A robust measure of complexity*

Egor Bronnikov† Elias Tsakas‡

Maastricht University

First draft: January 2025; This draft: June 2025

Latest version of the paper.

Abstract

We introduce a robust belief-based measure of complexity. The idea is that task A is classified as more complex than task B if the probability of correctly solving A is smaller than the probability of correctly solving B, regardless of the reward. We provide a full characterization of the incomplete order over the set of tasks that this measure induces. This characterization allows us to identify the degree of (prior) uncertainty as a novel dimension of complexity, i.e., in order for task A to be more complex than task B, it does not suffice that A is more difficult than B; it should also not be the case that the agent has much more prior information about A than about B. Then, using a lab experiment, where we can exogenously control both difficulty and uncertainty, we corroborate our theoretical predictions. Thus, the recently surging use of expected accuracy as a good measure of complexity is well warranted, as long as accuracy is elicited for multiple different rewards using the strategy method.

Keywords: complexity, measure, difficulty, uncertainty, incomplete relation, attention.
JEL codes: D83, D90.

1. Introduction

Complexity is a fundamental concept across numerous scientific domains, including computer science, cognitive sciences, neuroscience, etc. More recently, its importance has also been

*We are indebted to Duarte Goncalves, Hans-Theo Normann, Martin Strobel, Georg Weizsäcker and the audience in Berlin (Behavioral Economics Seminar Series) and Maastricht (BEELab meeting) for very helpful comments and valuable suggestions.

†Homepage: egorbronnikov.github.io; E-mail: egor.bronnikov@maastrichtuniversity.nl

‡Homepage: www.elias-tsakas.com; E-mail: e.tsakas@maastrichtuniversity.nl

recognized by (behavioral) economists, as a key determinant of decisions in many settings, with the potential to explain mistakes that people systematically make in such decisions (e.g., Banovetz and Oprea, 2023; Enke *et al.*, 2024a; Oprea, 2024b).

At the same time, it is also the case that complexity has been typically used in a casual way, without consensus on what exactly it means. This lack of a widely accepted precise definition can possibly explain why scholars have also focused on measurements of complexity, which can in turn serve as proxies for different types of complexity (Oprea, 2024a).[1] Common ways to measure complexity include direct metrics (Oprea, 2020), behavioral metrics (Banovetz and Oprea, 2023), and most importantly for this paper the increasingly popular belief-based metrics (Enke and Graeber, 2023; Enke *et al.*, 2024a,b; Agranov *et al.*, 2025).

The key idea within this last stream of literature is to use an agent's belief about their own accuracy as a proxy for complexity. In simple terms, without needing to explicitly specify which notion of complexity we have in mind, it is reasonable to assume that the chances of solving a more complex task are lower than the chances of solving a simpler one. The appeal of this approach is twofold: on the one hand it makes a lot of intuitive sense, and on the other hand it is practically quite easy to measure and compare expected accuracy across tasks, even in cases where the tasks are not that similar to one another.

However, as appealing as this approach is, there is a serious caveat. Namely, expected accuracy depends both on complexity as well as on the attention which is put towards solving the task. And since attention depends non-linearly on the underlying reward (for correctly solving the task), it often happens that expected accuracy is sensitive to the reward. That is, simply speaking, it will often be the case that the chances for solving task A are higher than the chances for solving task B if the reward is small, and vice versa if the reward is large. And this naturally gives rise to the question: when we observe such reversal, which of the two tasks should we take as the more complex one?

In this paper, we take a robust approach in ranking tasks. In particular, we say that task A is deemed more complex than task B, whenever the chances of correctly solving A are smaller than the chances of correctly solving B, *for every reward*. Obviously, this is a rather conservative criterion, as it imposes a strong dominance condition. But at the same time, whenever satisfied, it provides a quite compelling argument that the tasks are indeed ranked in this way.

In the first part of the paper, we provide decision theoretic foundations of our measure within a standard rational inattention framework. Our first main result provides a full characterization of the complexity order that the aforementioned criterion induces (Theorem 1). Not surprisingly, this complexity order is incomplete. But they key insight is that it depends on

[1] Measurements and formal definitions of complexity are closely linked with each other, in an analogous way to how belief elicitation mechanisms (Brier, 1950; Savage, 1971) are linked with definitions of subjective probability (Savage, 1954; Anscombe and Aumann, 1963).

two distinct parameters, viz., the difficulty of the task and the degree of (ex ante) uncertainty. More specifically, higher difficulty is only a necessary condition for higher complexity. In order to also become sufficient, the agent cannot be ex ante much more uncertain about the state realization. For example, in order for a student to find exam A more complex than exam B, it must be both the case that A is more difficult than B, and moreover that the student is not much better prepared about B than about A.[2]

The fact that complexity has these two distinct dimensions allows us to establish a link with the decision-theoretic literature on incomplete preferences. In particular, we show that our complexity order is represented by a vector (utility) function, like in Ok (2002).

Then, we focus on tasks that are not ranked by our complexity measure. By the previous theorem one task is more likely to be solved for some rewards, while the other is more likely to be solved with the rest of the rewards. In our second main result, we identify the rewards that make each of the two tasks more likely to be solved, viz., we prove that the more difficult and less uncertain task is more likely to be solved for small rewards, whereas the easier and more uncertain task is more likely to be solved for large rewards (Theorem 2). The intuition is quite clear: for small rewards she relies more on her prior knowledge, whereas for larger rewards she relies more on the attention she currently pays. A final implication of this result is that, even when we cannot order tasks with our complexity measure, we can still tell which task dominates in difficulty and which dominates in uncertainty, based only on the elicited chances of solving each task.

In the second part of the paper, we test the predictions of our theoretical model in a lab experiment, where we can exogenously control and manipulate the two dimensions of complexity that we previously identified.

Our experiment will consist of two stages. In the first stage —which is merely auxiliary for the second stage— the setting is similar to the one Dean and Neligh (2023), i.e., subjects see a screen with blue and red balls scattered across, and they are asked to guess the dominant color.[3] Difficulty is determined by the total number of balls on the screen, and uncertainty is determined by the probability of the screen they see being blue dominant. A combination of difficulty and uncertainty characterizes a task. We run two treatments (viz., high and low stakes) between subjects, which only differ in the bonus the subjects receive for guessing correctly. Then, in the second —and main— stage of the experiment we ask another group of subjects from the same pool to guess, for each task and each reward, the proportion of first-stage subjects that

[2]In the main body of the paper, we also allow the agent to derive different levels of intrinsic satisfaction from solving each task, and therefore a satisfaction parameter also appears in our characterization result. However, the aforementioned general intuition remains.

[3]In recent work, Goncalves *et al.* (2024) also employ a similar design to study the sensitivity of choices to incentive changes. We further elaborate on their work later on.

answered correctly.[4]

Both hypotheses that are derived from our two main theorems are confirmed. Starting with Theorem 1, we find that for any pair of tasks, A and B, for which A dominates B both in difficulty and uncertainty, second-stage participants reported significantly lower probability that the dominant color was guessed correctly in task A than in task B, in both treatments. Then, moving to Theorem 2, we consider pairs of tasks, A and B, such that A dominates B in difficulty, and B dominates A in uncertainty. Then, as predicted by Theorem 2, we find that whenever second-stage participants reported that A is more likely to be solved than B in the high-stakes treatment, they also reported it to be the case in the low-stakes treatment. Likewise, whenever second-stage participants reported that B is more likely to be solved than A in the low-stakes treatment, they also reported it to be the case in the high-stakes treatment. The aforementioned results essentially constitute a validation of our measure as a good proxy for complexity.

Summarizing, the overall contribution of the paper is twofold. First, we identify the degree of uncertainty as a novel dimension of complexity. While intuitively appealing, this channel of complexity has been previously overlooked, as complexity has been typically identified as mere difficulty (Oprea, 2024a, and references therein). Yet, our results are not entirely at odds with the existing literature in the sense that, although difficulty is not the only dimension of complexity, it is still the primary one. And crucially, this is not something we assume, but rather a property that we naturally follows from our measure of complexity.

The second main contribution of the paper is to provide foundations for expected accuracy as a measure of complexity. Expected accuracy has been recently surging in the experimental literature (Agranov *et al.*, 2025; Enke and Graeber, 2023; Enke *et al.*, 2024a,b), as it is both simple and intuitive, but unfortunately microeconomic foundations have been missing.[5] Therefore, by filling this gap, we clarify what exactly expected accuracy measures. Moreover, our results suggest that in order to use expected accuracy as a reliable measure of complexity, one needs to elicit it for multiple different rewards.

The literature on complexity is vast, and as such we are de facto forced to make a selection of what in our view is the most relevant subset. We will not even attempt to touch the related literatures within other disciplines, such as computer science or cognitive sciences.

[4]Note that we use as a proxy for their own expected accuracy, their beliefs about the expected accuracy of other individuals similar to them. This idea is similar to the one employed in the literature on Bayesian markets (Baillon, 2017). We further elaborate on practical issues that pertain to elicitation in Section 4.

[5]The only theoretical paper that shows (actual) accuracy to be monotonic with respect to complexity is the one of Goncalves (2024). However, he does so using a very specific definition of complexity, viz., the signal-to-noise ratio, as opposed to our paper where we remain agnostic as to how complexity is actually defined. Moreover, in his paper, Goncalves (2024) exogenously shuts down uncertainty as a dimension of complexity, as he assumes the prior to be the same for all tasks.

Early work focused primarily on the role of strategy complexity within game theory (e.g., for early contributions, see Rubinstein, 1986; Abreu and Rubinstein, 1988). More recently, the focus has shifted towards explaining mistakes and irrationalities, e.g., Oprea (2024b) study the effect of complexity on risk preferences, and Enke *et al.* (2024a) the respective effect on time preferences. As a response to these links, there are several attempts to formalize a definition of complexity, e.g., Gabaix and Graeber (2024) build a general model of production within a cognitive economy in order to operationalize complexity, whereas Oprea (2024a) borrows insights from computer science to introduce a framework within which complexity reflects the cost for handling a task. Others, define it as the signal-to-noise ratio (Callander, 2011; Fehr and Rangel, 2011; Goncalves, 2024), similarly to what is often done in psychometrics. Alternative approaches use tradeoffs (Shubbat and Yang, 2024) or degrees of contingent reasoning (Nagel and Saitto, 2025) to define complexity. The common denominator throughout these papers is that the respective definitions of complexity are input-based, i.e., complexity is defined by means of some characteristics of the environment.

What is closer to our work is the literature on measuring complexity. As Oprea (2024a) elegantly points out, this literature can be classified into three large streams, depending the measurement tool. Within the first stream, we encounter measurement with direct metrics, such as willingness to pay in order to avoid dealing with a certain task (Oprea, 2020), response times (Gill and Prowse, 2023; Goncalves, 2024), and biometrics (van der Wel and van Steenbergen, 2018). The second stream leverages behavioral metrics, such as procedural measurements (Banovetz and Oprea, 2023), and choice inconsistencies (Woodford, 2020).

Finally, the third stream —within which our paper belongs— uses belief-based metrics. These include subjective rankings, like for instance in Gabaix and Graeber (2024) where subjects are simply asked to rank tasks with respect to complexity, and most commonly beliefs about one's own accuracy (Agranov *et al.*, 2025; Enke and Graeber, 2023; Enke *et al.*, 2024a,b). Accuracy has also been shown to be (negatively) correlated with complexity, when the latter is defined as the signal-to-noise ratio and priors are exogenously assumed to be the same for all tasks (Goncalves, 2024). Then, using this structural framework, Goncalves *et al.* (2024) conduct an experiment —using a design similar to the one we have implemented in this paper— where they leverage (small) changes in incentives for a different purpose, viz., to infer the signal-to-noise ratio in a setting with fixed priors. In all other aforementioned papers, rewards are fixed. The effects of varying rewards are discussed in (Alaoui and Penta, 2022).

Somewhere in between the input-based and the measure-based approach one finds a stream of literature that uses lotteries characteristics as proxies for complexity (Huck and Weizsäcker, 1999; Fudenberg and Puri, 2022; Enke and Shubatt, 2024; Hua Hu, 2023).

This entire literature is part of the surging field of Cognitive Economics (Caplin, 2025; Enke, 2024), which also incorporates topics such as rational inattention, cognitive uncertainty,

etc. Specifically related to our work, within this broader literature, are the papers that study preference for simplicity (Puri, 2025; de Clippel *et al.*, 2025; Mononen, 2025, and references therein) and model uncertainty (Mussolf and Zimmermann, 2025, and references therein).

The paper is structured as follows: In Section 2 we introduce our theoretical framework. 3 we introduce our measure of complexity and prove our main characterization results. In Section 4 we study questions that pertain to elicitation of beliefs about expected accuracy. In Section 5 we present our experiment. In Section 6 we study how the relationship between our complexity measure and exogenously provided information. In Section 7 we discuss potential extensions and limitations of our theoretical model. Section 8 concludes. All proofs are relegated to the Appendix.

2. Guessing tasks

2.1. Task characteristics

Consider a binary state space $S = \{s_0, s_1\}$ and a (female) agent. A task requires the agent to guess the state realization.

Definition 1. Formally, a task is identified by the state space S itself. The set of all possible tasks is denoted by $\mathcal{S}_0$. ◁

Let $Z = \{0, 1\}$ denote the set of possible scores that the agent can potentially receive. In particular, it will be the case that $z = 1$ (resp., $z = 0$) whenever the examiner announces that the agent has correctly (resp., wrongly) guessed the state. Let $X := [0, \infty)$ be the set of monetary rewards that the agent can potentially receive. An outcome is a pair (x, z).

An act is a mapping $f : S \to X \times Z$ that induces an outcome at each state. For a task $S \in \mathcal{S}_0$ and a reward $x \in X$ for guessing the actual state in S, making a guess corresponds to choosing an act from menu $\{r_0^x, r_1^x\}$, where

$$r_k^x(s) = \begin{cases} (x, 1) & \text{if } s = s_k, \\ (0, 0) & \text{if } s \neq s_k. \end{cases}$$

The agent has SEU preferences over the set of all acts, i.e., there is a Bernoulli utility function $u_S : X \times Z \to \mathbb{R}$ and a belief which is identified by the probability $\mu_S \in (0, 1)$ of s_1 occurring, such that the preferences over acts are represented by

$$\mathbb{E}_{\mu_S}(u_S(f)) = (1 - \mu_S)u_S(f(s_0)) + \mu_S u_S(f(s_1)).$$

Throughout the paper, we assume that u_S is continuously increasing and unbounded, meaning that the agent prefers more money over less money, and a positive over a negative score.

Furthermore, for each $x \in X$, we define the net utility

$$v_S(x) := u_S(x,1) - u_S(0,0).$$

In general, we allow v_S to be task-specific, in the following sense: there is a task-independent $v : X \times Z \to \mathbb{R}$ so that, for every $S \in \mathcal{S}_0$ there is some $\beta_S > 0$ such that for all $x \in X$,

$$v_S(x) = \beta_S v(x). \tag{1}$$

The underlying idea is that risk preferences do not depend on the task, whereas the agent's satisfaction from answering correctly may be task-dependent. In particular, the larger β_S is, the more satisfied the agent will be from answering correctly task S.

The agent's indirect expected utility, as a function of the probability $q \in [0,1]$ that she attaches to s_1, is given by

$$g_S(q) := v_S(x)\max\{q, 1-q\} = \beta_S \underbrace{v(x)\max\{q,1-q\}}_{g(x)}, \tag{2}$$

which is obviously proportional to the probability she attaches to her best guess being correct.

The prior belief μ_S that the agent assigns to s_1 reflects her prior knowledge/experience about S. Thus, her degree of uncertainty about S is proportional to the Shannon entropy of μ_S, i.e., we define

$$\eta_S := \frac{1}{\log 2}\Big(\underbrace{\mu_S \log \mu_S + (1-\mu_S)\log(1-\mu_S)}_{H(\mu_S)}\Big). \tag{3}$$

That is, η_S takes values in $(0,1]$, and it is strictly decreasing with respect to the distance from the uniform belief. This notion of uncertainty has solid foundations within information theory (Cover and Thomas, 2006). Importantly, for the purposes of our paper, the only thing that matters is the ordinal relation that it induces over the set of all probability distributions, and in this sense it is without loss of generality to instead take any strictly increasing transformation of η_S.

Before making a guess, the agent decides how much attention to pay. Attention is modelled with a Bayesian signal, which is uniquely identified by a mean-preserving distribution of posterior probabilities (Kamenica and Gentzkow, 2011). The set of all signals is denoted by

$$\Pi_S = \Big\{ \pi \in \Delta\big([0,1]\big) : \mathbb{E}_\pi(q) = \mu_S \Big\}.$$

For any signal $\pi \in \Pi_S$, define the agent's ex ante indirect expected utility,

$$G_S(\pi) := \mathbb{E}_\pi\big(g_S(q)\big). \tag{4}$$

As usual, we assume that information acquisition is costly. The cost function is assumed to be uniformly posterior separable (Caplin *et al.*, 2022; Tsakas, 2020), i.e., there is a strictly convex function $c : [0,1] \to \mathbb{R}$ such that the cost of signal $\pi \in \Pi_S$ is given by

$$C_S(\pi) = \kappa_S\Big(\mathbb{E}_\pi\big(c(q)\big) - c(\mu_S)\Big), \tag{5}$$

where $c(q)$ represents the agent's marginal cost for acquiring information, and $\kappa_S > 0$ is a parameter of the task's difficulty. Note that consistently with the complexity literature (Oprea, 2024a), the cost consists a task characteristic (viz., the parameter κ_S) and an individual characteristic (viz., the function c). Posterior-separable costs functions have solid foundations (Denti, 2022) and are supported by experimental evidence (Dean and Neligh, 2023). Throughout the paper, we will focus on symmetric cost functions, which include the Shannon entropy (Sims, 2003), the Shorrocks entropy (Shorrocks, 1980), the Tsallis entropy (Caplin *et al.*, 2022) as special cases. Formally, this means that for every $q \in [0,1]$, we have $c(q) = c(1-q)$. For an axiomatization of symmetric cost functions, see Hébert and Woodford (2021). We further discuss the symmetry assumption in Section 7.2.

Remark 1. Overall, there are three parameters that are task-specific: the difficulty (κ_S), the satisfaction (β_S), and the degree of uncertainty (η_S). Among these parameters, difficulty is objective (i.e., it can be assumed to be common for all agents), whereas satisfaction and uncertainty are subjective (i.e., they typically vary across agents). Crucially, the reward x is not a task characteristic, but rather a directly observable parameter which can be controlled by the analyst, and varied arbitrarily within the same task S. ◁

2.2. Optimal attention

Throughout this section, fix a task $S \in \mathcal{S}_0$ and a reward $x \in X$. The agent faces a tradeoff, in that more informative signals help her to achieve higher expected utility, but at the same time are also more costly. That is, formally speaking, the agent solves the following optimization problem:

$$\max_{\pi \in \Pi_S}\Big(G_S(\pi) - C_S(\pi)\Big). \tag{6}$$

It is not difficult to verify that there is a unique optimal signal, henceforth denoted by π_S^x. By a standard concavification argument (e.g., Kamenica and Gentzkow, 2011; Matějka and McKay, 2015), there exists some threshold $q_S^x \in \big[0, 1/2\big]$ such that, for small degrees of uncertainty (i.e., whenever $\mu_S \leq q_S^x$ or $\mu_S \geq 1 - q_S^x$) the agent does not put any attention to the task and maintains her prior beliefs, whereas for large degrees of uncertainty (i.e., whenever $q_S^x < \mu_S < 1 - q_S^x$) her optimal signal mixes between the posteriors q_S^x and $1 - q_S^x$.

In other words, the agent will acquire information if and only if $\eta_S > H(q_S^x)/\log 2$. In the limit, when the reward vanishes, the information acquisition threshold becomes

$$\bar{\eta}_S := \frac{H(q_S^0)}{\log 2}. \tag{7}$$

That is, whenever $x = 0$, the agent will acquire information if and only if $\eta_S > \bar{\eta}_S$. It follows from standard arguments that $\bar{\eta}_S$ is a continuously increasing function of the difficulty-to-satisfaction ratio

$$\lambda_S := \frac{\kappa_S}{\beta_S}, \tag{8}$$

i.e., $\lambda_S \geq \lambda_{S'}$ implies $\bar{\eta}_S \geq \bar{\eta}_{S'}$. Furthermore, it is the case that $\bar{\eta}_S \to 0$ as $\lambda_S \to 0$, and $\bar{\eta}_S \to 1$ as λ_S grows arbitrarily large. The difficulty-to-satisfaction ratio will play an important role in our characterization of complexity later in the paper.

3. A belief-based measure of complexity

3.1. Definition of complexity

Assuming that the agent is rational (in the sense that she picks the optimal signal), her expected accuracy is given by the total probability of guessing correctly:

$$P(S, x) := \frac{G_S(\pi_S^x)}{\beta_S v(x)}. \tag{9}$$

In Figure 1, we illustrate it as a function of the prior belief μ_S. Whenever the degree of uncertainty is large (i.e., $q_S^x < \mu_S < 1 - q_S^x$), her expected accuracy will be $P(S, x) = 1 - q_S^x$. On the other hand, whenever the degree of uncertainty is small (i.e., $\mu_S \leq q_S^x$ or $\mu_S \geq 1 - q_S^x$), her expected accuracy will be $P(S, x) = \max\{\mu_S, 1 - \mu_S\}$. Thus, all together, we obtain

$$P(S, x) = \max\{\mu_S, 1 - q_S^x, 1 - \mu_S\}. \tag{10}$$

The latter is illustrated in Figure 1.

In the literature on belief-based complexity measures, expected accuracy is increasingly being used as a measure of task complexity (Agranov *et al.*, 2025; Enke and Graeber, 2023; Enke *et al.*, 2024a,b; Oprea, 2024a). This is largely due to the fact that it combines intuitive appeal and simplicity. At the same time, it is also consistent with theoretical results that show certain definitions of complexity to be positively correlated with accuracy (Goncalves, 2024).

However, there are is an important caveat: Expected accuracy depends both on task characteristics and on an experimental parameter, viz., the reward $x \geq 0$. This leads to an undesirable

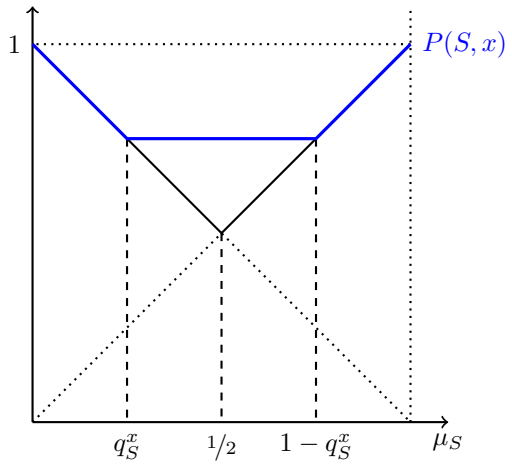


Figure 1: The blue piecewise linear function is the expected accuracy (as a function of the agent's prior), assuming that the optimal signal π_S^x has been used.

situation, viz., task A is more likely than task B to be solved correctly under a small reward, while at the same time task A is less likely than task B to be solved correctly under a large reward. And this naturally makes us wonder: should we classify A or B as more complex?

In this paper, we address this question by proposing a robust belief-based measure of complexity, which labels A more complex than B whenever the expected accuracy of A is always smaller than the expected accuracy of B, for every reward. This idea is formalized in the following definition.

Definition 2. Task $S \in \mathcal{S}_0$ is more complex than task $S' \in \mathcal{S}_0$ if

$$P(S,x) \leq P(S',x) \tag{11}$$

for all $x \in X$. In this case, we write $S \succeq S'$. ◁

The asymmetric and the symmetric parts of $\succeq$ are defined as usual. That is, we have $S \succ S'$ whenever $S \succeq S'$ and $S \not\preceq S'$, and respectively $S \sim S'$ whenever $S \succeq S'$ and $S \preceq S'$.

3.2. Characterization of complexity order

A task $S \in \mathcal{S}_0$ is said to be trivial if the optimal signal π_S^x reveals the true state with certainty (i.e., $q_S^x = 0$) for every $x \geq 0$. Obviously, if a task S is trivial, then it is simpler than any other task, as it will satisfy $P(S,x) = 1$ for every $x \in X$. The set of non-trivial tasks is henceforth denoted by $\mathcal{S} \subseteq \mathcal{S}_0$, and it is characterized by a difficulty-to-satisfaction threshold.

Proposition 1. *There is some $\lambda \geq 0$ such that,*

$$S \in \mathcal{S} \;\Leftrightarrow\; \lambda_S > \lambda \text{ and } \eta_S > 0. \tag{12}$$

The idea is quite simple: in order for a task to be non-trivial, the information costs must be sufficiently large to guarantee that the intrinsic incentives alone are not strong enough to always lead to a perfectly informative signal. In our main result below, we characterize how non-trivial tasks are ranked if we use our robust belief-based measure.

Our characterization of $\succeq$ within $\mathcal{S}$ has the same structure as the vector-valued utility representation of incomplete preference relations in Ok (2002). In particular, let us first define the vector-valued function $\boldsymbol{\phi} : \mathcal{S} \to \mathbb{R}^2$, where

$$\boldsymbol{\phi}_1(S) := \lambda_S \text{ and } \boldsymbol{\phi}_2(S) := \min\{\bar{\eta}_S, \eta_S\}. \tag{13}$$

Let $\geq$ be the usual order over $\mathbb{R}^2$, i.e., we will write $\boldsymbol{\phi}(S) \geq \boldsymbol{\phi}(S')$ if and only if $\boldsymbol{\phi}_1(S) \geq \boldsymbol{\phi}_1(S')$ and $\boldsymbol{\phi}_2(S) \geq \boldsymbol{\phi}_2(S')$. Then, we are ready to obtain our main characterization result.

Theorem 1 (Main characterization result)**.** *For any pair $S, S' \in \mathcal{S}$:*

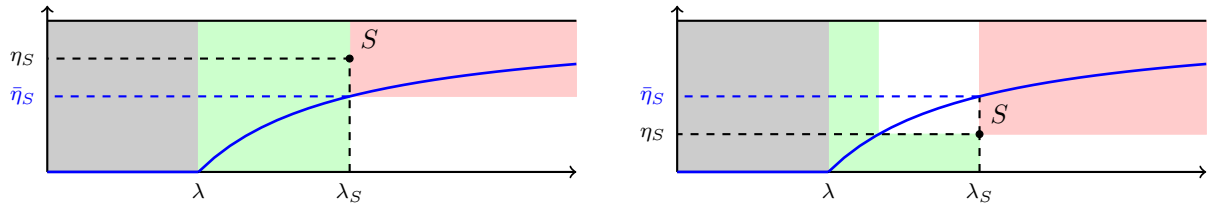
$$S \succeq S' \;\Leftrightarrow\; \boldsymbol{\phi}(S) \geq \boldsymbol{\phi}(S'). \tag{14}$$

Graphically, the previous result is illustrated in Figure 2 below. The two subfigures correspond to the two values that $\boldsymbol{\phi}_2(S)$ can potentially take, i.e., the one where $\boldsymbol{\phi}_2(S) = \bar{\eta}_S$ and therefore the information-acquisition constraint holds (i.e., $\eta_S > \bar{\eta}_S$), and the one where $\boldsymbol{\phi}_2(S) = \eta_S$ and therefore the constraint does not hold (i.e., $\eta_S \leq \bar{\eta}_S$).

In both cases, the grey area contains the trivial tasks. The red region contains the tasks $S' \in \mathcal{S}$ that are deemed more complex than S. On the other hand, the green region contains the tasks $S' \in \mathcal{S}$ that are deemed simpler than S. How we obtained the red region is obvious given our previous theorem. So, let us elaborate on how the green region arises. In both cases, S' is obviously simpler than S whenever $\lambda_{S'} \leq \lambda_S$ and $\eta_{S'} \leq \eta_S$. So let focus on tasks $S' \in \mathcal{S}$ such that $\lambda_{S'} < \lambda_S$ and $\eta_{S'} > \eta_S$. In the first case, by $\bar{\eta}_S$ being increasing in λ_S, it follows that $\bar{\eta}_{S'} < \bar{\eta}_S < \eta_S$, implying that every such S' is simpler than S. In the second case, it must also be the case that $\bar{\eta}_{S'} \leq \eta_S$. Otherwise, S and S' will be incomparable via $\succeq$.

Unlike most of the existing literature, which takes complexity essentially as a synonym of difficulty, the previous result identifies the degree of uncertainty as a novel dimension of complexity. This is consistent with the idea that a task may be deemed complex because it has not been previously encountered by the agent. For instance, students often complain that they did not have enough practice questions similar to the ones they faced in their final exam, even though their actual exam was not particularly difficult.

Nevertheless, despite identifying this new dimension of complexity, the traditional characteristics of a task (viz., difficulty and satisfaction) remain a primary channel of complexity. To see this, notice that although $\lambda_S > \lambda_{S'}$ is not sufficient for $S \succeq S'$, it is still necessary. In addition, if we keep the degree of uncertainty constant (like for instance in Goncalves, 2024), complexity is trivially reduced to difficulty-to-satisfaction.



(a) Large degree of uncertainty ($\eta_S > \bar{\eta}_S$).

(b) Small degree of uncertainty ($\eta_S \leq \bar{\eta}_S$).

Figure 2: In both cases, the red area contains the tasks that are deemed more complex than S, and the green area are the tasks that are deemed simpler than S. Finally, the grey area contains the trivial tasks, meaning that they are deemed simpler than every other task, including S. Finally, the white area contains the tasks that are not $\succeq$-comparable to S.

The fact that $\succeq$ is incomplete also distinguishes our measure of complexity from most definitions in the literature which typically induce a complete order, e.g., the signal-to-noise ratio (Callander, 2011; Fehr and Rangel, 2011; Goncalves, 2024) or willingenss to pay for avoiding a task (Oprea, 2020) or the subjective-ranking metric (Gabaix and Graeber, 2024). Nevertheless, the way in which $\succeq$ is incomplete is not arbitrary, i.e., every pair of tasks that are not $\succeq$-comparable is characterized by two threshold which determine which in which task the agent is expected to be more accurate for each reward.

Theorem 2 (Single-crossing)**.** *Suppose that $S, S' \in \mathcal{S}$ are not $\succeq$-comparable, in that $\boldsymbol{\phi}_1(S) > \boldsymbol{\phi}_1(S')$ and $\boldsymbol{\phi}_2(S) < \boldsymbol{\phi}_2(S')$. Then, there exist two thresholds $x_2 > x_1 > 0$ such that:*

(i) $P(S, x) > P(S', x)$ *for all* $x < x_1$,

(ii) $P(S, x) < P(S', x)$ *for all* $x_1 < x < x_2$,

(iii) $P(S, x) = P(S', x)$ *for all* $x \geq x_2$.

The intuition is quite simple. Two tasks are not ranked by $\succeq$ when one dominates in the difficulty-to-satisfaction ratio and the other one dominates in uncertainty. For small rewards, the agent is unlikely to pay attention to the problem, and will therefore rely on her prior knowledge. As a result, it is more likely that she guesses correctly in the task that she is more certain about, i.e., in the task that dominates in the difficulty-to-satisfaction ratio. On the other hand, for large rewards, she will primarily rely on the attention she will put. Thus, she is more likely to guess correctly in the easy task , i.e., in the task that dominates in uncertainty. And of course, for very large rewards, she will certainly guess correctly in both of them.

The previous result has strong empirical content in the following sense: whenever two tasks are not $\succeq$-comparable, we can always pinpoint which of the two tasks will be dominant in one dimension of complexity (viz., difficulty-to-satisfaction) and which one will be dominant in the

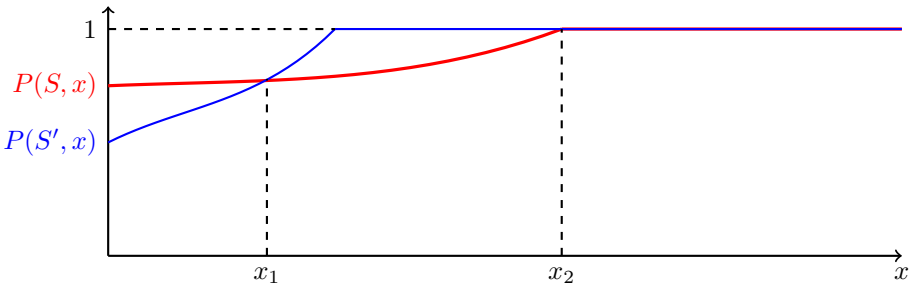


Figure 3: For illustration purposes, suppose that $\beta_S = \beta_{S'}$, i.e., two tasks yield the same satisfaction. The red task S is more difficult and more certain. This is why, for small rewards (viz., $x < x_1$) the agent relies more on her prior knowledge and is therefore more likely to solve S. On the other hand, the blue task S' is easier and less certain. Thus, for large rewards (viz., $x_1 < x < x_2$) she relies more on her attention, and hence she is more likely to solve S'. Finally, for very large rewards (viz., $x > x_2$) she will solve both tasks with certainty.

other dimension of complexity (viz., degree of uncertainty). It is exactly this property that will allow us later on to complete $\succeq$ by means of providing information about the task that involves high degree of uncertainty (Section 6).

4. Elicitation of the complexity order

Of course, any practical use of the aforementioned results fully relies on our ability to elicit the agent's complexity order. The obvious way to do this is by eliciting her belief $P(S,x)$ about guessing correctly for each task $S \in \mathcal{S}$ and each reward $x \geq 0$. This would be a standard belief elicitation exercise over the state space $Z := \{0,1\}$, where 0 and 1 respectively correspond to failing and succeeding in S when the reward is x.

The question that naturally arises then is whether there exists an incentive compatible mechanism that elicits $P(S,x)$. Obviously the agent cares about the state realization. This is because there are state-dependent side payoffs equal to $(0,0) \in X \times Z$ at state 0 and $(x,1) \in X \times Z$ at state 1. These side payoffs are on top of the prize that the belief elicitation mechanism yields. Hence, it is well-known from the literature on state-dependent SEU that $P(S,x)$ cannot be identified using traditional choice data, meaning that standard belief elicitation mechanisms —like binarized scoring rules— would not elicit the agent's actual expected accuracy (Tsakas, 2025, and references therein).

At the same time, it is not $P(S,x)$ per se that we would like to elicit, but rather the relation between $P(S,x)$ and $P(S',x)$ for each pair $S, S' \in \mathcal{S}$. This means that, even if we do not manage to truthfully elicit the agent's actual beliefs, it suffices to elicit how she ranks different tasks in terms of expected accuracy. As it turns out, this is always possible using a stochastic

elicitation mechanism, e.g., a binarized scoring rule (Hossain and Okui, 2013).

Before presenting this result, let us recall how the binarized scoring rule works. For notation simplicity, whenever it is obvious from the context, we omit reference to the task S and the reward x, and we simply write $P := P(S, x)$. The key idea that the agent reports a probability $R := R(S, x)$ of state $z = 1$. Then, based on how accurate R is in relation to the realized state in Z, she receives a number of lottery tickets for a fixed prize $y > 0$. In particular, the chances that she wins the prize depend on her reported probability R and the eventual realization of Z, as shown below:

	Wrong guess ($z = 0$)	Correct guess ($z = 1$)
R	$1 - \gamma R^2$	$1 - \gamma(1 - R)^2$

The parameter $\gamma \in (0, 1)$ determines the strength of the incentives provided by the scoring rule. It is without loss of generality to set $\gamma = 1$. As a result, the total probability of guessing wrongly and winning the prize is $(1 - P)(1 - R^2)$, and the corresponding outcome is $(0, y)$. Likewise, the total probability of guessing correctly and winning the prize is $P(1 - (1 - R)^2)$, and the corresponding outcome is $(1, x + y)$. And finally, the total probability of guessing correctly and losing the prize is $P(1 - R)^2$, and the corresponding outcome is $(1, x)$.

In this sense, by reporting R the agent chooses from a menu of acts. As usual, we assume that she will submit a report that maximizes her overall expected utility. In the standard setting, where the agent does not have any stakes in the state realization, the binarized scoring rule guarantees that P is the only optimal report, i.e., truth telling is strictly incentive compatible. Here, it is no longer the case, as there are side payoffs on top of the payment y that the scoring rule may yield. Nonetheless, as we have already mentioned, the optimal reports suffice for eliciting the complexity order.

Proposition 2 (Elicitation result)**.** *For every pair $S, S' \in \mathcal{S}$ and every $x \geq 0$:*

$$P(S, x) \geq P(S', x) \;\Leftrightarrow\; R(S, x) \geq R(S', x). \tag{15}$$

Interestingly, the previous result holds even though there are two important distortions. Namely, on the one hand, there are potential side payments for each task $S \in \mathcal{S}$ and each reward $x \geq 0$. On the other hand, these side payoffs are evaluated differently conditional on each task $S \in \mathcal{S}$, because β_S is typically not constant across $\mathcal{S}$.

Despite the positive message given by the previous result, there are still two question marks regarding implementation. First, if we pay both for the actual task and for elicitation, hedging opportunities will arise for the subject (Blanco *et al.*, 2010). Fortunately, this can be dealt with, by randomly paying only for one of the two. Second, there is an issue with the timing

of elicitation. In particular, if $R(S, x)$ is reported after the agent has undertaken the task, like in Enke and Graeber (2023), we will no longer be able to vary the reward, and therefore our measure will no longer be robust. Hence, the only option would be to elicit expected accuracy before the task is undertaken, using the strategy method across different rewards.

An alternative approach that would deal with both aforementioned issues would be to rely on an idea similar to the one Bayesian markets (Baillon, 2017), where we use beliefs about other individuals to proxy one's own beliefs about themselves. Accordingly, in our setting, we can elicit our subject's beliefs about the expected accuracy of another individual for different reward levels. Then, under the assumption that the subject considers this other individual being similar to themselves, the elicited accuracy can be used to construct our complexity measure.

5. Proof of concept

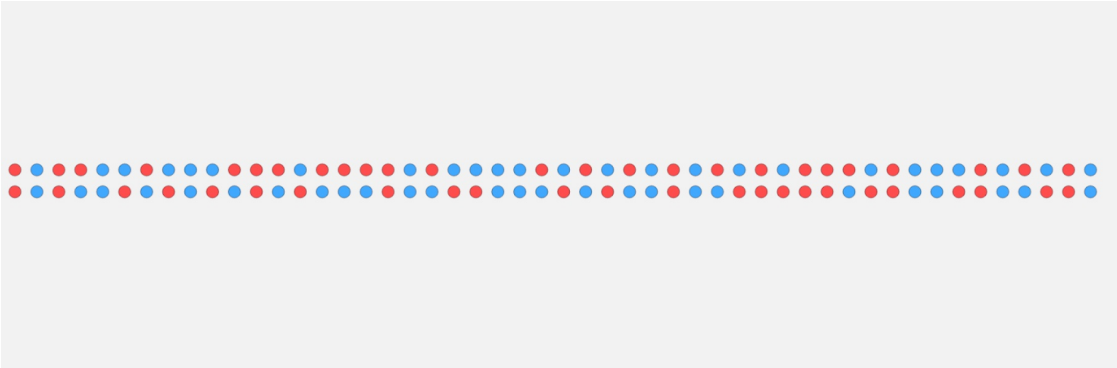
Our theory characterizes our complexity measure in terms of the three task-specific parameters: difficulty (κ_S), satisfaction (β_S), uncertainty (η_S). Thus, it provides testable hypotheses as to which tasks can be ranked by $\succeq$. In this section, we are going to test these hypotheses in a lab setting where we can observe and exogenously manipulate these parameters.

5.1. Experimental design

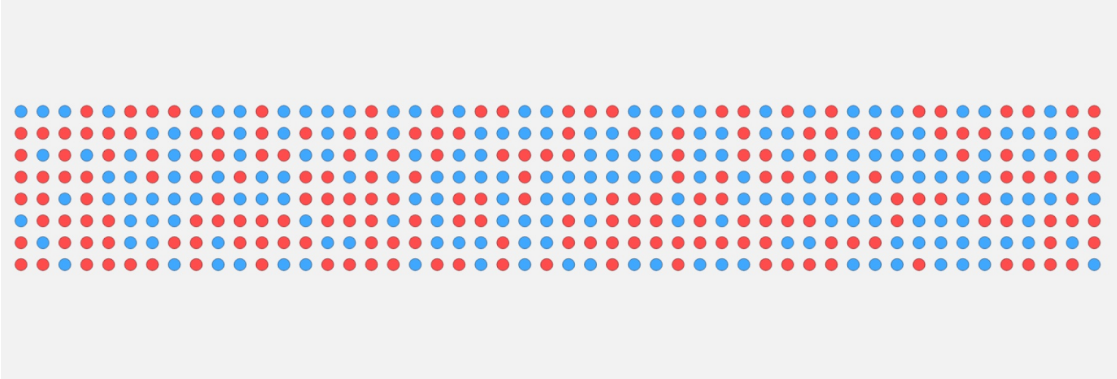
The experiment is divided in two stages. Both stages were run in the Behavioral and Experimental Economics Lab (BEELab) in Maastricht University, with subjects recruited from the same pool of students.

In the first stage each participant would face a sequence of tasks, which we describe below in detail, and are similar to the ones in Dean and Neligh (2023) and Goncalves *et al.* (2024). One of the tasks would be randomly drawn in the end, and if the participant answered correctly in this task, this participant would receive a bonus payment (€x), on top of the show-up fee (€5). There were two treatments in the first stage which would only differ in the size of the bonus, i.e., the low-stakes treatment where the bonus was $x = 0.50$, and the high-stakes treatment where the bonus was $x = 10$. That is, the tasks that participants would face in each of the two treatments were identical.

The building blocks of a task are panels with blue and red marbles randomly scattered on the screen. There are four types of panels, differing in the color that the majority of the marbles has (viz., the state) and the total number of marbles on the panel (viz., the difficulty). In particular, a panel contains either more red marbles (i.e., it is a red panel) or more blue marbles (i.e., it is a blue panel). Moreover, a panel is either easy (viz., it contains 100 marbles



(a) Easy blue-dominant panel: it contains 49 red and 51 blue marbles.



(b) Difficult red-dominant panel: it contains 201 red and 199 blue marbles.

Figure 4: Examples of two panels that are seen by participants in the first stage.

in total: 51 of one color and 49 of the other color) or (viz., it contains 400 marbles in total: 201 of one color and 199 of the other color). Examples of an easy and a difficult panel are shown in Figure 4.

At the beginning of each task, there is a pool of 10 panels. The participant knows whether the task is easy (i.e., all 10 panels in the pool are easy) or difficult (i.e., all 10 panels in the pool are difficult). Moreover, the participant knows the proportion of red/blue panels in the pool. In particular, in each task there will be either high degree of uncertainty (i.e., 5 red and 5 blue panels) or low degree of uncertainty (i.e., 8 panels of one color and 2 panels of the other color). Then, in each task, a panel is randomly chosen from the respective pool and the participant is asked to guess the state, i.e., whether there are more red or blue marbles in this panel.

Thus, there are 4 tasks in total, identified by points in Figure 5. Note that on the horizontal axis we measure difficulty, rather than the difficulty-to-satisfaction ratio. This is because throughout the experiment we assume participants to derive the same satisfaction from solving

an easy and a difficult task. This is a rather natural assumption, based on the fact that the tasks are neutral and most likely do not interact with ones own self-image. The advantage to assuming the same β_S for every $S \in \{EL, EH, DL, DH\}$ is that we can exogenously manipulate the two remaining task-specific parameters, viz., difficulty and uncertainty, and therefore to derive testable hypotheses.

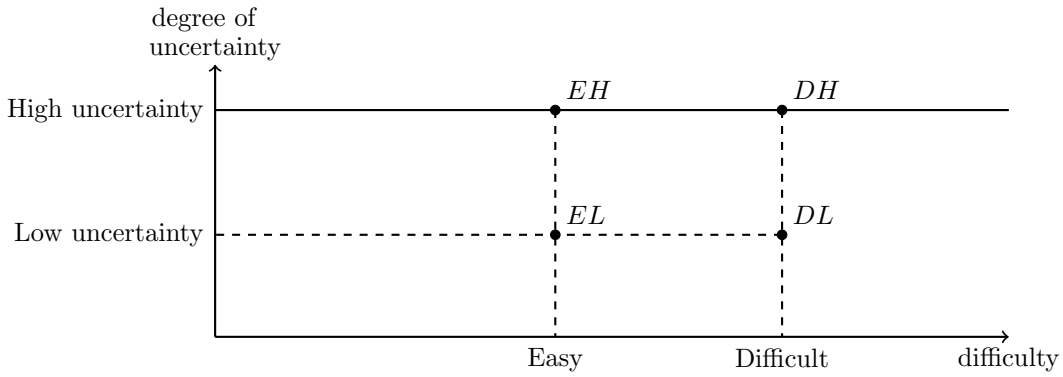


Figure 5: The tasks that the participants face in the first stage of the experiment.

The actual accuracy of participants in the first stage is irrelevant for our hypotheses testing. It will only be used to avoid deception in the second stage, which is the main part of the experiment. As we have already briefly discussed in Section 4, instead of eliciting participants' beliefs about their own accuracy, we will elicit others' beliefs about the accuracy of participants in the aforementioned first stage. This is exactly what we do in the second stage of our experiment.

The second stage begins by explaining to participants the first stage of the experiment, while explicitly telling them that the first stage was recently conducted in the same lab with participants drawn from the same pool as themselves. Then, for each task $S \in \{EL, EH, DL, DH\}$ and each reward $x \in \{0.5, 10\}$, they were asked to guess the proportion of first-stage participants that picked the correct color.

Formally speaking, what we elicit is the mean of the subjective distribution over the percentages of correct answers $\{0\%, 1\%, \ldots, 100\%\}$, which is henceforth denoted by $P(S, x)$. To guarantee incentive compatibility, we use a standard binarized quadratic scoring rule for eliciting distribution means (Schlag and van der Weele, 2013). The prize of this scoring rule is €12. Notice that, being aligned with Danz *et al.* (2022), the belief elicitation mechanism is shown to the participants only upon request (by clicking on a special button), and otherwise they are simply told that it is in their best interest to report truthfully.

5.2. Hypotheses and results

Recall that each second-stage participant reports 8 different percentages $P(S,x)$: one for each task $S \in \{EL, EH, DL, DH\}$ and each reward $x \in \{0.5, 10\}$. Our first hypothesis states that participants believe that accuracy is increasing in the incentives.

Hypothesis 1. *For each $S \in \{EL, EH, DL, DH\}$ we hypothesize the following:*

$$P(S,10) \geq P(S,0.5). \tag{$H.1$}$$

This is a sanity check, consistent with the findings in Dean and Neligh (2023).

Our second hypothesis follows directly from Theorem 1, according to which tasks that are simultaneously more difficult and involve more uncertainty are expected to always be less likely to solve, and a fortiori more complex.

Hypothesis 2. *For each $x \in \{0.5, 10\}$ we hypothesize the following:*

$$P(EL,x) \geq P(EH,x) \geq P(DH,x), \tag{$H.2a$}$$
$$P(EL,x) \geq P(DL,x) \geq P(DH,x). \tag{$H.2b$}$$

Based on our theory the previous hypothesis should hold irrespective of the attention constraint $\bar{\eta}_S$, in the sense that all tasks that are located north east of S are more complex than S according to our robust measure (Figure 2). The previous hypotheses are also consistent with the findings of Dean and Neligh (2023).

Our third hypothesis focuses on the comparison of two tasks, EH and DL, where DL dominates in one dimension (viz., difficulty) and EH dominates in the other dimension (viz., uncertainty). As a result, it is not obvious whether they are even comparable in the first place: this would depend on the attention constraint $\bar{\eta}_{EH}$, which is of course unobservable. Nonetheless, relying on the single-crossing condition of Theorem 2, we can formulate the following hypothesis.

Hypothesis 3. *We hypothesize the following:*

$$P(EH,0.5) \geq P(DL,0.5) \;\Rightarrow\; P(EH,10) \geq P(DL,10), \tag{$H.3a$}$$
$$P(EH,10) \leq P(DL,10) \;\Rightarrow\; P(EH,0.5) \leq P(DL,0.5). \tag{$H.3b$}$$

Of course, if EH and DL are $\succeq$-comparable, then the hypothesis follows directly from Theorem 1. If on the other hand, they are not $\succeq$-comparable, then in the context of Figure 3 we can set $S = DL$ and $S' = EH$. In this case, $P(EH,0.5) \geq P(DL,0.5)$ implies that $x_1 \leq 0.5$, and therefore it must also be the case that $x_1 < 10$. Hence, according to Theorem 2, we should

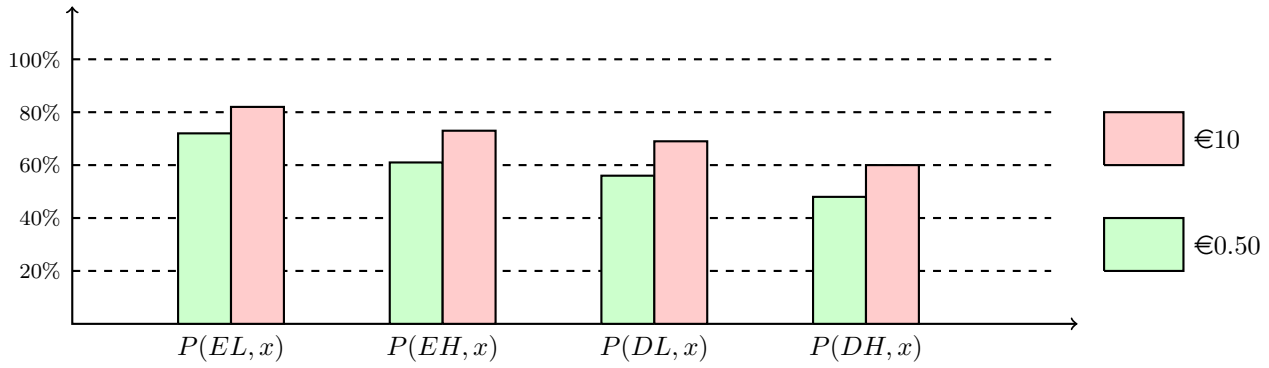


Figure 6: Summary of the expected accuracy reported in the second stage about the performance of participants in the first stage for each task and each reward.

also obtain $P(EH, 10) \geq P(DL, 10)$. The argument is identical for the second part of the hypothesis.

In total, we recruited 56 participants. Their average reported guesses (for the different tasks and rewards) are summarized in Figure 6.

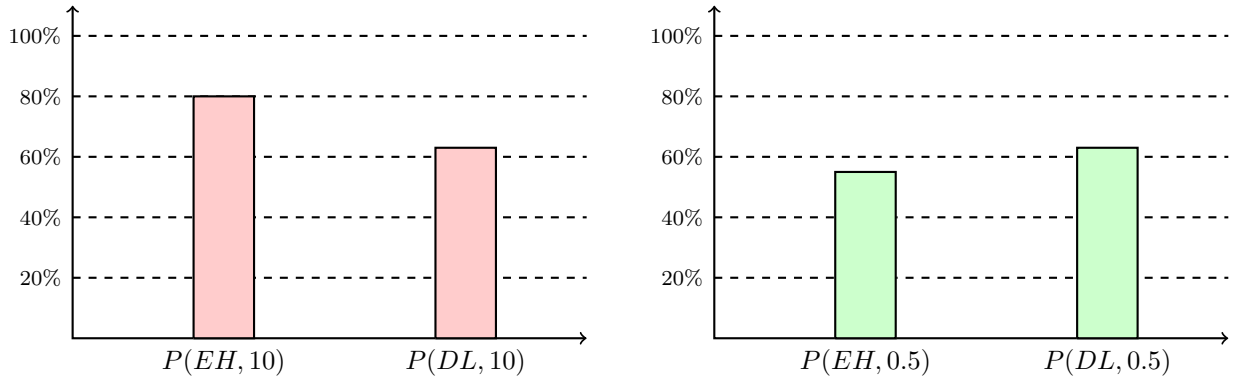
The first observation is that both in Hypothesis 1 and Hypothesis 2, the direction is the one we hypothesized, i.e., higher reward leads on average to higher expected accuracy. Then, using the Wilcoxon signed-rank test, we find that these differences are statistically significant (with $p < 0.01$). This implies that both hypotheses are corroborated.

Then, we turn to Hypothesis 3. To test it, we need to restrict attention to those participants who reported $P(EH, 0.5) \geq P(DL, 0.5)$ for $H.3a$, and respectively to those who reported $P(EH, 10) \leq P(DL, 10)$ for $H.3b$. In each of the two groups, the average accuracies that we respectively observed are the ones depicted depicted in Figure 7.

Obviously, the differences are in the direction that we hypothesized. Then, once again, using Wilcoxon signed-rank test, we find that these differences are statistically significant both in $H.3a$ (with $p < 0.01$) and in $H.3b$ (with $p < 0.03$). Hence, our last hypothesis is corroborated too.

6. Complexity and information

What primarily distinguishes our measure from other definitions of complexity in the literature is the role of the degree of uncertainty. In particular, the reason why two tasks are not $\succeq$-comparable is that they involve different degrees of uncertainty. In other words, if we take $\eta_S = \eta_{S'}$, it will necessarily be the case that S and S' are $\succeq$-comparable (Theorem 1). But then again, differences in the degree of uncertainty reflect differences in prior information about the tasks. This suggests that, since information has rendered the tasks incomparable, information



(a) Given $P(EH, 0.5) \geq P(DL, 0.5)$

(b) Given $P(EH, 10) \leq P(DL, 10)$.

Figure 7: Expected accuracy reported in the second stage about the performance participants in the first stage, conditional on the antecedents of Hypothesis 3.

can also render them comparable, i.e., we conjecture that by providing extra information to the agent, she will eventually be able to rank the tasks with respect to complexity. Interestingly, this will be the case without needing to assume much about the underlying information stream.

Suppose that we feed the agent information in the form of outcomes from some experiment about task S. Examples of such experiments could be for instance some past realizations of S, or recommendations of some trusted expert about the best guess in S.

As usual, the agent interprets the experiment as a stochastic mapping

$$\sigma : S \to \Delta(T)$$

where T contains the possible outcomes of the experiment. Without loss of generality, we consider binary experiments, i.e., we will have $T = \{t_0, t_1\}$, where t_0 is evidence in favor of s_0, and t_1 is evidence in favor of s_1. Formally, we have

$$\frac{\sigma(t_0|s_1)}{\sigma(t_0|s_0)} < 1 \quad \text{and} \quad \frac{\sigma(t_1|s_1)}{\sigma(t_1|s_0)} > 1.$$

Our model is flexible enough to allow for the possibility that the agent's perceived likelihood $\sigma(t|s)$ is misspecified. For example, what the agent believes is the reliability of some expert does not necessarily coincide with the actual reliability of this expert. Importantly, we remain agnostic on the agent's specification of the experiment. Furthermore, we remain agnostic as to how the agent interprets the experiment, i.e., we do not need to know anything about σ, other than the fact that the agent finds it informative somehow.

From the agent's point of view, the likelihood of observing a realized sample $\boldsymbol{t} = (t^1, \ldots, t^n)$

given state s is equal to

$$\sigma(\boldsymbol{t}|s) = \prod_{k=1}^{n} \sigma(t^k|s).$$

Thus, her posterior belief of s_1 conditional on $\boldsymbol{t}$ is denoted by

$$\mu_S^{\boldsymbol{t}} = \frac{\mu_S \sigma(\boldsymbol{t}|s_1)}{(1-\mu_S)\sigma(\boldsymbol{t}|s_0) + \mu_S \sigma(\boldsymbol{t}|s_1)}$$

Let us denote her optimal signal in task S given her posterior belief by $\pi_S^x(\cdot|\boldsymbol{t})$. Then, her resulting expected accuracy becomes

$$P(S, x|\boldsymbol{t}) = \frac{G_S\big(\pi_S^x(\cdot|\boldsymbol{t})\big)}{\beta_S v(x)}. \tag{16}$$

We will henceforth say that task $S' \in \mathcal{S}$ is more complex than task $S \in \mathcal{S}$ given a sequence of conditionally independent outcomes $\boldsymbol{t}$ from some experiment on S, whenever

$$P(S, x|\boldsymbol{t}) \geq P(S', x) \tag{17}$$

for all rewards $x \geq 0$.

The fact that $\boldsymbol{t}$ is randomly drawn from some experiment obviously implies that "S' being more complex than S" is an event which occurs with some probability. Let us formalize this idea. Consider the space $\boldsymbol{T} := T^{\mathbb{N}}$ of all infinite sequences $(t^1, t^2, \dots)$ of outcomes, and let

$$[\![t^1, \dots, t^n]\!] := \{t^1\} \times \cdots \times \{t^n\} \times T \times \cdots$$

be the event that the first n observations are given by $(t^1, \dots, t^n)$. Define the partition of $\boldsymbol{T}$,

$$\mathcal{T}_n := \big\{ [\![t^1, \dots, t^n]\!] \;\big|\; (t^1, \dots, t^n) \in T^n \big\},$$

and the generated σ-algebra $\mathcal{F}_n := \sigma(\mathcal{T}_n)$. As we have already mentioned, the agent's interpretation of the signal might be misspecified. Let $q \in (0,1)$ denote the actual data-generating process for each single outcome, i.e., formally, this is the actual probability of outcome t_1 being realized. Importantly, q is not necessarily known by us, and we remain agnostic as to where it is coming from. Nevertheless, by Kolmogorov extension theorem (Aliprantis and Border, 1994, Theorem 15.23), it is guaranteed that there is a unique probability measure over the set infinite sequences of outcomes

$$P_q \in \Delta(\boldsymbol{T}, \mathcal{F}),$$

where $\mathcal{F} := \bigcup_{n\in\mathbb{N}} \mathcal{F}_n$ is the limiting σ-algebra. This will be treated as the actual data-generating process of experimental outcomes.

Now, we are ready to define the event that S' is more complex than S given $(t^1, \dots, t^n)$:

$$[\![S' \succeq S | t^1, \dots, t^n]\!] := \bigcap_{x \geq 0} \Big\{ (t^1, t^2, \dots) \in \boldsymbol{T} : P(S, x | t^1, \dots, t^n) \geq P(S', x) \Big\}. \tag{18}$$

Clearly, this event is $\mathcal{F}_n$-measurable, and a fortiori $\mathcal{F}$-measurable, meaning that it admits some probability by P_q.

This means that whether S is more complex than S' given a sample $(t^1, \dots, t^n)$ is an event in $\boldsymbol{T}$, which admits some probability. Then, as we show below, this probability approaches 1 as the sample becomes large.

Proposition 3 (Completion of the complexity order)**.** *Let $\boldsymbol{\phi}_1(S) < \boldsymbol{\phi}_1(S')$. Then, for every $\varepsilon > 0$ and every $q \in (0,1)$, there is some $N > 0$ such that*

$$P_q\Big([\![S' \succeq S | t^1, \dots, t^n]\!] \Big) > 1 - \varepsilon, \tag{19}$$

for every $n > N$.

Remarkably, the previous result requires very little in terms of assumptions, i.e., regardless which experiment generates the outcomes, regardless which sequence of outcomes is realized, regardless how she interprets these outcomes, according to the result we can be almost sure that the agent will eventually rank the two tasks in terms of our complexity measure.

The proof of the previous result relies on the Central Limit Theorem. In particular, it is only a P_q-null subset of sequences of outcomes in $\boldsymbol{T}$ for which the degree of uncertainty η_S does not converge to 0. So, no matter what $\eta_{S'}$ is equal to, we will eventually obtain $\boldsymbol{\phi}_2(S) < \boldsymbol{\phi}_2(S')$, meaning that S will be classified as simpler than S'.

Note that Theorem 2 already tells us which task to collect information about. If we instead collect information about the other task, we will might be able to spontaneously rank the tasks for some finite samples $(t^1, \dots, t^n)$, but as the sample size increases, the tasks will become almost surely incomparable. This means that we need to rely on being lucky in order to rank the two tasks.

7. Extensions and limitations

7.1. State-dependent utilities

Suppose that the agent's utility function v_S is not just task dependent, but also state-dependent within S (Tsakas, 2025, and references therein). Consider, for instance, a task which asks whether there the economy was better under Biden (s_0) or under Trump (s_1), and for the sake of simplicity let the agent be a Republican partisan who intrinsically prefers s_1 to be true.

Assume that the latter is reflected in her utility function in the following way: we have $u_S^0 := \beta_S^0 u$ and $u_S^1 := \beta_S^1 u$ where $\beta_S^0 < \beta_S^1$, and her preferences are then represented by the State-dependent Subjective Expected Utility (abbrev., SDSEU),

$$\mathbb{E}_{\mu_S}(u_S(f)) = (1-\mu_S)u_S^0(f(s_0)) + \mu_S u_S^1(f(s_1)),$$

where again μ_S is the probability she actually assigns to s_1.

The fact that utility is state-dependent implies that the net expected utility function g_S will now become

$$g_S(q) = v(x)\max\left\{\beta_S^0(1-q), \beta_S^1 q\right\}.$$

Thus, the belief $\bar{q}_S$ that makes the agent indifferent between guess s_0 and guessing s_1 will be smaller than $1/2$. As a result, the posteriors $q_{s_0}^x < \bar{q}_S$ and $q_{s_1}^x > \bar{q}_S$ that determine the attention region will now be such that $q_{s_1}^x < 1 - q_{s_0}^x$, as opposed to the state-independent case where they are symmetrically located around $1/2$.

Hence, there will be tasks S such that the expected accuracy $P(S,x)$ is not increasing in the reward x. Let us illustrate this in Figure 8. Suppose that the intrinsic incentives are weak enough so that $q_{s_1}^0 < 1/2$. This means that for a prior belief μ_S which is smaller but close to $\bar{q}_S^x$, there will be intermediate rewards $x > 0$ such that $P(S,x) < 1 - \mu_S$. In the figure, this is when $q_{s_0}^x < \mu_S < \bar{q}_S$. But then again, when the reward becomes small enough $x' < x$, the agent will not pay any attention, and therefore we will have $P(S,x') = 1 - \mu_S$, meaning that $P(S,x') > P(S,x)$.

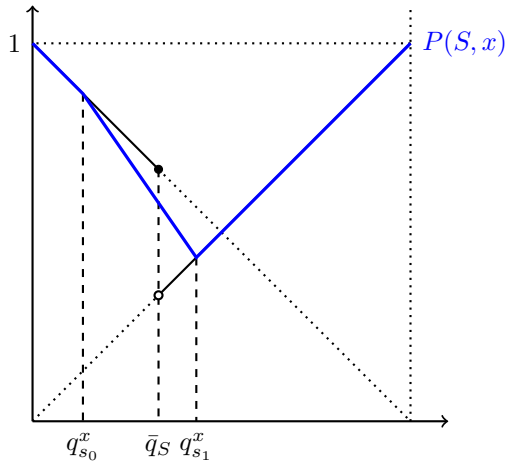


Figure 8: Given the task S and the reward x, the blue piecewise linear function is the expected accuracy (as a function of the agent's prior), assuming that the optimal signal π_S^x has been used.

Similarly, expected accuracy will not be decreasing with respect to difficulty. In particular, take another task S' which is identical to S in every aspect other than difficulty, i.e., we have

$\mu_S = \mu_{S'}$ and $u_S = u_{S'}$, but at the same time $\kappa_{S'} > \kappa_S$. Then, using a similar argument as the one above, we will obtain $P(S', x) > P(S, x)$, i.e., the agent has better chances at solving the more difficult task.

The reason why state-dependent preferences yield such counterintuitive predictions is that agent's expected benefit $g_S(q)$ is not monotonic in the probability of being correct. This is because $P(S, x)$ is discontinuous at $\bar{q}_S$. Intuitively, because of the agent's intrinsic preferences for state s_1, she will not necessarily strive to maximize the probability of answering correctly. Hence, in the presence of stakes about the state space, expected accuracy is not always a good proxy.

7.2. Cost specification

Throughout the paper, we have assumed that the cost function c is symmetric. How restrictive is this?

First of all, conceptually, the main criticism that symmetry has received in the literature is based on the idea that distinguishing between two states might be more difficult than distinguishing between two other states (Hébert and Woodford, 2021). In other words, asymmetries should enter the picture primarily in cases where the state space has some underlying distance, and more similar states are harder to tell apart. However, given that throughout the paper we focus on binary tasks, it is reasonably justified to maintain symmetric costs.

Second, if we nonetheless decide to relax the symmetry assumption, our main theorem will still partially hold. In particular, it will be the case that $\lambda_S \geq \lambda_{S'}$ together with $1/2 \leq \mu_S \leq \mu_{S'}$ (or together with $1/2 \geq \mu_S \geq \mu_{S'}$) implies $S \succeq S'$. Notice that these conditions are satisfied by the tasks that we use in our experiment in Section 5. In this sense, the intuition that complexity has two dimensions (viz., difficulty and uncertainty) still holds, meaning that symmetry is not that essential.

8. Conclusion

In this paper, we proposed a robust version of a belief-based measure of complexity that has recently surged (Enke and Graeber, 2023; Enke *et al.*, 2024a,b; Agranov *et al.*, 2025). Our starting point was that task A is classified as more complex than task B if the chances of A being solved are lower than the chances of B being solved, *irrespective of the reward*. The theoretical foundations that we provided for this criterion uncovered a novel dimension of complexity which was previously overlooked in the literature, viz., uncertainty. In particular, we showed that task A is more complex than task B (according to our criterion) if A is both more difficult and the agent knows ex ante less about it than B. Using a lab experiment, where we

could exogenously control difficulty and uncertainty, we managed to validate our measure. So, practically speaking, the conclusion is that expected accuracy is a good measure of complexity, as long as we elicit it for multiple different rewards.

A. Proofs

Proof of Proposition 1. By a standard concavification argument, there exists some $q_S^x \in \left[0, 1/2\right]$ such that the optimal signal π_S^x distributes all probability between q_S^x and $1-q_S^x$ whenever $q_S^x < \mu_S < 1 - q_S^x$, and it is completely uninformative otherwise. Define

$$\lambda := \sup \left\{ \lambda_S \;\middle|\; S \in \mathcal{S}_0 \text{ such that } q_S^0 = 0 \right\}. \tag{A.1}$$

($\Rightarrow$) : Take an arbitrary $S \in \mathcal{S}_0$. First, if $\eta_S = 0$, then $P(S, x) = 1$ for all $x \geq 0$. Second, if $\lambda_S \leq \lambda$ then $q_S^0 = 0$,which together with $q_S^x \leq q_S^0$ implies $q_S^x = 0$, and therefore $P(S, x) = 1$ for all $x \geq 0$. Putting the two together completes this part of the proof.

($\Leftarrow$) : For an arbitrary $S \in \mathcal{S}_0$, suppose that $\lambda_S > \lambda$ and $\eta_S > 0$. Then, by (A.1), we have $q_S^0 > 0$, and therefore $P(S, 0) < 1$, which completes the second part of the proof. □

Proof of Theorem 1. ($\Leftarrow$) : By $\min\{\bar{\eta}_S, \eta_S\} \geq \min\{\bar{\eta}_{S'}, \eta_{S'}\}$, we obtain $\eta_S \geq \min\{\bar{\eta}_{S'}, \eta_{S'}\}$. Then, consider two cases:

1) $\eta_S \geq \eta_{S'}$: This means $\max\{\mu_S, 1 - \mu_S\} \leq \max\{\mu_{S'}, 1 - \mu_{S'}\}$. Moreover, by q_S^x being increasing in λ_S, we have $1 - q_S^x \leq 1 - q_{S'}^x$. Hence, by Equation (10), we obtain $P(S, x) \leq P(S', x)$ for all $x \geq 0$.
2) $\eta_S < \eta_{S'}$: This means that $\eta_S \geq \bar{\eta}_{S'}$, and a fortiori $\max\{\mu_S, 1 - \mu_S\} \leq 1 - q_{S'}^0$. Moreover, by $q_{S'}^x$ being decreasing in x, we have $1-q_{S'}^x \geq 1-q_{S'}^0$. Hence, we obtain $\max\{\mu_S, 1-\mu_S\} \leq 1 - q_{S'}^x$. Then, together with $1 - q_S^x \leq 1 - q_{S'}^x$, we obtain

$$\max\{\mu_S, 1 - \mu_S, 1 - q_S^x\} \leq 1 - q_{S'}^x \leq \max\{\mu_{S'}, 1 - \mu_{S'}, 1 - q_{S'}^x\},$$

which, by Equation (10), implies $P(S, x) \leq P(S', x)$ for all $x \geq 0$.

Putting the two cases together directly yields $S \succeq S'$.

($\Rightarrow$) : Sufficiency follows directly from Theorem 2, which we prove below. □

Proof of Theorem 2. By $\lambda_S = \boldsymbol{\phi}_1(S) > \boldsymbol{\phi}_1(S') = \lambda_{S'}$, we obtain for all $x \geq 0$

$$q_{S'}^x \leq q_S^x \tag{A.2}$$

with equality holding if and only if $q_S^x = 0$. Applying the latter for $x = 0$, together with the fact that S and S' are non-trivial, yields

$$\bar{\eta}_{S'} < \bar{\eta}_S. \tag{A.3}$$

By combining (A.3) with our second hypothesis $\min\{\bar{\eta}_S, \eta_S\} = \boldsymbol{\phi}_2(S) < \boldsymbol{\phi}_2(S') = \min\{\bar{\eta}_{S'}, \eta_{S'}\}$, we obtain the following two inequalities:

$$\eta_S < \eta_{S'}, \tag{A.4}$$

$$\eta_S < \bar{\eta}_{S'}. \tag{A.5}$$

By $q_{S'}^x$ being strictly decreasing in x, there exists a unique $x_1 > 0$ such that $\mu_S \in (0, q_{S'}^x) \cup (1 - q_{S'}^x, 1)$ for all $x < x_1$. Thus, by combining (A.4) and (A.5) with (10), we obtain $P(S, x) > P(S', x)$. On the other hand, for all $x > x_1$, we have $\mu_S \in (q_{S'}^x, 1 - q_{S'}^x)$. Hence, by (A.4), (A.5) and (10), we obtain $P(S, x) \leq P(S', x)$, with equality holding if and only if $x \geq x_2 := \inf\{x > 0 : P(S, x) = 1\}$. The latter completes the proof. □

Proof of Proposition 2. Fix task $S \in \mathcal{S}$ and reward $x \geq 0$. Then, the overall expected utility from reporting R is equal to

$$\begin{aligned}\mathbb{E}_P(U(R)) &= (1-P)(1-R^2)u_S(0,y) + P(1-(1-R)^2)u_S(1,x+y) + P(1-R)^2 u_S(1,x) \\ &= \beta_S\Big((1-P)(1-R^2)u(0,y) + P(1-(1-R)^2)u(1,x+y) + P(1-R)^2 u(1,x)\Big).\end{aligned}$$

Then, the first order condition yields

$$\frac{\partial \mathbb{E}_P(U(R))}{\partial R} = -(1-P)Ru(0,y) + P(1-R)u(1,x+y) - P(1-R)u(1,x) = 0,$$

which is in turn equivalent to

$$\frac{1-R}{R} = \frac{1-P}{P} \cdot \frac{u(0,y)}{u(1,x+y) - u(1,x)}.$$

Since the second fraction on the righthand side is task independent and strictly positive, it follows that R is strictly increasing in P. □

Proof of Proposition 3. Let $\boldsymbol{t} = (t^1, t^2, \dots) \in \boldsymbol{T}$ be an infinite sequence of Bernoulli realizations. Let

$$\gamma_{\boldsymbol{t}} := \lim_{n\to\infty} \frac{\big|\{k = 1, \dots, n : t^k = t_1\}\big|}{n}$$

be the limiting frequency of t_1 in $\boldsymbol{t}$. Moreover, by $\sigma_0 := \sigma(t_0|s_1)$ and $\sigma_1 := \sigma(t_1|s_1)$ denote the perceived likelihoods. Then, the posterior odds can be written as

$$\frac{\mu_S^{\boldsymbol{t}}}{1-\mu_S^{\boldsymbol{t}}} = \frac{\mu}{1-\mu} \cdot \lim_{n\to\infty} \left(\left(\frac{\sigma_0}{1-\sigma_0}\right)^{1-\gamma_{\boldsymbol{t}}} \left(\frac{\sigma_1}{1-\sigma_1}\right)^{\gamma_{\boldsymbol{t}}} \right)^n.$$

Then, we have $\mu_S^{\boldsymbol{t}} \in (0,1)$ if and only if

$$(1-\gamma_{\boldsymbol{t}}) \log \frac{\sigma_0}{1-\sigma_0} = \gamma_{\boldsymbol{t}} \log \frac{\sigma_1}{1-\sigma_1}.$$

Notice that the last equation has a unique solution $\gamma^* \in (0,1)$. Then, by the Central Limit Theorem, the event $\{\boldsymbol{t} \in \boldsymbol{T} : \gamma_{\boldsymbol{t}} = \gamma^*\}$ is null, regardless of q, which completes the proof. □

B. Experimental results

For the first stage of the experiment we invited 20 subjects who were randomly placed in the two treatments (6 participants in the low-stakes and 14 participants in the high-stakes treatment).[6] The descriptive statistics of the first stage looked as follows. Obviously, with the number of observations that we have, it would be meaningless to do any kind of statistical analysis. Thus, as explained in the main body of the paper, we use the first stage only as an auxiliary treatment, which is run with the sole purpose of not deceiving participants in the second stage.

	Low Stakes	**High Stakes**
EL	66.7	92.9
EH	50.0	100.0
DL	16.7	0.0
DH	50.0	42.9

Table 1: Percentage of correct answers for each task and each treatment.

For the second stage, we preregistered 100 participants. In the end, we only managed to recruit 57 subjects (30 females/27 males) from our pool. The experiment was conducted in May 2025 in the Behavioral and Experimental Economic Laboratory (BEELab) in Maastricht University.

Their average guesses for the percentage of first-stage participants that answered correctly are depicted in Table 2. This data is depicted graphically in Figure 6.

[6] Due to random assignment the two treatments were not balanced. This is not a concern for our experiment as the first stage was merely used to calculate the payments in the second stage without deceiving the participants of the second stage.

	Mean	St. Dev.
$P(EL, €10)$	82.6	15.7
$P(EL, €0.5)$	72.3	19.8
$P(EH, €10)$	72.6	18.9
$P(EH, €0.5)$	61.9	18.1
$P(DL, €10)$	68.9	19.5
$P(DL, €0.5)$	56.1	21.6
$P(DH, €10)$	59.7	16.5
$P(DH, €0.5)$	48.0	13.7

Table 2: The average guess for the expected accuracy of first-stage participants for each task and each reward.

In order to test Hypothesis 1, we ran a Wilcoxon signed rank test for each of the four tasks. The results are shown in Table 3.

Hypothesis	z	$Prob > \|z\|$	Exact Prob
$P(EL, €10) = P(EL, €0.5)$	4.662	0.0000	0.0000
$P(EH, €10) = P(EH, €0.5)$	4.968	0.0000	0.0000
$P(DL, €10) = P(DL, €0.5)$	5.008	0.0000	0.0000
$P(DH, €10) = P(DH, €0.5)$	5.569	0.0000	0.0000

Table 3: The results of the Wilcoxon signed rank test for Hypothesis 1.

In order to test Hypothesis 2, we ran a Wilcoxon signed rank test for each inequality and for each reward. The results are shown in two separate tables, viz., in Table 4 for $H.2a$ and Table 5 for $H.2b$.

Hypothesis	z	$Prob > \|z\|$	Exact Prob
$P(EL, €10) = P(EH, €10)$	5.438	0.0000	0.0000
$P(EL, €0.5) = P(EH, €0.5)$	4.650	0.0000	0.0000
$P(EH, €10) = P(DH, €10)$	5.370	0.0000	0.0000
$P(EH, €0.5) = P(DH, €0.5)$	5.244	0.0000	0.0000

Table 4: The results of the Wilcoxon signed rank test for Hypothesis $H.2a$.

In order to test Hypothesis 3, we ran two Wilcoxon signed rank tests. For Hypothesis $H.3a$, the number of participants who guessed $P(EH, 0.5) \geq P(DL, 0.5)$ was 32. For Hypothesis

Hypothesis	z	$Prob > \|z\|$	Exact Prob
$P(EL, €10) = P(DL, €10)$	4.534	0.0000	0.0000
$P(EL, €0.5) = P(DL, €0.5)$	4.621	0.0000	0.0000
$P(DL, €10) = P(DH, €10)$	3.714	0.0002	0.0001
$P(DL, €0.5) = P(DH, €0.5)$	3.171	0.0015	0.0012

Table 5: The results of the Wilcoxon signed rank test for Hypothesis $H.2b$.

$H.3b$, the number of participants who guessed $P(EH, 0.5) \geq P(DL, 0.5)$ was 28. The results are shown below in two separate tables, viz., in Table 6 for $H.3a$ and Table 7 for $H.3b$.

Hypothesis	z	$Prob > \|z\|$	Exact Prob
$P(EH, €10) = P(DL, €10)$	3.696	0.0002	0.0001

Table 6: The results of the Wilcoxon signed rank test for Hypothesis $H.3a$. That is, the hypothesis is tested conditional on $P(EH, €0.5) \geq P(DL, €0.5)$.

Hypothesis	z	$Prob > \|z\|$	Exact Prob
$P(EH, €0.5) = P(DL, €0.5)$	-2.186	0.0288	0.0280

Table 7: The results of the Wilcoxon signed rank test for Hypothesis $H.3b$. That is, the hypothesis is tested conditional on $P(EH, €10) \leq P(DL, €10)$.

C. Experimental instructions

The complete set of instructions for the first stage of the experiment can be found by clicking on this link and for the second stage on this link.

For the first stage, these are the instructions for the high stakes treatment. The only difference in the low stakes treatment is that the bonus of €10 is replaced by €0.50.

References

Abreu, D. and Rubinstein, A. (1988). The structure of nash equilibria in repeated games with finite automata. *Econometrica*, **56**, 1259–1281.

Agranov, M., Schotter, A. and Trevino, I. (2025). Complex for whom? an experimental approach to subjective complexity. *Working Paper*.

Alaoui, L. and Penta, A. (2022). Cost-benefit analysis in reasoning. *Journal of Political Economy*, **130**, 4.

Aliprantis, C. and Border, K. (1994). *Infinite dimensional analysis*. Springer.

Anscombe, F. and Aumann, R. (1963). A definition of subjective probability. *Annals of Mathematical Statistics*, **34**, 199–205.

Baillon, A. (2017). Bayesian markets to elicit private information. *Proceedings of the National Academy of Sciences*, **114**, 7958–7962.

Banovetz, J. and Oprea, R. (2023). Complexity and procedural choice. *American Economic Journal: Microeconomics*, **15**, 3913–3951.

Blanco, M., Engelmann, D., Koch, A. and Normann, H.-T. (2010). Belief elicitation in experiments: is there a hedging problem? *Experimental Economics*, **13**, 412–438.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3.

Callander, S. (2011). Searching and learning by trial and error. *American Economic Review*, **111**, 2277–2308.

Caplin, A. (2025). *An introduction to cognitive economics: The science of mistakes*. Palgrave.

—, Dean, M. and Leahy, J. (2022). Rationally inattentive behavior: Characterizing and generalizing Shannon entropy. *Journal of Political Economy*, **130**, 1676–1715.

Cover, T. and Thomas, J. (2006). *Elements of Information Theory*. Wiley-Interscience.

Danz, D., Vesterlund, L. and Wilson, A. (2022). Belief elicitation and behavioral incentive compatibility. *American Economic Review*, **112**, 2851–2883.

de Clippel, G., Moscariello, P., Ortoleva, P. and Rozen, K. (2025). Caution in the face of complexity. *Working Paper*.

Dean, M. and Neligh, N. (2023). Experimental tests of rational inattention. *Journal of Political Economy*, **131**, 3415–3461.

Denti, T. (2022). Posterior separable cost of information. *American Economic Review*, **112**, 3215–3259.

Enke, B. (2024). The cognitive turn in behavioral economics. *Working Paper*.

— and Graeber, T. (2023). Cognitive uncertainty. *Quarterly Journal of Economics*, **138**, 2021–2067.

—, — and Oprea, R. (2024a). Complexity and time. *Journal of the European Economic Association*.

—, —, — and Yang, J. (2024b). Behavioral attenuation. *Working Paper*.

— and Shubatt, C. (2024). Quantifying lottery choice complexity. *Working Paper*.

Fehr, E. and Rangel, A. (2011). Neuroeconomic foundations of economic choice – recent advances. *Journal of Economic Perspectives*, **25**, 3–30.

Fudenberg, D. and Puri, I. (2022). Simplicity and probability weighting in choice under risk. *American Economic Review: Papers and Proceedings*, **112**, 421–425.

Gabaix, X. and Graeber, T. (2024). The complexity of economic decisions. *Working Paper*.

Gill, D. and Prowse, V. (2023). Strategic complexity and the value of thinking. *Economic Journal*, **133**, 761–786.

Goncalves, D. (2024). Speed, accuracy, and complexity. *Working Paper*.

—, Nunnari, S. and Zarate-Pina, J. (2024). Revealing complexity. *Extended Abstract*.

Hébert, B. and Woodford, M. (2021). Neighborhood-based information costs. *American Economic Review*, **111**, 3225–3255.

Hossain, T. and Okui, R. (2013). The binarized scoring rule. *Review of Economic Studies*, **80**, 984–1001.

Hua Hu, E. (2023). A procedural model of complexity under risk. *Working Paper*.

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

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, **101**, 2590–2615.

Matějka, F. and McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, **105**, 272–298.

Mononen, L. (2025). On preference for simplicity and probability weighting. *Working Paper*.

Mussolf, R. and Zimmermann, F. (2025). Model uncertainty. *Working Paper*.

Nagel, L. and Saitto, R. (2025). A measure of complexity for strategy-proof mechanisms. *Working Paper*.

Ok, E. (2002). Utility representation of an incomplete preference relation. *Journal of Economic Theory*, **104**, 429–449.

Oprea, R. (2020). What makes a rule complex? *American Economic Review*, **110**, 3913–3951.

— (2024a). Complexity and its measurement. *Handbook of Experimental Methods in the Social Sciences*.

— (2024b). Decisions under risk are decisions under complexity. *American Economic Review*, **114**, 3789–3811.

Puri, I. (2025). Simplicity and risk. *Journal of Finance*, **80**, 1029–1080.

Rubinstein, A. (1986). Finite automata play the repeated prisoner's dilemma. *Journal of Economic Theory*, **39**, 83–96.

Savage, L. (1954). *The foundations of statistics*. Wiley.

— (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, **66**, 783–801.

Schlag, K. and van der Weele, J. (2013). Eliciting probabilities, means, medians, variances and covariances without assuming risk neutrality. *Theoretical Economics Letters*, **3**, 38–42.

Shorrocks, A. (1980). The class of additively decomposable inequality measures. *Econometrica*, **48**, 613–625.

Shubbat, C. and Yang, J. (2024). Tradeoffs and comparison complexity. *Working Paper*.

Sims, C. (2003). Implications of rational inattention. *Journal of Monetary Economics*, **50**, 665–690.

Tsakas, E. (2020). Robust scoring rules. *Theoretical Economics*, **15**, 955–987.

— (2025). Belief identification by proxy. *Review of Economic Studies*.

van der Wel, P. and van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin and Review*, **25**, 2005–2015.

Woodford, M. (2020). Modeling imprecision perception, valuation and choice. *Annual Review of Economics*, **12**.