Journal of Behavioral and Experimental Economics 122 (2026) 102576

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

journal homepage: www.elsevier.com/locate/jbee

The impact of misinformation on consumer choices☆

Boris Knapp [a],*, Dominik Stelzeneder [b]

[a] Institute of Economics, Corvinus University of Budapest, fövam ter 8, Budapest, 1093, Hungary
[b] Vienna Graduate School of Economics, University of Vienna, Oskar-Morgenstern-Platz 1, Vienna, 1090, Austria

ARTICLE INFO

JEL classification:
C91
D84
G14
J16

Keywords:
Experiment
Misinformation
Beliefs
Bayesian updating

ABSTRACT

Misinformation – such as fake reviews or biased recommendations – poses a challenge for consumers that can be conceptualized as a two-step problem: First, they need to judge the credibility of the signal. Second, they need to update their prior beliefs conditional on their judgment and the signal realization. This paper reports on an experiment that disentangles both steps utilizing a novel approach that requires neither belief elicitation nor structural estimation of utility parameters. We utilize a mixed design, comparing subjects' decisions in one of three treatments to their individually optimal decision. The sequential nature of our experiment allows us to design a task that elicits the individually optimal decision directly. While we find no overall treatment effects, we uncover substantial heterogeneity: the interventions improved decision accuracy for male participants and for those with higher cognitive ability and education, but had no significant effect on others.

1. Introduction

Misinformation is increasingly polluting our information channels. People often rely on signals, such as ratings or recommendations, to reduce uncertainty about an unknown state of the world. Yet these signals are not always trustworthy. This poses a challenge that can be conceptualized as a two-step problem: first, a decision maker needs to form a belief about a signal's credibility, and then incorporate its content into a belief about the underlying state.

A classic example in consumer markets are online reviews. Designed as a means to reduce the information asymmetry between consumers and firms, review platforms like Amazon or Yelp are increasingly plagued by fake reviews.[1] Extracting information from reviews thus requires knowledge about the share of fake ones and the ability to correctly update beliefs.

Similar situations are ubiquitous: recruiters use letters of reference to infer candidates' abilities without knowing whether these letters are honest or overly positive. Investors get investment recommendations from platforms like *Stocktwits* but holding a position in a stock may compromise the objectivity of such a recommendation. Bond markets rely heavily on credit rating agencies (CRAs) which have been found to sometimes inflate ratings in order to foster business relations.[2]

In all of these examples, the pollution of information channels renders the signals less informative. Moreover, it complicates the information extraction process itself. First, a decision maker might have incorrect beliefs about the signal's credibility. If consumers incorrectly assess the prevalence of fake reviews, their inferences about product quality are likely to suffer. Second, the additional layer of uncertainty renders interpreting signals a more complicated Bayesian updating problem. Errors in Bayesian reasoning may thus also lead to wrong inference. The question we address in this paper is how misinformation affects consumer choices via mistakes in each of these two steps of the information extraction process.

To this end, we conduct an online experiment based on a simple cheap talk game we call *Trading Envelopes*: a sender observes the content of an envelope, either $5 or $0. Before sending it to a receiver, they label the envelope with one of two messages: "This envelope contains $5" or "This envelope is empty". Initially, the receiver only knows the prior probability that the envelope is empty. Based on the message and their belief about the share of dishonest senders, β, they can update their belief about the envelope's content. We elicit the receivers' willingness to accept (WTA) for the envelope across four treatments, varying the additional information provided to receivers.

☆ We thank the participants of seminars at the University of Vienna and the Corvinus University of Budapest for helpful comments. We gratefully acknowledge technical support from the *Vienna Center for Experimental Economics (VCEE)* that made their server available to us. Boris Knapp is a recipient of a DOC Fellowship of the Austrian Academy of Science at the Vienna Graduate School of Economics (VGSE). Dominik Stelzeneder gratefully acknowledges the financial support by the Vienna Graduate School of Economics funded by the Austrian Science Fund (FWF): W 1264 Doktoratskollegs (DKs).

* Corresponding author.
E-mail addresses: boris.knapp@uni-corvinus.hu (B. Knapp), dominik.stelzeneder@univie.ac.at (D. Stelzeneder).

[1] See He et al. (2022) or the report by SafetyDetectives research lab (2021).
[2] Jiang et al. (2012) find that issuer-paid ratings tend to be inflated compared to those paid for by investors.

https://doi.org/10.1016/j.socec.2026.102576
Received 21 February 2025; Received in revised form 23 April 2026; Accepted 24 April 2026
Available online 27 April 2026

In treatment T1, receivers are informed about the share of dishonest senders β. In treatment T2, receivers may use a "slider" that maps the message and any prior belief they might hold about β into a posterior probability that the envelope is empty. Treatment T3 combines both pieces of information while the baseline treatment T0 includes no additional information.

The treatments in our experiment are inspired by simple informational policies that can be implemented relatively easily and already have been to some extent. An example of this is *Fakespot.com*, where users can obtain an estimate of the proportion of fake reviews for a given listing together with an adjusted rating based on this estimate.

Two features of our experiment allow us to assess the effects of misinformation on the receivers' choices. First, due to the sequential nature of the *Trading Envelopes* game, we conduct the sender part of the study first. This allows us to inform receivers about the *actual* β. Second, we elicit receivers' *true valuation* for the envelope in a separate task in which they submit their WTA for a lottery ticket. Using the information on β, we designed the lottery to be *outcome equivalent* to the decision problem in the *Trading Envelopes* game. We thus interpret the WTA for the lottery ticket as the WTA a receiver would have in Task 1 *if* they had the correct belief *and if* they were able to update perfectly. The difference between a subject's WTA for the envelope and the lottery ticket then defines the *decision error*.

We find sizable and significant decision errors, however, none of our treatments could reduce them significantly when looking at the pooled sample of 365 receivers. Breaking down the sample by demographic variables and cognitive ability measures, a more nuanced picture emerges: some of our treatments significantly reduced decision errors in receivers that were male, highly educated, or scored highly on a cognitive reflection test while none of the treatments had a significant effect in the other subgroups.

While the effects of misspecified beliefs and incorrect updating on consumer choices are interesting in their own right, they also have important implications in many markets. Addressing the problem of misinformation directly has been of moderate success so far. Dealing with the consequences is therefore imperative. The simplicity of our treatments means that similar interventions can be implemented in most markets relatively easily. The heterogeneity of our treatment effects suggests that such interventions might be impactful in certain markets despite not being a silver bullet against the negative effects of misinformation.

1.1. Related literature

We contribute to the literature on the role of beliefs in games, which can be broadly divided into two strands. One strand infers beliefs from observed choices (e.g., Aguirregabiria, 2021; Schneider, 2019), while the other elicits beliefs directly and relates them to subsequent decisions (e.g., Costa-Gomes & Weizsäcker, 2008; Nyarko & Schotter, 2002). Our approach differs by using a treatment that directly induces correct beliefs and comparing the resulting decisions to those of an untreated control group. This design allows us to estimate the effect of misspecified beliefs on behavior without eliciting beliefs explicitly.

We further contribute to a growing literature that quantifies individual decision errors and the resulting losses in consumer surplus. Carpenter et al. (2021), Harrison, Morsink et al. (2020), and Harrison, Martínez-Correa et al. (2020) were among the first to measure such losses in expected consumer surplus. In their experiments, participants' pre-treatment choices were used to structurally estimate the parameters of individual utility functions, and post-treatment decisions were then compared to the benchmark choices that maximize those estimated utilities. Gao et al. (2023) advance this line of research by developing a Bayesian hierarchical framework for welfare evaluations of risky choices, which refines the estimation of individual risk preferences used in normative analysis. In contrast, our approach avoids structural estimation and assumptions about the exact functional form of utility. Instead, we derive each participant's optimal benchmark decision directly: a subject's choice in Task 2 serves as the benchmark for their optimal decision in Task 1. This allows us to quantify decision errors and welfare losses without imposing parametric restrictions on preferences.

Our study can also be interpreted through the lens of complexity aversion (Huck & Weizsäcker, 1999; Mador et al., 2000).[3] From this perspective, our Task 1 treatments correspond to different levels of complexity. While most prior work examining how lottery complexity affects choice behavior (e.g., Butler & Loomes, 2007; Moffatt et al., 2015; Sonsino et al., 2002), in our setting subjects evaluate the lotteries independently. Oberholzer et al. (2024) show that the effect of complexity depends on whether subjects make comparative or evaluative judgments. More recently, Georgalos and Nabil (2025) extend this line of research by testing alternative models of complexity aversion using data from Moffatt et al. (2015), concluding that a simple toolbox model provides the best fit.

Lastly, a related line of research investigates behavioral interventions aimed at mitigating the spread or impact of misinformation. Andı and Akesson (2021) show that a social-norm-based nudge can reduce individuals' intentions to share false news stories on social media platforms. Pennycook et al. (2020) study the effect of fact-checking labels on the perceived accuracy of news articles and the willingness to share them online. The most closely related study is Akesson et al. (2023), who provide the first experimental estimates of how fake online reviews affect individual demand and welfare. In an incentive-compatible online experiment, they show that fake reviews lead consumers to choose lower-quality products and reduce welfare by about 12 cents per dollar spent, but that an educational intervention reduces this welfare loss by nearly half. On the theoretical side, Gesche (2021) and Knapp (2025) examine how information policies – such as raising consumer awareness of deception risks or mandating the disclosure of conflicts of interest – can protect naïve consumers, albeit at the expense of informed ones.

1.2. Outline

The remainder of the paper is structured as follows. In the following Section 2, we analyze the *Trading Envelopes* game on which the experiment is based. Section 3 explains our experimental design. In Section 4, we introduce the concept of decision error and show how it relates to consumer surplus. In Section 5, we develop our hypotheses. In Section 6, we present preliminary findings and discuss issues related to the robustness of our experimental design. Our main findings are presented in Section 7. Section 8 concludes the paper.

2. Trading envelopes game

The experiment is based on the following sender–receiver game. There is an unknown state of the world, $\omega \in \{v_R, 0\}$, with the commonly known probability distribution $Pr(\omega = v_R) = Pr(\omega = 0) = \frac{1}{2}$. In our setting this state corresponds to the content of an envelope which can either contain a monetary payoff of v_R or be empty. There is a sender (he) who observes ω and sends a binary message $m \in \{A, B\}$ to an uninformed receiver (she). The messages have natural interpretations such that $m = A$ corresponds to $\omega = v_R$ and $m = B$ to $\omega = 0$.[4] There is no *monetary* cost associated with sending any of the messages

[3] We hesitate to relate our study to the literature on ambiguity aversion because we believe that this would conflate the concepts of strategic uncertainty and ambiguity.

[4] On the one hand, this is true because of the statements themselves: A : "This envelope contains \$5." and B : "This envelope is empty." On the other hand, these interpretations are facilitated additionally by two restrictions that we introduce below.

but we assume that senders are averse to lying. The receiver reads the sender's message and then takes a binary action $a \in \{1, 0\}$ where $a = 1$ corresponds to opening the envelope, receiving its content, and $a = 0$ corresponds to foregoing opening the envelope, taking the outside option instead. The outside option is drawn uniformly from the interval $[0, v_R]$.

The sender receives a payoff of $v_S > 0$ if the receiver opens the envelope and nothing otherwise. On top of that he bears a *psychological* cost associated with lying about the state of the world. We assume that the sender's utility function is additively separable, i.e.,

$$u_S(a, \omega, m) = au(v_S) - c(\omega, m). \tag{1}$$

As described above, the receiver's payoff is either the envelope's content or her outside option, depending on the action taken:

$$u_R(a, \omega, y) = au(v_R) + (1 - a)u(y). \tag{2}$$

The function $u : \mathbb{R} \mapsto \mathbb{R}$ with $u(0) = 0$ and $u'(x) > 0 \forall x \in \mathbb{R}$ denotes a subjects vNM-utility function from receiving monetary payoffs. The function $c_i : \{v_R, 0\} \times \{A, B\} \mapsto \mathbb{R}$ captures subject i's lying aversion, i.e., their disutility from sending message m when the state of the world is ω.

We impose two restrictions on the players' strategies. First, the sender automatically sends $m = A$ when $\omega = v_R$. That is, he only makes a decision when the envelope is empty in which case he can either send the truthful message B or lie by sending message A. Second, the receiver automatically rejects the envelope ($a = 0$) upon reading $m = B$. This means that she only makes a decision in case the message is A.

One consequence of these restrictions is that they simplify the game such that the receiver's problem boils down to estimating the probability that the sender sends $m = A$ when the envelope is empty. This *lying probability* plays a central role and we denote it by $\beta = Pr(m = A|\omega = 0)$ Likewise, the sender cares about the receiver's belief about β when deciding whether to lie or not.

Another consequence is that a sender's lying aversion can be captured by a single parameter. The only decision a sender faces is whether to tell the truth or lie when $\omega = 0$. Therefore they only care about the difference in lying costs in these two cases, $c_i(0, A) - c_i(0, B)$. We denote this difference by $c_i \equiv c_i(0, A) - c_i(0, B)$ and assume that it is drawn from a commonly known, continuous, and atomless distribution F with support $[0, \bar{c}]$.[5]

The game as described above is illustrated in Fig. 1 and has a unique Perfect Bayesian Equilibrium (equilibrium henceforth) which we now derive. To save on notation, we denote $\bar{v}_S = u(v_S)$ and $\bar{v}_R = u(v_R)$.

Suppose that the sender lies with some probability β. The receiver then compares her outside option with her expected utility from opening the envelope, given β. Consequently, she opens the envelope if and only if $u(y) < \bar{v}_R/(1 + \beta)$.[6] The probability of this is

$$p(\beta) = \frac{1}{v_R} u^{-1}\left(\frac{\bar{v}_R}{1 + \beta}\right), \tag{3}$$

which is strictly decreasing in β due to our assumption that $u' > 0$. Moreover, $p(0) = 1$.

Given $\omega = 0$, the sender's expected utility from lying is $\bar{v}_S\, p(\beta) - c_i(0, A)$ while the expected utility from telling the truth is $-c_i(0, B)$. The sender will thus lie as long as $c_i < \bar{v}_S\, p(\beta)$.[7] The equilibrium lying probability is determined by the equation

$$\beta = F\left(\bar{v}_S\, p(\beta)\right). \tag{4}$$

Since $F\left(\bar{v}_S\, p(0)\right) > 0$, F and u are strictly increasing, and p is strictly decreasing, (4) always has a unique solution.[8]

3. Experimental design

Our experiment is based on the *Trading Envelopes* game presented in the previous section. Each sender has private information about the content of an envelope, either \$0 or \$5, and must send one of two messages to the receiver of the envelope:

A: "This envelope contains \$5."
B: "This envelope is empty."

Subsequently, the receiver decides whether to open it, receiving its content, or sell it for an amount that has been randomly drawn from [\$0.00, \$5.00]. The sender receives \$3 if the envelope is opened and \$0 otherwise. Both senders and receivers are informed about this structure of the game which is illustrated in Fig. 2.

3.1. Senders

We distribute 200 envelopes – 100 empty, 100 containing \$5 – randomly among 200 senders. Before observing their envelope's content, the senders have to choose between messages A and B for the case that it is empty. In the case that the envelope contains \$5, message A is sent automatically. After they make a choice, they are informed about their envelope's content and have to fill out a short demographic questionnaire. The envelope, labeled with one of the two messages, is then sent to a randomly selected receiver to be either opened or sold unopened. If the envelope is opened, the sender earns a bonus of \$3 on top of a \$2 participation fee. Otherwise, they receive the participation fee but no bonus.

3.2. Receivers

The receiver part of the experiment consists of six tasks in total. After reading the instructions and seeing the graphical illustration of the experiment in Fig. 2, participants answer six comprehension questions about the experiment. Immediately thereafter, they receive feedback along with a brief explanation why the correct answers were correct. This comprehension quiz not only allows us to screen participants based on their understanding of the instructions but it also reinforces the key points before starting the experimental tasks.

Task 1 corresponds to the *Trading Envelopes* game. Each receiver is randomly assigned one of 200 envelopes. They know that 100 of them contain \$5 while the remaining ones are empty, and that a sender labeled each of the envelopes according to the rules explained in Section 3.1. Thus, the envelope is labeled either with message A ("This envelope contains \$5.") or with message B ("This envelope is empty."). Before finding out which of the two messages is written on their envelope, they have to name their (willingness to accept) WTA for the envelope for the case that it is message A. That is, they must specify a minimum selling price between \$0.00 and \$5.00.[9] Only later are the participants informed about the label of their envelope. If the envelope is indeed labeled with message A, the computer generates a random amount from [\$0.00, \$5.00] and offers it in exchange for the envelope. If that offer is at least as high as the stated WTA, the envelope is sold and the receiver receives the offered amount. They do not learn about the content of the envelope in this case. Otherwise, if the offer is below the WTA, the envelope is not sold but opened instead. In this case, the receiver gets its content, either \$0 or \$5. If, however, the message

[5] This implies $c_i(0, A) \geq c_i(0, B)$, i.e., no subject enjoys lying. This assumption is not crucial and we could instead assume the more general support $[\underline{c}, \bar{c}]$.

[6] Note that the exact way in which ties are broken is irrelevant because the distribution of outside options is atomless. We assume for simplicity that ties are broken in favor of the outside option.

[7] Again, the tie-breaking rule is irrelevant.

[8] Note that the equilibrium value of β increases with v_S. We used $v_S = 5$ in our pilot study and $v_S = 3$ in the final experiment resulting in $\beta = 0.73$ and $\beta = 0.57$, respectively. While this difference is not statistically significant, it has the sign predicted by this model.

[9] They do not have to specify their WTA for the case that it is message B because a letter labeled with message B is sold automatically, which is equivalent to a WTA of 0.

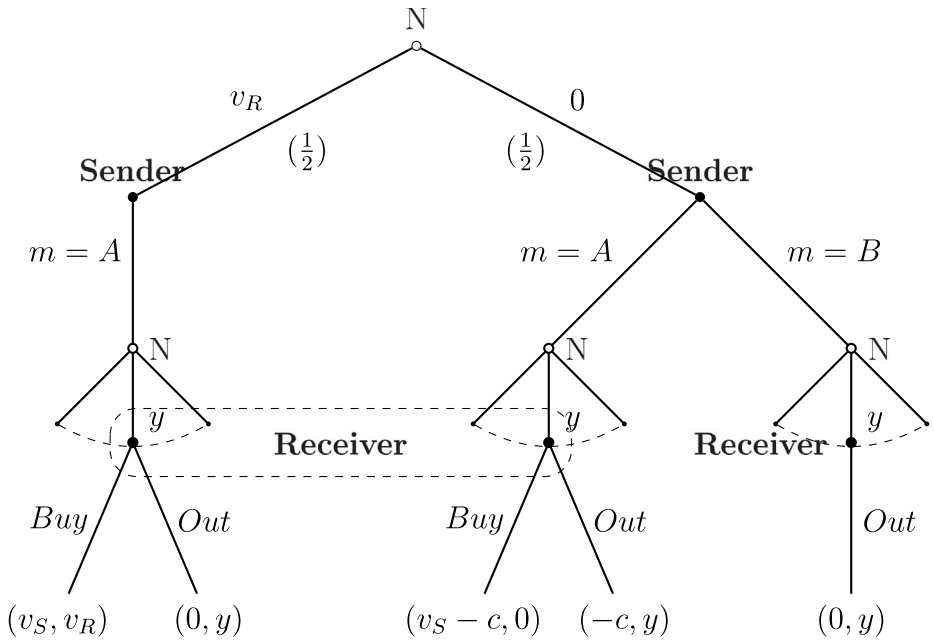

Fig. 1. Game tree of the Trading Envelopes game.

is B, then the envelope is sold automatically for the offered amount. We thus elicited valuations using the Becker–DeGroot–Marschak (BDM) mechanism, a standard incentive-compatible procedure in experimental valuation tasks (Becker et al., 1964).[10]

In Task 2, each receiver is given a lottery ticket with a winning probability based on β, the share of senders who lied. It is calculated to be equal to the posterior probability that an envelope showing message A contains \$5. The outcome of the lottery has two components, a payoff to the receiver and a payoff of \$3 to one sender. If the ticket wins, the receiver gets \$5 and the sender who gets the payoff of \$3 is randomly chosen among those whose envelopes contained \$5. If it loses, the receiver gets \$0 and the sender is randomly selected among those whose envelopes contained \$0 and who chose to label them with message A. With the same probability that the envelope in Task 1 shows message B, the lottery ticket is sold automatically. Thus, the lottery is *outcome equivalent* to Task 1. As in Task 1, we elicit subjects' WTA for the lottery ticket via the BDM method. We do not inform participants about the outcome of a lottery if the ticket is sold. The purpose of Task 2 is to elicit receivers' valuation for the envelope absent strategic uncertainty. We elaborate on this in Section 4.

Task 3 consists of a lottery similar to the one in Task 2, with the only difference that no sender receives any payoff. This lottery is thus equivalent to the lottery in Task 2 except for the social externality and allows us to measure the extent to which receivers' preferences are other-regarding. As in Tasks 1 and 2, no feedback is given before the end of the experiment.

Task 4 is a cognitive ability test consisting of two parts. The first part is the *Cognitive Reflection Test* introduced by Frederick (2005), which contains three questions with intuitively appealing but incorrect answers.[11] Arriving at the correct answers typically requires more deliberate reasoning. Participants earn 0.40 for each correct response. The second part consists of 34 *Raven's Progressive Matrices* (Raven et al., 1998). Each matrix displays six or nine figures arranged according to a pattern, with one figure missing. The participants must select the figure that correctly completes the matrix. They have three minutes to solve as many as possible and earn 0.30 per correct solution. Before the timer starts, we explain the task with an example. Task 4 provides a measure of receivers' cognitive ability, which we expect to be an important factor in their decisions.

In Task 5, we use the frequency method (Schlag & Tremewan, 2021) to elicit subjects' beliefs about the share of senders who lied. They receive a payoff of \$1 if their answer deviates by less than three percentage points from the true β and nothing otherwise.

Task 6 consists of 20 decision problems. For each of these problems, the participants choose between a risky lottery and a safe option. The lottery is the same throughout, yielding \$1 with a probability of 75% and \$0 otherwise. The safe option is a payoff of 5 cents in the first decision problem and increases by 5 cents for every subsequent problem. Hence, the twentieth problem is a choice between the lottery and a payoff of \$1. Of these 20 decisions, one is randomly selected for payment.

The BDM mechanism used in Tasks $1-3$ is incentive compatible only if preferences are monotone (Azrieli et al., 2018). Task 6 therefore serves as a monotonicity test, allowing us to identify and exclude participants whose choices violate this assumption. To do so, we count the number of times a participant switches between the lottery and the safe option. Under monotonicity, a participant can have at most one switching point, and it must be from the lottery to the safe option. The reason is that after switching from the lottery to the safe option, all subsequent safe options yield larger payoffs and are thus strictly preferred under monotonicity.

[10] While the BDM mechanism has an established theoretical rationale and extensive empirical use in experimental valuation tasks, prior work has emphasized practical caveats — notably that subjects' misunderstanding or uncertainty about their valuation can introduce additional noise and, in some settings, weaken the mechanism's incentive-compatibility (Horowitz, 2006; Plott & Zeiler, 2005).

[11] Haigh (2016) raises concerns about repeated exposure due to the test's popularity. More recently, Bialek and Pennycook (2018), Meyer et al. (2018), and Stagnaro et al. (2018) show that the CRT retains its validity despite subjects' prior exposure to it.

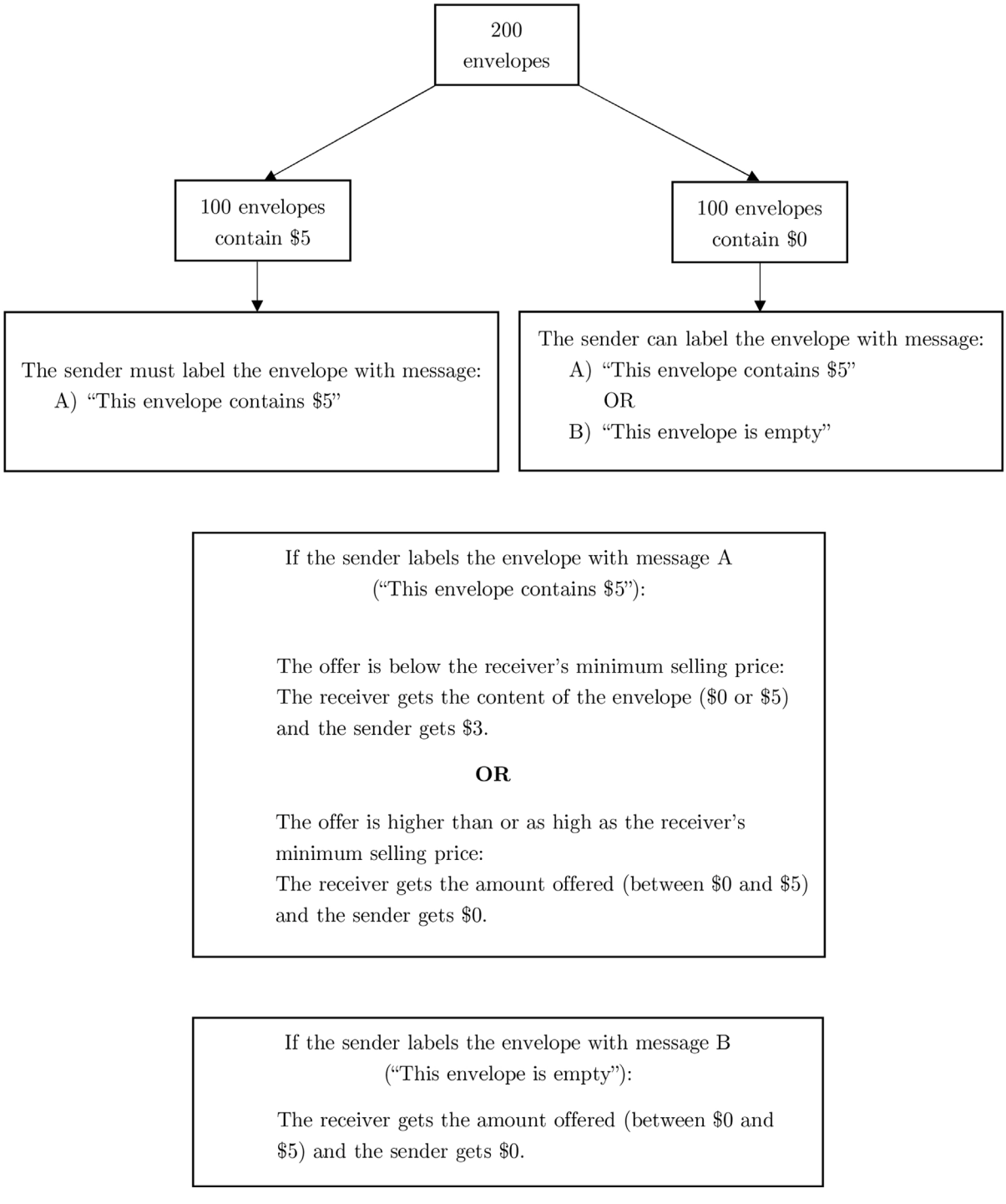

Fig. 2. Graphical illustration of the sender–receiver game.

Table 1
Treatments.

		Correct β	
		No	Yes
Slider	No	T0 (*baseline*)	T1 (*beta*)
	Yes	T2 (*slider*)	T3 (*full info*)

Notes. The table shows the 2 × 2 between-subject experimental design for Task 1.

Finally, participants have to fill out a short demographic questionnaire before we inform them about the outcomes and payoffs in a randomly chosen payoff-relevant task. In addition to what they earn in this task, they receive a participation fee of $4.[12]

3.3. Treatments

Our experiment employs a 2 × 2 between-subject design, varying the conditions in Task 1. The motivating idea is that for choosing optimally in Task 1, participants need to solve a two-steps problem. First, they must form an accurate belief about β, the share of senders who lied. Second, based on this belief, they have to calculate the posterior probability that an envelope with message A contains $5. Our design varies the information provided in Task 1 with the baseline treatment ($T0$) providing no additional information beyond the prior probability that any given envelope contains $5.

In treatment T1, we inform receivers about the exact share of senders who lied. To this end, we exploit the experiment's sequential nature, running the sender part first to obtain the realized value of β before running the receiver part.

In treatment T2, participants are given a tool specifically designed to help them calculate the posterior probability that their envelope is full or empty, based on their belief about β. This tool is presented as a slider, which can be adjusted from 0% to 100%, corresponding to different beliefs about β. The slider calculates the posterior probability that the envelope contains $5, given message A and the participant's belief about β, based on Eq. (5). It uses Eq. (6) to calculate the probability that the envelope is empty under the same conditions. The outcomes of these calculations are displayed below the slider, allowing participants to see the probabilities that the envelope is full or empty, based on their chosen belief.

$$\pi(\$5|A) = \frac{1}{1+\beta} \tag{5}$$

$$\pi(\$0|A) = \frac{\beta}{1+\beta} \tag{6}$$

Treatment T3 combines both pieces of information from T1 and T2. Participants are informed about the exact value of β *and* are provided with a slider to map their beliefs into posterior probabilities. Thus, T3 eliminates the strategic uncertainty and computational complexity of Task 1, making it equivalent to the lottery in Task 2 in terms of payoff and information. Table 1 summarizes the treatment design.

3.4. Implementation

The experiment was programmed in *oTree* (Chen et al., 2016), run on the servers of the *Vienna Center for Experimental Economics*, and carried out via *Prolific* in January 2022.[13] It was not pre-registered. The experiment consisted of two parts, one for the senders and one for the receivers. As explained in Section 3.3, we ran the two parts separately, starting with the senders. Our samples consisted of US residents representative of the US population in age, ethnicity, and sex, based on US Census data.[14]

We recruited 200 senders as follows. Of the 325 participants who accessed our study, 302 completed it. These 302 participants had to label their envelope and complete an additional task that consisted of the same *Cognitive Reflection Test* as in receivers' Task 4. The sole purpose of this additional task was to increase the number of recruited senders to 300, which allows access to *Prolific*'s representative sample option. Senders were randomly assigned a type in $\{0, 1, 2\}$, which determined the task relevant for them. For senders of types 0 and 1, the envelope-labeling task was relevant and they were allocated empty and full envelopes, respectively. For type 2 senders, the additional task was relevant. Consequently, they were not part of the subsequent stage of the experiment. There were 102 senders of type 2, leaving us with 200 senders that were part of the following stage. On average, senders took 6.2 min to complete the study and earned $3.16.

For the second part of our experiment, we recruited 400 receivers. Out of 446 participants who accessed our study 400 finished it. One participant did not give consent for data processing and 45 dropped out before reaching the end. At the beginning of the study, they were assigned to one of our four treatments and had to complete all the tasks described in Section 3.2. At the end of the study, just before their payoffs were displayed, they were matched with a sender. This matching simultaneously determined whether their envelope was empty or not, which label was on it, and which of the Tasks 1-4 was selected to be payoff relevant. Tasks 5 and 6 were payoff relevant for all receivers. Finally, participants filled out a short demographic questionnaire and had the possibility to give feedback. On average, receivers took 18.5 min to complete the study and earned $9.72.

The precise matching mechanism was implemented as follows. Each sender was matched with two receivers. One of these receivers was assigned to either Task 1 or Task 2, and the other to either Task 3 or Task 4. This procedure ensured that every sender was paired with exactly one receiver whose payoff depended on the sender's envelope label. It also guaranteed that behind every lottery ticket in Task 2 there was a sender who actually received the external payoff. This matching protocol allowed us to keep our instructions relatively simple while avoiding any misleading or incomplete descriptions of the experimental environment.

Matching proceeded sequentially in the order participants finished the study. Specifically, receivers 1 and 3 were matched with sender 1, receivers 2 and 4 with sender 2, and so on. The task assigned to each receiver followed a repeating sequence determined by their rank modulo 4: receiver 1 had Task 1, receiver 2 Task 2, receiver 3 Task 3, receiver 4 Task 4, receiver 5 Task 1, and so forth.

4. Decision error and welfare

Our main focus is on the effect of our treatments on subjects' decision errors and how this in turn affects their consumer surplus (CS). To evaluate decision errors, we compare a participant's reported WTAs in Tasks 1 and 2. Recall that Task 2 is outcome equivalent to Task 1, differing only in what information is available and how it is presented. In particular, Task 2 is equivalent to Task 1 given that the receiver has the correct estimate of β and perfectly updates their beliefs given this estimate. Based on this consideration, we define the decision error as the absolute difference of the WTAs in the two tasks.

[12] The participation fee for receivers was higher than for senders because the experiment took longer for receivers.

[13] The standard caveat for online experiments applies: participants were not actively monitored. Compared to the dominant online platform, *MTurk*, participants on *Prolific* exhibit lower rates of inattention (Albert & Smilek, 2023; Douglas et al., 2023). In fact, experimental data collected via *Prolific* is of similar quality as data from established university labs (Gupta et al., 2025; Suri et al., 2025).

[14] *Prolific* provides sample stratification by age, ethnicity, and sex according to US or UK Census data. We selected the US option for greater diversity and broader relevance.

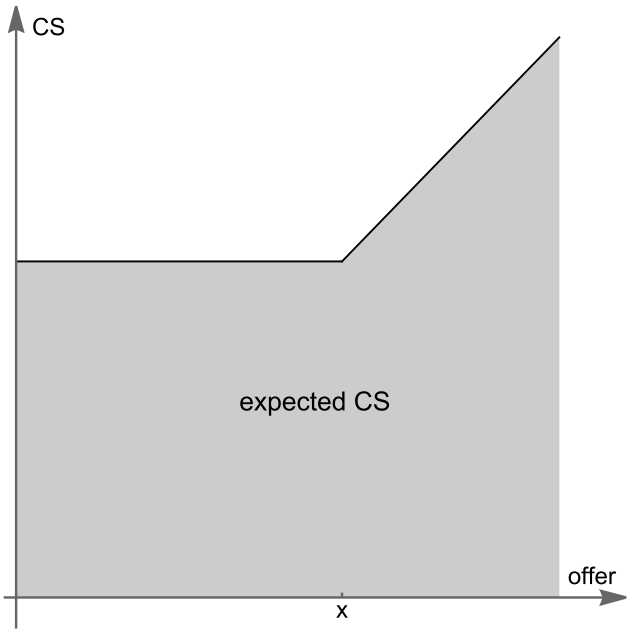

Fig. 3. Expected consumer surplus in the case where WTA1 is equal to the true valuation.

Our approach of defining the decision error this way does not come without assumptions. In particular, an underlying assumption is that a receiver's WTA in Task 2 perfectly reveals their valuation for the lottery ticket. In other words, we assume that an error can occur in Task 1 but not in Task 2. If errors occur in both tasks, our estimates will be biased. We discuss this possibility in Appendix A in detail but the main takeaway is as follows: First, if noise occurs in Task 2 as well as in Task 1, we *overestimate* the decision error. Second, if treatments T1–T3 do not increase the noise in Task 1 compared to the baseline treatment T0, we *underestimate* the treatment effects. The assumption that providing better information should not introduce more noise is reasonable in our opinion and also supported by our findings. This may in part explain why we find large and significant baseline decision errors, whereas the treatment effects in the pooled sample are not significant.

The benefit of defining the decision error based on WTAs is that, not only can we measure the decision error itself but also the associated loss in consumer surplus without the need for estimating utility functions. In the following analysis, we show how the magnitude of a participant's decision error directly affects their expected CS.

Let us first assume that the decision error is 0. This means that a subject's valuation of the envelope, x, and their WTA, y, are the same, i.e. $y = x$. In this case, the expected CS can be calculated as follows: They obtain a surplus of x whenever the randomly drawn offer in exchange for the envelope is below x and the offered money otherwise. Recall that the offer is a randomly drawn amount from the interval $[0, 5]$. Let us denote the offer by z and its cdf by $F(z) = \frac{z}{5}$. Then,

$$CS_{no\_error} = \int_0^x x\,dF(z) + \int_x^5 z\,dF(z) = F(x)x + \int_x^5 \frac{z}{5}dz = \frac{x^2}{5} + \frac{z^2}{10}\Big|_x^5$$
$$= \frac{2x^2}{10} + \frac{5^2}{10} - \frac{x^2}{10} = \frac{x^2 + 5^2}{10}$$

Given that a subject's valuation of the envelope lies in $[0, 5]$, CS_{no_error} takes its minimum at $x = 0$ where it is 2.5 and its maximum at $x = 5$ where it is 5. The shaded area in Fig. 3 illustrates the CS in the case where WTA and valuation coincide, i.e. when the decision error is zero.

Let us now turn to the case in which the subject makes a decision error by submitting a WTA that is not equal to their valuation of the envelope. As before, x denotes the subject's valuation for the envelope and y their WTA, only now $y \neq x$. Then, the expected CS in Task 1 can be computed as:

$$CS_{with\_error} = \int_0^y x\,dF(z) + \int_y^5 z\,dF(z) = F(y)x + \int_y^5 \frac{z}{5}dz = \frac{xy}{5} + \frac{z^2}{10}\Big|_y^5$$
$$= \frac{2xy}{10} + \frac{5^2}{10} - \frac{y^2}{10} = \frac{x^2 - (x-y)^2}{10} + \frac{5^2}{10}$$
$$= CS_{no\_error} - \frac{(x-y)^2}{10}$$

And the foregone consumer surplus is simply:

$$CS_{foregone} = \frac{(x-y)^2}{10} \tag{7}$$

Thus, the loss in consumer surplus is proportional to the decision error squared. Graphically, this can be illustrated as in Fig. 4. The left panel depicts the case where the WTA is below the actual valuation of the envelope. In this case, whenever the offer is between the WTA and the valuation, the participant receives y although they would have valued opening the letter with $x > y$. Mathematically, the loss in expected CS is the integral of the difference over all such offers, weighted by their probability densities. This corresponds to the shaded area. The case where the WTA is above the valuation is depicted in the right panel of Fig. 4. In both cases, it is thus the difference between valuation and WTA, i.e. the *decision error*, that determines the loss in expected CS.

5. Hypotheses

In this section, we develop our hypotheses. The main research question we ask is how (i) misspecified beliefs about β, and (ii) mistakes in Bayesian updating affect decisions in Task 1. In particular, we are interested in the effect of (i) and (ii) on the decision error and consumer surplus.

According to our model in Section 2, the receiver's expected utility from opening the envelope is $\bar{v}_R(1+\beta)$. A misspecified belief about β will therefore lead to over- or underestimation of the expected utility, resulting in a deviation from the *true* WTA. Because we define the decision error as the absolute difference between a subject's stated WTA (Task 1) and their *true* WTA (Task 2), misspecified beliefs unambiguously lead to an increase in decision error and decrease in CS. Therefore, treatment T1 should have a negative effect on decision error and foregone CS.

Mistakes in Bayesian updating that cause subjects to overstate or understate their WTA likewise unambiguously increase decision error and decrease CS.

The direction of the joint effect depends on several factors, including the types of updating mistakes participants make and the extent to which these mistakes correlate with their beliefs. In particular, it is possible that the two sources of error partially offset one another, producing a joint effect smaller than the sum of the individual effects. We do not have a strong prior regarding the precise nature of these interactions. Our only prediction is that the two types of mistakes do not fully cancel out and that the overall effect remains negative.

Hypothesis 1. *All three treatments, T1–T3, have negative effects on decision error and forgone consumer surplus.*

In an additional analysis, we look at the heterogeneity of treatment effects with respect to sex, cognitive ability, age, and education. This analysis is exploratory and we therefore abstain from formulating concrete hypotheses.

6. Descriptive statistics and preliminaries

Here we present the descriptive statistics for senders and receivers before discussing issues related to monotonicity and carry-over effects. Finally, we present balance checks to verify that our matching protocol successfully achieved randomization across treatments.

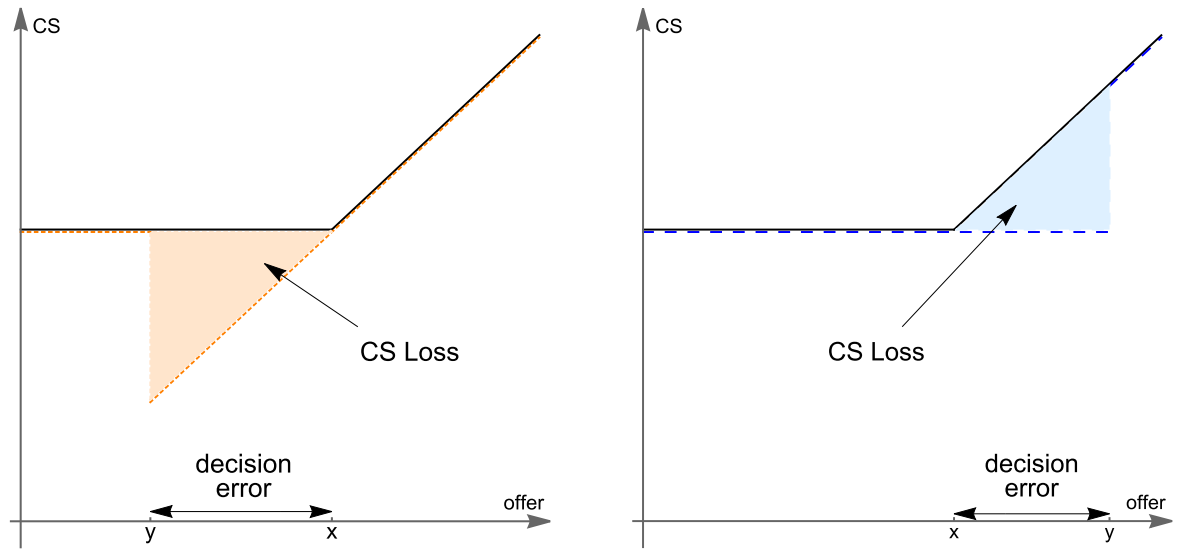

Fig. 4. The shaded areas illustrate the loss in CS due to decision error. The left (right) panel depicts the case where WTA1 is lower (higher) than the true valuation.

Table 2
Descriptive statistics.

Variable	Type (1)	Scale (2)	Mean (3)	Std. Dev. (4)	Min (5)	Median (6)	Max (7)
Task 1 *(Envelope)*	Cardinal	0.00 to 5.00	3.16	1.18	0.00	3.00	5.00
Task 2 *(Lottery 1)*	Cardinal	0.00 to 5.00	3.13	1.19	0.00	3.00	5.00
Task 3 *(Lottery 2)*	Cardinal	0.00 to 5.00	3.23	1.24	0.00	3.00	5.00
Task 4A *(CRT)*	Cardinal	0 to 3	1.92	1.14	0	2	3
Task 4B *(Raven)*	Cardinal	0 to 34	15.20	4.06	1	15	26
Task 5 *(Beliefs)*	Cardinal	0 to 100	55.72	20.74	0	57	100
Female	Binary	0(m), 1(f)	0.49	0.50	0	–	1
Age	Cardinal	1 to 99	45.26	15.90	19	45	79
Education	Ordinal	1 to 5	1.96	0.89	1	2	5
Payoff (in $)	Cardinal	4.00 to 22.40	9.72	2.29	4.00	10.00	16.10

Notes. Descriptive statistics of the selected sample of receivers, excluding participants with non-monotone preferences. N = 365.

6.1. Senders' decisions

In the first part of the experiment, the senders had to choose a label for their envelope. 57% of all senders who received an empty envelope labeled it with message A ("This envelope contains $5."). Envelopes that contained $5 were automatically labeled with message A. This means that with a probability of $0.5 \times (1 - 0.57) = 0.215$ the envelope in Task 1 was sold automatically because it showed message B ("This envelope is empty."). Moreover, the share of senders that labeled an empty envelope with message A determined the probability that an envelope with message A contained $5 according to Eq. (5). Thus, an envelope that showed message A contained $5 with a probability of 0.64.

Our experimental design requires outcome equivalence between Task 1 and the lottery in Task 2. We obtained outcome equivalence by setting the probability that the lottery ticket was sold automatically to 0.215, and the winning probability to 0.64.

6.2. Receivers' decisions

In this section, we present descriptive statistics and preliminary analyses of the receiver part of the experiment. In what follows, participants whose choices in Task 6 violated monotonicity are already excluded (see Section 6.3).

Table 2 shows summary statistics of the outcomes in Tasks 1-5 and the demographic variables we use for the heterogeneity analysis. Fig. 5 shows the average WTAs in Task 1 across treatments. The treatment effects on WTA1 are insignificant for treatments T2 (0.083, $p = 0.622$) and T3 (−0.005, $p = 0.979$) and negative for T1 (−0.399, $p = 0.030$). As we will see in Section 6.4, potential carry-over effects in T1 warrant caution in interpreting this effect. Figure B.6 in Appendix B shows the distributions of WTAs in Task 1 for each of our 4 treatments.

6.3. Monotonicity

We assessed the monotonicity of our subjects' preferences using Task 6 as described in Section 3.2. It consisted of 20 choices between a constant risky lottery and a strictly increasing safe outcome. A participant with monotone preferences switches at most once from the lottery to the safe outcome. In our experiment, 35 participants switched more than once, violating monotonicity. They were thus excluded from the analysis.[15]

[15] Note that subjects whose choices in Task 6 were consistent with monotonicity might not in fact have monotone preferences. This may occur, for example, if their choices in Task 6 were random and looked monotone 'by chance'. Likewise, the fact that a subject failed to display monotonicity in Task 6, does not necessarily mean that their preferences are non-monotone. It may be, for instance, that they got bored by the time of Task 6 and started choosing randomly. This limitation is common to all experiments where monotonicity is elicited in a separate task.

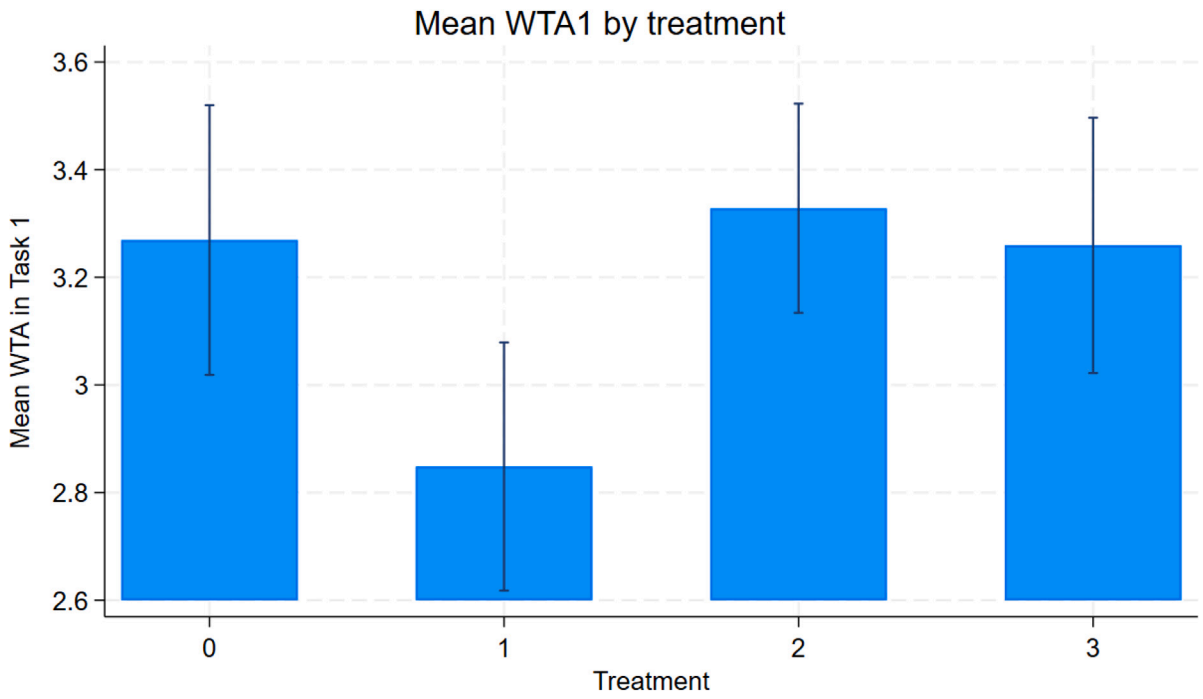

Fig. 5. Means of WTAa in Task 1 across treatments in the pooled sample with their 95% confidence intervals.

6.4. Carry-over effects

Next, we turn to carry-over effects which occur when a treatment in one task affects the decision of a participant in a subsequent task (Charness et al., 2012; Greenwald, 1992). In our experiment, this means that the treatment in Task 1 could have affected decisions in Task 2. In this case, our judgment about participants' decision errors could be biased. However, since we are interested in how decision errors vary across treatments our main concern is that carry-over effects differ across treatments. Therefore, we test for *carry-over imbalances*, i.e., whether carry-over effects on Task 2 vary across treatments. We do this in two ways.

First, we regress the decisions in Task 2 on the categorical treatment variable. If the carry-over effects are balanced, we should not find significant differences in means across the four treatments. The results of this first test are shown in column (3) of Table 3. We see that treatment T1 (*correct beta*) had a marginally significant effect on decisions in Task 2 (−0.34 compared to treatment T0, $p = 0.071$). Decisions did not significantly differ between T0 (baseline), T2 (slider), and T3 (full info). Although the joint *F*-test does not reject the *null* ($p = 0.212$), it seems possible that treatment T1 affected decisions in Task 2 differently than the other treatments.[16] This potential carry-over imbalance in treatment T1 could bias our treatment effect estimates. Therefore, all estimates for treatment T1 have to be interpreted with caution.

Second, we test whether our treatments affected the share of participants who made two identical decisions in Tasks 1 and 2 due to a "taste for consistency" (Falk & Zimmermann, 2012). In our opinion, this is the most likely mechanism for carry-over effects. In particular, subjects might have adjusted their decision after recognizing that Tasks 1 and 2 were identical in terms of outcomes. Column (5) of Table 3 reports the share of decisions that coincide by treatment. These shares do not differ significantly across treatments. Neither the joint test ($p = 0.532$, Fisher's exact test) nor *Wald* tests that compare all treatments pairwise are significant at conventional significance levels. Thus, while subjects seem to have a "taste for consistency", with approximately 1/3 of them choosing the same WTA in both tasks, we do not find evidence for imbalances between treatments.[17]

Table 3
Carry-over imbalances.

Treatment	N	Mean WTA2	Treatment effects	T*i* = T0	WTA1=WTA2
	(1)	(2)	(3)	(4)	(5)
T0 (*baseline*)	93	3.26	–	–	32%
T1 (*beta*)	95	2.92	−0.34	$p = 0.071$	42%
T2 (*slider*)	91	3.10	−0.16	$p = 0.349$	34%
T3 (*full info*)	86	3.24	−0.02	$p = 0.902$	36%
All	365	3.13	–	Joint F-test: p = 0.212	Fisher's exact: p = 0.532

Notes. The table reports the analysis of carry-over imbalances. Rows (1) and (2) show the number of participants and the average WTA in Task 2 for each treatment, respectively. Row (3) reports the treatment effects of Task 1 treatments on WTAs in Task 2, with corresponding p-values shown in row (4). Row (5) reports the share of respondents who reported the same WTA in Tasks 1 and 2.

6.5. Balance checks

Table 4 shows the mean values of six control variables across the four treatments in columns (1)–(4). These include our demographic controls (Female, Age, Education), our measures of cognitive ability

[16] Results of our balance checks (see Table 4) do not suggest that a randomization issue caused the observed differences.

[17] A common remedy for carry-over effects in within design is to randomize the order of treatment exposure and to test for order effects. In our experiment, carry-over effects potentially occur between two *tasks*, not between two *treatments*. In fact, we expose each subject to only one treatment. Since we are interested in *carry-over imbalances* between treatments rather than *carry-over effects*, randomization is not helpful in our case. While it would allow us to identify carry-over effects on Task 2 it would also introduce new carry-over effects on Task 1. However, since we vary Task 1 across treatments we cannot test for carry-over imbalances for that task. Thus, randomization would do more harm than good.

Table 4
Balance checks.

Variable	T0 (1)	T1 (2)	T2 (3)	T3 (4)	Fisher's exact (5)
Female	0.53	0.52	0.42	0.49	p = 0.446
Comprehension	4.98	5.09	5.02	5.16	p = 0.265
CRT	1.83	1.96	1.89	2.00	p = 0.869
Raven	14.39	15.17	15.68	15.62	p = 0.107
Age	45.75	49.08	42.81	43.08	p = 0.128
Education	2.01	1.88	1.91	2.02	p = 0.879
Joint F-test:					p = 0.531

Notes. The table displays mean values across all four treatments. Fisher's exact test is based on median splits of non-binary variables (all except Sex). The joint F-test is based on a multivariate regression of the variables on treatment indicators.

(CRT, Raven), and the number of correctly answered comprehension questions (Comprehension).

Column (5) shows the p-values from Fisher's exact test for association between each of the control variables and the categorical treatment variable. We transformed the non-binary controls into binary variables using a median split. The median values for Age, Education, CRT, and Raven can be found in Table 2. The median value for Comprehension was 5 (out of 6).

We also performed a multivariate regression of treatment on the original non-transformed control variables and report the p-value of the joint F-test in the last row of column (5). None of the tests provides evidence for an association between treatment status and our six control variables, suggesting our randomization into treatments was successful.

7. Results

7.1. Specification

In this section, we present our main results, the effect of our treatments on subjects' decision error.[18] Recall that the decision error is defined as the absolute difference between a receiver's WTAs in Tasks 1 and 2. The data are thus censored from below at 0. Because the range for the WTA was \$0.00 to \$5.00, they are also censored from above at 5. In practice, the upper threshold does not seem to be binding.[19] The estimates presented here are based on a simple Tobit model with the decision error as the dependent variable and a categorical treatment indicator as the independent variable. The regression equation reads

$$y_i^* = \alpha_0 + \alpha_1 \cdot T1 + \alpha_2 \cdot T2 + \alpha_3 \cdot T3 + \varepsilon_i, \qquad i = 1, \dots, N, \tag{8}$$

where y^* is the unobserved latent variable and relates to our observed variable y (decision error) as follows:

$$y_i = \begin{cases} 0 & \text{if } y_i^* \leq 0 \\ y_i^* & \text{if } 0 < y_i^* < 5 \\ 5 & \text{if } y_i^* \geq 5 \end{cases} \tag{9}$$

As an alternative specification, we also estimate a two-part model which we present in Appendix C. The main conceptual difference between the two models is the underlying assumption about the data generating process (DGP) of the *zeros*. The Tobit model assumes that all observed values result from the same DGP while the two-part model assumes that the zeros and the positive values result from two different DGPs.

In the context of our study, the Tobit model assumes the existence of some underlying variable, say, *mistakability*, which determines decision error. Because the decision error cannot become negative, we only observe a censored version of this relationship. One possible interpretation of the assumption of the two-part model is that the zeros result from participants who 'gave up thinking' and simply repeated their answer whereas positive values result from participants who 'were thinking'. While we acknowledge the merit of both models, we chose the Tobit model as our main specification.

It is important to note that both models assume that the error term ε is homoskedastic and normally distributed. However, while both assumptions are necessary for consistently estimating the Tobit model, neither assumption is necessary for consistency in the two-part model.

Following Cameron and Triverdi (2009), we test both assumptions in the Tobit specification and reject both null hypotheses of homoskedasticity and normality (both $p < 0.001$). We also test and reject these assumptions in the two-part model using the Breusch–Pagan/Cook–Weisberg test for heteroskedasticity ($p = 0.002$) and The Shapiro–Wilk test for normality ($p < 0.001$).

Here, we only present results from our main specification. In Appendix C, we report results from the two-part model as well and find that they are largely consistent with the findings presented here. For the primary treatment effects of treatments T1 – 3 in the Tobit specification, a simulation-based post-hoc power calculation yields power of approximately 0.37, 0.19 and 0.30 at the 5% significance level, respectively.[20]

7.2. Main findings

Table 5 shows how each treatment affected the participants' decision errors compared to the baseline treatment in column (1). First, the decision error in the baseline treatment was both substantial (\$0.77) and significant ($p < 0.001$). Second, none of our treatments had a significant effect on decision error. Even though all three treatments reduced it, the point estimates are not significantly different from zero ($T1$: −0.177, $p = 0.108$; $T2$: −0.121, $p = 0.281$; $T3$: −0.162, $p = 0.149$).

Table 5 also shows the effects of our treatments on the foregone consumer surplus in column (5). A similar picture emerges: the foregone consumer surplus in the baseline treatment is significant ($T0$: \$0.14, $p < 0.001$) while our treatment effects are not ($T1$: \$ − 0.041, $p = 0.131$; $T2$: \$ − 0.040, $p = 0.133$; $T3$: \$ − 0.042, $p = 0.124$).

7.3. Heterogeneity analysis

As we saw in Section 7.2, our informational treatments did not significantly reduce receivers' decision errors in the pooled sample. We explore this issue further in this section by studying whether there is treatment effect heterogeneity with respect to our demographic control variables (Female, Age, Education) and cognitive ability measures (Raven, CRT). To this end, we split our sample according to the median of each of these variables and estimate (8) separately for each of the subsamples. We also estimate difference-in-differences specifications to see whether any treatment effect heterogeneity is significant. In Appendix B, we examine whether the share of coinciding WTAs in Tasks 1 and 2 differs across treatments and across subgroups defined by the control variables. For every control variable, we conduct separate Fisher's exact tests within each treatment group. The results indicate that treatment effect heterogeneity is unlikely to be explained by differences in these shares.

[18] We report treatment effects on WTA in Appendix B.

[19] The upper and lower bounds are also implied by our conceptual framework of expected utility theory because no degree of risk aversion/affinity rationalizes WTAs below \$0.00 or above \$5.00.

[20] Because analytical power formulas are not available for the Tobit specification, post-hoc power is computed via simulation using the estimated parameters from the main specification and refers to the treatment effects in column (1) of Table 5 only. Detailed information about the post-hoc power calculations can be found in the appendix.

Table 5
Unconditional marginal effects on decision error and foregone consumer surplus by sex.

	Decision error				Foregone CS			
	All (1)	Men (2)	Women (3)	DiD (4)	All (5)	Men (6)	Women (7)	DiD (8)
T1 (*beta*)	−0.177 (0.110)	−0.221 (0.165)	−0.140 (0.145)	0.065 (0.235)	−0.041 (0.027)	−0.075* (0.045)	−0.012 (0.030)	0.052 (0.062)
T2 (*slider*)	−0.121 (0.113)	−0.298* (0.155)	0.092 (0.168)	0.390* (0.223)	−0.040 (0.027)	−0.095** (0.042)	0.016 (0.034)	0.109** (0.055)
T3 (*full info*)	−0.162 (0.112)	−0.287* (0.161)	−0.043 (0.156)	0.239 (0.229)	−0.042 (0.027)	−0.091** (0.044)	0.004 (0.032)	0.090 (0.058)
T0 (*baseline*)	0.766*** (0.095)	0.852*** (0.163)	0.689*** (0.097)	−0.153 (0.178)	0.142*** (0.034)	0.194*** (0.068)	0.095*** (0.019)	−0.069 (0.050)
N	365	187	178		365	187	178	

Notes. Rows T1–T3 report unconditional average marginal treatment effects relative to the baseline treatment T0; row T0 reports mean outcomes in the baseline treatment with bootstrapped standard errors. Columns (1)–(4) report effects on decision error for the pooled sample, males only, females only, and the difference between male and female treatment effects, respectively. Columns (5)–(8) report the corresponding effects for consumer surplus. Standard errors are reported in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

7.3.1. Differences between male and female subjects

Table 5 shows the results for male and female subjects, along with those of the pooled sample. Columns (2) and (3) show the treatments effects on decision error for male and female subjects, respectively, and column (4) shows the difference-in-differences. Columns (6)–(8) show the corresponding effects on forgone consumer surplus.

Female subjects' decision errors did not seem to decrease due to our treatments. Subjects who received full information in treatment T3 did not make significantly better decisions than those who did not receive any additional information (−0.043, $p = 0.781$). The point estimates for the effects of treatments T1 (*correct beta*) and T2 (*slider*) are insignificant as well (T1: −0.140, $p = 0.334$; T2: 0.092, $p = 0.587$).

Male subjects, on the other hand, could improve their decisions in the full information treatment T3 (−0.287, $p = 0.075$). The average gap between decisions in Tasks 1 and 2 decreased by 34% relative to the gap in the baseline treatment T0. It seems that this improvement was largely driven by the effect of the slider, which, on its own, had an effect of almost identical size on the decision error (treatment T2: −0.298, $p = 0.055$). Treatment T1 did not reduce male subjects' decision error significantly.[21]

When we compare the treatment effects on male and female subjects, we see that the slider has a significantly different effect for the two groups. Male subjects benefited more from this treatment than female ones (0.390, $p = 0.079$). The effect of treatment T3 (full info) is not significantly different for the two groups (0.239, $p = 0.298$).

Turning to consumer surplus, we see that male subjects were able to significantly decrease foregone consumer surplus by \$0.095 ($p = 0.025$) in treatment T2 and by \$0.091 ($p = 0.038$) in treatment T3. These are each reductions of almost half of foregone consumer surplus compared to male subjects' baseline level in treatment T0.[22] The reduction of \$0.075 ($p = 0.098$) in treatment T1 has to be interpreted with a grain of salt due to the observed carry-over imbalance.

Female subjects, in contrast, do not seem to benefit from the treatments: all three estimates are insignificant (T1: \$−0.012, $p = 0.697$; T2: \$0.016, $p = 0.638$; T3: \$0.004, $p = 0.894$). The availability of the slider in treatment T2 helped males to reduce foregone consumer surplus by \$0.109 more than females ($p = 0.049$).

We do not observe a significant difference between male and female subjects in the effect of treatments T1 (0.052, $p = 0.405$) or T3 (0.090, $p = 0.118$). The results for consumer surplus therefore mirror those for decision errors, indicating that subjects' mistakes directly translated into losses in consumer surplus.

For the remainder of this section, we focus on the effects on decision errors; results for consumer surplus are reported in Appendix B.3 for brevity.

7.3.2. Differences by cognitive ability

The effects with respect to cognitive ability are shown in Table 6. While the point estimates of the treatment effects are all negative (except the effects of T2 and T3 on subjects with a low score in the CRT), none of them are significantly different from 0.

Nevertheless, we observe two interesting effects. First, subjects with a high CRT score had much higher decision errors at baseline than those with low scores (0.936 vs. 0.653). This difference is only marginally significant (0.326, $p = 0.065$) but certainly surprising. One possible explanation is that low scoring subjects 'gave up thinking' more often, choosing coinciding WTAs in Tasks 1 and 2 at higher rates. However, the shares of coinciding WTAs in Tasks 1 and 2 are neither substantially different (low CRT: 36%, high CRT: 38%) nor significantly different (Fisher's exact test: $p = 0.383$).

Looking at the difference in differences estimates, we see no significant differences with respect to our cognitive ability measure based on Raven's matrices. However, there are significant differences with respect to our other measure based on the CRT. Most notably, T3 had a substantially and significantly larger effect on high scoring subjects compared to low scoring ones (−0.694, $p = 0.002$). This difference is about twice as large as the baseline difference (0.326, $p = 0.065$), meaning that treatment T3 more than compensated for the poor baseline performance of receivers with a high CRT score.

The other treatments also had larger effects on high scoring receivers but these differences were smaller and only marginally significant (T1: −0.391, $p = 0.096$; T2: −0.409, $p = 0.066$).

7.3.3. Differences by age and education

Table 7 shows our findings with respect to Age (columns (1)–(3)) and to educational attainment (columns (4)–(6)).

Age does not seem to play an important role. The decision errors at baseline are similar and not significantly different, all treatment effects are insignificant, and there is no significant difference in treatment effects either.

Education, on the other hand, seems more important. In the baseline treatment, subjects with a highschool diploma or higher had higher decision errors compared to subjects with no highschool diploma (0.888 vs. 0.533). This difference, however, is only marginally significant (0.311, $p = 0.094$). Furthermore, treatments T1 and T3 significantly

[21] Remember that treatment T1 is potentially compromised due to carry-over imbalance.

[22] Note that male subjects' foregone consumer surplus in the baseline treatment T0 is slightly larger in our sample than that of female subjects. The difference is, however, not statistically significant ($p = 0.165$).

Table 6
Unconditional marginal effects on decision error by cognitive ability.

Decision error	Raven ≤ 15 (1)	Raven > 15 (2)	DiD (3)	CRT ≤ 2 (4)	CRT = 3 (5)	DiD (6)
T1 (*beta*)	−0.167	−0.184	−0.027	−0.011	−0.421	−0.391*
	(0.153)	(0.159)	(0.238)	(0.146)	(0.162)	(0.235)
T2 (*slider*)	−0.194	−0.058	0.144	0.053	−0.378	−0.409*
	(0.157)	(0.162)	(0.226)	(0.151)	(0.165)	(0.223)
T3 (*full info*)	−0.118	−0.188	−0.075	0.141	−0.553	−0.694***
	(0.168)	(0.155)	(0.233)	(0.158)	(0.156)	(0.225)
T0 (*baseline*)	0.791***	0.729***	−0.005	0.653***	0.936***	0.326*
	(0.119)	(0.156)	(0.182)	(0.120)	(0.146)	(0.177)
N	183	182		206	159	

Notes. Rows T1–T3 report unconditional average marginal treatment effects relative to the baseline treatment T0; row T0 reports mean decision errors in the baseline treatment with bootstrapped standard errors. Columns (1)–(3) report treatment effects on decision errors for subjects with a low Raven score, a high Raven score, and the differences in treatment effects between low and high scoring subjects, respectively. Columns (4)–(6) report the corresponding treatment effects based subjects' CRT scores.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 7
Unconditional marginal effects on decision error by age and education.

Decision error	Age ≤ 45 (1)	Age > 45 (2)	DiD (3)	Education ≤ 1 (4)	Education > 1 (5)	DiD (6)
T1 (*beta*)	−0.127	−0.212	−0.096	0.219	−0.364***	−0.568**
	(0.183)	(0.131)	(0.236)	(0.189)	(0.135)	(0.248)
T2 (*slider*)	−0.249	0.018	0.276	−0.003	−0.193	−0.179
	(0.166)	(0.155)	(0.225)	(0.169)	(0.145)	(0.236)
T3 (*full info*)	−0.141	−0.199	−0.070	0.053	−0.281**	−0.328
	(0.177)	(0.139)	(0.230)	(0.180)	(0.142)	(0.244)
T0 (*baseline*)	0.801***	0.735***	−0.102	0.533***	0.888***	0.311*
	(0.144)	(0.123)	(0.179)	(0.107)	(0.130)	(0.186)
N	183	182		116	249	

Notes. Rows T1–T3 report unconditional average marginal treatment effects relative to the baseline treatment T0; row T0 reports mean decision errors in the baseline treatment with bootstrapped standard errors. Columns (1)–(3) report treatment effects on decision errors for subjects below 45 years of age, above that age, and the differences in treatment effects between these age groups, respectively. Columns (4)–(6) report treatment effects on decision errors for subjects with low education, high education, and the differences in treatment effects between these groups, respectively.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

reduced decision errors in high-education receivers (T1: −0.364, $p = 0.007$; T2: −0.281, $p = 0.047$), while having no effect on the remaining ones. Only the difference for T1 is significant (−0.568, $p = 0.022$).

8. Conclusion

This paper examined how misinformation affects decision-making in environments where individuals rely on potentially unreliable signals to form beliefs about an uncertain state. Using a controlled online experiment based on the *Trading Envelopes* game, we isolated and quantified two key mechanisms: misperception of signal credibility and incorrect Bayesian updating. Our measure of *decision error*, defined as the deviation between a receiver's valuation of an envelope and an outcome-equivalent lottery, captures the cumulative impact of these mechanisms on individual choice.

Our results reveal that even in a relatively simple environment (compared to real-world investment or purchase decisions), participants exhibit sizable decision errors. Informing participants about the true share of dishonest senders or providing computational assistance for Bayesian updating did not reduce these errors on average. However, treatment effects are heterogeneous: interventions improved decision accuracy among men and participants with higher cognitive ability and education, while leaving others unaffected.

Examining our study through the lens of complexity aversion, one can view our treatments as varying the complexity of Task 1. The baseline treatment is the most complex because participants evaluate the lottery given their subjective beliefs about the lying probability. As long as these beliefs are not degenerate, this results in a compound lottery with high complexity. Compared to that, in treatment T1 the true lying probability is given, resulting in a simple lottery with lower complexity. In treatment T2, the lottery's effective complexity is reduced by a *slider* that facilitates the calculation of posterior probabilities. Treatment T3 combines the features of T1 and T2, resulting in the lowest complexity. Interestingly, we did not find evidence for complexity aversion, i.e. subjects did not systematically penalize more complex lotteries. However, our finding that decisions became noisier as complexity increased is consistent with the noise hypothesis of Oberholzer et al. (2024).

These findings contribute to a broader understanding of how misinformation distorts decision-making in consumer, labor, and financial markets. While efforts to directly eliminate misinformation have achieved only limited success, our study indicates that targeted informational interventions – especially when tailored to specific user groups – can mitigate some of its consequences. Future research should explore how to design such interventions to accommodate cognitive heterogeneity and how these effects scale in real-world market settings where incentives and reputation mechanisms are more complex.

CRediT authorship contribution statement

Boris Knapp: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation,

Conceptualization. **Dominik Stelzeneder:** Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.socec.2026.102576.

Data availability

Data will be made available on request.

References

Aguirregabiria, V. (2021). Identification of firms' beliefs in structural models of market competition. *Canadian Journal of Economics/Revue Canadienne D'Économique*, *54*(1), 5–33. http://dx.doi.org/10.1111/caje.12503.

Akesson, J., Hahn, R. W., Metcalfe, R. D., & Monti-Nussbaum, M. (2023). *The impact of fake reviews on demand and welfare*: *Technical Report*, National Bureau of Economic Research.

Albert, D. A., & Smilek, D. (2023). Comparing attentional disengagement between prolific and mturk samples. *Scientific Reports*, *13*(1), 20574.

Andı, S., & Akesson, J. (2021). Nudging away false news: Evidence from a social norms experiment. *Digital Journalism*, *9*(1), 106–125.

Azrieli, Y., Chambers, C. P., & Healy, P. J. (2018). Incentives in experiments: A theoretical analysis. *Journal of Political Economy*, *126*(4), 1472–1503. http://dx.doi.org/10.1086/698136.

Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, *9*(3), 226–232.

Bialek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods*, *50*(5), 1953–1959.

Butler, D. J., & Loomes, G. C. (2007). Imprecision as an account of the preference reversal phenomenon. *American Economic Review*, *97*(1), 277–297.

Cameron, A. C., & Triverdi, P. K. (2009). *Microeconometrics using stata*. College Station, TX: Stata Press.

Carpenter, J., Huet-Vaughn, E., Matthews, P. H., Robbett, A., Beckett, D., & Jamison, J. (2021). Choice architecture to improve financial decision making. *The Review of Economics and Statistics*, *103*(1), 102–118. http://dx.doi.org/10.1162/rest_a_00881.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1–8. http://dx.doi.org/10.1016/j.jebo.2011.08.009.

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Costa-Gomes, M. A., & Weizsäcker, G. (2008). Stated beliefs and play in normal-form games. *Review of Economic Studies*, *75*(3), 729–762.

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *Plos One*, *18*(3), Article e0279720.

Falk, A., & Zimmermann, F. (2012). A taste for consistency and survey response behavior. *CESifo Economic Studies*, *59*(1), 181–193. http://dx.doi.org/10.1093/cesifo/ifs039.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. http://dx.doi.org/10.1257/089533005775196732.

Gao, X. S., Harrison, G. W., & Tchernis, R. (2023). Behavioral welfare economics and risk preferences: A Bayesian approach. *Experimental Economics*, *26*(2), 273–303.

Georgalos, K., & Nabil, N. (2025). Testing models of complexity aversion. *Journal of Behavioral and Experimental Economics*, *116*, Article 102354.

Gesche, T. (2021). De-biasing strategic communication. *Games and Economic Behavior*, *130*, 452–464.

Greenwald, A. G. (1992). Within-subjects designs: To use or not to use? In *Methodological issues & strategies in clinical research* (pp. 157–167). American Psychological Association, http://dx.doi.org/10.1037/10109-021.

Gupta, N., Rigotti, L., & Wilson, A. (2025). The experimenters' dilemma: Inferential preferences over populations. *Working Paper*.

Haigh, M. (2016). Has the standard cognitive reflection test become a victim of its own success? *Advances in Cognitive Psychology*, *12*(3), 145.

Harrison, G., Martínez-Correa, J., Morsink, K., Ng, J. M., & Swarthout, T. (2020). Compound risk and the welfare consequencesof insurance. *Working Paper*.

Harrison, G., Morsink, K., & Schneider, M. (2020). Do no harm? The welfare consequences of behavioural interventions. *Working Paper*.

He, S., Hollenbeck, B., & Proserpio, D. (2022). The market for fake reviews. *Marketing Science*, *41*(5), 896–921.

Horowitz, J. K. (2006). The Becker-DeGroot-Marschak mechanism is not necessarily incentive compatible, even for non-random goods. *Economics Letters*, *93*(1), 6–11.

Huck, S., & Weizsäcker, G. (1999). Risk, complexity, and deviations from expected-value maximization: Results of a lottery choice experiment. *Journal of Economic Psychology*, *20*(6), 699–715.

Jiang, J. X., Stanford, M. H., & Xie, Y. (2012). Does it matter who pays for bond ratings? Historical evidence. *Journal of Financial Economics*, *105*(3), 607–621.

Knapp, B. (2025). Fake reviews and naive consumers. *Working Paper*, http://dx.doi.org/10.1093/jleo/ewag010.

Mador, G., Sonsino, D., & Benzion, U. (2000). On complexity and lotteries' evaluation–three experimental observations. *Journal of Economic Psychology*, *21*(6), 625–637.

Meyer, A., Zhou, E., & Frederick, S. (2018). The non-effects of repeated exposure to the cognitive reflection test. *Judgment and Decision Making*, *13*(3), 246–259.

Moffatt, P. G., Sitzia, S., & Zizzo, D. J. (2015). Heterogeneity in preferences towards complexity. *Journal of Risk and Uncertainty*, *51*(2), 147–170.

Nyarko, Y., & Schotter, A. (2002). An experimental study of belief learning using elicited beliefs. *Econometrica*, *70*(3), 971–1005. http://dx.doi.org/10.1111/1468-0262.00316.

Oberholzer, Y., Olschewski, S., & Scheibehenne, B. (2024). Complexity aversion in risky choices and valuations: Moderators and possible causes. *Journal of Economic Psychology*, *100*, Article 102681.

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, *66*(11), 4944–4957.

Plott, C. R., & Zeiler, K. (2005). The willingness to pay–willingness to accept gap, the "endowment effect," subject misconceptions, and experimental procedures for eliciting valuations. *American Economic Review*, *95*(3), 530–545.

Raven, J. C., Raven, J., & Court, J. H. (1998). *A manual for Raven's progressive matrices and vocabulary scales*. London: H. K. Lewis.

SafetyDetectives research lab (2021). Amazon fake reviews scam exposed in data breach. https://web.archive.org/web/20210525123756/https://www.safetydetectives.com/blog/amazon-reviews-leak-report/. [Accessed 31 May 2021].

Schlag, K., & Tremewan, J. (2021). Simple belief elicitation: An experimental evaluation. *Journal of Risk and Uncertainty*, *62*(2), 137–155. http://dx.doi.org/10.1007/s11166-021-09349-6.

Schneider, U. C. (2019). Identifying and estimating beliefs from choice data - an application to female labor supply. *Working Paper*.

Sonsino, D., Benzion, U., & Mador, G. (2002). The complexity effects on choice with uncertainty–experimental evidence. *The Economic Journal*, *112*(482), 936–965.

Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2018). Performance on the cognitive reflection test is stable across time. *Judgment and Decision Making*, *13*(3), 260–267.

Suri, D., Kube, S., & Schultz, J. (2025). Evaluating online data collection platforms using a simple rule-following task. *Economics Letters*, Article 112509.