

# Willingness to Verify News Headlines: Experimental Evidence from Mexico\*

## ([Check the latest version here.](#))

By DARIO TRUJANO-OCHOA<sup>†</sup> AND JOSE GLORIA<sup>‡</sup>

*This study explores the determinants of individuals' demand for verification. We measure this demand in a laboratory setting by eliciting participants' willingness to pay (WTP) to verify whether a headline is accurate or contains misinformation. The experimental design holds constant the payoffs from correct classification, the accuracy of the verification signal, and the baseline probability of misinformation. By abstracting from these elements—which jointly determine the value of verification in the field—we isolate the effects of feedback, political content, and participants' political positions on verification behavior. We analyze how both WTP and confidence respond to feedback and to the political content of headlines, and we compare observed verification demand to a theoretical benchmark in which verification is purely accuracy-motivated.*

*JEL: C93, D83, D91*

*Keywords: Misinformation, Feedback, Verification, Willingness to Pay, Confidence*

When we read information online, there is a decision problem: should we invest time and effort to verify its accuracy? This decision should depend on the cost of verification, the cost of relying on false information to form our opinions and make decisions, and how confident we are in our own initial criteria. This paper experimentally explores the determinants of how much people are willing to spent verifying headlines.

The spread of misinformation through digital platforms is faster and more profound than factual information (Vosoughi et al., 2018), and it is primarily people (not bots) who share misinformation inadvertently (Arin et al., 2023). False information, prevents making decisions based on true information; and at the society level, it erodes trust in institutions, and fuels polarization (Kavanagh and Rich, 2018). Verification by individuals is vital to curve the dangerous effects of mis-

\* The authors want to thank the professors who discussed the early versions of the present project: Gary Charness, Cesi Cruz, Daniel Martin, Ignacio Esponda, Ryan Oprea, Sevgi Yuksel, Erik Eyster, and to the specialists and professors in Mexico who increased the discussion and understanding of the relevant problem of misinformation: Grisel Salazar, Daniel Moreno, Horacio Larreguy, Antonio Arechar, Pablo Soto, and Arturo Bouzas. We also thank AlianzaMX for the fellowship that allowed the authors to travel to Mexico to develop this research. This research was approved the the UCSB Human Subjects Committee with the number 1-24-0532.

<sup>†</sup> UCSB, USA, [dariotrujanoochoa@ucsb.edu](mailto:dariotrujanoochoa@ucsb.edu).

<sup>‡</sup> UCLA, USA, [josegloria@ucla.edu](mailto:josegloria@ucla.edu).

information and is complemented by other efforts including debunking, platform fact-checking and the use of platform users to rate content are present.<sup>1</sup>

In this paper, we measured the demand for news verification.<sup>2</sup> Previous research mostly focuses on the effects of exogenous interventions and overlooks the value individuals place on verifying the accuracy of the news they receive. In this context, verification is a signal that informs if a statement is true or false.<sup>3</sup> In the field, there is large degree of heterogeneity in the value of having accurate information, which we control in the laboratory by paying a fixed amount for correct classification. Other studies have focus on the effects of different level of accuracy or confidence of the fact-checking process, we implemented a perfectly informative signal. This allows us to rule out individual differences in understanding of the accuracy of the signals in Bayesian updating.

Intuitively, the willing to pay to verify (WTP) should be higher when people are less confident in the veracity of the information they read. If someone knows that a headline they read is true (or false), verifying it would only make lose money, but if they consider that a headline is true or false with the same likelihood, the value of verification is the highest. We formalize this reasoning with an explicit framework for the decision problem of information classification, where individuals highly confident in classifying information correctly are predicted to reduce their demand for verification. This is consistent with a accuracy-motivated benchmark. However, we found evidence contradicting this prediction. Participants were more confident classifying political headlines and also more willing to pay to verify this headlines. Also, participants expressing support for the government in Mexico were more confident and willing to pay more. Analyzing in detail the relation between confidence and WTP we didn't find the negative relation expected in theory.

One of the objectives was to assess if a feedback intervention could impact individuals' value on verification. While interventions like debunking, inoculation, and digital literacy provide indirect feedback on accuracy when classifying headlines, this research directly measures the demand for verification in an experiment, examining how feedback on individual performance and on group performance alters this valuation. In general, the literature does not have a consensus on the impact of feedback on overconfidence. We found weak evidence of the effects of feedback on the willingness to pay for verification.

We designed the classification-verification task to measure individuals' demand

<sup>1</sup>The announcement from Meta at the beginning of 2025 about replacing fact-checkers with a model similar to Community Notes used in X, increases the relative importance of studying the users willingness to verify news headlines. Previously, Facebook (Meta) collaborated with fact-checking organizations to tag false information (<https://transparency.meta.com/policies/community-standards/misinformation>), and X uses Community Notes (<https://x.com/communitynotes?lang=en>) to let the users decide through an algorithm over their ratings.

<sup>2</sup>Fact-checking as a top-down policy aimed at identifying and disseminating information about fake news, whereas verification involves an individual's proactive choice to seek additional details.

<sup>3</sup>This paper focuses on headlines classified as true or false by professional fact-checkers. However, not every statement can be classified this way. For example, normative statements (e.g., "Headlines should prioritize impactful language to capture readers' attention.").

for verifying news headlines and their accuracy rate. The participants classified 50 headlines as 'true' or 'false' across five blocks, and reported their willingness to pay (WTP) to verify a headline's accuracy using a perfect signal. This allowed us to control for individual differences in belief updating (Trujano-Ochoa, 2024). Participants were randomly assigned to one of three groups: *control* (no feedback), *individual* feedback, and *others* feedback. Under the benchmark model, normative willingness to pay for verification is linked to the probability of a correct classification (confidence), while observed verification demand may reflect additional motives. The methodology isolated the effects of feedback on participants' accuracy, demand for verification, and confidence. The task was incentivized to ensure engagement and minimize differences in the importance people put on recognizing true and false news headlines. All participants were invited from two of the largest universities in Mexico City. The headlines had political and non-political content, and participants' attitudes about the current ruling party in Mexico were measured.<sup>4</sup>

To our knowledge, this is the first study that elicits the value of verifying individual headlines. It contributes to the literature by examining the impact of feedback on both confidence levels and the perceived value of verification, providing evidence that feedback on the accuracy of other participants can reduce this value. Also, by focusing on headlines from Mexico—the largest Spanish-speaking country where misinformation and polarization have recently increased—this research addresses a critical gap in the literature, which has predominantly concentrated on interventions studied in the U.S. and Europe (Kozyreva et al., 2024; Bateman and Jackson, 2024). The study's experimental design controls for participant heterogeneity, ensuring that its results are broadly applicable to diverse populations. Moving beyond the typical focus on exogenous interventions, this paper innovatively explores how feedback mechanisms influence individuals' active demand for verification, offering insights into how such mechanisms can enhance the effectiveness of fact-checking in combating misinformation.

## I. Previous Literature

Fact-checking and verification behaviors have been found effective in combating misinformation. Martel and Rand (2024) found that fact-checking reduced the belief in false headlines, analyzing 21 experiments; although the effect was larger among people with more trust in the fact-checkers. There is a possibility of fact-checking backfiring; however, the general effect is positive (Velez and Liu, 2025; Swire-Thompson et al., 2022), and it is recommended as an effective tool to fight false beliefs. Reviews by Kozyreva et al. (2024) and Bateman and Jackson (2024) recognized that promoting media literacy, fact-checking, and labeling content

<sup>4</sup>Political leaning in Mexico is better described by the attitude people have towards MORENA, the party that was reelected in 2024. This contrasts with liberal (Democrat)-conservative (Republican) politics in the US, since parties from different political perspectives formed coalitions against the incumbent party, MORENA.

are vital tools to counter misinformation. Other interventions, such as accuracy prompts, encourage users to pause and assess content accuracy before sharing. For instance, [Pennycook et al. \(2021\)](#) introduced an “accuracy nudge” intervention that reduced fake news spread by prompting users to reflect on accuracy. Also, [Pennycook and Rand \(2022\)](#) focused on interventions promoting cognitive engagement to reduce misinformation dissemination. These studies mostly focus on the exogenous exposure to fact-checking; however, this exposure is frequently endogenous. We refer to verification as endogenous demand for fact-checking, although these two concepts have been used interchangeably.

In a context of misinformation, overconfidence could be behind a low level of verification by discouraging further research once a headline has been decided to be considered true or false. People’s overconfidence has been shown to impact their ability to discern fake news, leading to greater engagement with false information. According to [Lyons et al. \(2021\)](#), individuals tend to overestimate their ability to distinguish between real and fake news. Similarly, [Pennycook and Rand \(2020\)](#) found that overconfidence in one’s cognitive abilities correlates with a higher likelihood of accepting false claims as accurate. [Ortoleva and Snowberg \(2015\)](#) found that overconfidence due to correlation neglect leads to higher polarization.

Beyond the domain of verification, the literature shows mixed results on feedback’s effects on overconfidence. While some incentivized studies have found feedback reduces overconfidence ([Ferraro, 2005](#); [Eberlein et al., 2011](#); [Kogelnik, 2022](#)), others in educational contexts report no effects ([Pulford and Colman, 1997](#); [Erat et al., 2022](#)). Moreover, some studies reveal feedback’s asymmetric effects on motivated beliefs. [Oprea and Yuksel \(2022\)](#) observed that feedback increases subjective probabilities of outperforming others, while [Thaler \(2024\)](#) noted that individuals update beliefs asymmetrically when feedback reinforces ego-related worldviews. [Kartal and Tyran \(2022\)](#) showed that overconfident participants continue voting despite low-accuracy signals. Finally, [Moore and Healy \(2008\)](#) distinguished between overestimation, overplacement, and overprecision, noting that feedback often has minimal effects on recalibrating overconfidence, particularly for difficult tasks where individuals overestimate their performance.

Two working papers are most closely related to the present project. First, the study by [List et al. \(2024\)](#) investigates the demand side of misinformation by testing whether critical-thinking interventions—videos and personality-based feedback—reduce individuals’ vulnerability to false content. Its contribution lies in showing that prompting individuals to “slow down” can decrease belief in fake headlines and modestly increase willingness to report misinformation. These results highlight the cognitive determinants of misinformation demand. The present paper differs from [List et al. \(2024\)](#) as it incorporates monetary incentives, measures willingness to verify information, and examines the role of confidence and political attitudes in shaping verification behavior. Second, the study by [Assenza et al. \(2024\)](#) focuses on belief formation, information processing, and how indi-

viduals update or misinterpret signals under uncertainty. This work is related because it provides a theoretical and empirical foundation for understanding why people might avoid or misweight corrective information, a core mechanism underlying misinformation verification decisions. In contrast, the present study extends beyond belief-updating frameworks by introducing incentivized choices and headline-level confidence elicitation, allowing a direct test of whether confidence, political alignment, and payoff structures causally affect willingness to pay for verification. Together, both papers are related to the current project but differ from it in methodological design, incentive structure, and the specific behavioral mechanisms under study.

## II. Hypotheses on the Effects of Feedback

This study is structured around three main hypotheses that explore the impact of feedback on willingness to pay (WTP) to verify the accuracy of the headlines and the confidence. Two other secondary hypotheses regarding the effects on accuracy are included. Under the assumption of rationality, participants' WTP should be proportional to the value they give to information. These hypotheses are integrated into the methodology to test their validity in a controlled environment.<sup>5</sup>

**HYPOTHESIS 1:** *Participants are generally overconfident and will expect better performance in classifying headlines than what is reflected in the feedback they receive.*

Following the literature on overconfidence, this hypothesis suggests that participants tend to overestimate their classification accuracy before receiving feedback. In this experiment, participants' predictions about their classification accuracy are expected to be higher than the accuracy indicated by their actual performance, as revealed by the feedback.

**HYPOTHESIS 2:** *Participants' willingness to pay (WTP) for verification will be higher when they receive feedback on others' classification accuracy than when they receive individual accuracy rate feedback.*

According to previous results on the asymmetric effect of feedback, participants who receive feedback on the performance of others could perceive this feedback as more informative and objective. As a result, they will place a larger value on the verification process and be willing to pay more to ensure the accuracy of the headlines they classify. The expectation is that group feedback, being less affected by motivated reasoning, will lead to a higher WTP as participants seek to mitigate the perceived difficulty of the task.

<sup>5</sup>This study was preregistered in OSF: <https://osf.io/jxr82>. We changed the order of the hypothesis for exposition purposes, leaving the hypothesis about accuracy at the end.

**HYPOTHESIS 3:** *Participants’ willingness to pay (WTP) for verification is influenced by the political content of the headlines, with differing effects depending on whether the headline favors or opposes the current government.*

- 1) *Supporters of the Government: Lower WTP for verification of favorable headlines; higher WTP for verification of unfavorable headlines.*
- 2) *Opponents of the Government: Higher WTP for verification of favorable headlines; lower WTP for verification of unfavorable headlines.*

When participants are presented with headlines that contain political content, it is hypothesized that their WTP for verification will vary depending on whether the headline aligns with their political beliefs. Specifically, if a headline is favorable to the current government, participants who support the government will likely have lower WTP for verification. This is because they are more inclined to accept information that aligns with their pre-existing beliefs without seeking further verification. Conversely, participants who oppose the current government may exhibit higher WTP for verification of favorable headlines, as they may be more skeptical of information that contradicts their beliefs and, thus, more motivated to confirm its accuracy.

On the other hand, for headlines unfavorable to the current government, supporters of the government may demonstrate higher WTP for verification, driven by a desire to challenge or disprove information that opposes their political views. Opponents of the government, however, may show lower WTP for verification of unfavorable headlines, as they may be more likely to accept information that aligns with their negative views of the government without the need for additional confirmation.

#### *A. Hypothesis on Accuracy*

**HYPOTHESIS 4:** *Feedback on individual performance improves the accuracy of headline classification compared to no feedback.*

**HYPOTHESIS 5:** *Feedback on others’ classification accuracy improves headline classification accuracy compared to no feedback.*

According to these hypotheses, giving participants feedback on their performance in classifying headlines will lead to higher accuracy in future classification tasks. The expectation is that personal feedback will encourage participants to adjust their behavior, thereby improving accuracy. This is related to previous studies showing that people improve when prompted on the importance of accuracy (Pennycook et al., 2021).

### **III. Theoretical Framework for Decision-Making in the Verification Decision**

This section describes the agent’s decision-making problem of classifying, verifying, and reclassifying a headline as accurate (true) or fake (false). Classifying

is a signal detection problem, and purchasing additional signals for this decision requires calculating the expected value of sample information. The instrumental value of information is assumed to be equal to the willingness to pay for verification. The preset setting provides a framework for understanding the decision-making process when deciding whether to verify headlines.

#### A. Problem Setup Without Purchasing a Signal

Consider an agent tasked with classifying headlines as accurate ( $t$ ) or fake ( $f$ ) ( $c \in \{t, f\}$ ). The state of the world is  $\omega \in \Omega = \{T, F\}$ , with the prior probability of encountering a fake headline denoted by  $p_f = P(\omega = F)$ .<sup>6</sup> Consequently, the prior probability of encountering an accurate headline is  $1 - p_f$ .

The agent's utility for correctly classifying a headline as accurate is  $U_T$ , and for correctly classifying a headline as fake is  $U_F$ . The utility for misclassifying a fake headline as accurate is  $U_{TF} < U_T$ , and for misclassifying an accurate headline as fake is  $U_{FT} < U_F$ . The condition is case-insensitive for evaluating the correct classification (i.e.,  $c = \omega$  means a correct classification).

The probability of correctly classifying the headline is determined by  $P(c = t|T)$  and  $P(c = f|F)$ . We assume that  $1 < \frac{P(c=t|T)}{P(c=t|F)}$  and  $1 < \frac{P(c=f|F)}{P(c=f|T)}$  to assure that the initial classification  $c$  is informative in the sense that the initial classification  $c$  gives information relative to the prior probability of each state  $\omega$ .<sup>7</sup> This is stated formally in the following proposition.

**PROPOSITION 1:** *Informativeness of the initial classification  $c$ .*

$$1 < \frac{P(c=\omega|\omega)}{P(c=\omega|\bar{\omega} \neq \omega)} \text{ if and only if } P(\omega|c \neq \omega) < P(\omega) < P(\omega|c = \omega)$$

Notice that a commonly found assumption  $0.5 < P(s = t|T) = q_t$  and  $0.5 < P(s = f|F) = q_f$  is sufficient to make the signal  $S$  informative according to proposition 1. The proof of this proposition can be found in the appendices.

The expected utilities when a headline is classified as accurate,  $EU_{\text{no signal}}(t)$ , and as fake,  $EU_{\text{no signal}}(f)$ , are given respectively by the equations:

$$EU_{\text{no signal}}(t) = P(T|t) \cdot U_T + P(F|t) \cdot U_{TF}$$

$$EU_{\text{no signal}}(f) = P(F|f) \cdot U_F + P(T|f) \cdot U_{FT}$$

The previous equations clearly show that if the headline's classification is informative, reading it is valuable because the expected utility is higher than when considering the prior probabilities alone.

<sup>6</sup>We simplify  $P(\omega = F)$  to  $P(F)$ . For  $c, s \in \Omega$ , we specify.

<sup>7</sup>We are assuming here that the classification is a signal to the same agent without considering the content of a headline  $h$  which is most likely multidimensional. This classification process also follows an optimization process where  $c = \omega \iff \frac{P(\omega|h)}{P(\omega|\bar{\omega} \neq \omega|h)} > \frac{U_T - U_{FA}}{U_F - U_{TF}}$ . However, in line with the objectives of the present research, we focus on analyzing the informativeness of the initial classification  $c$  in proposition 1, without examining the properties of the headlines or the payoffs.

### B. Conditional WTP Analysis

This section showed the optimal willingness to pay (WTP) for signal  $S$  after the initial classification. The WTP is the maximum amount an agent would be willing to pay to observe signal  $S$ .

The decision to purchase information is made after the agent observes a headline, once the signal has been classified. Therefore, the signal's value depends on  $c$ . We are assuming that sequential information acquisition is optimal. The problem of sequential decision-making was stated in general by Wald (1947), and Arrow et al. (1949) analyzed how to learn from sequential information.

This section presents the condition that makes verifying the initial classification valuable. After initially classifying the headline as accurate ( $c = t$ ) or false ( $c = f$ ), the agent can reclassify the headline  $r \in \{t, f\}$  based on the signal's realization  $s \in \{t, f\}$ . Let's consider first a valuable signal  $S$  with the conditions in proposition 2.

**DEFINITION 1:** *A signal is valuable if  $EU(r = s) \geq EU(r = c)$*

The informativeness of the signal is determined by  $P(s = t|T) = q_t$ . We need a strong enough signal  $S$  so that it is valuable and the optimal decision is to reclassify according to the signal ( $r = s \in \{t, f\}$ ). Also, we assume that the signal realization  $s$  is independent of the previous classification  $c$  conditional on the state of the world  $\omega \in \{T, F\}$  (i.e.  $P(s|\omega, c) = P(s|\omega)$ ).

**PROPOSITION 2:** *Conditions for Valuable Signal*

*A signal  $S$  is valuable if and only if*

$$\frac{P(\omega|s = \omega, c \neq \omega)}{P(\tilde{\omega}|s = \omega, c \neq \omega)} < \frac{U_\omega - U_{\tilde{\omega}\omega}}{U_{\tilde{\omega}} - U_{\omega\tilde{\omega}}} \equiv U_\omega$$

*with  $\tilde{\omega} \in \Omega, \tilde{\omega} \neq \omega$ .*

We are also assuming that the initial classification is valuable and therefore follows the analogous condition  $\frac{P(\omega|c=\omega)}{P(\tilde{\omega}|c=\omega)} < U_\omega$ .

By allowing the agent to update its classification based on the signal, we account for dynamic decision-making. The WTP to verify is derived by comparing the expected utility with the signal (considering reclassification) to the expected utility without the signal. This approach shows how additional information improves decision-making accuracy. The detailed mathematical steps and proofs are provided in the appendix. The expected utility of reclassification is calculated by updating the agent's posterior beliefs using Bayes' rule and comparing the expected utilities with and without reclassification.

For an agent tasked with classifying headlines as accurate or fake, a signal  $S$  indicating the state of the world must be sufficiently strong to ensure that the agent reclassifies based on this signal.

## WTP EQUATION

Initially, the agent classifies a headline as either accurate ( $t$ ) or fake ( $f$ ). The agent updates their beliefs upon receiving a signal  $s$ , which can either confirm or contradict the initial classification. The posterior probabilities are calculated using Bayes' rule. For example, the posterior probability of the headline being accurate given the signal  $s = t$  and the initial classification  $c$  is:

$$P(T|s = t, c) = \frac{q_t \cdot P(T|c)}{q_t \cdot P(T|c) + (1 - q_f) \cdot P(F|c)}$$

Similarly, the posterior probability of the headline being fake given the signal  $s = f$  and the initial classification  $c$  is:

$$P(F|s = f, c) = \frac{q_f \cdot P(F|c)}{(1 - q_t) \cdot P(T|c) + q_f \cdot P(F|c)}$$

The agent's decision to reclassify based on the signal depends on the expected utilities. The expected utility of reclassification given the signal  $s = t$ , or  $s = f$ , are respectively:

$$EU_{\text{new classification}}(s = t, c) = P(T|s = t, c) \cdot U_T + P(F|s = t, c) \cdot U_{TF}$$

$$EU_{\text{new classification}}(s = f, c) = P(F|s = f, c) \cdot U_F + P(T|s = f, c) \cdot U_{FT}$$

The combined expected utility of updating the signal, considering both possible signals, is:

$$\begin{aligned} EU_{\text{signal}}^{\text{update}}(c) &= P(s = t|c) \cdot EU_{\text{new classification}}(s = t, c) + \\ &\quad P(s = f|c) \cdot EU_{\text{new classification}}(s = f, c) \\ &= [q_t \cdot P(T|c) + (1 - q_f) \cdot P(F|c)] \cdot [P(T|s = t, c) \cdot U_T + P(F|s = t, c) \cdot U_{TF}] + \\ &\quad [(1 - q_t) \cdot P(T|c) + q_f \cdot P(F|c)] \cdot [P(F|s = f, c) \cdot U_F + P(T|s = f, c) \cdot U_{FT}] \end{aligned}$$

The WTP to verify the headline is calculated by comparing the expected utility with the signal to the expected utility without the signal:

$$V(c) = EU_{\text{signal}}^{\text{update}}(c) - EU_{\text{no signal}}(c)$$

## C. Simplifying Assumptions

Let's assume equal prior probabilities  $p_f = 0.5$  and equal utilities  $U_T = U_F = 1$  and  $U_{TF} = U_{FT} = 0$ . Also, assume that the prevalence of fake and accurate news is the same  $P(T) = P(F) = p_f = 0.5$ . These assumptions about the payoffs allow us to interpret the signal's value purely in probability terms related to its informativeness. Substituting these assumptions into the expected utility

equations, we get:

$$EU_{\text{no signal}}(t) = P(T|t) \cdot 1 + P(F|t) \cdot 0 = P(T|t) = \frac{P(t|T)}{P(t|T) + P(t|F)}$$

$$EU_{\text{no signal}}(f) = P(F|f) \cdot 1 + P(T|f) \cdot 0 = P(F|f) = \frac{P(f|F)}{P(f|T) + P(f|F)}$$

And the expected utilities simplify to:

$$EU_{\text{new classification}}(s = t, c) = P(T|s = t, c)$$

$$EU_{\text{new classification}}(s = f, c) = P(F|s = f, c)$$

Finally, the combined expected utility of updating the signal, considering both possible signals, is:

$$EU_{\text{signal}}^{\text{update}}(c) = [q_t \cdot P(T|c) + (1 - q_f) \cdot P(F|c)] \cdot P(T|s = t, c) + [(1 - q_t) \cdot P(T|c) + q_f \cdot P(F|c)] \cdot P(F|s = f, c)$$

#### PERFECT SIGNAL

Here, we calculate the WTP considering the condition  $q_f = q_t = 1$ ; perfect signal. This assumption ensures that the signal is strong enough to follow even without the other simplifying assumptions and substantially simplifies the interpretation of  $V(c)$ . For the case  $c = f$  and  $s = t$ :  $q_t \cdot P(T|f) > (1 - q_f) \cdot P(F|f) \iff P(T|f) > 0$ . The case  $c = s$  and  $s = f$  requires  $P(F|t) > 0$ . Both conditions are satisfied by the assumption. The perfect signal assumption simplifies the expected utility of observing the signal  $S$ . Thus, the combined expected utility with the signal is:

$$EU_{\text{signal}}^{\text{update}}(c) = P(T|c) \cdot 1 + P(F|c) \cdot 1 = P(T|c) + P(F|c) = 1$$

Therefore,

$$(1) \quad V(c) = \begin{cases} 1 - P(T|t), c = t \\ 1 - P(F|f), c = f \end{cases}$$

The value of the signal  $S$  is the difference between the posterior probability of reclassifying correctly after observing the signal and the posterior probability of initially classifying correctly. Therefore, we can interpret the signal value as a function of the probability of a correct initial classification. Notice that if we change the payoff of a correct answer such that  $U_T = U_F > U_{TF} = U_{FT}$ , we only have to multiply the posterior probabilities difference by  $\pi = U_T - U_{TF}$  to get

$V(c)$ . Therefore, the willingness to pay to verify should be:

$$(2) \quad WTP(c) = \pi V(c) = \pi(1 - P(\omega|c = \omega))$$

Under the benchmark model, normative WTP is fully determined by the agent's posterior probability of a correct classification  $P(\omega|c = \omega)$ .

#### *D. Theoretical Framework as a Benchmark for the Empirical Results*

The theoretical framework developed in this section provides a normative benchmark for verification demand under the assumption that individuals value information solely for its contribution to classification accuracy. Under this benchmark, verification is purchased only to the extent that it improves decision quality, and the willingness to pay for verification is fully determined by the decision-maker's posterior probability of having classified a headline correctly. In this setting, the model implies that, under purely accuracy-motivated verification, higher confidence reduces the expected informational gain from verification and therefore lowers the normative willingness to pay.

Importantly, the framework is not intended as a maintained hypothesis about observed behavior, but rather as a reference point against which empirical verification demand can be evaluated. Relative to this benchmark, systematic deviations in elicited willingness to pay—such as weak, null, or positive relationships between confidence and verification demand—indicate that verification may carry value beyond its instrumental role in improving accuracy. The empirical analysis that follows uses this benchmark to assess the extent to which observed verification behavior reflects additional motives, including psychological, identity-related, or context-specific considerations.

To interpret verification decisions in the experiment, it is useful to distinguish between normative willingness to pay for verification and elicited willingness to pay. Normative willingness to pay (WTP) refers to the value of verification implied by the theoretical framework under purely accuracy-motivated decision making. With a perfectly informative signal, normative WTP is fully determined by the posterior probability of a correct classification (Confidence). By contrast, elicited WTP refers to the verification demand observed in the experiment, measured as the amount participants state they are willing to pay to verify a headline. In particular, observed verification demand may reflect psychological reassurance, preferences for confirmation, identity-related payoffs, or other non-instrumental motives, and is therefore treated throughout the paper as a behavioral outcome whose determinants are examined empirically rather than imposed by the model.

#### IV. Experimental Design: Classification-Verification Game

Participants are tasked with categorizing headlines as accurate or fake and reporting their WTP for verification. They evaluated 50 headlines in five blocks. The first three blocks contained non-political headlines, and the last two blocks contained political headlines. Feedback was provided according to the experimental condition the participant was randomly assigned to: *control* (no feedback), *individual* feedback, and *others* feedback. The design measures participants' classification accuracy, confidence, and willingness to pay for verification. All this is framed in terms of the decision-making problem presented in section III with a perfect signal. After the main task, participants complete a demographics and political orientation survey.

##### A. Experimental Blocks and Feedback Treatments

Participants completed five blocks of 10 headlines. Each block contained the same set of headlines and was presented randomly within the block for each of the 14 sessions. Each headline corresponds to one instance of the classification-verification game (round); they were instructed to classify as either *true* ( $t$ ) or *false* ( $f$ ), with an equal prior probability of each state ( $P(T) = P(F) = 0.5$ ), and reported their willingness to pay to verify the headline. To incentivize participants to classify accurately, participants earned 10 Mexican Pesos (MXN) in that block for each correctly classified headline regardless of whether the classification was *true* or *false* ( $U_T = U_F = 10$  MXN). Conversely, misclassifications, whether mistakenly classifying an accurate headline as *false* or a fake headline as *true*, yielded a payoff of 0 MXN ( $U_{FT} = U_{TF} = 0$  MXN). The classification and WTP for each headline decision had a 20-second time limit.

For each headline, participants also indicated their willingness to pay (WTP) to access a perfect signal ( $S$ ) that could reveal the headline's actual status. The perfect signal, available for purchase, would reveal the actual state of each headline with certainty ( $P(s = t|T) = q_t = 1$  and  $P(s = f|F) = q_f = 1$ ). The actual state of the headline was never revealed; if the signal was purchased, it added 10 MXN to the payoffs of the block independently of the initial classification made.

After completing the ten classifications and WTP decisions in a block, participants reported their estimated probability of correctly classifying the headlines. They indicated the probability of correctly classifying a headline that they reported as *true*, and *false*. Participants also reported these probabilities for others' classifications. There was no time limit when participants reported their confidence. This provided a direct self-assessed confidence measure at the block level.

After estimating their probabilities at the end of each block, participants received feedback according to the treatment group to which they were randomly assigned. Feedback treatments were designed to inform participants about their classification performance, either individually or relative to others, with the con-

### Headline Number 18

Time left to complete this page: 0:01

Please classify the following headline: (If your classification is correct, you could earn an extra 10 MXN.)

#### Iran Censored the Olympics; All Women Appear with Rectangles or Asterisks Covering Them

Your Classification:

☐ The information is accurate ☒ Contains false information

How much are you willing to pay to verify this news?

1.5

Next

FIGURE 1. SCREENSHOT OF THE TRANSLATED CLASSIFICATION-VERIFICATION GAME AS SEEN BY THE PARTICIPANTS.

trol group as a reference. The feedback types and descriptions are presented in table 1. By block, all treatment groups were shown a summary of the times they classified a headline as *true* or *false*, and the feedback treatments were shown the accuracy rates conditional on the headlines classified as *true* or *false*.

TABLE 1—FEEDBACK TREATMENTS

Treatment Group	Feedback at the End of the Block
Control Group	No feedback on accuracy was given.
Individual Feedback	Personal accuracy rate for the block, conditional on the headlines participants classified as <i>accurate</i> or <i>fake</i> .
Others Feedback	Average accuracy rate of other participants conditional on the headlines others classified as <i>accurate</i> or <i>fake</i> .

This structure allowed researchers to observe how different feedback types influenced participants' accuracy, confidence, and valuation of the verification signal throughout the experiment.

### B. Experimental Procedures

The participants were 192 undergraduate students in Mexico. The average age was 20, and 55% were women. They were recruited from UNAM (National Autonomous University of Mexico) and IPN (National Polytechnic Institute),

the first and second most important public schools in Mexico<sup>8</sup>. The experiment occurred at the schools where participants were studying in September 2024.

Participants receive a utility of 10 Mexican Pesos (MXN) for each correctly classified headline, whether it is accurate or fake ( $U_T = U_F = 10$  MXN). Conversely, they receive a utility of 0 MXN for misclassifications, whether they mistakenly classify an accurate headline as fake or a fake headline as accurate ( $U_{TF} = U_{FT} = 0$  MXN). This setup incentivizes participants to classify accurately and to value the signal appropriately based on its accuracy-assurance potential.<sup>9</sup>

After all blocks, participants complete a survey that collects demographic data and assesses their support for the current government. At the end of the experiment, one block is randomly selected for payment, and participants receive 10 MXN for each correctly classified headline in that block. Thus, final earnings are linked directly to classification accuracy.

## V. Methodological Considerations

This experiment controls for variables relevant to verification demand. The key parameters and assumptions—such as the equal prior probabilities, equal utilities, and a perfect signal—simplify the problem and allow for a focus on the probabilistic aspects of classification and verification. This section presents the description of the methods used in the experiment.

### A. Willingness to Pay for Verification and Confidence

We used the BDM mechanism to measure participants' WTP for verification (Becker et al., 1964). This was presented as a second-price auction against a computer randomly choosing numbers from 0 to 5. To participate in the auction, participants chose a number between 0 and 5 in increments of 0.5 from a drop-down menu. The upper bound of 5 MXN was selected because this is the expected value of the perfect signal,<sup>11</sup> only considering the prior we also tried to minimize the range of the potential values since large intervals could reduce effort in the task (Mamadehussene and Sguera, 2023). If they win the auction, the signal is

<sup>8</sup>In the national ranking considering all universities, UNAM is the most important university, and IPN can be ranked third (<https://www.usnews.com/education/best-global-universities/mexico>) or forth (<https://www.topuniversities.com/university-rankings-articles/world-university-rankings/best-universities-mexico>), depending of the ranking.

<sup>9</sup>The incentive amounts were significant for the participants. They were paid 10 Mexican Pesos (MXN) per correctly classified headline from a randomly selected block. This is approximately 0.5 USD per headline. One participant could make between 0 and 100 MXN from the classification-verification game. In Mexico the average salary of a profesionist is 53.5 MXN.<sup>10</sup> Then, given that this task took less than an hour (including instructions and a survey) and that the participants were undergraduates, the money they received was enough.

<sup>11</sup>Five MXN is the expected value of classifying randomly; either because the headline is uninformative, or there was no effort in thinking about the headline. Risk aversion increases the value in this case; however, around 60% of participants never choose this option, and only 16% of all rounds had a WTP equal to 5.

verified, and they win 10 MXN independently of the classification they made of the headline. If they lose the auction by betting less than the computer, their payoff will depend only on their initial classification.

Cason and Plott (2014) found that the BDM mechanism can fail to elicit the value of \$2. However, this mechanism has been shown to be effective for measuring WTP in the field when providing comprehension tests (Burchardi et al., 2021). To reduce the possibility of misunderstandings, participants completed a comprehension test reporting their payoffs across different scenarios, and they couldn't continue with the experiment until they answered all questions correctly. They also had four training sessions where the WTP process was explained, and they received feedback on whether they managed to purchase the verification or not (no feedback was provided in the following rounds).

### B. Perfect Verification

The methodology presented here simplifies the news verification problem, enabling us to explore the causal effects of feedback on verification's value. The perfect signal is a deliberate abstraction to isolate valuation rather than belief updating.

We present a perfect signal to avoid motivated misinterpretation of the signal's likelihoods and to make the WTP a function of confidence levels only. The results on motivated inference from Thaler (2024) are significant, as a partially informative signal can lead to biased updating and decrease WTP for the signal. Even online searching of news, which is a common practice promoted in digital literacy programs, could backfire (Aslett et al. (2024); Hoes et al. (2023)). Finally, it has been shown that higher levels of conservatism predict lower WTP for signals (Trujano-Ochoa, 2024), which complicates the interpretation of individual differences in their WTP if the signal is partially informative.

### C. Headlines Selection

In the experiment, participants classified 50 headlines separated into five blocks of 10 headlines. We tested 60 headlines on Prolific to select 50 that generated similar accuracy rates between the five blocks and to provide the feedback in the *Others Feedback* treatment.

To select the headlines to test, we used the two most recognized fact-checking efforts in Mexico (Sánchez and Pereyra-Zamora, 2022). The online publication AnimalPolitico<sup>12</sup> and VerificadoMX<sup>13</sup> were used as the sources to find relevant fake news circulating in Mexico. The methodology they used can be reviewed on their websites. In summary, they have a team of journalists searching for popular false publications on social media and publishing evaluations of the content on

<sup>12</sup><https://animalpolitico.com/verificacion-de-hechos>

<sup>13</sup><https://verificado.com.mx/>

their websites. NewsGPT<sup>14</sup> was used to find headlines that were real but difficult to classify. All the headlines generated were verified independently by the authors. At this stage, we first selected 60 headlines: 30 political and 30 non-political, half true and the other half false. Also, based on the political headlines, 15 were classified as information favoring the government, and 15 as critics of the government.

To choose these headlines, we ran a study in Prolific among Mexicans whose first language was Spanish. We asked for the classification of the 60 headlines with the same structure and incentives as in the final experiment, and estimated the rate of each headline being classified correctly. The headline composition of the blocks was made so that they have similar difficulty levels. The list of 50 headlines used in the experiment can be found in the appendix table B.B2. The political headlines were separated from the non-political headlines to evaluate the possible effect of the headlines' content on confidence levels.

#### D. Political position

In the exit survey, two questions were used to determine the participant's political position: 1) Who did you vote for? and 2) if the approved work made by AMLO<sup>15</sup> as president. If a participant answered "Morena"<sup>16</sup> to the first question and "Agree" to the second, that person was classified as a "supporter of the government". If a participant voted for any party other than the government and disagreed with the statement, they were classified as "opposing the government". Polarization in Mexico has increased, just as in the US. However, the primary division is in options about AMLO and the political party he founded to run for president (Morena)<sup>17</sup>. This is an opinion shared in traditional media<sup>18</sup>, academics, and journalists<sup>19</sup>. The experiment took place in September 2024, more than three months after the Mexican Presidential election.

#### E. Feedback by Block

The confidence elicitation and feedback provision was designed to occur at the end of each block of 10 headlines to avoid a loss of confidence in the signal and to provide feedback immediately after the confidence was elicited. If the signals contradicted prior beliefs, participants could have lost confidence in their accuracy.

<sup>14</sup>The request was made in August, three weeks before the start of the first session: <https://chatgpt.com/g/g-NnU2wmnZ5-news-gpt-chat-with-hundreds-of-news-sources/c/7e750031-b534-481c-83cf-2dc6917d98b4>

<sup>15</sup>This is a common way that people and media use to refer to Andres Manuel Lopez Obrador. The president of Mexico from 2018 to 2024.

<sup>16</sup>This is the name of the incumbent party during the federal elections in 2024.

<sup>17</sup><https://apnews.com/article/mexico-election-polarized-divided-heat-violence-4d5f620f0f8f9b7ef6efa8b3083561a8>

<sup>18</sup><https://apnews.com/article/mexico-election-polarized-divided-heat-violence-4d5f620f0f8f9b7ef6efa8b3083561a8>

<sup>19</sup><https://www.eluniversal.com.mx/tendencias/la-reflexion-de-denise-maerker-sobre-las-elecciones-2024-que-se-volvio-viral/>

Motivated inference among participants, as reported by [Thaler \(2024\)](#) and [Oprea and Yuksel \(2022\)](#), could lead them to infer that the signal was wrong. The bias should be extreme to affect the results since the signal is told to be perfect, but providing feedback at the aggregate level prevented the development of disbelief in the signal when negative feedback was provided on political headlines. Finally, eliciting confidence before providing feedback at the end of each block facilitated the comparison between expected and actual accuracy rates; both were at the block level, and one immediately after the other.

Confidence is elicited at the block level, preventing headline-level analysis of confidence vs. WTP. Future research will elicit headline-level confidence to measure more precisely the relationship between confidence and verification.

## VI. Results

Participants were 195 undergrad students from Mexico City. In total, seven participants (3% of the original sample) were excluded from the following analysis. We excluded 5 participants (2.5% of the original sample) because they answered less than 25% of the rounds. Also, one participant was excluded because all the confidence reported was 50% in each block and answered 5\$ in every round, which proved a lack of understanding in the experiment. Table 2 shows the average values of the essential variables grouped by treatment. The first four variables are participants' characteristics; ages didn't change much across treatments, and the proportion of female participants was slightly higher in the Individual group, but overall, the groups were balanced. The results in the accuracy and correctness suggest underestimation and overplacement (following [Moore and Healy \(2008\)](#)'s classification). In aggregate, participants classified the headlines as *true* 50.1% of the rounds, and they classified 61.1% correctly. The proportion of classifying a headline correctly was 61.0% and 61.1% when they classified a headline as *false* and *true*, respectively.

To analyze the effects on confidence, willingness to pay (WTP), and accuracy, we ran a regression analysis. The standard errors were clustered at the level of participant to account for the within-participant correlation between their decisions. The effects on willingness to pay, and accuracy can be observed in table 3, and the effects on confidence 4. The regressions are separated because those in table 3 were run at the level of rounds, while the regressions in table 4 were run at the level of blocks. This difference is because the confidence was only measured at the end of a block of 10 rounds. In these regressions, the first block was dropped to analyze the effects after feedback. Finally, the number of observations is not a multiple of 5 because in some rounds, participants didn't provide an answer in the 20 seconds they had to.

### A. Willingness to Pay (WTP) for Verification

Regression 1 of table 3 shows the statistically significant positive effect of classifying a headline as *true* ( $p$ -value < 0.01). This indicates a possible verification

TABLE 2—SUMMARY BY TREATMENT OF THE MAIN VARIABLES. THE AVERAGE (PROPORTION) OF EACH VARIABLE IS PRESENTED FOR EACH TREATMENT.

Variable	Control	Individual	Others
Age	20	20.1	20.2
Male	0.516	0.319	0.435
Support Gov	0.203	0.232	0.161
Oppose Gov	0.156	0.203	0.194
Missing Headlines	0.033	0.026	0.055
Accuracy Estimate	0.574	0.525	0.542
Accuracy Estimate Others	0.526	0.509	0.495
Accuracy	0.618	0.603	0.594
Classification ( $c = a$ )	0.492	0.505	0.508
WTP	2.81	2.65	2.46
N Participants	64	69	62

bias when participants classify something as *true*. If the verification bias were balanced, and participants had the same bias when they believe a headline is *false* the coefficient would be zero. Therefore, this positive coefficient is evidence of an unbalanced confirmation bias towards positive statements. This result can also be observed in figure 2, where the distribution of WTP for the *false* classification first-order stochastically dominates the distribution of WTP in *true* classification in each treatment. Finally, this result is maintained when only considering the political headlines in blocks four to five (regressions 3 and 4 in table 3).

RESULT 1: *Participants exhibited a **greater** willingness to pay (WTP) to verify the information they classified as true.*

Also, in regression 1 of table 3, it can be observed that receiving feedback on **others' accuracy feedback** lowered participants' WTP to verify classifications at a 10% level of significance. This reduction in demand for verification suggests that collective feedback diminishes the perceived need for individual verification, possibly because people receive feedback on accuracy rates higher than they expected. This result can be observed in figure 3, where the distribution of WTP for the control group first-order stochastically dominates the distribution of WTP in the "Others" treatment. According to the Kolmogorov-Smirnov test, we can reject the hypothesis of identical distributions between the *Control* and *Others* treatments ( $p - value < 0.01$ ). There was no difference between the *Individual* and *Others* treatments ( $p - value = 0.3973$ ).

RESULT 2: *Participants who received feedback on the performance of others were **less** willing to pay to verify headlines.*

TABLE 3—REGRESSION ON THE ACCURACY AND WILLINGNESS TO PAY AMONG THE POLITICAL HEADLINES TO ANALYZE INTERACTIONS. THE SE WERE CLUSTERED AT THE INDIVIDUAL LEVEL AND THE FIRST BLOCK WAS EXCLUDED. THE LAST TWO REGRESSIONS ONLY CONSIDER THE POLITICAL HEADLINES IN THE LAST TWO BLOCKS

	<i>Dependent variable:</i>			
	WTP	Accuracy	WTP Pol.	Accuracy Pol.
	(1)	(2)	(3)	(4)
Individual Feedback	−0.194 (0.212)	−0.014 (0.014)	−0.213 (0.229)	−0.017 (0.017)
Others Feedback	−0.351* (0.204)	−0.011 (0.014)	−0.405* (0.223)	−0.034* (0.019)
Round	0.003 (0.002)	0.0001 (0.001)	0.004 (0.004)	−0.012*** (0.001)
Accurate	−0.057 (0.041)		−0.043 (0.057)	
'True' (c = t)	0.210*** (0.054)	0.001 (0.004)	0.204*** (0.066)	0.212*** (0.006)
Age	−0.027 (0.050)	0.004 (0.003)	−0.040 (0.051)	0.001 (0.004)
Male	−0.061 (0.176)	0.011 (0.012)	−0.028 (0.191)	0.015 (0.015)
Political	0.128** (0.054)	0.039* (0.022)		
Gov. Supporter	0.483** (0.210)	0.037** (0.015)	0.419* (0.248)	0.011 (0.023)
Favor Gov. News			−0.042 (0.044)	−0.207*** (0.016)
Gov. Critic	0.220 (0.230)	0.003 (0.014)	0.211 (0.248)	−0.067*** (0.023)
Supporter X Favor			0.074 (0.085)	0.008 (0.032)
Critic X Favor			0.089 (0.077)	0.099*** (0.031)
Constant	3.032*** (1.005)	0.506*** (0.064)	3.387*** (1.045)	1.099*** (0.099)
Observations	7,572	7,572	3,805	3,805
R <sup>2</sup>	0.030	0.003	0.028	0.091
Adjusted R <sup>2</sup>	0.029	0.002	0.025	0.088

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

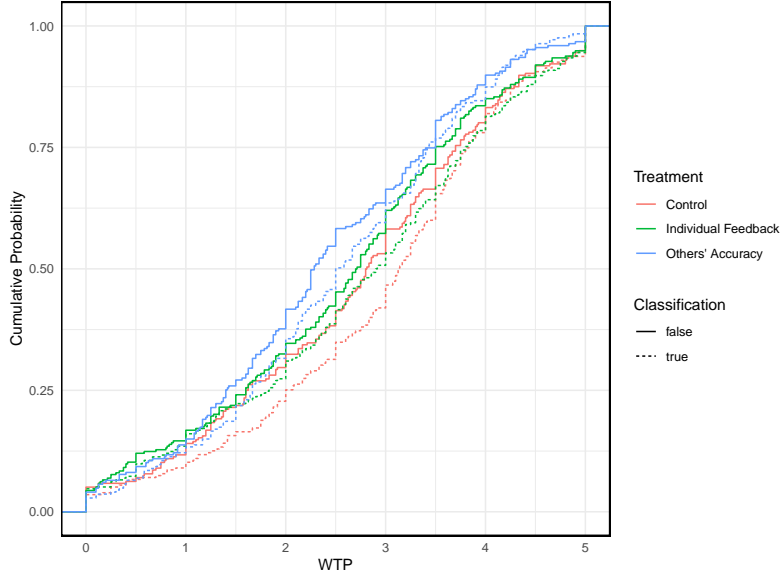


FIGURE 2. EMPIRICAL CDF OF THE WILLINGNESS TO PAY BY TREATMENT AND CLASSIFICATION OF THE HEADLINE. TO CREATE THIS GRAPH, THE AVERAGE WTP PER BLOCK WAS CALCULATED.

### B. Confidence and Overconfidence

While the normative benchmark, under purely accuracy-motivated verification, implies a negative relationship between confidence and verification demand, the empirical relationship is weak or positive. This divergence motivates the interpretation that verification carries value beyond accuracy improvement.

In regression 1 and 2 of table 4, the levels of confidence are explained by the treatments, participants' characteristics, and headline properties. Confidence was measured by the reported  $\hat{P}(T|t)$  and  $\hat{P}(F|f)$  at the end of each block. The classification as *true* was averaged by block, representing the proportion of headlines classified as *true*. Regression 2 and 3 excluded 17 participants who always chose 50% as their level of confidence.<sup>20</sup> Finally, regression 3 in table 4 shows the results for confidence in the accuracy of another participant.

The *Individual* treatment had a negative effect on confidence, statistically significant at 10%. *Others* treatment had a negative but not significant effect. If the headlines were political, and if the participant was a supporter of the current government, they would be clearer predictors of the level of confidence with effects statistically significant at 5%. All the effects were preserved when participants

<sup>20</sup>These participants explained around a third of the rounds where a confidence of 50% was indicated, which was chosen in 25% of all rounds.

TABLE 4—REGRESSION ON THE CONFIDENCE. THE SE WERE CLUSTERED AT THE INDIVIDUAL LEVEL AND THE FIRST BLOCK WAS EXCLUDED. REGRESSION 2 EXCLUDES THE PARTICIPANTS WHO ALWAYS CHOSE 0.5.

	<i>Dependent variable:</i>		
	Confidence	Confidence	Others
	(1)	(2)	(3)
Individual Feedback	−5.176* (2.982)	−5.300* (3.216)	−3.736 (2.999)
Others Feedback	−3.796 (2.983)	−3.999 (3.257)	−4.267 (3.167)
Block	−0.143 (0.737)	−0.150 (0.806)	−0.019 (0.653)
'True' (c = t) prop.	0.786 (0.973)	0.853 (1.066)	1.429 (0.899)
Age	0.990* (0.542)	0.980* (0.574)	0.397 (0.524)
Male	2.968 (2.419)	2.972 (2.655)	0.886 (2.545)
Political	3.251** (1.454)	3.559** (1.592)	1.380 (1.394)
Gov. Supporter	7.139** (3.269)	7.134** (3.487)	3.929 (3.058)
Gov. Critic	5.582* (3.118)	5.746* (3.373)	2.143 (3.250)
Constant	32.041*** (11.176)	32.428*** (11.861)	43.146*** (10.246)
Observations	1,553	1,418	1,418
R <sup>2</sup>	0.053	0.053	0.021
Adjusted R <sup>2</sup>	0.047	0.047	0.014

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

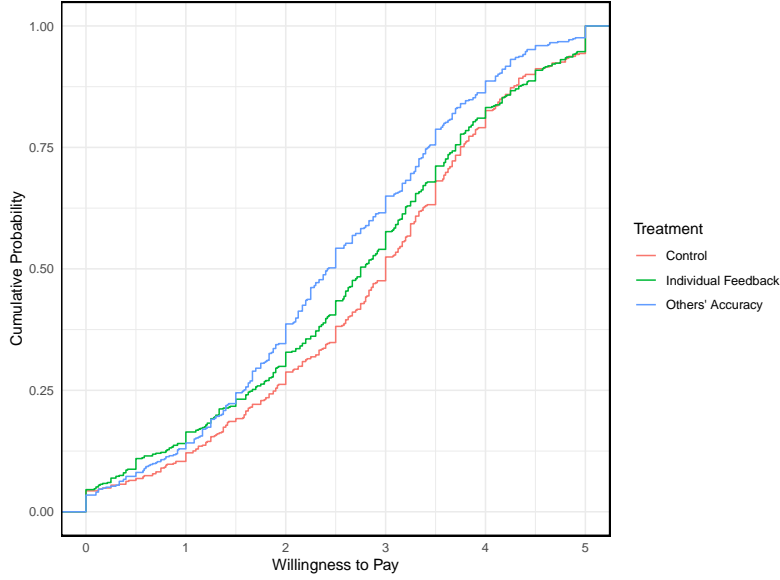


FIGURE 3. EMPIRICAL CDF OF THE WILLINGNESS TO PAY BY TREATMENT. TO CREATE THIS GRAPH, THE AVERAGE WTP PER BLOCK WAS CALCULATED.

who were always unconfident were excluded in regression 2. In the regression on the confidence in others' accuracy (regression 3, table 4), the direction of the effects was preserved, but no variable had a significant effect.

At the end of each block, participants were asked to report the probability they believe their classifications were correct when they classified a headline as true  $\hat{P}(T|t)$  and false  $\hat{P}(F|f)$ , which measures their confidence. Overconfidence was assessed by comparing the participants' reported probabilities of correct classification with the actual probabilities derived from the task. An overall measure of overconfidence was calculated as  $\hat{P}(c = \omega) - P(c = \omega)$ , where  $c$  is the classification and  $\omega$  is the true state of the world; a positive number of this measure means overconfidence and a negative number underconfidence. Considering the average overconfidence per participant, the hypothesis of no difference from zero is rejected ( $p$ -value  $< 0.001$ ) with a difference of  $-7.15$ . The same result is found when analyzing the differences by initial classification:  $\hat{P}(F|f) - P(F|f) = -4.51$  and  $\hat{P}(T|t) - P(T|t) = -8.57$  (both with  $p$ -value  $< 0.001$ ). These results align with the data in table 2 showing that the average confidence was below the accuracy rates.

Since underconfidence was more prevalent, we should expect that the feedback treatments increase confidence. However, figure 5 shows that confidence decreased in blocks 2 and 3 for the *Others* treatment. In the political headlines,

both feedback treatments had similar levels of confidence. However, the Control treatment had an upward trend in the last two blocks that included the political headlines. The second block was the most difficult, as can be observed in figure 4. Underconfidence and a negative effect of feedback suggest that participants were reacting asymmetrically; overreacting to negative feedback. This could be due to the avoidance of bad news about their performance, which occurs only in the **Individual** treatment.

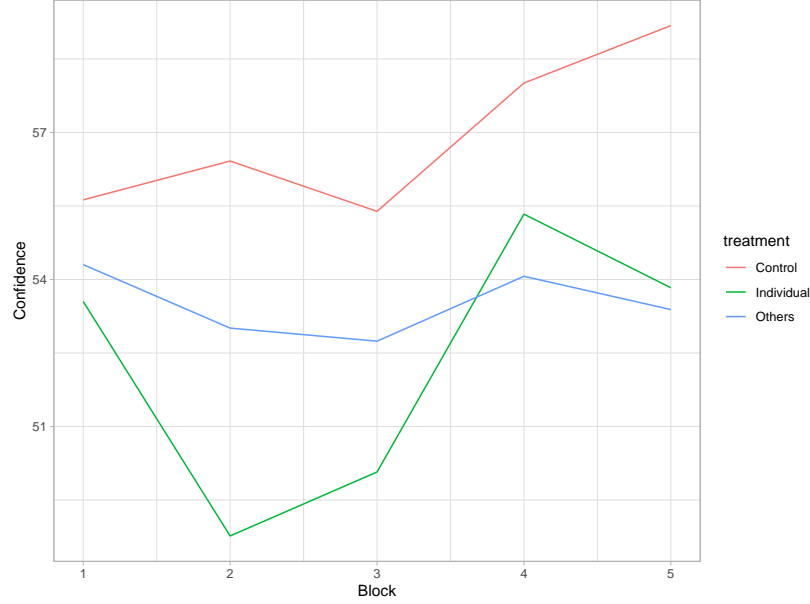


FIGURE 4. AVERAGE OVERCONFIDENCE IN EACH BLOCK BY TREATMENT.

**RESULT 3:** *Contrary to hypothesis 1, overconfidence in headline classification, participants were not overconfident.*

### C. Effects of Political Headlines and Political Position

Political content increases WTP and confidence, but political alignment does not significantly moderate WTP.

Figure 7 illustrates the effects of political variables on the average willingness to pay to verify (WTP) headlines. We can observe a larger WTP among participants who expressed support for the incumbent government, and a consistently larger WTP when the headline was political. The figure echoes the results found in the first regression of table C1.

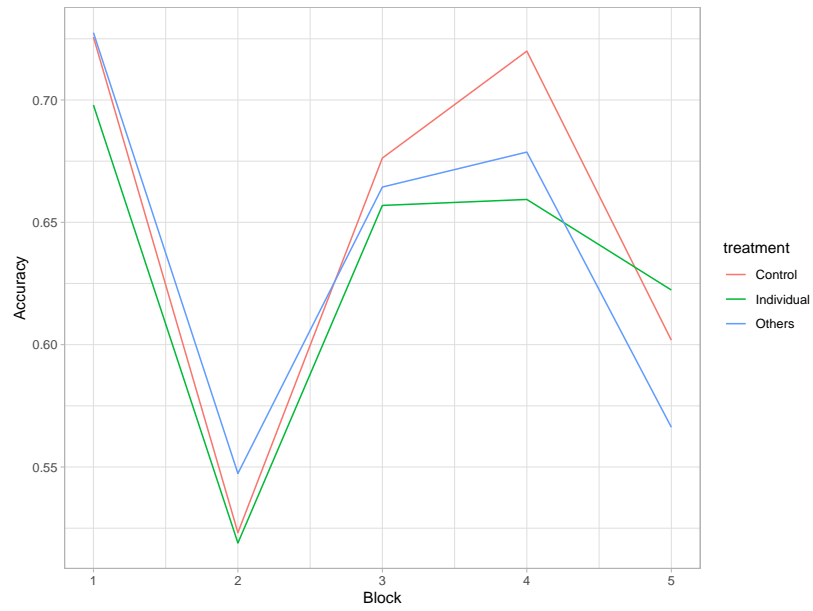


FIGURE 5. AVERAGE OVERCONFIDENCE IN EACH BLOCK BY TREATMENT.

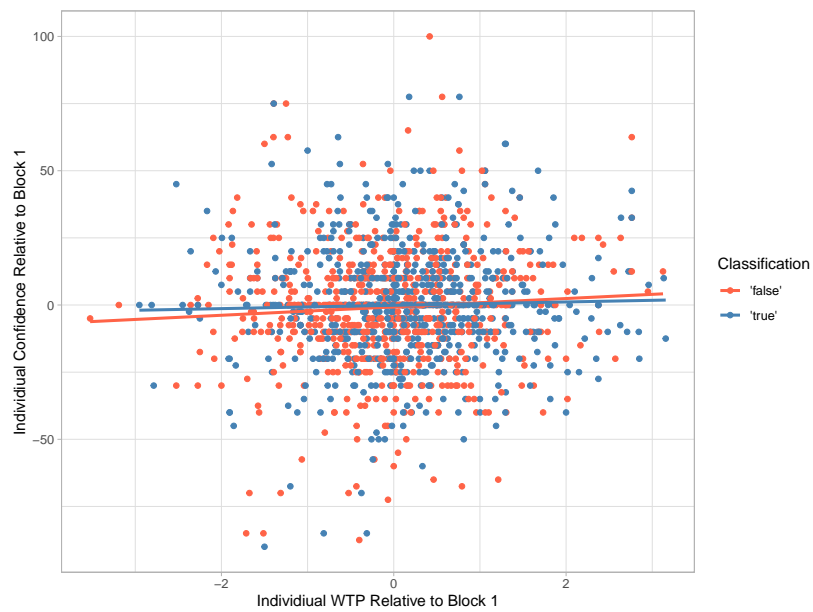


FIGURE 6. AVERAGE OVERCONFIDENCE IN EACH BLOCK BY TREATMENT.

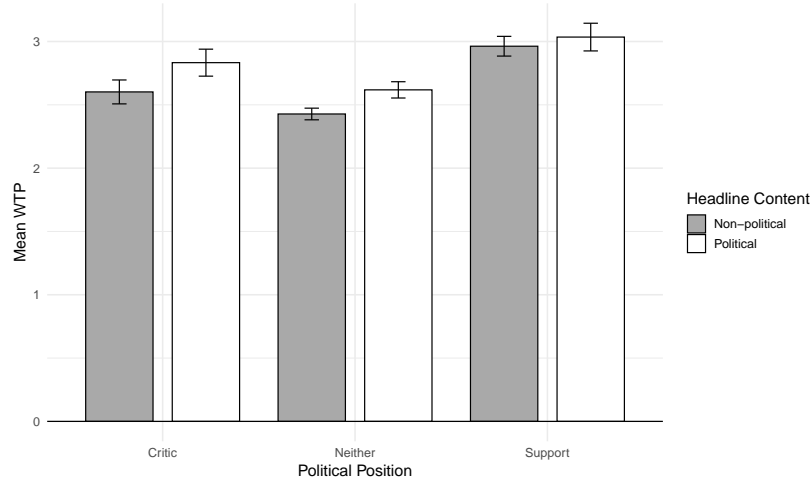


FIGURE 7. AVERAGE WILLINGNESS TO PAY FOR VERIFICATION BY HEADLINE POLITICAL CONTENT TYPE AND PARTICIPANTS' POLITICAL POSITION.

According to regression 1 of table 3, participants demonstrated an **increase in WTP for verifying political news**. This effect was amplified among **government supporters**, who displayed a higher demand for verification as well as a larger accuracy.

Regression 3 and 4 of table 3 analyze the interaction effects of the political content of the headlines (favorable to the government or not) and political position (government supporter, critic, or neither). The effects of the headline classification and *Others* treatment on WTP were preserved. However, in contrast to considering all the headlines **position towards the government did not significantly affect the WTP to verify political headlines**, and contrary to hypothesis 3, there were interaction effects of political content and political position.

In terms of accuracy, regression 4 shows that critics of the government had lower accuracy in classifying political headlines, especially when a headline favored the current government. Additionally, political headlines classified as *true* were more likely to be correct. Contrary to hypotheses 4 and 5, among political headlines, there is no positive effect of feedback treatments on task accuracy. Finally, the round number had a robust negative effect on accuracy. In regression 4, it can be seen that the accuracy was larger when the classification was *true* for political headlines.

RESULT 4: *WTP was **larger** when participants classified a political headline as *true*.*

RESULT 5: *Accuracy was **lower** among critics of the government and when the*

*headlines favored the government.*

Participants demonstrated higher confidence levels when classifying **political headlines**, and among **government supporters**. However, having a political position against the government was only significant at 10%.

RESULT 6: *Confidence in headline classification increased when the headline was political and when participants supported the government.*

#### D. Confidence and WTP

According to the benchmark model, the relationship between confidence and WTP should be negative, as stated by III.C. However, the direct comparison between these variables is not possible since WTP was measured by headline, and Confidence was an aggregated measure for all the headlines in a block of 10 headlines.

As an exercise, the average WTP per block was used to compare with the confidence; the correlation between confidence and average WTP was not significantly different from zero ( $p\text{-value} = 0.1081$ ) but positive (0.036). A confidence measure per headline is necessary to estimate the relation. It is left to future research to investigate more deeply if there is a positive relation between confidence and WTP. However, the some results in the present study points towards a positive relationship between confidence and WTP since the variables associated with higher confidence (political content of the headlines, and political position of the participants) are also associated with higher elicited WTP.

## VII. Discussion

We found a negative relationship between the feedback on others' performance and the willingness to pay for information. Underconfidence was a more prevalent trait of the participants and that feedback could reduce confidence, this pattern is consistent with a reduction in the perceived value of verification relative to the benchmark. Participants could be learning that the task is more manageable than expected, and consider that the probability of correctly classifying is higher, which reduces the value of new verifying the headline. However, there was only a significant effect of individual feedback on confidence.

#### A. Classification and Verification Behavior

Participants classified headlines as *true* ( $c = t$ ) approximately 50% of the time ( $P(c = t) = 0.5$ ), aligning with the known prior probabilities of each state being equal ( $P(T) = P(F) = 0.5$ ), which they were informed about. In terms of confidence, there was a slight asymmetry:  $P(T|c = t) = 54.7\%$  and  $P(F|c = f) = 53.6\%$ . This difference was minimal, suggesting that participants may adhere to the martingale property, as their classification accuracy remained consistent with the prior.

Interestingly, participants showed a greater willingness to pay (WTP) for verification when they classified a headline as *true* ( $c = t$ ), suggesting suggesting an asymmetric valuation of verification conditional on classification. This finding implies an inequality that can be expressed through the decision-making analysis:

$$\begin{aligned}
 WTP(c = t) &> WTP(c = f) \\
 \Leftrightarrow \\
 1 - P(T|c = t) &> 1 - P(F|c = f) \\
 \Leftrightarrow \\
 P(A|c = t) &< P(F|c = f)
 \end{aligned}$$

Additionally, participants were more accurate when classifying political headlines as *true* compared to when they classified the headline as *false*. This contradicts the expectation that people would exhibit more caution in marking political information as accurate. One possible explanation is that participants may have an internalized utility structure that penalizes certain types of misclassifications, such as inaccurately labeling false information as true (e.g.,  $U_{TF} < 0$ ). However, if such a utility structure were in effect, we might expect a lower overall proportion of headlines classified as accurate due to increased caution.

The best verification strategy available to people is an open question beyond the scope of this paper. People have imperfect verification practices and heterogeneous preferences for believing trustworthy news and rejecting false information. We abstract from this consideration by implementing a perfect signal.

### B. Incentives

If a participant decided not to pay attention to the task, they maximize their expected payoffs by selecting 5 MXN as their willingness to pay. Since this was a second-price auction against a random process, the expected value would be 75 MXN. However, not many rounds were selected with a WTP of 5 MXN. For the Control treatment, it was 18.02 %; for the Individual treatment, it was 13.23 %; and for the other treatment, it was 10.2 %. These results coincide with the general finding of a reduced demand for verification in the Others treatment. This suggest that participants inferred the task was easy after the feedback. Therefore, they prefer to verify the headlines and reduce the verification amount.

It is also possible that providing feedback changes the utility associated with each kind of mistake. This could be behind the effects of highlighting the importance of accuracy observed in [Pennycook et al. \(2021\)](#) and [Pennycook and Rand \(2022\)](#). In the present experiment, this could be behind the decrease in confidence when participants got individualized feedback. However, more research is needed to determine the lasting effects of feedback provision.

### C. Time

Participants had a limit of 20 seconds to classify the headlines and indicate their willingness to pay. As noted in table 2, the proportion of missing headlines was small.

In figure 8, we can see that there was no difference in the distribution of time spent on headlines that were classified as *true* and *false*, with the mode at 10 seconds. The 20-second limit seemed enough for participants in most of the headlines, and if there were participants waiting until the last second and potentially making wrong decisions, they represented a small proportion of the rounds.

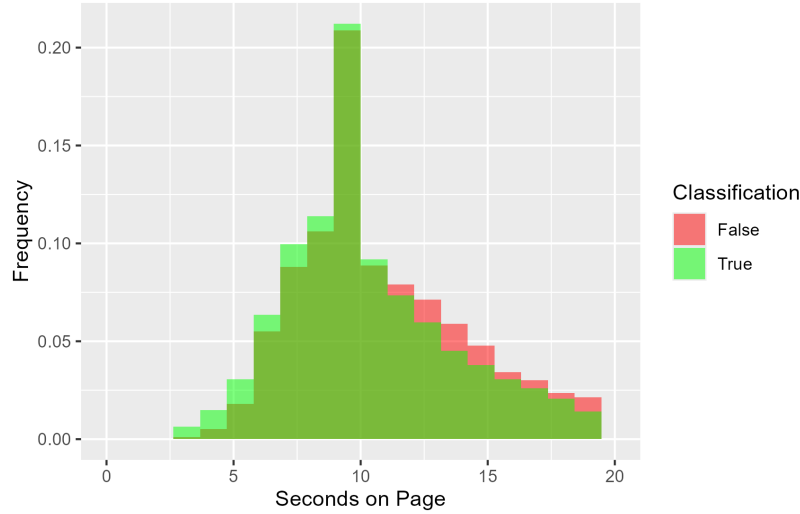


FIGURE 8. FREQUENCY OF TIME SPENT ON HEADLINES CLASSIFIED AS TRUE AND FALSE.

The headlines were also hard enough to classify, but not too difficult that the participants chose randomly. In figure 3, we can see that True headlines have a larger proportion of classifications as *true*. As a reference, the red line indicates an even classification among participants, which means that classifying the headline was difficult. Although some headlines are close to this line (between the blue lines in 0.4 and 0.6), the participants accurately classified most headlines. We expected that difficult headlines would take more time, and there is a group of headlines with this behavior. However, most headlines were classified in around 10 seconds, as also noted in figure 8. Further analyses of the time can be found in the appendix.

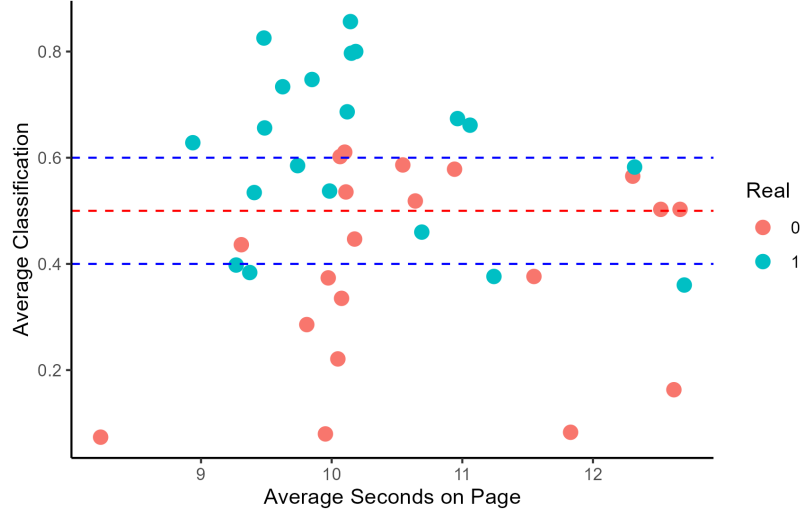


FIGURE 9. AVERAGE TIME SPENT ON EACH HEADLINE AND THE PROPORTION OF CORRECT CLASSIFICATIONS AGAINST THE AVERAGE CLASSIFICATION AS "TRUE."

## VIII. Conclusion

This study provides novel insights into how feedback influences individuals' willingness to pay for verifying news headlines (WTP), their confidence in classification, and their accuracy in identifying misinformation. By incorporating feedback mechanisms in a controlled experimental setting, we found weak evidence of the effect of feedback on the demand for verification.

Feedback on the accuracy of others seems to reduce the WTP for verification; although this was significant only at 10%. Participants in the "Others" feedback condition demonstrated a lower demand for verification than those receiving individual or no feedback. This reduction suggests that social feedback leads participants to infer that the classification task is easier than initially perceived, diminishing the perceived need for verification. However, there was no significant effect of feedback about individual performance on verification.

When examining political content, participants exhibited a higher WTP for verifying politically charged headlines, particularly among those supporting the government in Mexico. However, we did not find significant interactions between participants' political alignment and the leaning of the headlines on WTP. This suggests that while political content affects verification demand, it is not directly moderated by the alignment between participants' political positions and the headlines' content. Nonetheless, participants who supported the government were more confident when classifying headlines favorable to their political stance, demonstrating how political alignment amplifies perceived accuracy. These re-

sults underscore the role of political bias in shaping confidence and verification behaviors.

Interestingly, participants seem to have an unbalanced verification bias, being more willing to pay to verify headlines they initially classified as *true* compared to those classified as *false*. This suggests that interventions targeting verification behaviors should account for individuals' tendencies to confirm, rather than challenge, their initial judgments.

Participants were more accurate when classifying political headlines as *true*, reinforcing that individuals are more cautious in affirming rather than rejecting information in this context. Also, critics of the government had a lower accuracy. Also the accuracy was lower when the headlines favored the government. This evidence of confirmation bias shows that it affects the decisions participants made without an effect on the demand for information.

Future research should further investigate the relationship between confidence and verification at a finer level of measurement. However, the results of this study indicate that confidence alone does not account for individuals' demand for verification. In contrast to a benchmark model in which verification is purely accuracy-motivated and fully determined by confidence, observed willingness to pay is systematically shaped by factors such as political content, political identity, feedback, and the nature of the initial classification. These findings suggest that verification may carry value beyond its instrumental role in improving accuracy.

## REFERENCES

- Arin, K Peren, Deni Mazrekaj, and Marcel Thum**, “Ability of detecting and willingness to share fake news,” *Scientific Reports*, 2023, 7298.
- Arrow, K. J., D. Blackwell, and M. A. Girshick**, “Bayes and Minimax Solutions of Sequential Decision Problems,” *Econometrica*, 7 1949, 17, 213.
- Aslett, Kevin, Zeve Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler, and Joshua A. Tucker**, “Online searches to evaluate misinformation can increase its perceived veracity,” *Nature*, 1 2024, 625, 548–556.
- Assenza, Tiziana, Alberto Cardaci, and Stefanie J. Huber**, “Fake News: Susceptibility, Awareness and Solutions” Fake News: Susceptibility, Awareness and Solutions,” 2024.
- Bateman, Jon and Dean Jackson**, *Countering Disinformation Effectively: An Evidence-Based Policy Guide*, Carnegie Endowment for International Peace, 2024.
- Becker, G. M., M. H. DeGroot, and J. Marschak**, “Measuring utility by a single-response sequential method,” *Behavioral Science*, 1 1964, 9, 226–232.
- Burchardi, Konrad B., Jonathan de Quidt, Selim Gulesci, Benedetta Lerva, and Stefano Tripodi**, “Testing willingness to pay elicitation mechanisms in the field: Evidence from Uganda,” *Journal of Development Economics*, 9 2021, 152, 102701.
- Cason, Timothy N and Charles R Plott**, “Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing,” *Journal of Political Economy*, 12 2014, 122, 1235–1270.
- Charness, Gary, Uri Gneezy, and Vlastimil Rasocha**, “Experimental methods: Eliciting beliefs,” *Journal of Economic Behavior and Organization*, 2021, 189, 234–256.
- Danz, David, Lise Vesterlund, and Alistair J Wilson**, “Belief Elicitation and Behavioral Incentive Compatibility,” *American Economic Review*, 2022, 112, 2851–2883.
- Eberlein, Marion, Sandra Ludwig, and Julia Nafziger**, “The effects of feedback on self-assessment,” *Bulletin of Economic Research*, 2011, 63, 307–3378.
- Erat, Serhat, Kurtulus Demirkol, and M Eyyüp Sallabas**, “Overconfidence and its link with feedback,” *Active Learning in Higher Education*, 2022, 3, 173–187.

- Ferraro, Paul J**, “Know thyself: incompetence and overconfidence,” *Experimental Laboratory Working Paper Series No. 2003-001. Department of Economics, Andrew Young School of Policy Studies, Georgia State University*, 2005.
- Hoes, Emma, Brian Aitken, Jingwen Zhang, Tomasz Gackowski, and Magdalena Wojcieszak**, “Prominent misinformation interventions reduce misperceptions but increase scepticism,” *Nature Human Behavior*, 2023.
- Kartal, Melis and Jean Robert Tyran**, “Fake News, Voter Overconfidence, and the Quality of Democratic Choice,” *American Economic Review*, 10 2022, 112, 3367–97.
- Kavanagh, Jennifer and Michael D. Rich**, *Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life*, RAND Corporation, 1 2018.
- Kogelnik, Maria**, “Performance Feedback and Gender Differences in Persistence,” *SSRN Electronic Journal*, 12 2022.
- Kozyreva, Anastasia, Philipp Lorenz-Spreen, Stefan M. Herzog, Ullrich K.H. Ecker, Stephan Lewandowsky, Ralph Hertwig, Ayesha Ali, Joe Bak-Coleman, Sarit Barzilai, Melisa Basol, Adam J. Berinsky, Cornelia Betsch, John Cook, Lisa K. Fazio, Michael Geers, Andrew M. Guess, Haifeng Huang, Horacio Larreguy, Rakoen Maertens, Folco Panizza, Gordon Pennycook, David G. Rand, Steve Rathje, Jason Reifler, Philipp Schmid, Mark Smith, Briony Swire-Thompson, Paula Szewach, Sander van der Linden, and Sam Wineburg**, “Toolbox of individual-level interventions against online misinformation,” *Nature Human Behaviour* 2024 8:6, 5 2024, 8, 1044–1052.
- List, John A, Lina M Ramirez, Julia Seither, Jaime Unda, and Beatriz Vallejo**, “Toward an Understanding of the Economics of Misinformation: Evidence from a Demand Side Field Experiment on Critical Thinking,” 2024.
- Lyons, Benjamin A., Jacob M. Montgomery, Andrew M. Guess, Brendan Nyhan, and Jason Reifler**, “Overconfidence in news judgments is associated with false news susceptibility,” *Proceedings of the National Academy of Sciences of the United States of America*, 6 2021, 118.
- Mamadehussene, Samir and Francesco Sguera**, “On the Reliability of the BDM Mechanism,” *Management Science*, 2023, 69, 1166–1179.
- Martel, Cameron and David G. Rand**, “Fact-checker warning labels are effective even for those who distrust fact-checkers,” *Nature Human Behaviour* 2024 8:10, 9 2024, 8, 1957–1967.
- Moore, Don A. and Paul J. Healy**, “The Trouble With Overconfidence,” *Psychological Review*, 4 2008, 115, 502–517.

- Oprea, Ryan and Sevgi Yuksel**, “Social Exchange of Motivated Beliefs,” *Journal of the European Economic Association*, 2022, 20, 667–699.
- Ortoleva, Pietro and Erik Snowberg**, “Overconfidence in Political Behavior,” *American Economic Review*, 2015, 105, 504–535.
- Pennycook, Gordon and David G. Rand**, “Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking,” *Journal of Personality*, 4 2020, 88, 185–200.
- and —, “Nudging Social Media toward Accuracy,” *Annals of the American Academy of Political and Social Science*, 3 2022, 700, 152–164.
- , **Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand**, “Shifting attention to accuracy can reduce misinformation online,” *Nature* 592, 2021, 592, 590–595.
- Pulford, Briony D. and Andrew M. Colman**, “Overconfidence: Feedback and item difficulty effects,” *Personality and Individual Differences*, 7 1997, 23, 125–133.
- Swire-Thompson, Briony, Nicholas Miklaucic, John P. Wihbey, David Lazer, and Joseph DeGutis**, “The backfire effect after correcting misinformation is strongly associated with reliability,” *Journal of experimental psychology. General*, 2022, 151, 1655.
- Sánchez, Julio C Aguila and Pamela Pereyra-Zamora**, “Infodemics in Mexico: A look at the Animal Político and Verificado fact-checking platforms,” *Health Education Journal*, 2022, 81, 982–992.
- Thaler, Michael**, “The Fake News Effect: Experimentally Identifying Motivated Reasoning Using Trust in News,” *American Economic Journal: Microeconomics*, 2024, 16, 1–38.
- Trujano-Ochoa, Dario**, “Do Others Learn like Me? Higher Order Willingness to Pay for Information,” 2024.
- Velez, Yamil Ricardo and Patrick Liu**, “Confronting Core Issues: A Critical Assessment of Attitude Polarization Using Tailored Experiments,” *American Political Science Review*, 5 2025, 119, 1036–1053.
- Vosoughi, Soroush, Deb Roy, and Sinan Aral**, “The spread of true and false news online,” *Science*, 2018, 359, 1146–1151.
- Wald, Abraham**, “Foundations of a General Theory of Sequential Decision Functions,” *Econometrica*, 10 1947, 15, 279.
- Wilson, Alistair J and Emanuel Vespa**, “Paired-uniform scoring: Implementing a binarized scoring rule with non-mathematical language,” 2018.

## MATHEMATICAL APPENDIX

## A1. Expected Value of the Signal Considering Reclassification

## GENERAL SETUP

- 1) **Initial Classification:** The agent initially classifies the headline as  $c$  (either accurate ( $t$ ) or fake ( $f$ )).
- 2) **Receive Signal:** The agent receives a signal  $s$  which can either confirm or contradict their initial classification.
- 3) **Reclassification:** Based on the signal, the agent makes a new classification  $c'$ .

EXPECTED UTILITY WITH SIGNAL AND RECLASSIFICATION ( $EU_{\text{SIGNAL}}^{\text{UPDATE}}(c)$ )

The expected utility with the signal and reclassification is calculated by considering the updated posterior probabilities and the new classification based on the signal.

## STEP 1: DEFINE PROBABILITIES AND UTILITIES

• **Initial Posterior Probabilities:**

$$P(T|c) = \frac{P(c|T) \cdot (1 - p_f)}{P(c)}, \quad P(F|c) = \frac{P(c|F) \cdot p_f}{P(c)}$$

Where:

$$P(c) = (1 - p_f) \cdot P(c|T) + p_f \cdot P(c|F)$$

• **Signal Probabilities:**

$$q_t = P(s = t|T), \quad q_f = P(s = f|F)$$

• **Utilities:**

$$U_T, U_F, U_{TF}, U_{FT}$$

## STEP 2: DEFINE UPDATED POSTERIOR PROBABILITIES GIVEN SIGNAL

After observing the signal, the agent updates their beliefs:

• **Posterior Probabilities Given Signal  $s = t$ :**

$$P(T|s = t, c) = \frac{q_t \cdot P(T|c)}{q_t \cdot P(T|c) + (1 - q_f) \cdot P(F|c)}$$

$$P(F|s = t, c) = \frac{(1 - q_f) \cdot P(F|c)}{q_t \cdot P(T|c) + (1 - q_f) \cdot P(F|c)}$$

• **Posterior Probabilities Given Signal  $s = f$ :**

$$P(T|s = f, c) = \frac{(1 - q_t) \cdot P(T|c)}{(1 - q_t) \cdot P(T|c) + q_f \cdot P(F|c)}$$

$$P(F|s = f, c) = \frac{q_f \cdot P(F|c)}{(1 - q_t) \cdot P(T|c) + q_f \cdot P(F|c)}$$

STEP 3: CALCULATE EXPECTED UTILITY AFTER SIGNAL

We assume that the signal is strong enough (proposition 1) to ensure that the best reclassification is to follow the signal's indication of the state of the world. Otherwise, the signal would have no instrumental value; therefore,  $EVSI = 0$ .

The expected utility with the signal, considering the possibility of reclassification, is:

$$EU_{\text{signal}}^{\text{update}}(c) = P(s = t|c) \cdot EU_{\text{new classification}}(s = t, c) + P(s = f|c) \cdot EU_{\text{new classification}}(s = f, c)$$

Here,  $P(s = t|c)$  and  $P(s = f|c)$  are the probabilities of receiving the signals  $s = t$  and  $s = f$  given the initial classification  $c$ . These probabilities are determined by Bayes' rule, considering the agent's initial classification and the properties of the signal.

Given the initial classification  $c$ :

$$P(s = t|c) = q_t \cdot P(T|c) + (1 - q_f) \cdot P(F|c)$$

$$P(s = f|c) = (1 - q_t) \cdot P(T|c) + q_f \cdot P(F|c)$$

EXPECTED UTILITY AFTER SIGNAL  $s = t$

$$EU_{\text{new classification}}(s = t, c) = P(T|s = t, c) \cdot U_T + P(F|s = t, c) \cdot U_{TF}$$

EXPECTED UTILITY AFTER SIGNAL  $s = f$

$$EU_{\text{new classification}}(s = f, c) = P(F|s = f, c) \cdot U_F + P(T|s = f, c) \cdot U_{FT}$$

STEP 4: COMBINE EXPECTED UTILITIES

$$EU_{\text{signal}}^{\text{update}}(c) = [q_t \cdot P(T|c) + (1 - q_f) \cdot P(F|c)] \cdot [P(T|s = t, c) \cdot U_T + P(F|s = t, c) \cdot U_{TF}]$$

$$+ [(1 - q_t) \cdot P(T|c) + q_f \cdot P(F|c)] \cdot [P(F|s = f, c) \cdot U_F + P(T|s = f, c) \cdot U_{FT}]$$

#### STEP 5: EXPECTED VALUE OF THE SIGNAL (EVSI)

Finally, the EVSI is the difference between the expected utility with the signal and the expected utility without the signal.

$$EVSI = EU_{\text{signal}}^{\text{update}}(c) - EU_{\text{no signal}}(c)$$

#### CONCLUSION

By allowing the agent to update their classification based on the signal, we account for the dynamic decision-making process. The expected value of the signal (EVSI) is derived by comparing the expected utility with the signal (considering reclassification) to the expected utility without the signal. This approach shows the impact of additional information on improving decision-making accuracy.

#### A2. Proof of Proposition 2: Need for a Strong Enough Signal

To ensure that people follow the signal  $S$  for reclassification, we must prove that the expected utility of reclassifying based on the signal is higher than not reclassifying. Without loss of generality, we will first consider the case where the initial classification  $c = f$  and the signal  $s = t$ .

#### INITIAL SETUP

- 1) **Initial Classification:**  $c = f$  (classified as fake)
- 2) **Signal Received:**  $s = t$  (signal indicates true)

We must show that reclassifying the headline as true ( $c' = t$ ) based on the signal is optimal.

#### EXPECTED UTILITY OF NOT RECLASSIFYING

If the agent does not reclassify and sticks with the initial classification  $c = f$  but knows the signal  $s = t$ , the expected utility is:

$$EU_{\text{no reclassification}}(f, s = t) = P(T|s = t, f) \cdot U_{FT} + P(F|s = t, f) \cdot U_F$$

#### EXPECTED UTILITY OF RECLASSIFYING

If the agent reclassifies the headline based on the signal  $s = t$ , the expected utility is:

$$EU_{\text{reclassification}}(f, s = t) = P(T|s = t, f) \cdot U_T + P(F|s = t, f) \cdot U_{TF}$$

## POSTERIOR PROBABILITIES

The posterior probabilities given the signal  $s = t$  and initial classification  $c = f$  are:

$$P(T|s = t, f) = \frac{q_t \cdot P(T|f)}{q_t \cdot P(T|f) + (1 - q_f) \cdot P(F|f)}$$

$$P(F|s = t, f) = \frac{(1 - q_f) \cdot P(F|f)}{q_t \cdot P(T|f) + (1 - q_f) \cdot P(F|f)}$$

## CONDITION FOR RECLASSIFYING

To prove that reclassifying based on the signal is optimal, we need:

$$EU_{\text{reclassification}}(f, s = t) > EU_{\text{no reclassification}}(f, s = t)$$

Substituting the utilities, we get:

$$P(T|s = t, f) \cdot U_T + P(F|s = t, f) \cdot U_{TF} > P(T|s = t, f) \cdot U_{FT} + P(F|s = t, f) \cdot U_F$$

Given the simplifying assumptions:

$$U_T = 1, \quad U_F = 1, \quad U_{TF} = 0, \quad U_{FT} = 0$$

The inequality simplifies to:

$$P(T|s = t, f) \cdot 1 + P(F|s = t, f) \cdot 0 > P(T|s = t, f) \cdot 0 + P(F|s = t, f) \cdot 1$$

This reduces to:

$$P(T|s = t, f) > P(F|s = t, f)$$

## VERIFYING THE POSTERIOR PROBABILITIES

Substitute the posterior probabilities:

$$\frac{q_t \cdot P(T|f)}{q_t \cdot P(T|f) + (1 - q_f) \cdot P(F|f)} > \frac{(1 - q_f) \cdot P(F|f)}{q_t \cdot P(T|f) + (1 - q_f) \cdot P(F|f)}$$

Since the denominators are the same, we can simplify this to:

$$q_t \cdot P(T|f) > (1 - q_f) \cdot P(F|f)$$

Since  $P(F|f) = 1 - P(T|f)$ , we have:

$$q_t \cdot P(T|f) > (1 - q_f) \cdot (1 - P(T|f))$$

Expanding and rearranging terms, we get:

$$q_t \cdot P(T|f) > (1 - q_f) - (1 - q_f) \cdot P(T|f)$$

$$q_t \cdot P(T|f) + (1 - q_f) \cdot P(T|f) > (1 - q_f)$$

$$P(T|f) \cdot (q_t + 1 - q_f) > (1 - q_f)$$

Dividing both sides by  $(q_t + 1 - q_f)$ :

$$P(T|f) > \frac{1 - q_f}{q_t + 1 - q_f}$$

This shows that the signal needs to be strong enough such that  $q_t$  is sufficiently large compared to  $1 - q_f$ , ensuring that the agent reclassifies the headline as true based on the signal. This proves that a strong signal is necessary to ensure that people follow the signal  $S$  for reclassification.

TRIVIAL CASE:  $c = s = t$

If the initial classification  $c = t$  and the signal  $s = t$ , then reclassification is unnecessary because the initial classification is already true. The expected utility remains the same:

$$EU_{\text{reclassification}}(t, s = t) = P(T|s = t, t) \cdot U_T + P(F|s = t, t) \cdot U_{TF}$$

Given the simplifying assumptions, this reduces to:

$$EU_{\text{reclassification}}(t, s = t) = P(T|s = t, t) \cdot 1 + P(F|s = t, t) \cdot 0 = P(T|s = t, t)$$

The expected utility of not reclassifying is:

$$EU_{\text{no reclassification}}(t, s = t) = P(T|s = t, t) \cdot U_T + P(F|s = t, t) \cdot U_{TF}$$

Given the simplifying assumptions, this reduces to:

$$EU_{\text{no reclassification}}(t, s = t) = P(T|s = t, t) \cdot 1 + P(F|s = t, t) \cdot 0 = P(T|s = t, t)$$

Since both expected utilities are equal, reclassification is trivial in this case.

OTHER CASES

The same process follows for the cases  $c = t, s = f$  and  $c = s = f$ . For these cases, the conditions are as follows:

1) **Case**  $c = t, s = f$ :

$$EU_{\text{reclassification}}(t, s = f) > EU_{\text{no reclassification}}(t, s = f)$$

Substituting the utilities, we get:

$$P(F|s = f, t) \cdot U_F + P(T|s = f, t) \cdot U_{FT} > P(F|s = f, t) \cdot U_{TF} + P(T|s = f, t) \cdot U_T$$

Given the simplifying assumptions:

$$P(F|s = f, t) \cdot 1 + P(T|s = f, t) \cdot 0 > P(F|s = f, t) \cdot 0 + P(T|s = f, t) \cdot 1$$

This reduces to:

$$P(F|s = f, t) > P(T|s = f, t)$$

Verifying the posterior probabilities:

$$\frac{q_f \cdot P(F|t)}{q_f \cdot P(F|t) + (1 - q_t) \cdot P(T|t)} > \frac{(1 - q_t) \cdot P(T|t)}{q_f \cdot P(F|t) + (1 - q_t) \cdot P(T|t)}$$

Since the denominators are the same, we can simplify this to:

$$q_f \cdot P(F|t) > (1 - q_t) \cdot P(T|t)$$

Since  $P(T|t) = 1 - P(F|t)$ , we have:

$$q_f \cdot P(F|t) > (1 - q_t) \cdot (1 - P(F|t))$$

Expanding and rearranging terms, we get:

$$q_f \cdot P(F|t) > (1 - q_t) - (1 - q_t) \cdot P(F|t)$$

$$q_f \cdot P(F|t) + (1 - q_t) \cdot P(F|t) > (1 - q_t)$$

$$P(F|t) \cdot (q_f + 1 - q_t) > (1 - q_t)$$

Dividing both sides by  $(q_f + 1 - q_t)$ :

$$P(F|t) > \frac{1 - q_t}{q_f + 1 - q_t}$$

2) **Case**  $c = s = f$ : If the initial classification  $c = f$  and the signal  $s = f$ , then reclassification is unnecessary because the initial classification is already correct. The expected utility remains the same:

$$EU_{\text{reclassification}}(f, s = f) = P(F|s = f, f) \cdot U_F + P(T|s = f, f) \cdot U_{TF}$$

Given the simplifying assumptions, this reduces to:

$$EU_{\text{reclassification}}(f, s = f) = P(F|s = f, f) \cdot 1 + P(T|s = f, f) \cdot 0 = P(F|s = f, f)$$

The expected utility of not reclassifying is:

$$EU_{\text{no reclassification}}(f, s = f) = P(F|s = f, f) \cdot U_F + P(T|s = f, f) \cdot U_{TF}$$

Given the simplifying assumptions, this reduces to:

$$EU_{\text{no reclassification}}(f, s = f) = P(F|s = f, f) \cdot 1 + P(T|s = f, f) \cdot 0 = P(F|s = f, f)$$

Since both expected utilities are equal, reclassification is trivial in this case.

#### CONDITIONS

Therefore, to ensure that the signal is strong enough to prompt optimal reclassification in both cases, we need to satisfy two key conditions:

$$P(T|f) > \frac{1 - q_f}{q_t + 1 - q_f}$$

$$P(T|t) < \frac{q_f}{q_f + 1 - q_t}$$

Considering the conditions for proposition 1 we have that,

$$\frac{1 - q_f}{q_t + 1 - q_f} < P(T|f) < P(T|t) < \frac{q_f}{q_f + 1 - q_t}$$

*A3. Proof of Proposition 1: Sufficient and Necessary Condition for*

$$P(T|c = f) < P(T) < P(T|c = t) \text{ and } P(F|c = t) < P(F) < P(F|c = f)$$

*A4. Proof of Necessity and Sufficiency*

We will prove that  $1 < \frac{P(c=t|T)}{P(c=t|F)}$  and  $1 < \frac{P(c=f|F)}{P(c=f|T)}$  if and only if  $P(T|c = f) < P(T) < P(T|c = t)$  and  $P(F|c = t) < P(F) < P(F|c = f)$ .

#### DEFINITIONS AND SETUP

Let:

- $P(T)$  be the prior probability that the state is true.
- $P(F)$  be the prior probability that the state is fake.
- $P(c = t|T)$  be the probability of classifying a headline as true given it is true.

- $P(c = t|F)$  be the probability of classifying a headline as true given it is fake.
- $P(c = f|F)$  be the probability of classifying a headline as fake given it is fake.
- $P(c = f|T)$  be the probability of classifying a headline as fake, given it is true.

#### POSTERIOR PROBABILITIES

The posterior probabilities after observing the classification  $c$  are given by:

- Posterior probability of  $T$  given  $c = f$ :

$$P(T|c = f) = \frac{P(c = f|T) \cdot P(T)}{P(c = f|T) \cdot P(T) + P(c = f|F) \cdot P(F)}$$

- Posterior probability of  $T$  given  $c = t$ :

$$P(T|c = t) = \frac{P(c = t|T) \cdot P(T)}{P(c = t|T) \cdot P(T) + P(c = t|F) \cdot P(F)}$$

- Posterior probability of  $F$  given  $c = f$ :

$$P(F|c = f) = \frac{P(c = f|F) \cdot P(F)}{P(c = f|T) \cdot P(T) + P(c = f|F) \cdot P(F)}$$

- Posterior probability of  $F$  given  $c = t$ :

$$P(F|c = t) = \frac{P(c = t|F) \cdot P(F)}{P(c = t|T) \cdot P(T) + P(c = t|F) \cdot P(F)}$$

#### PART 1: SUFFICIENCY ( $\Rightarrow$ )

Assume that  $1 < \frac{P(c=t|T)}{P(c=t|F)}$  and  $1 < \frac{P(c=f|F)}{P(c=f|T)}$ . We want to show that this implies  $P(T|c = f) < P(T) < P(T|c = t)$  and  $P(F|c = t) < P(F) < P(F|c = f)$ .

#### ANALYZE THE POSTERIOR PROBABILITIES

- 1) **For  $P(T|c = f)$ :** Given the condition  $1 < \frac{P(c=f|F)}{P(c=f|T)}$ , we know that:

$$\frac{P(c = f|F)}{P(c = f|T)} > 1$$

This implies  $P(c = f|F) > P(c = f|T)$ . As a result, in the posterior probability expression:

$$P(T|c = f) = \frac{P(c = f|T) \cdot P(T)}{P(c = f|T) \cdot P(T) + P(c = f|F) \cdot P(F)}$$

The denominator  $P(c = f|T) \cdot P(T) + P(c = f|F) \cdot P(F)$  will be larger than the numerator  $P(c = f|T) \cdot P(T)$ , causing  $P(T|c = f)$  to be smaller than the prior  $P(T)$ . Therefore:

$$P(T|c = f) < P(T)$$

2) **For**  $P(T|c = t)$ : Given the condition  $1 < \frac{P(c=t|T)}{P(c=t|F)}$ , we know that:

$$\frac{P(c = t|T)}{P(c = t|F)} > 1$$

This implies  $P(c = t|T) > P(c = t|F)$ . As a result, in the posterior probability expression:

$$P(T|c = t) = \frac{P(c = t|T) \cdot P(T)}{P(c = t|T) \cdot P(T) + P(c = t|F) \cdot P(F)}$$

The numerator  $P(c = t|T) \cdot P(T)$  will dominate the denominator  $P(c = t|T) \cdot P(T) + P(c = t|F) \cdot P(F)$ , causing  $P(T|c = t)$  to be larger than the prior  $P(T)$ . Therefore:

$$P(T|c = t) > P(T)$$

3) **For**  $P(F|c = f)$  **and**  $P(F|c = t)$ : Similarly, the same reasoning applies to  $P(F|c = f)$  and  $P(F|c = t)$ , given that:

$$\frac{P(c = f|F)}{P(c = f|T)} > 1 \quad \text{and} \quad \frac{P(c = t|T)}{P(c = t|F)} > 1$$

This implies that:

$$P(F|c = t) < P(F) < P(F|c = f)$$

#### PART 2: NECESSITY ( $\Leftarrow$ )

Assume that  $P(T|c = f) < P(T) < P(T|c = t)$  and  $P(F|c = t) < P(F) < P(F|c = f)$ . We need to show that this implies  $1 < \frac{P(c=t|T)}{P(c=t|F)}$  and  $1 < \frac{P(c=f|F)}{P(c=f|T)}$ .

#### ANALYZING THE POSTERIOR PROBABILITIES

- **\*\*For**  $P(T|c = f) < P(T)$ :\*\*

Given the posterior probability expression:

$$P(T|c = f) = \frac{P(c = f|T) \cdot P(T)}{P(c = f|T) \cdot P(T) + P(c = f|F) \cdot P(F)}$$

If  $P(T|c = f) < P(T)$ , then the likelihood ratio  $\frac{P(c=f|F)}{P(c=f|T)}$  must be greater than 1. This is because the posterior  $P(T|c = f)$  being less than  $P(T)$  implies that the signal  $c = f$  is more likely to come from the fake state  $F$ , meaning:

$$\frac{P(c = f|F)}{P(c = f|T)} > 1$$

- \*\*For  $P(T|c = t) > P(T)$ :\*\*

Given the posterior probability expression:

$$P(T|c = t) = \frac{P(c = t|T) \cdot P(T)}{P(c = t|T) \cdot P(T) + P(c = t|F) \cdot P(F)}$$

If  $P(T|c = t) > P(T)$ , then the likelihood ratio  $\frac{P(c=t|T)}{P(c=t|F)}$  must be greater than 1. This is because the posterior  $P(T|c = t)$  being greater than  $P(T)$  implies that the signal  $c = t$  is more likely to come from the true state  $T$ , meaning:

$$\frac{P(c = t|T)}{P(c = t|F)} > 1$$

- \*\*For  $P(F|c = f) > P(F)$  and  $P(F|c = t) < P(F)$ :\*\*

By symmetry, the same reasoning applies for  $P(F|c = f) > P(F)$  and  $P(F|c = t) < P(F)$ . The likelihood ratios  $\frac{P(c=f|F)}{P(c=f|T)} > 1$  and  $\frac{P(c=t|T)}{P(c=t|F)} > 1$  are necessary conditions to satisfy these posterior inequalities.

#### CONCLUSION

Thus, we have shown that:  $1 < \frac{P(c=t|T)}{P(c=t|F)}$  and  $1 < \frac{P(c=f|F)}{P(c=f|T)}$  are necessary and sufficient conditions for:

$$P(T|c = f) < P(T) < P(T|c = t)$$

$$P(F|c = t) < P(F) < P(F|c = f)$$

## COROLLARY: INFORMATIVENESS OF THE SIGNAL

The same analysis can be applied to the signal. Therefore  $1 < \frac{P(s=t|T)}{P(s=t|F)}$  and

$$1 < \frac{P(s=f|F)}{P(s=f|T)} \iff$$

$$P(T|s=f) < P(T) < P(T|s=t)$$

$$P(F|s=t) < P(F) < P(F|s=f)$$

## MATERIALS

*B1. Confidence Elicitation***Confidence in Block Classification 3**

Answer the following questions with the probability in percentage terms.  
Where 100 means the event always occurs, 0 means it never occurs, and 50 means it occurs half of the time.

**Please consider the block of 10 news headlines that you just classified:**

You classified 5 headlines as "The information is accurate" and 5 as "Contains false information".

One of the 5 headlines you classified as accurate will be selected at random.  
What is the probability that the headline is actually accurate?

70 ▾

One of the 5 headlines you classified as false will be selected at random.  
What is the probability that the headline is actually false?

55 ▾

**Now, consider the classification that other participants made in this block of 10 news headlines:**

A headline classified as accurate by another participant will be selected at random.  
What is the probability that the headline is actually accurate?

50 ▾

A headline classified as false by another participant will be selected at random.  
What is the probability that the headline is actually false?

60 ▾

Next

FIGURE B1. SCREENSHOT OF THE TRANSLATED CONFIDENCE ELICITATION AS SEEN BY THE PARTICIPANTS.

*B2. Headlines Used in the Experiment*

Block	Real	Headline	Translated Headline
1	1	Se inaugura un nuevo museo en honor a Cantinflas en la Ciudad de México	A New Museum in Honor of Cantinflas is Inaugurated in Mexico City
1	1	El salario mínimo en México se incrementa 20% en 2024	Minimum Wage in Mexico Increases by 20% in 2024
1	1	La variante Ómicron es la única de preocupación que circula a nivel mundial; es más transmisible, aunque menos peligrosa que la variante Delta	The Omicron Variant is the Only Variant of Concern Circulating Worldwide; It Is More Transmissible, Though Less Dangerous than the Delta Variant

1	1	Se suspende programa humanitario para trabajar o solicitar asilo en Estados Unidos para Haití, Venezuela, Nicaragua y Cuba	Humanitarian program to work or apply for asylum in the United States for Haiti, Venezuela, Nicaragua, and Cuba is suspended
1	1	Incrementó en el uso de energías renovables en México	Increase in the Use of Renewable Energy in Mexico
1	0	Turismo internacional se desploma en 2024, México ya no es un destino atractivo	International Tourism Collapses, Mexico is No Longer an Attractive Destination
1	0	Luto en México por accidente aéreo de un avión de pasajeros. No hubo sobrevivientes	National Mourning in Mexico. Terrible Passenger Plane Crash in 2024. No Survivors
1	1	En 2024, se intensificaron los incendios forestales en México	In 2024, Wildfires Intensified in Various Regions of Mexico
1	1	Existen programas de apoyo a pequeñas empresas lanzados por el gobierno mexicano	There Are Support Programs for Small Businesses Launched by the Mexican Government
1	0	Se firma un tratado del Foro Económico Mundial que busca reconocer la pedofilia como orientación sexual	A World Economic Forum Treaty Is Signed to Recognize Pedophilia as a Sexual Orientation
2	1	El INAH cobrará \$60 por tomar fotografías para uso comercial en museos y sitios arqueológicos	INAH Will Charge \$60 for Taking Photos in Museums and Archaeological Sites for Commercial Use
2	0	Se publica la lista de apellidos que pueden solicitar la ciudadanía española	Spain Publishes a List of Surnames That Allow One to Apply for Spanish Citizenship
2	0	En Irán censuraron los Juegos Olímpicos; todas las mujeres aparecen con rectángulos o asteriscos cubriéndolas	Iran Censored the Olympics; All Women Appear with Rectangles or Asterisks Covering Them
2	0	Iniciará juicio en contra de la ministra presidenta de la SCJN por participar en el paro de trabajadores del Poder Judicial.	Trial Against the Chief Justice of the Supreme Court for Participating in the Judicial Workers' Strike to Begin
2	0	La Organización de Estados Americanos (OEA) sanciona a México por dar asilo a Jorge Glass en la embajada mexicana en Ecuador	The Organization of American States (OAS) Managed to Sanction Mexico for Granting Asylum to Jorge Glass in the Mexican Embassy
2	0	El hijo de Nicolás Maduro es captado en video manejando un Ferrari dorado	Nicolás Maduro's son is seen driving a golden Ferrari

2	1	El Ozempic, promovido en redes para bajar de peso, es un tratamiento controlado para la diabetes tipo 2	Ozempic Is Actually a Controlled Treatment for Type 2 Diabetes
2	1	México alcanza cifra récord en exportaciones agrícolas	Mexico Reaches Record High in Agricultural Exports
2	0	Atletas ucranianos portaron pulseras de tobillo con GPS para evitar su huida después de los Juegos Olímpicos de París 2024	Ukrainian Athletes Wore GPS Ankle Bracelets to Prevent Them from Fleeing After the Olympic Games
2	1	Ningún país ha declarado confinamiento por mpox tras la nueva emergencia sanitaria anunciada por la OMS	No country has declared a lockdown due to mpox following the new health emergency announced by the WHO
3	0	La cama "antisexo" siguió siendo utilizada en los Juegos Olímpicos de París 2024	The "Anti-Sex" Bed Will Continue to Be Used at the Paris 2024 Olympics
3	0	El aeropuerto de Suecia fue descontaminado debido a contagios de Mpox	Sweden's Airport Was Decontaminated Due to Mpox Infections
3	0	Miss Venezuela protestó ante las cámaras contra su gobierno en una alfombra roja	Miss Venezuela Protested Against Her Government on a Red Carpet
3	1	Avances en la investigación de nuevas vacunas desarrolladas en México	Advances in the Research of New Vaccines Developed in Mexico
3	0	Un estadounidense se suicidó saltando desde su habitación durante el Baja Beach Fest 2024 en México	An American Committed Suicide During the Baja Beach Fest 2024 in Mexico
3	1	Descubrimiento de nuevas ruinas mayas en la península de Yucatán en 2024	Discovery of New Mayan Ruins in the Yucatán Peninsula
3	1	México termina en el puesto 65 del medallero en los Juegos Olímpicos de París 2024	Mexico Finishes 65th in the Medal Table at the Paris 2024 Olympic Games
3	1	Mexico envió dos aviones en 2023 para rescatar connacionales varados en Israel por el conflicto en Gaza	Sedena and SRE Sent Two Planes to Rescue Mexicans Stranded in Israel
3	0	Consumir alimentos alcalinos ayuda a contrarrestar la variante Omicron del coronavirus	Maintaining a pH (Acidity Level) Above 5.5 Can Prevent Covid-19 Infection

3	0	El Consejo para Prevenir y Eliminar la Discriminación (COPRED) busca suspender la celebración del Día del Padre en los centros educativos	COPRED Urges Elementary Schools Not to Exclude Children from Non-Normative Families on Father's Day
4	1	México tiene la tasa más baja de desempleo de la OCDE	Mexico Has the Lowest Unemployment Rate in the OECD
4	0	Gobierno de México entrega el nuevo "Bono Mujeres" por 2 mil 700 pesos	Mexican Government Issues the New "Women's Bonus" for 2,700 Pesos
4	1	Se registró una disminución de 5.1 millones personas en pobreza en el actual gobierno	A Decrease of 5.1 Million People in Poverty Was Recorded During the Current Government
4	0	Tribunal Electoral encuentra irregularidades graves en el triunfo de Claudia Sheinbaum	Electoral Tribunal Finds Serious Irregularities in Claudia Sheinbaum's Victory
4	1	La presidenta electa Claudia Sheinbaum anuncia beca universal para estudiantes de nivel básico	President-Elect Claudia Sheinbaum Announces Universal Scholarship for Elementary School Students
4	0	El Tren Maya se completará sin impacto ambiental, según estudios científicos independientes	The Maya Train Will Be Completed Without Environmental Impact, According to Independent Scientific Studies
4	1	11 Ministros de la Suprema Corte de Justicia de la Nación ganan \$206,246 pesos mensuales netos	11 Supreme Court Justices Earn \$206,246 Pesos Monthly Net
4	1	México envió dos aviones a Israel para rescatar a las y los mexicanos varados por el conflicto con Palestina.	Mexico Sent Two Planes to Israel to Rescue Mexicans Stranded Due to the Conflict with Palestine
4	1	En solo ocho de cada 100 delitos en México se abre una carpeta de investigación	Only Eight Out of Every 100 Crimes in Mexico Lead to an Investigation Being Opened
4	0	México prepara una reunión con los presidentes de Rusia y Corea del Norte para comprar armas	Mexico Prepares a Meeting with the Presidents of Russia and North Korea to Buy Weapons
5	1	En el gobierno de AMLO se logró una reducción en la tasa de homicidios.	AMLO's Government Achieved a Reduction in the Homicide Rate
5	1	Primer director femenino de la CFE es nombrado en México	First Female Director of the CFE Is Appointed in Mexico
5	0	Durante el gobierno de Andrés Manuel López Obrador, la deuda pública subió 64%	During Andrés Manuel López Obrador's Government, Public Debt Increased by 64%

5	0	Poder Judicial de la Federación hay 53,737 personas que ganan más que el Presidente	In the Federal Judiciary, 53,737 People Earn More Than the President
5	1	La pobreza extrema en México incrementó de 2018 a 2022	Extreme Poverty in Mexico Increased from 2018 to 2022
5	1	Hay déficit presupuestario en el 2024 por parte del gobierno de México	Mexico's Government Faces a Budget Deficit in 2024
5	1	EU critica al Gobierno de AMLO por 'desacreditar a periodistas'	U.S. Criticizes AMLO's Government for 'Discrediting Journalists'
5	1	Sedena gastó más que el presupuesto autorizado por el Congreso	Sedena Spent More Than the Budget Authorized by Congress
5	0	México ya produce el 90% de la gasolina que consume, como afirma AMLO	Mexico Now Produces 90% of the Gasoline It Consumes, As Claimed by AMLO
5	0	Metro de CDMX dejará de ser gratis para adultos mayores	CDMX Metro Will No Longer Be Free for Senior Citizens

## REGRESSIONS APPENDIX

TABLE C1—REGRESSION ON THE CONFIDENCE AND WILLINGNESS TO PAY. THE SE WERE CLUSTERED AT THE INDIVIDUAL LEVEL AND THE FIRST BLOCK WAS EXCLUDED.

	<i>Dependent variable:</i>		
	Confidence	WTP	Accuracy
	(1)	(2)	(3)
Individual Feedback	−5.741* (2.996)	−0.187 (0.215)	−0.015 (0.014)
Others Feedback	−3.653 (3.027)	−0.347* (0.203)	−0.012 (0.014)
Block	−0.607 (0.815)	0.034 (0.028)	0.014 (0.010)
'True' (c = t)	1.052 (1.072)	0.209*** (0.054)	0.001 (0.004)
Accuracy	−0.399 (0.397)	−0.057 (0.041)	
Age	0.976* (0.571)	−0.028 (0.050)	0.004 (0.003)
Male	2.903 (2.482)	−0.064 (0.175)	0.012 (0.012)
Confidence		0.001 (0.003)	−0.0002 (0.0002)
Political	4.920*** (1.600)	0.114* (0.062)	0.014 (0.023)
Support Gov	6.891** (3.288)	0.475** (0.210)	0.039*** (0.015)
Against Gov	5.054 (3.098)	0.214 (0.228)	0.004 (0.014)
Constant	34.078*** (11.543)	2.972*** (1.017)	0.480*** (0.066)
Observations	7,572	7,572	7,572
R <sup>2</sup>	0.054	0.030	0.004
Adjusted R <sup>2</sup>	0.053	0.029	0.002

*Note:*

\*p&lt;0.1; \*\*p&lt;0.05; \*\*\*p&lt;0.01

TABLE C2—REGRESSION ON THE WILLINGNESS TO PAY AND ACCURACY. THE REGRESSION IS ON THE DATA AT THE BLOCK LEVEL TAKING THE AVERAGE OF THE VARIABLES IN EACH BLOCK. THE SE WERE CLUSTERED AT THE INDIVIDUAL LEVEL TO ACCOUNT FOR WITHIN-PARTICIPANT CORRELATION WHEN THERE ARE MULTIPLE OBSERVATIONS PER PARTICIPANT.

	<i>Dependent variable:</i>	
	WTP	Accuracy
	(1)	(2)
Individual Feedback	−0.205 (0.232)	−0.018 (0.017)
Others Feedback	−0.396* (0.224)	−0.035* (0.019)
Block	0.051 (0.048)	−0.140*** (0.015)
Confidence	0.002 (0.004)	0.00003 (0.0003)
'True' (c = t)	0.206*** (0.066)	0.211*** (0.006)
Accuracy	−0.044 (0.057)	
WTP		−0.004 (0.005)
Age	−0.042 (0.051)	0.001 (0.004)
Male	−0.034 (0.191)	0.015 (0.015)
Support Gov	0.406 (0.248)	0.014 (0.023)
News Favor Gov	−0.036 (0.045)	−0.223*** (0.016)
Against Gov	0.201 (0.247)	−0.068*** (0.023)
Support Gov X Favor Gov	0.074 (0.085)	0.005 (0.032)
Favor Gov X Against Gov	0.087 (0.077)	0.102*** (0.031)
Constant	3.284*** (1.073)	1.275*** (0.111)
Observations	3,805	3,805
R <sup>2</sup>	0.029	0.091
Adjusted R <sup>2</sup>	0.025	0.088
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01	

## C1. Time on Pages Figures

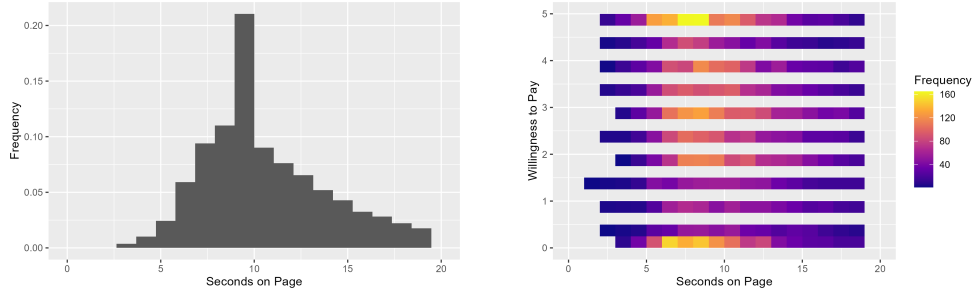


FIGURE C1. TIME SPENT ON EACH HEADLINE. ON THE LEFT SIDE, THERE IS THE WHOLE DISTRIBUTION, AND ON THE RIGHT SIDE, THE DISTRIBUTIONS BY WILLINGNESS TO PAY LEVEL.

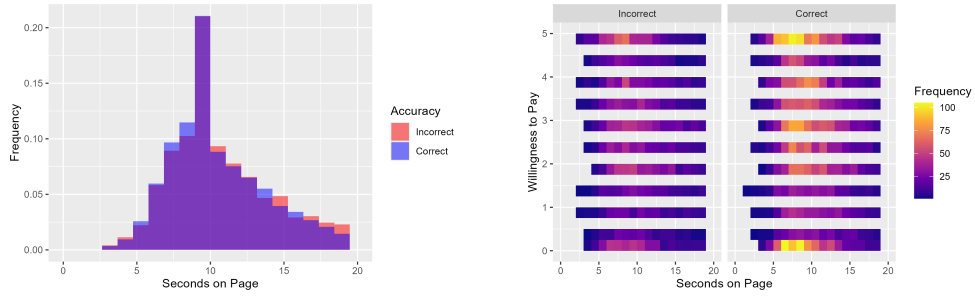


FIGURE C2. TIME SPEND ON EACH HEADLINE BY THE ACCURACY OF THE CLASSIFICATION. IN THE LEFT SIDE THERE IS THE WHOLE DISTRIBUTION, AND IN THE RIGHT SIDE THE DISTRIBUTIONS BY WILLINGNESS TO PAY LEVEL.

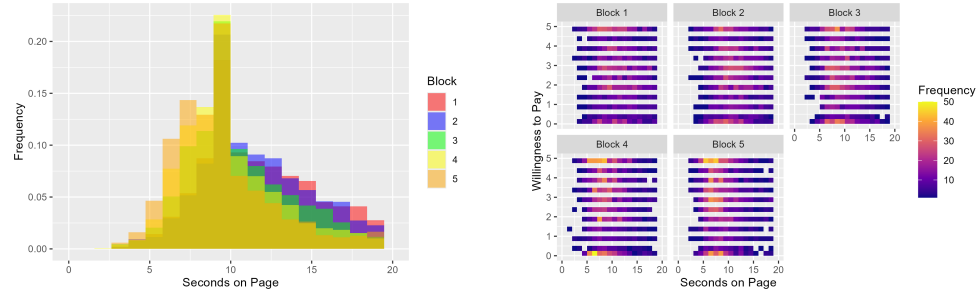


FIGURE C3. TIME SPEND ON EACH HEADLINE BY BLOCK. IN THE LEFT SIDE THERE IS THE WHOLE DISTRIBUTION, AND IN THE RIGHT SIDE THE DISTRIBUTIONS BY WILLINGNESS TO PAY LEVEL.

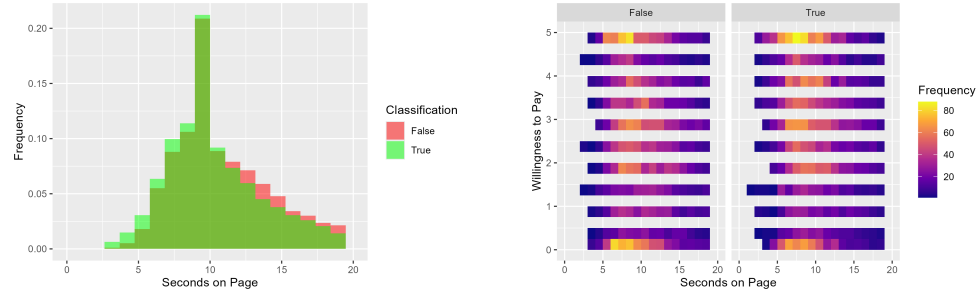


FIGURE C4. TIME SPEND ON EACH HEADLINE BY CLASSIFICATION. IN THE LEFT SIDE THERE IS THE WHOLE DISTRIBUTION, AND IN THE RIGHT SIDE THE DISTRIBUTIONS BY WILLINGNESS TO PAY LEVEL.

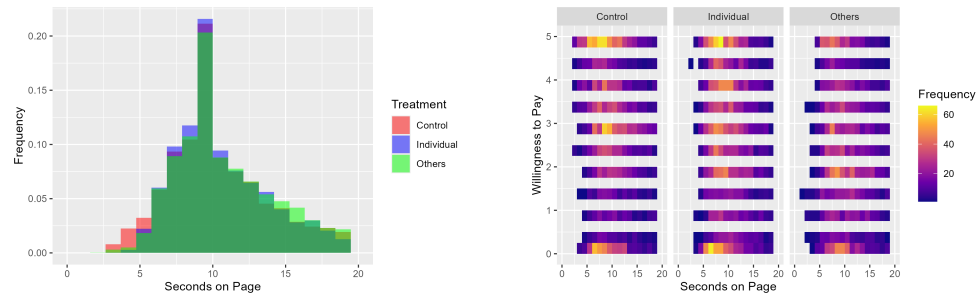


FIGURE C5. TIME SPEND ON EACH HEADLINE BY TREATMENT. IN THE LEFT SIDE THERE IS THE WHOLE DISTRIBUTION, AND IN THE RIGHT SIDE THE DISTRIBUTIONS BY WILLINGNESS TO PAY LEVEL.

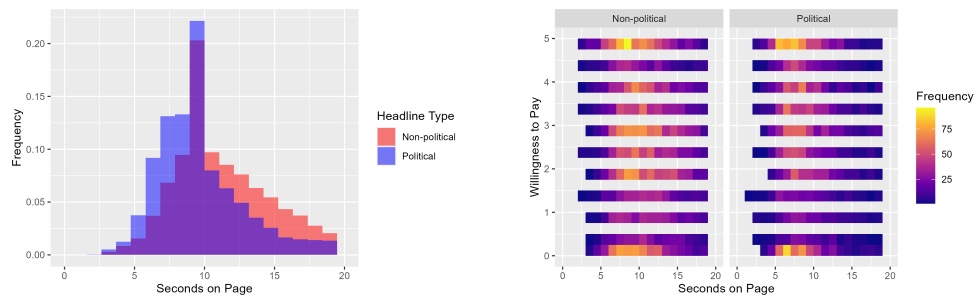


FIGURE C6. TIME SPEND ON EACH HEADLINE BY POLITICAL CONTENT. IN THE LEFT SIDE THERE IS THE WHOLE DISTRIBUTION, AND IN THE RIGHT SIDE THE DISTRIBUTIONS BY WILLINGNESS TO PAY LEVEL.