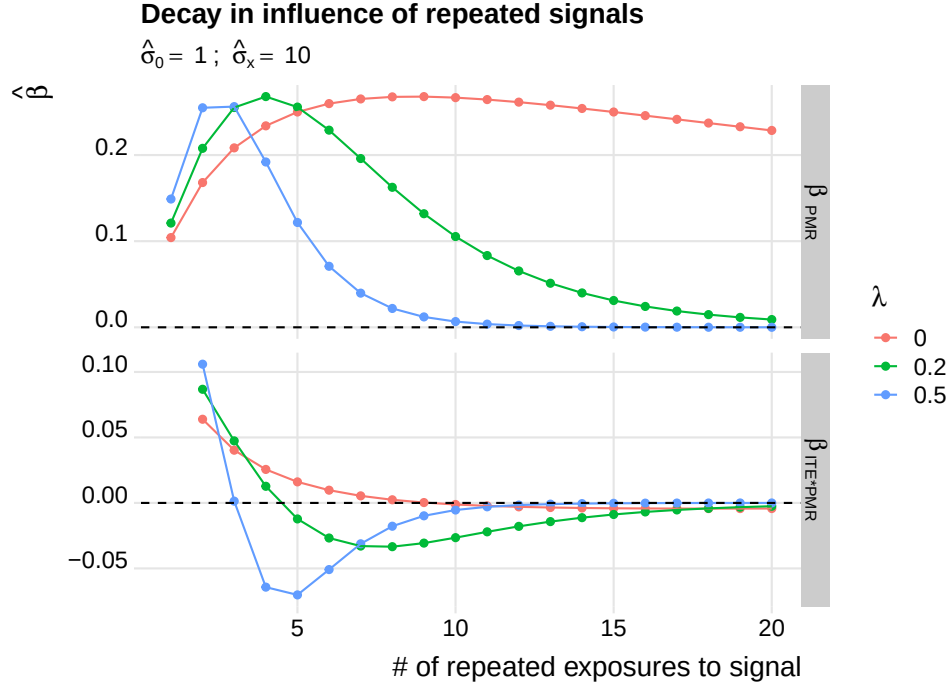


Figure 8: Difference in posteriors between politically concordant and discordant partisan sources (β_{PMR} , top row) and the interaction between repeated exposure and PMR ($\beta_{ITE*PMR}$, bottom row) by different numbers of repeated exposures (x-axes), across different values of λ (colors).

6 Regression Tables

This section presents numerical results for all graphs discussed in the paper. We start presenting the paper main results using only false headlines across studies 1-4. Then we present the same estimates for the true headlines, as discussed in the paper.

Table 7: Regression Table: Study 1 (false headlines only)

	T1 - ITE	T1 - PMR	T1 - ITE x PMR	T2 - ITE	T2 - PMR	T2 - ITE x PMR
Intercept	0.387*** (0.028)	0.322*** (0.030)	0.312*** (0.030)	0.387*** (0.026)	0.326*** (0.028)	0.316*** (0.028)
Prior Exposure (ITE)	0.069*** (0.011)		0.040*** (0.015)	0.046*** (0.011)		0.041*** (0.016)
Concordant (PMR)		0.164*** (0.009)	0.151*** (0.010)		0.151*** (0.009)	0.156*** (0.010)
ITE x PMR			0.053** (0.020)			-0.020 (0.021)
Num.Obs.	10 472	9296	9296	10 354	9066	9066
R2 Marg.	0.004	0.028	0.032	0.002	0.024	0.024
R2 Cond.	0.207	0.232	0.234	0.201	0.228	0.228
ICC	0.2	0.2	0.2	0.2	0.2	0.2

* p < 0.1, ** p < 0.05, *** p < 0.01

Outcome: accuracy assessment. Random intercepts at the headline and participant level.

Table 8: Regression Table: Study 2 (false headlines only)

	ITE	PMR	ITE x PMR
Intercept	0.496*** (0.049)	0.368*** (0.031)	0.451*** (0.052)
Seen Once (ITE)	-0.080 (0.060)		-0.124** (0.063)
Seen Twice (ITE)	-0.063 (0.061)		-0.124* (0.064)
Concordant (PMR)		0.152*** (0.009)	0.088*** (0.016)
ITE(Once) x PMR			0.088*** (0.020)
ITE(Twice) x PMR			0.122*** (0.027)
Num.Obs.	10 396	9354	9354
R2 Marg.	0.005	0.023	0.031
R2 Cond.	0.202	0.230	0.235
ICC	0.2	0.2	0.2

* p < 0.1, ** p < 0.05, *** p < 0.01

Outcome: accuracy assessment. Random intercepts at the headline and participant level.

Table 9: Regression Table: Study 3 (false headlines only)

	ITE	PMR	ITE x PMR
Intercept	0.268*** (0.031)	0.178*** (0.043)	0.167*** (0.044)
Prior Exposure (ITE)	0.034*** (0.010)		0.024 (0.018)
Concordant (PMR)		0.196*** (0.013)	0.197*** (0.019)
Non-Political		0.133* (0.072)	0.117 (0.074)
ITE x Concordant			-0.002 (0.026)
ITE x Non-Political			0.032 (0.026)
Num.Obs.	6024	5460	5460
R2 Marg.	0.001	0.032	0.034
R2 Cond.	0.225	0.267	0.269
ICC	0.2	0.2	0.2

* p < 0.1, ** p < 0.05, *** p < 0.01

Outcome: accuracy assessment. Random intercepts at the headline and participant level.

Table 10: Regression Table: Study 4 (false/fake headlines only)

	T1(False) - ITE	T1(False) - PMR	T1(False) - ITE x PMR	T2(Fake) - ITE	T2(Fake) - PMR	T2(Fake) - ITE x PMR
Intercept	0.426*** (0.046)	0.267*** (0.039)	0.274*** (0.040)	0.312*** (0.028)	0.257*** (0.027)	0.245*** (0.028)
Prior Exposure (ITE)	-0.012 (0.011)		-0.015 (0.016)	0.003 (0.010)		0.024 (0.015)
Concordant (PMR)		0.302*** (0.012)	0.295*** (0.016)		0.116*** (0.011)	0.143*** (0.015)
ITE x PMR			0.013 (0.023)			-0.054** (0.021)
Num.Obs.	6520	5840	5840	6568	5968	5968
R2 Marg.	0.000	0.094	0.094	0.000	0.015	0.016
R2 Cond.	0.146	0.226	0.226	0.195	0.209	0.210
ICC	0.1	0.1	0.1	0.2	0.2	0.2

* p < 0.1, ** p < 0.05, *** p < 0.01

Outcome: accuracy assessment. Random intercepts at the headline and participant level.

Table 11: Regression Table: Study 1 (true headlines only)

	T1 - ITE	T1 - PMR	T1 - ITE x PMR	T2 - ITE	T2 - PMR	T2 - ITE x PMR
Intercept	0.583*** (0.043)	0.514*** (0.045)	0.509*** (0.046)	0.582*** (0.043)	0.520*** (0.045)	0.514*** (0.045)
Prior Exposure (ITE)	0.034*** (0.011)		0.019 (0.015)	0.013 (0.011)		0.023 (0.015)
Concordant (PMR)		0.180*** (0.009)	0.175*** (0.010)		0.168*** (0.009)	0.172*** (0.010)
ITE x PMR			0.019 (0.020)			-0.018 (0.021)
Num.Obs.	10 473	9297	9297	10 356	9067	9067
R2 Marg.	0.001	0.034	0.034	0.000	0.029	0.029
R2 Cond.	0.221	0.246	0.246	0.236	0.249	0.249
ICC	0.2	0.2	0.2	0.2	0.2	0.2

* p < 0.1, ** p < 0.05, *** p < 0.01

Outcome: accuracy assessment. Random intercepts at the headline and participant level.

Table 12: Regression Table: Study 2 (true headlines only)

	ITE	PMR	ITE x PMR
Intercept	0.526*** (0.058)	0.504*** (0.037)	0.462*** (0.061)
Seen Once (ITE)	0.093 (0.070)		0.067 (0.075)
Seen Twice (ITE)	0.091 (0.071)		0.051 (0.076)
Concordant (PMR)		0.172*** (0.009)	0.130*** (0.016)
ITE(Once) x PMR			0.058*** (0.020)
ITE(Twice) x PMR			0.080*** (0.027)
Num.Obs.	10 399	9355	9355
R2 Marg.	0.008	0.030	0.039
R2 Cond.	0.194	0.227	0.231
ICC	0.2	0.2	0.2

* p < 0.1, ** p < 0.05, *** p < 0.01

Outcome: accuracy assessment. Random intercepts at the headline and participant level.

Table 13: Regression Table: Study 3 (true headlines only)

	ITE	PMR	ITE x PMR
Intercept	0.640*** (0.081)	0.517*** (0.089)	0.525*** (0.089)
Prior Exposure (ITE)	0.016 (0.011)		-0.015 (0.019)
Concordant (PMR)		0.149*** (0.014)	0.133*** (0.020)
Non-Political		0.260* (0.153)	0.237 (0.154)
ITE x Concordant			0.032 (0.027)
ITE x Non-Political			0.047* (0.028)
Num.Obs.	6024	5460	5460
R2 Marg.	0.000	0.049	0.049
R2 Cond.	0.249	0.264	0.264
ICC	0.2	0.2	0.2

* p < 0.1, ** p < 0.05, *** p < 0.01

Outcome: accuracy assessment. Random intercepts at the headline and participant level.

Table 14: Regression Table: Study 4 (true headlines only)

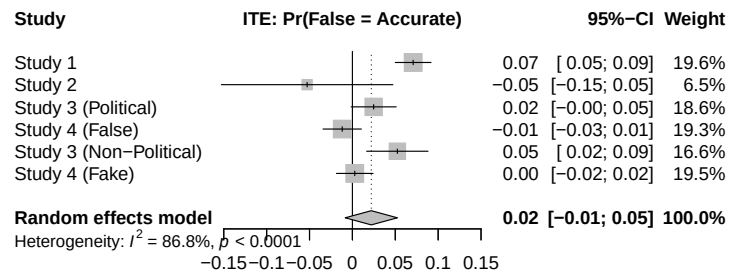
	T1 - ITE	T1 - PMR	T1 - ITE x PMR	T2 - ITE	T2 - PMR	T2 - ITE x PMR
Intercept	0.602*** (0.032)	0.486*** (0.033)	0.490*** (0.034)	0.591*** (0.036)	0.480*** (0.036)	0.487*** (0.037)
Prior Exposure (ITE)	-0.012 (0.011)		-0.008 (0.016)	-0.012 (0.011)		-0.014 (0.016)
Concordant (PMR)		0.228*** (0.012)	0.231*** (0.017)		0.215*** (0.012)	0.214*** (0.016)
ITE x PMR			-0.005 (0.023)			0.003 (0.023)
Num.Obs.	6520	5840	5840	6568	5968	5968
R2 Marg.	0.000	0.054	0.054	0.000	0.047	0.048
R2 Cond.	0.134	0.185	0.185	0.152	0.205	0.205
ICC	0.1	0.1	0.1	0.2	0.2	0.2

* p < 0.1, ** p < 0.05, *** p < 0.01

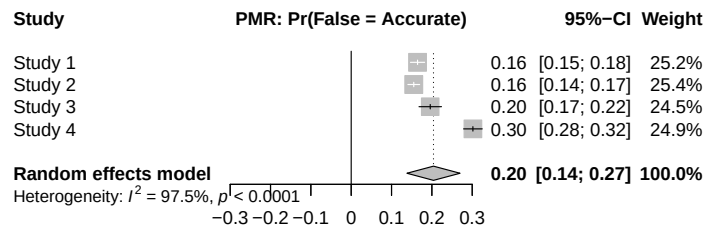
Outcome: accuracy assessment. Random intercepts at the headline and participant level.

7 Meta Analytical Results

In this section, we estimate, for each study, the marginal effect of belief in false headlines on prior exposure (ITE) and partisan concordance (PMR), and combine the study-level estimates using a random-effects meta-analysis estimator. This approach allows us to pool all four results and estimate the robustness of the claims in the main paper. For Study 3, we estimate the ITE split by political and non-political headlines, and for Study 4, we separate the false and wholly fabricated (“fake”) treatment arms. Figures 9 present the results. The meta-analytic estimate for Partisan motivated reasoning is strong, consistent across studies, and statistically significant ($M_{PMR} = 0.20$, 95% CI [0.14, 0.27]). Meanwhile, the estimate for ITE is small and not statistically different from zero ($M_{ITE} = 0.02$, 95% CI [−0.03, 0.06]). These findings align with the broader findings described in the paper, based on individual estimates across the four studies.



(a) Meta-Analysis ITE



(b) Meta-Analysis PMR

Figure 9: Meta Analytical Results Across Four Studies for the ITE and PMR Treatment Effects

8 Robustness Checks: Four Point Continuous Outcome

In the main paper, we report results using a binary response measuring respondents' assessment of the accuracy of a false headline. This measure recodes our original four-point scale outcomes (*Not at all accurate*, *Not accurate*, *Somewhat accurate*, *Very Accurate*). Table 15 re-estimates the models comparing the binary and four-point scale. Of the 20 coefficients reported in the main manuscript, only two change substantively; first, the ITE coefficient for false political headlines in study 3 changes and is estimated more precisely, achieving statistical significance at conventional levels ($\beta = 0.069$, p-value = 0.007), second the ITE coefficient for true political headlines in study 1 is less precisely estimated, and loses statistical significance ($\beta = 0.026$, p-value = 0.19).

Table 15: Marginal effects of ITE and PMR on perceived accuracy: binary versus 4-point outcome.

Condition	Veracity	Binary (0–1)	4-point (1–4)
ITE			
Study 1	False	0.071***	0.143***
	True	0.029**	0.026
Study 2	False	-0.053	-0.128
	True	0.048	0.136
Study 3 (Non-Political)	False	0.052**	0.104**
	True	0.039*	0.066*
Study 3 (Political)	False	0.025+	0.069**
	True	0.005	0.029
Study 4 (Fake)	False	0.003	0.031+
	True	-0.012	-0.022
Study 4 (False)	False	-0.012	0.005
	True	-0.012	-0.017
PMR			
Study 1	False	0.165***	0.362***
	True	0.180***	0.382***
Study 2	False	0.155***	0.329***
	True	0.168***	0.359***
Study 3	False	0.196***	0.519***
	True	0.150***	0.381***
Study 4	False	0.302***	0.588***
	True	0.228***	0.426***

Note:

Cells are marginal effects (average slopes) on the respective outcome scale. ITE is the slope of prior exposure; PMR is the Concordant minus Discordant contrast. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

9 Deviations from the Pre-Analysis Plan

In this section, we report deviations between the main paper and our pre-registration plans. First, the pre-analysis plan for Study 1 included an initial stage that randomly assigned respondents to complete the survey on either a mobile device (smartphone or tablet) or a desktop or laptop computer. We never implemented this randomization because of technical issues with the Qualtrics platform. We therefore removed the device experiment from the study entirely before starting the survey field, and participants completed the survey on the device of their choice. Since device assignment was never implemented, this change does not affect any of the pre-registered tests reported here. Second, our manuscript reorders and groups the hypotheses presented, and repeated across the pre-registrations. Because in all papers, we test for ITE and PMR effects, and their interactions, these are entered in every pre-registration. We grouped specific hypotheses related to modifications in the experimental design (duration, non-political, fake headlines, and true headlines) into a single hypothesis 3 in the paper. This purely changes the presentation of the hypotheses and organizes these studies under a unified theoretical framework. We also slightly edited the wording of the hypotheses to improve readability. Third, the pre-registered hypotheses for warning labels focus solely on moderating ITE effects; for completeness, we present in the paper both the warning labels' effects on ITE and PMR. In fact, the results are only in the direction of the pre-registered ITE effects, and null for the attenuation effect of warning labels on PMR.

10 Research Ethics

Approval, consent, and privacy. All four studies were reviewed and approved under a single protocol by the institutional review board at [REDACTED] (IRB-FY2024-8231), as an expedited review of survey research. Studies 3 and 4 were approved as amendments to that protocol. Participants were U.S. adults (18+) drawn from commercial online panels: Qualtrics (Studies 1–2) and Connect by CloudResearch (Studies 3–4). At the start of each survey, participants saw an IRB-approved consent form and chose to consent or decline; those who declined did not proceed. The form told participants that participation was voluntary, that they could stop at any time without penalty, and that some of the news they would see was true and some false. Participants could also withdraw their data after submission by contacting the panel provider, which would remove their responses before we received them. Consent was documented by the click-through response rather than a signature, which the IRB approved because the research posed no more than minimal risk and involved no procedures for which written consent is normally required. The researchers never interacted with participants and received no personally identifiable information. Recruitment and payment were handled entirely by the panel providers, and the Qualtrics surveys used the setting that prevents collection of IP addresses. The data were stored on encrypted, password-protected university servers and are shared only in de-identified form.

Deception and debriefing. The studies involve incomplete disclosure. Participants were told that some headlines were false but not which ones, and this could not be disclosed in advance because the studies measure belief in true and false headlines. In Studies 1–3, the false headlines came from stories rated false by a fact-checking organization (Snopes), and we edited the images to attribute them to real partisan outlets (e.g., Fox News or MSNBC) in order to manipulate the source cue. In Study 4, the false headlines were “placebo” fakes: we prompted a large language model with real news events from the month before fielding and asked it to reverse their veracity. The research team reviewed all headlines and removed any that might expose participants to risk. The IRB approved the deception and the waiver of the full disclosure of consent elements on the grounds that the research posed no more than minimal risk, that the waiver would not adversely affect participants’ rights and welfare, and that the research could not practicably be carried out without it. At the end of the survey, participants were debriefed. The debrief listed the headlines that were true and pointed to fact-checking resources, and it stated that all other headlines were false. We deliberately

did not restate the false headlines, in order to avoid re-exposing participants to misinformation. Study 1 participants were debriefed only at the end of the Study 2 wave, when they were re-contacted one day later, so Study 1 participants who did not return for Study 2 were not debriefed.

Risks. We judged the risks to be minimal: exposure to true and false headlines is comparable to what people encounter in daily life on social media and news sites, and the main foreseeable risk was a breach of data confidentiality, which the procedures above are designed to prevent. Participants who saw the same headline twice, or saw false headlines, were exposed only briefly and in an experimental setting, followed by the debrief. The IRB classified the study as minimal risk.

Compensation. Panel providers paid participants directly, and we did not set or see individual payments. Participants were paid standard market prices as set by each platform. Qualtrics panelists received the compensation they had agreed to with the panel before entering the survey, in forms such as points, gift cards, or cash. CloudResearch Connect participants were paid at rates no lower than the platform's recommended minimum. Under the protocol, all participants who completed the survey were paid; no one was excluded from payment by attention checks or other quality criteria.

Participant pool and vulnerable groups. Samples were quota-based on age, gender, race, and partisanship (Studies 1–2 also on region), and Table 1 in the Supplementary Information reports the demographic composition of each sample. Opt-in panels differ from the general population: the Connect samples in Studies 3–4 skew younger, more educated, and more Democratic than the Qualtrics samples, and panels underrepresent people without reliable internet access, so the results should not be read as representative of all U.S. adults. We did not recruit or target any group typically considered vulnerable (e.g., minors, prisoners, or people with impaired decision-making capacity), and everyone in the samples was a consenting adult. Materials were in English only.

Differential benefits and harms. Because assignment to conditions was random, no demographic or partisan group was systematically more likely to receive more false content or repeated exposure. Partisan balance was built into the design, with equal numbers of headlines favorable to each party. Warning labels were shown to one treatment arm only, and control participants never saw the same headline twice. Participants benefited from compensation and from the debrief's directions to fact-checking resources. We

do not expect the research to benefit or harm any group differentially. The findings may help platforms and policymakers evaluate warning labels and other interventions, which would benefit the public generally.

References

- Coppock, Alexander. 2023. *Persuasion in parallel: How information changes minds about politics*. University of Chicago Press.
- Dechêne, Alice, Christoph Stahl, Jochim Hansen and Michaela Wänke. 2010. “The truth about the truth: A meta-analytic review of the truth effect.” *Personality and Social Psychology Review* 14(2):238–257.
- Druckman, James N and Mary C McGrath. 2019. “The evidence for motivated reasoning in climate change preference formation.” *Nature Climate Change* 9(2):111–119.
- Fazio, Lisa K, David G Rand and Gordon Pennycook. 2019. “Repetition increases perceived truth equally for plausible and implausible statements.” *Psychonomic bulletin & review* 26:1705–1710.
- Fazio, Lisa K, Nadia M Brashier, B Keith Payne and Elizabeth J Marsh. 2015. “Knowledge does not protect against illusory truth.” *Journal of experimental psychology: general* 144(5):993.
- Flynn, Daniel J, Brendan Nyhan and Jason Reifler. 2017. “The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics.” *Political Psychology* 38:127–150.
- Ghanem, Dalia, Sarojini Hirshleifer and Karen Ortiz-Becerra. 2023. “Testing attrition bias in field experiments.” *Journal of Human resources* .
- Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson and David Lazer. 2019. “Fake news on Twitter during the 2016 US presidential election.” *Science* 363(6425):374–378.
- Guay, Brian and Christopher D Johnston. 2022. “Ideological asymmetries and the determinants of politically motivated reasoning.” *American Journal of Political Science* 66(2):285–301.
- Guess, Andrew, Jonathan Nagler and Joshua Tucker. 2019. “Less than you think: Prevalence and predictors of fake news dissemination on Facebook.” *Science advances* 5(1):eaau4586.
- Holyoak, Keith J and Robert G Morrison. 2012. *The Oxford handbook of thinking and reasoning*. Oxford University Press.

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

Lodge, M. 2013. *The rationalizing voter*. Cambridge University Press.

Pennycook, Gordon, Tyrone D Cannon and David G Rand. 2018. "Prior exposure increases perceived accuracy of fake news." *Journal of experimental psychology: general* 147(12):1865.

Stroud, Natalie Jomini. 2011. *Niche news: The politics of news choice*. Oxford University Press.