# Complex for Whom?
# An Experimental Approach to Subjective Complexity

Marina Agranov* Andrew Schotter† Isabel Trevino‡

August 13, 2026

**Abstract**

How people perceive the complexity of a decision problem is a key ingredient in their behavior. Since perceptions are heterogeneous, so is behavior. In this paper, we present a set of tools to elicit subjective perceptions of choice problems. Our objective is to map decision problems into the distribution of subjective perceptions and then map these perceptions to choices to better understand what drives the results established in the literature. We find that aggregate behavior that looks anomalous can be driven by one subset of subjects who engage with the problem in a particular way (have a particular "cognitive signature") and hence is not universal, while aggregate behavior that appears to be consistent with theory can be determined by the composition of the different behaviors of people with different perceptions of the problem's complexity.

## 1 Introduction

The failure of economic agents to behave rationally when faced with a single-person or multiple-person decision problem (game) has often been attributed to the problem's complexity. However, that explanation raises the question of how to measure complexity. We can identify two common approaches in the literature. The first is what we can call "Objective Complexity" in that it theoretically defines the complexity of a problem based on its properties. For example, scholars have defined the notion of the State Space Complexity of a game as the number of legal game positions reachable from the initial position of the

*Caltech and NBER, magranov@caltech.edu.
†New York University, andrew.schotter@nyu.edu.
‡University of California San Diego, itrevino@ucsd.edu.

game (Allis (1994)).[1] An alternative empirical approach is to look at the aggregate behavior of people who have faced a decision problem. For example, Oprea (2020) measures the relative complexity of two different decision rules by recording how much, on average, people would need to be paid to use one rule rather than the other.[2] When a problem has an objective and correct solution, one can simply count the average number of errors made when dealing with these tasks to get a measure based on the accuracy of the aggregate responses, which can correlate with the complexity of the problem. We refer to these types of measurements as "aggregate revealed complexity" because, while empirical, they rely on aggregate results. Both of these approaches attempt to define the complexity of a problem. However, how people behave when they face a problem is a function of how they perceive it, and different people will perceive problems differently. Hence, if we want to understand aggregate behavior, we must understand the underlying heterogeneity, which is determined by the different ways people perceive the same problem.

In this paper, we propose a methodology to study the heterogeneity of subjective perceptions of a problem and its effects on behavior. We present a set of tools that allow us to observe the previously unobserved distribution of subjective assessments or perceptions of a problem. Our approach is necessarily empirical, as our objective is to uncover the heterogeneity of individual assessments related to the complexity of a decision problem and to observe how these assessments affect choices via two mappings. The first mapping (Mapping 1) is from the description of a decision problem to the distribution of subjective perceptions it induces. This illustrates how different tasks and different specifications of a task can be perceived differently by subjects, showing that the distributions of subjective assessments can be task-dependent. The second mapping (Mapping 2), from distributions of cognitive signatures to final choices, studies choices conditional on subjective assessments. This allows us not only to understand differences in behavior, but also to identify whether aggregate results established in the literature are driven by a subset of subjects who perceive the problem in a specific way. In this sense, the approach we follow in this paper is one of revealed *subjective* complexity.[3]

We classify the perception of a subject about a given decision problem through the notion of what we call the subject's "cognitive signature" for that problem. To define this notion we combine the Choice Process Protocol (CPP) of Caplin et al. (2011) with the

[1] Alternatively, one can define the complexity of a game by its game tree size, i.e., the number of leaf nodes in the game tree rooted at the game's initial position, representing the total number of possible games that can be played. In computer science, Kolmogorov complexity in information theory states that the complexity of an object (or problem) is the length of the shortest computer program (in a predetermined programming language) that produces the object (or solution) as output.

[2] Another example is Enke and Graeber (2023) who use aggregate measures of reported confidence in individual choices as a way to measure the complexity of different decision tasks.

[3] In economics we typically focus on behavior and not on subjective perceptions about the decision problem. The revealed preference approach suggests we pay attention to what people choose rather than how they interact with the problem they face. In this paper we claim that perception is a key ingredient in choice.

notion of ex-post confidence in individual choices (Boldt et al., 2019; Enke and Graeber, 2023). The CPP gives us an incentivized measure of the deliberation process that a subject follows when solving a problem, while confidence measures how certain she is about her solution.[4]

On the basis of these two tools, we, as outside observers, classify a person's cognitive signature of a problem based on the way she reaches her final decision and how she views it ex-post. We say a person's approach to a problem is Fast and Confident if the subject reached her final decision in less time than the median decision time of all subjects facing any problem in the same problem class (i.e., when choosing between two lotteries, faster than the median time of all subjects who chose between all lottery pairs presented to them) and, at the same time, reported a higher confidence in her decision than the median confidence of all subjects in this problem class Similarly, an approach is Slow and Confident if the person exhibits a longer consideration time and is confident in her final choice, Slow and Not Confident if the person exhibits a longer consideration time and at the end of the process is still unsure of her decision (lower than the median confidence), and finally, Fast and Not Confident if the person makes a quick decision and expresses low confidence in it. [5] It is the distribution across these four categories that will serve as our main analytical tool.

It is important to point out that the aim of our measure is not to classify the complexity of decision problems, but rather to understand the heterogeneity of subjective assessments of this complexity based on the effort people put into solving a problem and the confidence they have in their final decisions. As outside observers, we then investigate the behavior of people we placed in our four assessment categories in order to get an insight into how aggregate behavior is determined.

Our discussion focuses on three main types of results. First, we characterize the two mappings defined above to understand whether distinct decision problems generate different distributions across our four-way classification of subjective assessments (Mapping 1), and then we track, via Mapping 2, how people behave, conditional on their Mapping 1 assessments. Focusing on these two mappings helps us understand whether the difference in behavior between two specific decision problems is more likely to be determined by differences in the distributions of individual assessments of each problem or on the behavior related to a specific assessment.

[4]The CPP incentivizes subjects to select their preferred choice at every point during the consideration period, thus providing several measures of effort, such as time to first and final choices and number of choice revisions.

[5]There are different interpretations of these categories in terms of what they imply about the complexity of a decision problem. For example, one could interpret fast and confident decisions as coming from a person who perceives the problem as easy, slow and not confident decisions as reflecting that a problem is perceived as hard, or fast and not confident decisions as stemming from people who choose not to actively engage with the problem, whether because they believe the problem is too hard or because they are just not interested in engaging with it.will focus on the descriptive nature of our measures, in terms of contemplation time and reported confidence, and not push for a specific interpretation

Second, by studying canonical decision problems in behavioral and experimental economics, we investigate the relationship between well-known observed aggregate behavior and the heterogeneity that underlies it. In doing this, we come across instances where seemingly anomalous aggregate behavior is driven by the behavior of one of our four assessment classes, while in other instances, where aggregate behavior appears to be consistent with theory, we find that this result is actually due to the aggregation of the divergent behaviors of subjects with different cognitive signatures and hence is not universal.

Lastly, we compare subjective assessments that result from our data to different objective notions of complexity that have been proposed in the literature to see if these notions correlate with our findings.

Our paper makes a number of relevant contributions. First, we provide a systematic way of approaching heterogeneity based on objective features of the choice process and subjective assessments that is widely transferrable across decision tasks and easily implementable. Second, by investigating the behavior associated with our four types of cognitive signatures, we can judge when observed aggregate behavior is universal or driven by a particular type or subset of agents. This has implications for economic theory, since it suggests that anomalous behavior on the aggregate level may not be universal but rather driven by a subset of people with a particular cognitive signature. In this case, theory may actually be better than it is given credit for while in other cases, where its predictions are good, such predictions may be an artifact of a composition effect of the behaviors of different classes of people canceling each other out.

Third, our results are relevant for the design of economic mechanisms, and behavioral mechanism design theory in general, since if we can attribute deviations from predicted or desired behavior to a subset of people, say people whom we designate as Fast and Not Confident, then we can attempt to simplify the mechanism or its description so as to allow such users to better understand its tradeoffs. In addition, our results suggest that it may be important to match the type of mechanism used with the population designated to use it.

Finally, as our title suggests, our paper implies an important qualification to the common practice of defining complexity objectively, since the important question is not only whether a problem is complex but rather for whom it is complex. Focusing on this question may allow us to avoid the idea that some of the anomalous behavior we observe and characterize as universal may be less so.

We will proceed as follows. In Section 2 we review the literature related to our investigation. In Section 3 we dive into our methodological approach and characterize our four categories of cognitive signatures. Section 4 presents our experimental design with a careful discussion of the decision tasks used in the experiment. Section 5 presents our results by first characterizing the distributions of cognitive signatures across decision tasks and then by illustrating the type of insight that our methodology can produce to better understand how cognitive signatures affect behavior. Section 6 concludes.

## 2 Literature Review

There have been several approaches in the literature to try to measure complexity (see Oprea (2024a) for a recent survey). Response time (RT) has been studied as a proxy for effort, with the intuition that people take longer to make a decision in tasks where it is harder to find the optimal choice. Some examples of papers that use RT in this way include Wilcox (1993) for lottery choices, Rubinstein (2007) who uses RT to differentiate between cognitive and intuitive choices, or Gill and Prowse (2023) in strategic interactions (see Spiliopoulos and Ortmann (2018) for a survey on the use of RT in economics). Goncalves (2024) challenges the idea that this type of effort measure can capture complexity by showing theoretically a non-monotonic relationship between stopping time and problem complexity in a Wald optimal-stopping model. This non-monotonicity reflects the fact that once tasks become too complex, consideration times might become low again if people choose not to engage with the task. We find evidence supporting this prediction of Goncalves (2024), suggesting that effort measures alone cannot distinguish different perceptions of complexity.

Other approaches to elicit perceptions of complexity include Oprea (2020) who measures the subjective cost of making a decision by asking subjects their willingness to pay to avoid it, using rules that are algorithmically complex vs simple rules. Gabaix and Graeber (2024) ask people directly to rank the complexity of a series of tasks that range from lottery choices to intertemporal consumption and forecasting. Unlike these papers, we do not ask subjects any statements that are explicitly related to the difficulty of the task.

A series of papers have studied complexity of lottery choices. Armantier and Treich (2016) show that people's valuations of similar lotteries depend on the complexity with which the events are presented. They find similarities in attitudes towards their complex bets and compound and ambiguous lotteries. Puri (2024) presents an axiomatic characterization and evidence for preferences for simplicity, based on the size of the lottery support, in choice under risk. Enke and Shubatt (2023) develop an index of objective complexity for lottery choices based on an algorithm that predicts the error rate when looking for the lottery with the highest expected value in the choice set, which is mainly based on the dissimilarity between lotteries. They find that this measure explains choice errors and is predictive of attenuation in a large data set.

The role of confidence in decision making has been studied extensively in psychology and neuroscience (see Grimaldi et al. (2015) for a survey of human and animal studies in psychology, De Martino et al. (2012); da Silva Castanheira et al. (2021); Boldt et al. (2019); Rollwage et al. (2020) for studies that use confidence measures in value-based choices and Luttrell et al. (2013) for the neuroscientific approach to metacognitive confidence). In economics, Enke and Graeber (2023) ask people to give an estimate of the optimality of their choice and propose cognitive uncertainty (the inverse of confidence) as a measure of complexity and show that this measure correlates with behavioral anomalies such as biases in belief formation and the probability weighting function in risky choices (Tversky and

Kahneman, 1992). Other papers that have used confidence as a measure of complexity in economics include Enke et al. (2025) and Hu (2024). Retrospective confidence in individual choices, unlike effort measures, involves introspection once decisions have been made. In this sense, it can be thought of as an index that reflects complexity, rather than a direct measure of it.

Finally, our aim of understanding how subjective perceptions of a task's complexity affect final choices is related to the model of Bordalo et al. (2025) where decision makers construct heterogeneous representations of a task, based on its salient features, before making a decision. Although our specific objectives differ, we share the foundational principle that decisions are determined by heterogeneous perceptions of the task. In the case of Bordalo et al. (2025), individual representations of a task are shaped by the features that are salient to each decision maker.[6] This distribution of representations, via salience, leads to a distribution of final choices across people.

# 3 Methodology

Before presenting our experimental design, we pause to introduce the method we use to elicit the subjective assessments (cognitive signatures) of our participants. The method itself relies on two established tools that have each been employed separately in the literature. The novelty in our analysis is the combination of these tools, which enables us to obtain a richer understanding of how participants approach decision tasks and revise their choices.

The first component of our measure is incentivized response time, measured using the choice process protocol (CPP) developed by Caplin et al. (2011) and applied in Agranov et al. (2015) and Kessler et al. (2023). In the CPP, each round begins with task instructions, after which participants see a screen with a series of buttons corresponding to each possible choice. Participants select one button at a time and can switch freely throughout the known duration of each round. Incentives in the CPP are tied to the choice selected at a *randomly determined second* within a round, which is drawn by the computer after the round. At the start of the experiment, if no button is selected, the payoff to the subject is zero, creating an incenitve to make a quick first choice.

The CPP encourages participants to select the option they think is the best decision at every point in time because any interim choice can determine final payoffs. The CPP generates rich data that can shed light on how people reach their final choices in a task by providing the whole dynamic path of their deliberation process. Rather than utilizing all this richness, in our main analysis we focus on final choices and their *incentivized response time*, i.e., time to last click. This serves as a proxy for subjects' contemplation time. Without the CPP protocol, standard response time can still proxy for contemplation time, though less reliably, since participants lack incentives to record their final choice

[6] See Bordalo et al. (2022) for an overview of salience in decision making.

immediately once they feel they have reached it. We show in Appendix B.1 that the CPP does not alter final choices by studying behavior in additional sessions in which payoffs were based on final rather than random-second choices. Comparing these with CPP sessions, we find no significant differences in the distribution of final choices across any of the tasks (see Figure 9).[7]

The second tool that comprises our proposed measure reflects how confident participants are in their final choices. At the end of each round, participants report on a 0–100 scale how certain they are that the final option they selected was the best option for them. This measure parallels the cognitive uncertainty measure of Enke and Graeber (2023) and resembles confidence assessments widely used in psychology (e.g., De Martino et al. (2012)).[8]

Taken together, the two tools—contemplation time from the CPP and ex-post confidence—allow us to classify how participants approach tasks and evaluate their decisions. Specifically, they capture heterogeneity both in the choice process (how quickly decisions are reached) and in a retrospective self-assessment (how participants reflect on their performance after making a choice). These two complementary measures allow us to better understand how a subject interacts with a decision problem and define the subject's cognitive signature with that specific problem. We are interested in classifying people by their cognitive signatures and understanding how different signatures relate to behavior.

As a first step in this direction, we propose a simple and portable classification, based on whether the contemplation time and confidence of a choice are above or below the corresponding median normalized values.[9] Specifically, a decision is classified as *slow* (*fast*) if its normalized contemplation time exceeds (falls below) the median normalized contemplation time, and as *confident* (*not confident*) if its normalized confidence exceeds (falls below) the median normalized confidence.

The normalization can be implemented in several ways. Under *population-level* normalization, contemplation time (confidence) values are normalized relative to the median values observed across all subjects in the population facing the same family of tasks. Alternatively, under *individual-level* normalization values are normalized relative to each subject's individual observations, computed across the set of comparable tasks or even all tasks. The appropriate approach depends on the research question and the available data; population-level normalization is particularly useful when each subject completes a single task of a given type, as it enables comparisons across subjects facing the same decision problem. When subjects complete multiple rounds of similar tasks, for example several rounds of

[7] We have also conducted robustness checks of our main results using active consideration time, which is the difference between the time to last click and the first click, and find similar results. This analysis is available from the authors upon request.

[8] See Bernheim et al (2026) for a discussion of how to interpret what this question measures and its use in different experimental tasks.

[9] Using the median rather than the mean is a robust choice because response-time distributions are typically right-skewed.

binary lottery choices, either normalization can be used: normalizing at the population level (pooling the data from all subjects in all decision tasks) emphasizes differences across individuals, whereas normalizing at the individual level (pooling the data of all decision tasks for each subject separately) focuses on within-subject variation in different versions of a decision task, while accounting for persistent differences in decision-making styles.

Throughout this paper, we use the population-level normalization among similar groups of tasks to emphasize the inter-person comparisons. Appendix B discusses in detail the advantages and disadvantages of using different types of normalizations and shows that our main findings are robust to the alternative normalization scheme where we use individual-level medians for all tasks completed by each subject in the experiment. Our objective is not to advocate for a particular normalization method, but rather to demonstrate the flexibility and portability of the proposed classification across a wide range of experimental settings and to illustrate the types of research questions it can be used to address.

With this in mind, we classify the cognitive signature of a decision maker in a task into four categories, according to their decision time and confidence, as follows:

1. **FAST and CONFIDENT** decisions: below-median contemplation time and above-median confidence.

2. **SLOW and CONFIDENT** decisions: above-median contemplation time and above-median confidence.

3. **SLOW and NOT CONFIDENT** decisions: above-median contemplation time and below-median confidence.

4. **FAST and NOT CONFIDENT** decisions: below-median contemplation time and below-median confidence.

This classification will be the workhorse for our analysis. We are interested in understanding how these four subjective assessments are distributed between the different decision problems encountered by our subjects. These distributions will characterize the heterogeneity of cognitive signatures on the same decision problem in an organized way. Ultimately, our goal is to understand how different problems lead to different distributions over cognitive signatures and how these distributions shape observed choices. Although conventional work in the past has typically ascribed choice anomalies and observed biases to preferences and perceptions of probabilities, we attribute them, at least in part, to the heterogeneity in people's subjective assessments of the tasks they face.

Our categorization can be used to shed light on how the decision process might influence subjective perceptions of a task's complexity. We can think of this as a two-stage process. First, the decision maker observes the task and decides whether to spend time actively thinking about it or not. Quick decisions can arise either because the task is viewed as easy, and hence requires little thought, or the task appears difficult or uninteresting, and

the decision maker opts to not engage with it and devotes little time thinking about it. For other tasks, the decision maker thinks the problem is tractable but not obvious and decides to spend more time to find the best decision. At some point in this process, the decision maker stops either because she arrives at a satisfactory decision or because she realizes that it is not worth spending more time trying to find the best decision because it is a hard problem. After the final decision is submitted, the decision maker engages in a retrospective evaluation of whether she arrived at the best choice for her by expressing her ex post confidence in that choice.

This particular interpretation of our categories suggests that our classification can help understand subjective perceptions of task complexity by incorporating elements of ex ante judgment (whether to engage thoughtfully with the problem or not), the endogenous evaluation of when to stop thinking about the problem, and the ex post evaluation of final choices.

Regardless of the interpretation, our two-by-two classification is rich yet manageable, providing a new tool to understand the heterogeneity of cognitive signatures in a task and their relationship with final choices in a variety of decision problems.

### 3.1 Discussion of our methodology

It is important to establish why we need both contemplation time and confidence to assess decisions rather than just one of them. Intuitively, contemplation time alone might not distinguish between tasks that are very straight-forward with intuitive responses, and tasks where people give up on finding their preferred choice. Goncalves (2024) makes this point theoretically using the drift-diffusion model to illustrate that response time cannot be used to measure the complexity of a problem because quick choices may be due to decision makers who find the problem easy and solve it correctly very fast, or so hard that they do not even attempt to solve it accurately. Similarly for confidence, it is hard to compare the confidence statement of a subject who thinks long and hard about a problem and discovers a solution that she is confident (not confident) with, to a subject who does not actively engage with the problem and comes to the same conclusion.[10]. Such people lack the meta-cognition necessary to evaluate their situation.

To illustrate the limitations associated with inferring a subjective assessment related to complexity using only contemplation time or confidence, we focus on five tasks in our experiment that have an objectively correct answer: a binary lottery choice with one lottery first-order stochastically dominating the other (FOSD henceforth), two different binary choices between two deterministic analogs of lotteries that we call mirrors (similar to Oprea (2024b)), and two tasks where subjects face situations of pivotality, one that requires participants to engage in contingent reasoning and one that does not (see Esponda

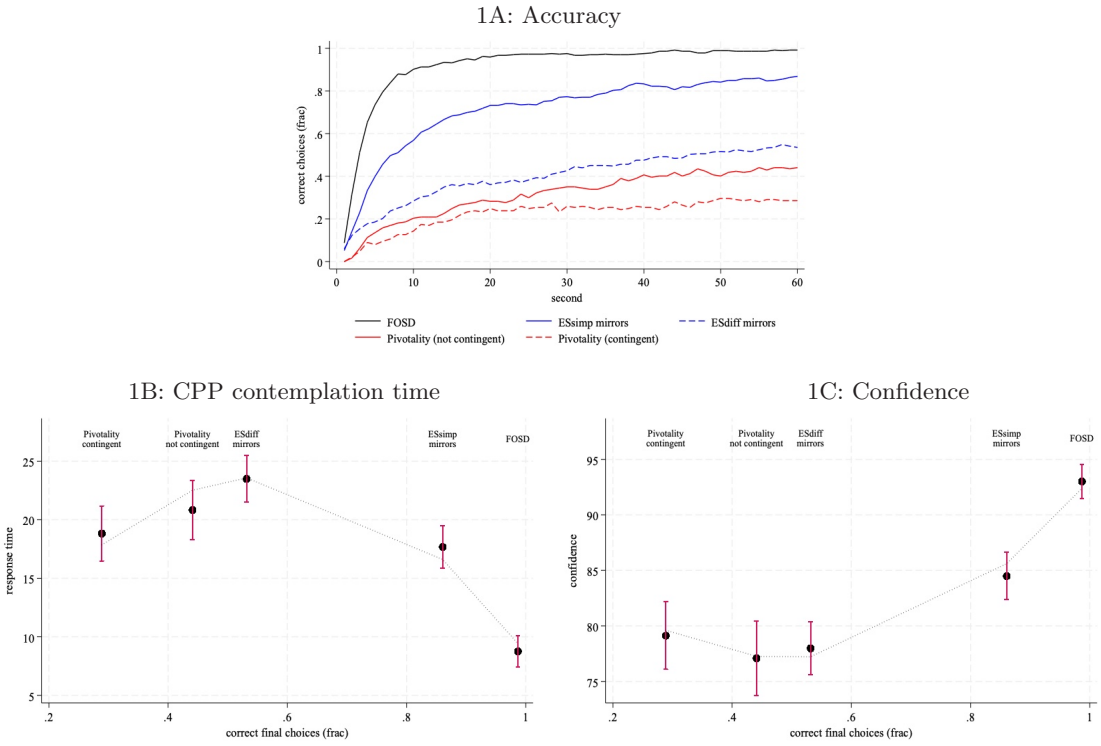
[10]Experimental research has demonstrated that, in some tasks, people are adept at accurately judging their own performance, while in others, they succumb to biases and behave non-optimally without realizing it, see (Grimaldi et al., 2015)

and Vespa (2014)). See Section 3 for a detailed description of these tasks.

For the purposes of argument here, for each decision problem we look at the proportion of subjects who arrive at the correct final choice at the end of a round as a proxy for the objective difficulty of the problem .

Panel 1A of Figure 1 presents the fraction of correct choices in a given problem as a function of time spent on it for tasks with objectively correct answers. As we can see, we find a natural ordering of our five problems according to this objective measure, with FOSD having the highest accuracy (lowest difficulty), followed by the two mirrors tasks, followed by pivotality (non-contingent), followed by pivotality (contingent), and this ordering holds across the whole deliberation period. We use this ordering of task difficulty to assess whether contemplation time or confidence alone can track choice accuracy. Panels 1B and 1C illustrate, respectively, average response time (across people) and average reported confidence, for each of these five tasks, ordered by aggregate choice accuracy (following Panel 1A), from the lowest (pivotality that requires contingent reasoning) to the highest (FOSD lottery choice).

Figure 1: Tasks with Correct Answers in Binary treatments




Consistent with Goncalves (2024), Panel 1B of Figure 1 illustrate a non-monotonic

relationship between contemplation time and task difficulty, measured by choice accuracy, suggesting that fast decisions can be due either to people finding the task easy and solving it quickly and accurately, or to them not engaging with the task by exerting effort and making a quick, wrong choice. Similarly, Panel 1C portrays a non-monotonic relation between confidence and task difficulty, showing that subjects might report similar ex post confidence for tasks where the accuracy of their choices is very dissimilar, suggesting that people might not realize when their choices are wrong. This suggests that neither contemplation time nor confidence alone provide a good assessment of the subjective assessment or cognitive signature of a task.[11]

# 4 Experimental Design

We designed our experiment to capture the thinking process of subjects (using the CPP), their final decisions, and their confidence in a series of standard tasks and games, many of which are associated with a vast literature. We use well-known tasks with the objective of studying how subjective assessments shape established results in the literature. We also study tasks that have gained attention in recent years to relate how their a priori notions of complexity and findings correlate with ours.

**Subject Pool.** We conducted our experiment in Prolific with a total of 976 participants between the ages of 18 and 65, living in the United States, fluent in English, and with a high approval rating in Prolific. For each treatment, an equal number of men and women were recruited. The experiments were carried out in March - May 2024.

**Implementation.** The experiment was programmed in oTree Chen et al. (2016).[12] We used recorded video instructions to explain the structure of the experiment and the task rules. The video instructions were accompanied by slides with written instructions to accommodate differences in preferred learning styles. Participants had to complete several comprehension quizzes to demonstrate their understanding of how objects are presented on the screen and the CPP methodology. The complete instructions and screenshots can be found in the Online Appendix.

**Treatments.** We ran four treatments that differ in two dimensions. The first dimension is the size of the choice set, that is, the number of options that subjects can choose from in a single task. For simplicity, we refer to these as either **Binary** treatments (2 options, e.g., the choice between two lotteries), and **Non-Binary** treatments (101 options, e.g., the bid

[11] See Appendix D.1 for additional tests of the predictions of Goncalves (2024) regarding contemplation time of subjects with different cognitive abilities.

[12] The experiment was approved by Caltech (IR23-1365) and pre-registered on aspredicted.org (AsPredicted #112194).

in an auction where the bids are integer numbers between 0 and 100 inclusive). Subjects had 60 seconds to respond to each task in the binary treatments and 90 seconds in the Non-Binary treatment.[13] We ran two versions of the binary and non-binary treatments that differed in the specific tasks faced by subjects. We first explain the different tasks in our experiment and then list which of these correspond to each treatment variation.

- **Binary Lottery Choices.** In all Binary treatments, each participant completed 6 rounds where subjects had to choose between different sets of two lotteries, detailed in Table 1. Each participant faced the following binary lottery choices: a choice involving first-order stochastic dominance between lotteries L5 and L6 (FOSD hereafter), a choice involving a mean-preserving spread between lotteries L5 and L7 (MPS hereafter), what we call an Enke-Shubatt simple lottery choice between lotteries L1 and L2 (ESsimp lottery task), an Enke-Shubatt difficult choice between lotteries L3 and L4 (ESdiff lottery task hereafter), and either two questions eliciting the Common Ratio effect, L8 vs L9 and L10 vs L11 (CR1 and CR2 hereafter), or two questions eliciting the Common Consequence effect of the Allais Paradox, L8 vs L12 and L10 vs L11, (CC1 and CC2 hereafter).

Table 1: Lottery Rounds in Binary Treatments

| | | | |
|---|---|---|---|
| FOSD | L5: ($15,$7; 0.50,0.50) | vs | L6: ($3,$1; 0.50,0.50) |
| MPS | L5: ($15,$7; 0.50,0.50) | vs | L7: ($21,$1; 0.50,0.50) |
| ESsimp lottery task | L1: ($20,$10; 0.50,0.50) | vs | L2: ($12,$11; 0.20,0.80) |
| ESdiff lottery task | L3: ($25,$2; 0.60,0.40) | vs | L4: ($30,$7; 0.25,0.75) |
| CR1 | L8: ($12; 1.00) | vs | L9: ($30,$0; 0.50,0.50) |
| CR2 | L10: ($12,$0; 0.20,0.80) | vs | L11: ($30,$0; 0.10,0.90) |
| CC1 | L8: ($12; 1.00) | vs | L12: ($30,$12,$0; 0.10,0.80,0.10) |
| CC2 | L10: ($12,$0; 0.20,0.80) | vs | L11: ($30,$0; 0.10,0.90) |

Notes: We depict lotteries using the following notation: the lottery ($x,$y,$z; m,n,p) implies prize $x with probability m, $y with probability n, and $z with probability p.

The FOSD choice is simple and is included, in part, to make sure that the instructions and presentation of the lotteries were clear to subjects. The other lottery choices are less straightforward; for example, in the mean-preserving spread question, the choice depends on risk attitudes: risk-averse participants are expected to choose L5 over L7.

The ESsimp and ESdiff lottery tasks were selected based on the index developed by Enke and Shubatt (2023), which relates the complexity of a choice between lotteries

[13] These durations were calibrated based on preliminary pilots to balance two forces. On the one hand, rounds should be long enough so that participants do not experience time pressure and can finish their thinking process naturally. On the other hand, if each task lasts too long, this unnecessarily prolongs the experiment, increases its costs, and runs the risk of participants losing their attention span.

to the excess dissimilarity of the cumulative distribution functions of the probabilities in the choice set. According to this measure, if the probabilities in lottery pair A have an excess dissimilarity larger than that of the lotteries in lottery pair B, then the choice in lottery pair A is more difficult than in lottery pair B. We calculate this index for our lottery choices and refer to the choice between L1 and L2 as Enke-Shubatt simple (ESsimp) and to the choice between L3 and L4 as Enke-Shubatt difficult (ESdiff).[14]

Finally, each participant in the binary treatments completed two questions that elicited the common ratio or common consequence effects. We refer to these questions as CR1 and CR2 for common ratio questions and CC1 and CC2 for common consequence questions. The common ratio and common consequence effects were originally proposed by Allais (1953) and are known in decision theory as primary deviations from expected utility.[15] The common ratio effect is the empirical observation that when people choose between a smaller, more probable amount and a larger, less probable amount, reducing the probabilities by a constant factor makes them more likely to opt for the riskier choice. The common consequence effect suggests that people's preferences between two lotteries change when a common consequence is added to both options[16]

- **Binary Choices of Deterministic Mirrors.** In addition to the lottery choices we just described, we included a set of questions that involve similar binary choices but are not lotteries. The deterministic mirror of a lottery, introduced by Oprea (2024b), has an objective value corresponding to the expected value of a lottery but, unlike lotteries, involves no risk at all. As an example of a mirror, imagine a set of 100 boxes, 50 of which contain \$20 each and the other 50 contain \$10 each, and the payoff to a person that chooses this object is the average amount of the money contained in all the boxes. This collection of boxes has an objective value of \$15, which is the same as the expected value of lottery L1 depicted in Table 1, but has no uncertainty. We call this object the deterministic mirror of L1 and denote it by M1. In our binary treatments, each participant had to choose between two different collections of boxes, which we call mirrors. We chose parameters for the mirrors that mimic the prize structure and relative frequencies of L1 vs L2 (ESsimp), which we refer to as M1 vs M2 (ESsimp mirror) and of L3 vs L4 (ESdiff) which we refer to as M3 vs M4 (ESdiff). This allows us to analyze subjective assessment of tasks

[14] Using the calculator tool provided by Enke and Shubatt (2023), we calculate that the objective problem complexity of L1 vs L2 is 0.17, while it is 0.27 for L3 vs L4.

[15] These effects have been extensively documented in experiments (Blavatskyy et al., 2023) and serve as the basis for new theories (Gul, 1991; Cerreia-Vioglio et al., 2015; Loomes and Sugden, 1982; Bordalo et al., 2012).

[16] The parameters chosen for these questions follow McGranaghan et al. (2024a), who explore the connection between the two effects.

across different parametrizations of mirrors and, for a given parametrization, across a lottery and its deterministic mirror.

- **Contingent Reasoning in Pivotality.** Anticipating the consequences of one's actions is an integral part of the analysis of decision making in economics. There is a growing experimental literature documenting the difficulty in performing contingent reasoning and how this tendency translates into mistakes in individual and strategic settings (Esponda and Vespa, 2014, 2023; Dal Bo et al., 2018; Martinez-Marquina et al., 2019; Ali et al., 2021; Ngangoue and Weizsacker, 2021). In our binary treatments, we elicit the ability to think contingently using the simplified design of Esponda and Vespa (2014). Each participant is paired with two computers and the three of them vote for either a red ball or a blue blue. The majority vote determines the group's decision. Initially, the state is drawn from a known prior distribution (70% red and 30% blue in our experiment). Computers are programmed to vote as follows: if the state is red, they always vote red; if the state is blue, they vote blue with probability 50% and red otherwise. We implemented two versions of this task, one that requires contingent reasoning (where we do not tell the subjects what the computers have chosen) and one that does not (where we tell the subjects what the computers have chosen). Since the group's decision is determined by the majority of votes, contingent reasoning reveals that the only situation in which one's vote is pivotal is when the two computers cast different votes. This happens only if the state is blue. Therefore, in the presence of contingent reasoning, pivotality logic prescribes always voting blue. In the version that does not require contingent reasoning, we simply asked subjects to choose how they would vote if one computer voted blue and another voted red.[17]

- **Certainty Equivalents of Lotteries.** As part of the non-binary treatments, we elicited certainty equivalents of three lotteries: L1 and L3, presented in Table 1, as well as lottery L13, which pays \$22, \$15, and \$5 with probabilities 40%, 40%, and 20%, respectively. We chose these lotteries to facilitate comparisons with the index of complexity for individual lotteries developed by Enke and Shubatt (2023) and Puri (2024).[18] According to the complexity index of Enke and Shubatt (2023), lotteries L3 and L13 share similar levels of complexity, while both are more complex than lottery L1.[19] On the other hand, Puri (2024) suggests that decision makers may

[17] If this task was selected for payment, then the computers' votes were simulated using the same rule as in the contingent version of the game and if both computers voted the same color, this color was recorded as the group's decision.

[18] Enke and Shubatt (2023) develop both a complexity measure of binary lottery choices and a complexity measure of individual lotteries. To distinguish between these two, we use the ESsimp lottery label to indicate that a *binary lottery choice* is simple and use the ESsimp lottery label to indicate that the *valuation of a lottery* is simple.

[19] According to Enke and Shubatt (2023), the complexity index of L1 is 2.57, while L3 has 4.168 and L13 has 4.176.

perceive lotteries that contain more outcomes as more complex, which may lead to differences in the valuations of L3 and L13. Certainty equivalents were elicited using a standard multiple price list (MPL) in which participants specified the switch point from choosing the lottery to choosing the monetary amounts presented in ascending order. The amounts ranged from \$0 to \$25 in increments of 25 cents.

- **Simplicity Equivalents of Deterministic Mirrors.** We also elicited the simplicity equivalents of the three mirrors that mimic the prize structure and relative frequencies of lotteries L1, L3, and L13, but do not involve risk. We call these mirrors M1, M3, and M13, respectively. Eliciting the simplicity equivalent of these mirrors is the analogue of eliciting certainty equivalents for lotteries. Following (Oprea, 2024b), the objective value of a mirror simply requires that subjects multiply the prizes and frequencies and aggregate these into a single value, similar to computing the expected value of the corresponding lottery.

- **Belief-Updating Tasks.** We asked participants to complete six belief updating tasks using the standard binary state, binary signal neutral paradigm extensively studied in the literature (Benjamin, 2019; Enke and Graeber, 2023; Esponda et al., 2024; Augenblick et al., 2025; Ba et al., 2023; Agranov and Reshidi, 2026). In each task there are 100 hypothetical projects, $p$ of which are Successes and the remaining $100 - p$ are Failures. The computer randomly selects one of these projects and runs a test on the selected project. The test accuracy is $q$, that is, if the project is a Success (Failure), the test result is positive (negative) with probability $q$ and negative (positive) with probability $1 - q$. Participants observe the prior $p$, the accuracy of the test $q$, and the realization of a signal, and are asked to state their posterior belief that the selected project is a success. Table 2 contains the exact parameters we use. Participants were incentivized to reveal their posteriors using a standard incentive-compatible BDM mechanism[20] One set of parameters ($p = 15$ and $q = 0.80$) corresponds to the classical parameterization of Kahneman and Tversky (1972), which became the standard for studying base rate neglect (Benjamin, 2019; Esponda et al., 2024; Gneezy et al., 2023).

- **Auctions.** We consider four formats of independent private value auctions with two bidders looking to buy a single unit of an object. Private values for the object are drawn uniformly and independently between zero and a hundred experimental points. The four formats are First-Price, Dutch, Second-Price, and English auction. In the First-Price (Second-Price) auction, we tell participants that the winner of the auction is the highest bid among the two and pays the price equal to her bid (the second

[20]The BDM is theoretically an incentive-compatible method for eliciting truthful responses regardless of participants' risk attitudes (Becker et al. (1964)). To help participants understand this method, we stated that they had no incentive to report beliefs falsely if they wanted to win the \$10 prize if one of these rounds was selected for the bonus payment (see Danz et al. (2021) for belief elicitation methods).

highest bid). For the Dutch auction, we tell subjects that the auction starts at the highest price of 100 and gradually decreases as the auction proceeds. Subjects are asked to submit their bid, which represents the 'freezing' price. The highest freezing price wins the objects and pays her bid. For the English auction, we tell subjects that the price starts at the lowest value of 0 and increases as the auction proceeds. Subjects submit the value of the bid at which they want to drop out of the auction. The person who submits the highest dropout price wins the auction and pays the second-highest dropout price. Theoretically, the first price and the Dutch auctions are strategically equivalent, and so are the second price and the English auction.

- **Public Good Games.** We implemented two versions of the standard linear public good game in our Non-Binary treatments (Ledyard, 1995). Participants play in groups of five, each participant is endowed with 100 points that they can choose to keep or to allocate to a group project. The total number of points allocated to the group project is multiplied by a known constant $\alpha$ and returned in equal shares to all group members, regardless of their contribution. That is, participants get a return of 1-to-1 from the points kept for private consumption and a return of $\alpha$-to-1 from any point contributed to a public project, where $\alpha < 1$ is referred to as the marginal per capita return (MPCR). We implemented two versions of this game, one with a high MPCR of 0.75, and one with a low MPCR of 0.25.

**Summary of Treatments.** Table 2 displays all tasks and games performed in each treatment, which we call Binary 1, Binary 2, Non-Binary 1 and Non-Binary 2. Subjects participated in one of our four treatments. The order of blocks and the order of rounds within a block were randomized at the subject level. The only exception is Blocks A and B in the Binary 1 and Non-Binary 1 treatments, which appeared next to each other (in a random order across subjects) since they share part of the instructions. For the Non-Binary treatments, only one auction format was implemented for each subject: in Non-Binary 1, it was either a First-Price or a Dutch auction (randomly selected); in Non-Binary 2, it was either a Second-Price or an English auction (randomly selected). Finally, to reduce fatigue, at the end of each block (or every three rounds in blocks with more than 3 rounds), we presented participants with an unincentivized visual puzzle in which they were asked to find a hidden animal in a nature picture. We present an example of this 'brain break' in the Online Appendix.[21]

**Payments.** All participants received a participation fee upon completion: \$5 in binary treatments and \$7 in non-binary treatments. In addition, each participant had a 20% chance to be selected into a bonus group where a randomly selected round would provide additional payment. Following to the CPP, the choice selected in the chosen round at a

[21] This technique was first used in McGranaghan et al. (2024b).

Table 2: Experimental Design

| | BINARY 1 | BINARY 2 | NON-BINARY 1 | NON-BINARY 2 |
|---|---|---|---|---|
| | **Lottery Choices** | | **Certainty Eq of Lotteries** | **Belief updating** |
| | L5 vs L6 | L5 vs L6 | L1 | (15%, 70%) |
| | L5 vs L7 | L5 vs L7 | L3 | (15%, 80%) |
| Block A | L1 vs L2 | L1 vs L2 | L13 | (30%, 75%) |
| | L3 vs L4 | L3 vs L4 | | (80%, 65%) |
| | L8 vs L9 | L8 vs L12 | | (80%, 85%) |
| | L10 vs L11 | L10 vs L11 | | (90%, 75%) |
| | **Mirror Choices** | | **Simplicity Eq of Mirrors** | |
| Block B | M1 vs M2 | M1 vs M2 | M1 | |
| | M3 vs M4 | M3 vs M4 | M3 | |
| | | | M13 | |
| | **Contingent reasoning** | | **Public Good game** | |
| Block C | Pivotality task contingent | Pivotality task not contingent | high MPCR | low MPCR |
| | | | **Auctions** | |
| Block D | | | First-Price or Dutch | Second-Price or English |
| nb of subjects | 194 | 186 | 295 | 301 |

Notes: The exact parameters of lotteries are described in Table 1. For the belief-updating tasks, the pair $(x, y)$ represents the parameters where $x$ depicts the prior and $y$ depicts the precision of a binary signal.

randomly selected second determined the bonus. Binary treatments lasted approximately 30 minutes and participants earned, on average, $7. The Non Binary treatments lasted approximately 40 minutes and participants earned, on average, $8.5.

# 5 Results

In this section we present a collection of facts in order to illustrate the usefulness and portability of our method across decision domains. The data derived from our experiment is very vast and rich, so these results are by no means exhaustive, instead, our purpose is to offer the reader a glimpse into the possibilities and types of insights that our method conveys. We first present the distributions of cognitive signatures that arise for all the different decision tasks in our experiment. We refer to this as Mapping 1: how different tasks give rise to different distributions of subjective assessments. We then discuss Mapping 2, which characterizes behavior, conditional on a cognitive signature. To do this, we present a series of examples where particular cognitive signatures are responsible for driving aggregate behavior that is typically documented in the literature. This illustrates the importance of the heterogeneity analysis that is possible with our measure, since aggregate results may be misleading by either suggesting that behavior is anomalous when, in fact,

only a subset of subjects are responsible for the observed anomalies, or by suggesting that behavior is in line with economic theory when, in fact, there could be a composition effect where different types deviate from the theory but these deviations might offset each other when decisions are aggregated across types. Throughout our discussion of Mappings 1 and 2 we also compare our classification of cognitive signatures to existing measures of objective task complexity proposed in the literature.

### 5.1 Distributions of Cognitive Signatures across Tasks

Figure 2 presents the distribution of cognitive signatures (Fast and Confident, Slow and Confident, Slow and Not Confident, and Fast and Not Confident) across the decision problems in our experiment. The top panel reports results for binary choice tasks, while the bottom panel reports results for tasks with non-binary choice sets.

As discussed in Section 3, we normalize contemplation time and confidence using population medians computed within groups of comparable tasks. That is, in the Binary treatment all lottery and mirror binary choice tasks are normalized jointly. In the Non-Binary treatments all lottery and mirror valuation tasks are normalized jointly. We also normalize all belief-updating tasks jointly. Finally, we normalize the two pivotality tasks jointly, the four auction formats jointly, and the two public-good parameterizations jointly.[22]

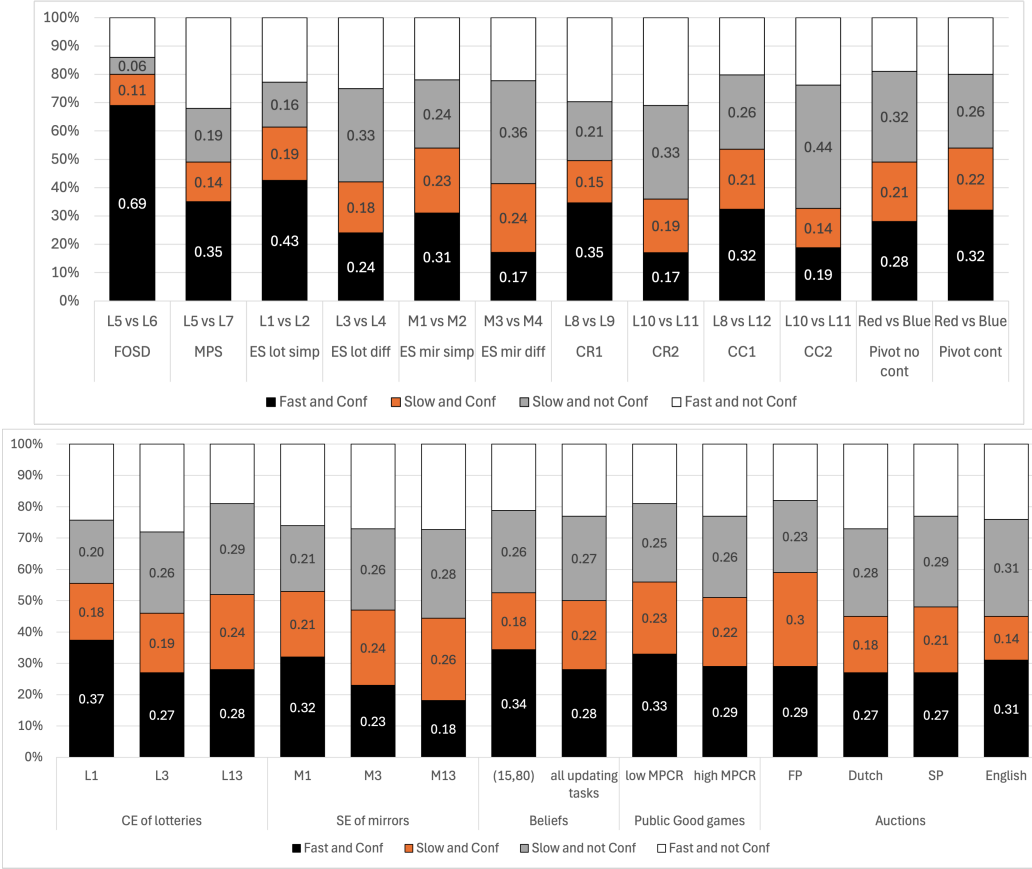
Something to notice in Figure Figure 2 is that the proportion of Fast and Not Confident types is relatively stable across tasks; it is about 20-30 % in both binary and non-binary treatments. Fast decisions that are accompanied by low ex-post confidence in performance can arise for many reasons, such as subjects finding the task too hard to even attempt it or simply choosing not to engage with it. With our classification we can shed light on the proportion of subjects who, for one reason or another, choose not to actively engage with the decision task, which is identified as people who make a quick choice and anticipate their poor performance. Experimental economists have been aware of the possibility that a fraction of subjects might behave this way and our results show that this proportion of subjects is relatively stable across different types of decision tasks.[23]

Looking at the top panel, we see a great heterogeneity in the assessment of binary lottery choice problems, which is indicative of how different specifications of the same type of task can give rise to different subjective assessments about them (Mapping 1). As expected, presenting subjects with a choice between lottery L5 and lottery L6, where L5 first order stochastically dominates L6, was the comparison that led to the highest proportion of fast and confident choices. Using our cognitive signature categorization as a metric for how subjects subjectively perceive a problem, we could say that this decision task was seen

[22] Appendix B.1 presents the distribution of cognitive signatures with an alternative normalization for confidence and consideration time to show that the main patterns discussed in this section remain intact. This alternative normalization uses individual medians across all tasks completed by a subject in a session instead of population medians for comparable tasks.

[23] See Bernheim et al (2026) for a technology in identifying improvable choices, i.e., choices made by subjects that have not internalized the main features of the environment they are facing.

Figure 2: Individual Assessments of Tasks and Games




Notes: Assessments are defined relative to the population medians for the following groups: all lotteries and mirrors in the Binary treatment, all lotteries and mirrors in the Non-Binary treatment, all belief updating tasks in the Non-Binary treatment, both versions of pivotality tasks, four versions of auctions, and two versions of public good games. The top panel corresponds to the Binary treatment, the bottom one to the Non-Binary treatment.

as the simplest when compared to all binary and non-binary problems. This comparison serves as a "first pass" for our procedures, given its simplicity. As we can see, different specifications of lotteries and mirror binary tasks are more likely to give rise to a higher proportion of fast and confident choices, which could be interpreted as easier choices, while others give rise to a larger proportion of slow and non-confident choices, which could be interpreted as harder decision tasks.

Enke and Shubatt (2023) propose an index of objective complexity based on the excess

dissimilarity of the cumulative distribution functions of the lotteries in the choice set. L1 vs L2 (ES lot simp) refers to a "simple" choice and L3 vs L4 (ES lot diff) refers to a "difficult" binary choice, according to this index. The distributions of cognitive signatures for these problems are consistent with this interpretation, since fast and confident decisions comprise 43% of all decisions in the simple problem, compared to 24% for the difficult problem ($p < 0.01$), while 16% of decisions in the simple problem were slow and not confident, compared to 33% in the difficult problem ($p < 0.01$). Using the same approach to compare the subjective assessments of choices between deterministic mirrors, we observe a similar pattern (31% vs 17% of fast and confident choices in M1 vs M2, $p < 0.01$, and 24% vs 36% of slow and not confident choices in M3 vs M4, $p < 0.01$). Notice, however, that there is a larger proportion of slow choices in mirrors than in their associated lotteries, which could be interpreted as people being more likely to make intuitive choices in lotteries than in the corresponding mirrors.[24] Furthermore, there is a systematic difference in individual assessments between the first and second questions in both the Common Ratio and Common Consequences problems. The first question elicits a significantly higher fraction of fast and confident decisions than the second (35% vs. 17% for CR, $p < 0.01$; 32% vs. 19% for CC, $p < 0.01$). In both variants of the Allais paradox, the first question contrasts a sure outcome with a non-degenerate lottery, whereas the second compares two non-degenerate lotteries.[25]

Comparing the distributions of cognitive signatures for the valuation of individual lotteries and mirrors in the non-binary tasks (lower panel of Figure 2), we see similar, but less stark patterns.[26] Just as in binary choices, our measure is consistent with existing objective measures of complexity for both lotteries and mirrors, i.e., there is a larger fraction of fast and confident choices for L1 than L3 (37% vs. 27%, $p = 0.01$, consistent with Enke and Shubatt (2023)) and L1 than L13 (37% vs. 28%, $p = 0.02$, consistent with Puri (2024)). The same is true for M1 and M3 (32% vs 23%, $p = 0.02$) and M1 and M13 (32% vs 18%, $p < 0.01$).[27] However, perceptions are similar between L3 and L13 ($p > 0.10$ for

[24] This evidence is at odds with the following decision-making process: when comparing two lotteries, economic theory suggests that one has to compare their expected values and in addition account for differences in risk. The first part of the process is the same as in evaluating mirrors, while the second part is unique for lotteries. The described process would suggest that the evaluation of lotteries is more intricate than the evaluation of mirrors because of the additional dimension of risk, which would lead to slower choices, but this is not what we find.

[25] We use a non-parameteric Wilcoxon SignRank Test of matched observations for all pairwise comparisons.

[26] This is understandable since the cognitive exercise of coming up with a valuation, whether it is for a certainty or simplicity equivalent, is essentially the same, i.e., finding a one-dimensional equivalent (money) for a multi-dimensional object (a lottery or mirror). To do this, it is natural to expect the subject to integrate or aggregate values and probabilities (frequencies) and engage in a calculation of either expected utility or expected value. Choosing between two lotteries or mirrors, on the other hand, involves comparing two multidimensional objects.

[27] Recall that the parametrization of L1 and M1 corresponds to what Enke and Shubatt (2023) refer to as a less complex lottery, according to their index of objective complexity for individual lotteries, L3 and

all assessments), consistent with Enke and Shubatt (2023).[28] The same is true for the valuation of mirrors ($p > 0.10$ for all assessments).

For the rest of the non-binary problems, as shown in the lower panel of Figure 2, it is interesting to see that, unlike binary lottery choices, different specifications of a decision task are sometimes assessed similarly, in the sense that the distributions of cognitive signatures across, for example, the four auction formats, are not statistically different. This is also the case for the two specifications of the public good games and for belief updating tasks. As we will see below, the fact that these different specifications give rise to similar subjective assessments about them (Mapping 1) does not mean that the behavior of subjects, conditional on their assessments, is the same (Mapping 2). This suggests that if we detect differences in behavior across these problems, such differences are not a function of how the problems are perceived but rather a function of how people behave, conditional on their perceptions.

Table 3 presents statistical significance for pairwise differences between the distributions of assessments in Figure 2.

## 5.2 Cognitive Signatures and Behavior

We now shift our attention to the behavior of subjects, conditional on their cognitive signature, for each decision problem they faced (Mapping 2). This heterogeneity analysis illustrates the importance of understanding the subjective assessments of tasks in order to better interpret aggregate results.

As mentioned above, the analysis that follows is by no means exhaustive, instead, our objective is to illustrate the types of insights that our methodology can produce that go beyond the typical results presented in the literature for each of the tasks (or group of tasks). Specifically, we focus on the role that certain cognitive signatures can have in driving classic results that have been established in the literature for these different decision problems.

### 5.2.1 Auctions

One robust finding of experimental auction studies is the consistent overbidding in First-Price, Dutch, and Second-Price auctions, along with equilibrium (dominant strategy) bidding in English auctions. Since auctions are mechanisms that have both theoretical and real-world significance, it is important to understand where these deviations come from.

To investigate this, Figure 3 presents the distributions of deviations from equilibrium bidding behavior for each cognitive signature in each of the four auction formats. These

M3 correspond to a difficult lottery according the same index, and L13 and M13 correspond to what Puri (2024) refers to as a more difficult valuation than L1 (M1) and L3 (M3) because L13 (M13) has a larger support (see Table 2).

[28] The comparison of L3 and L13 is of interest because both of these lotteries have the same expected value, with L3 having a higher variance but L13 a higher number of prizes in its support.

Table 3: Differences in Distributions of Individual Assessments across Tasks

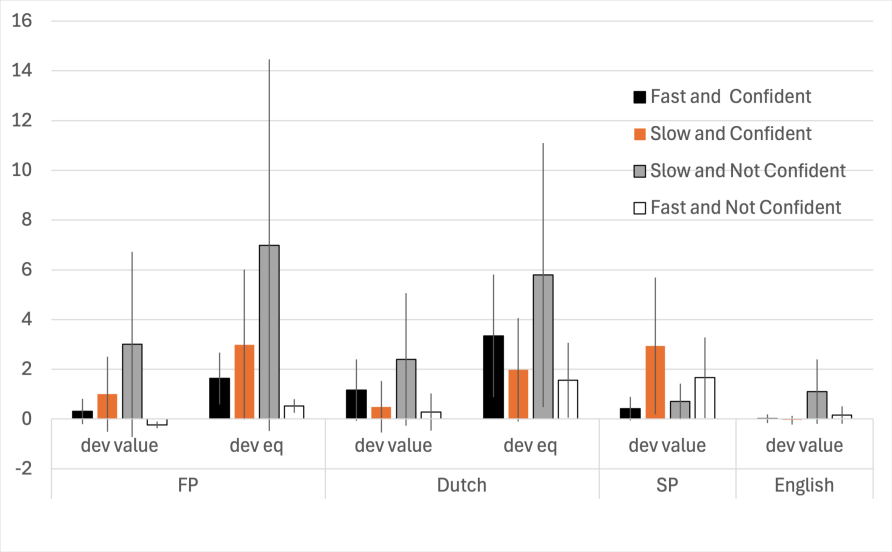
| Distributions of individual assessments are **different** | | Distributions of individual assessments are **similar** | |
|---|---|---|---|
| Pairs of Tasks | Chi-sq (p-value) | Pairs of Tasks | Chi-sq (p-value) |
| BINARY treatments | | | |
| FOSD vs MPS | 91.25 ($p < 0.01$) | Pivotality cont vs non-cont | 1.71 ($p = 0.64$) |
| ESsimp lotteries vs ESdiff lotteries | 41.08 ($p < 0.01$) | | |
| Common Consequence: CC1 vs CC2 | 17.36 ($p < 0.01$) | | |
| Common Ratio: CR1 vs CR2 | 17.27 ($p < 0.01$) | | |
| ESsimp mirrors vs ESdiff mirrors | 23.63 ($p < 0.01$) | | |
| ESsimp lotteries vs ESsimp mirrors | 14.42 ($p < 0.01$) | | |
| ESdiff lotteries vs ESdiff mirrors | 8.92 ($p = 0.03$) | | |
| NON-BINARY treatments | | | |
| ESsimp lottery vs Puri lottery | 11.77 ($p = 0.01$) | ESsimp lottery vs ESdiff lottery | 6.88 ($p = 0.08$) |
| | | ESdiff lottery vs Puri lottery | 6.26 ($p = 0.10$) |
| Puri lottery vs Puri mirror | 9.19 ($p = 0.03$) | ESsimp lottery vs ESsimp mirror | 1.95 ($p = 0.58$) |
| | | ESdiff lottery vs ESdiff mirror | 3.23 ($p = 0.36$) |
| | | First-price vs Dutch | 6.92 ($p = 0.07$) |
| | | First-price vs Second-price | 3.36 ($p = 0.34$) |
| | | Second-price vs English | 2.65 ($p = 0.45$) |
| | | Public Goods: low vs high MPCR | 2.17 ($p = 0.54$) |
| | | Beliefs: contra vs confirm prior signals | 4.01 ($p = 0.26$) |

deviations are measured as the average distance from the Risk Neutral Nash Equilibrium (RNNE) bid in the First-Price (FP) and Dutch auctions and the average distance from bidding one's value, the dominant action, in Second-Price (SP) and English auctions, as is typically done in the literature (see Kagel (1995)). On top of this, we also present the distribution of deviations from valuations for the First-Price and Dutch auctions.

Figure 3 reveals several notable patterns. First, in all auction formats, participants who make quick decisions tend to bid close to their valuations, regardless of their ex-post confidence. This suggests that individuals who rely on fast and intuitive decision-making might follow a simple and salient heuristic of bidding their own valuation and that acting quickly offers no time for a subject to think twice and contradict this salient bid.[29] In FP and Dutch auctions, this heuristic leads to systematic overbidding relative to the RNNE predictions (as extensively documented in the literature by Kagel (1995)) because valuations are typically higher than the RNNE predictions. But for the the SP and English auctions this intuitive heuristic coincides with equilibrium predictions, suggesting that some of the observed equilibrium behavior might not reflect the cognitive understanding of the solution to the problem.

[29]The only exception is overbidding in the SP auction among participants who made fast but low-confidence choices.

Figure 3: Auctions: Observed Bids relative to RNNE and Own Value for FP and Dutch and relative to Dominant strategy of bidding own value for SP and English.




Notes: We report average percentage deviation of observed bids from RNNE behavior in the First-price and Dutch auctions and from the dominant strategy (bidding one's value) in the Second-price and English auctions. For the First-price and Dutch auctions we also report average percentage deviation of observed bids from one's value. Positive values correspond to over-bidding and negative values correspond to under-bidding.

Note that identifying this behavioral pattern—and distinguishing intuitive heuristic bidding from deliberate strategic bidding—would not be possible without our methodology for classifying final bids. It is precisely this approach that allows us to uncover how different decision processes, or cognitive signatures, translate into distinct bidding behaviors across auction formats.

Second, overbidding in SP auctions is primarily driven by slow and confident decisions. On average, this group of subjects bid three times the valuation of the object and there is a high dispersion of choices among this group. Contrary to this, overbidding is essentially not present in the English auction, regardless of cognitive signatures, illustrating the robustness of the English auction format.

Third, for the FP and the Dutch auctions, the main contributors of average overbidding are participants who make thoughtful slow decisions, but who at the end of their assessment are still unsure of their final bid. These subjects bid on average six to seven times higher than the RNNE prediction prescribes and drive the average overbidding up.

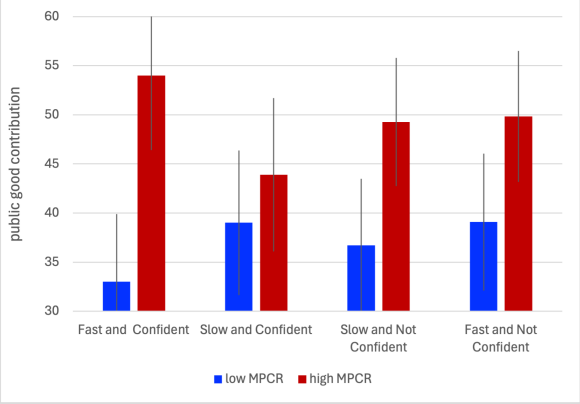
It is interesting to point out that, as we saw in Figure 2, the four different auction

formats we study are not associated with significantly different distributions of cognitive signatures, but we observe important differences in their behavior, conditional on these assessments, that allows us to better understand the long experimental literature in auctions. As a result, behavioral differences across these auctions cannot be attributed to Mapping 1, which maps decision problems into assessment distributions, but rather to Mapping 2, which maps assessments into behavior.

### 5.2.2 Public Goods

The study of public goods in the form on the voluntary contribution game is one of the oldest topics in experimental economics. We vary the marginal per capita return (MPCR) that determines the positive externality of contributions to the public good from 0.25 to 0.75. In theory, as long as the MPCR is less than 1, not contributing is a dominant strategy, however, experimental evidence shows a strong sensitivity to the MPCR. Our methodology allows us to study whether this result depends on the subjects' cognitive signatures. Figure 4, plots the average contribution to the public account for high and low MPCR and for each subject category.

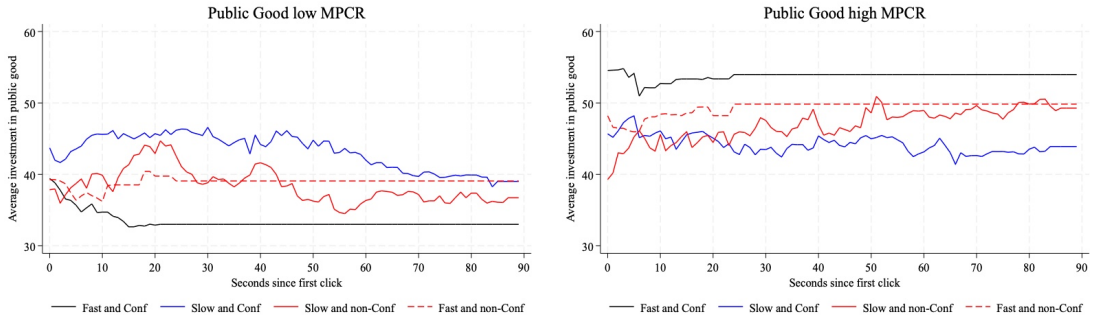
Figure 4: Average Contributions in Public Good Games




As we can see, subjects who are Fast and Confident have the greatest sensitivity in their contributions to differences in the MPCR, while those who respond to the problem in a Slow and Confident manner are the least sensitive.

We know from the experimental public goods literature that contributions decrease as the game is repeated. Subjects play this game only once in our experiment, but we can observe the dynamics of their response during the 90 seconds they have for deliberation by looking at their choice revisions, using the Choice Process Protocol. This allows us to observe how their introspection evolved over time.

Figure 5: Evolution of Choices in Public Good Games




Notes: Evolution of contributions to the public good as a function of thinking time. For each subject, the thinking time starts at the time of the first click and progresses onward.

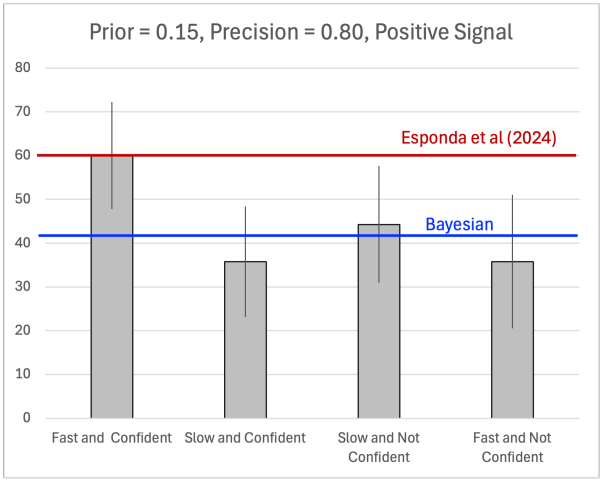
Figure 5 illustrates the dramatic change in behavior of the Fast and Confident players across the two treatments, which is consistent with our observations from Figure 4. From the very beginning, Fast and Confident subjects contribute the least in the 0.25 treatment and the most in the 0.75 treatment and maintain these levels through their 90 seconds of introspection. Subjects who take a longer time to make a final decision decrease their contributions as they contemplate their actions in the 0.25 treatment, yet increase them in the 0.75 treatment. Note, of course, that there is no feedback here, just internal contemplation.

### 5.2.3 Belief updating

**Base Rate Neglect.** Deviations from Bayesian updating have been widely studied in the literature. A well established bias is base-rate neglect, a phenomenon in which people place too much weight on a signal and ignore the underlying base-rate, or prior, from which that signal was drawn. Our subjects faced a number of probability updating problems, among which was the classic parametrization used by Kahneman and Tversky (1973), who first identified base-rate neglect. Their results were replicated by numerous studies, and its persistence was recently characterized by Esponda et al. (2024). In this parametrization there is a prior probability of 0.15 about a given state of the world being true and a signal with 80% accuracy which offers consistency with the state. For this problem, the Bayesian posterior is 0.42 after a signal consistent with the true state, while complete base-rate neglect would yield a posterior of 0.80.

Figure 6 presents the average reported posteriors for each cognitive signature. For three of the four cognitive signatures, average reported posteriors are not statistically different from the Bayesian benchmark. The remaining group (subjects who make fast and confident decisions) exhibits systematic base-rate neglect, placing excessive weight on the precision

Figure 6: Posteriors as a function of assessments, base-rate neglect specification




of the signal relative to the prior. This group constitutes 34% of our sample, the largest share.[30]

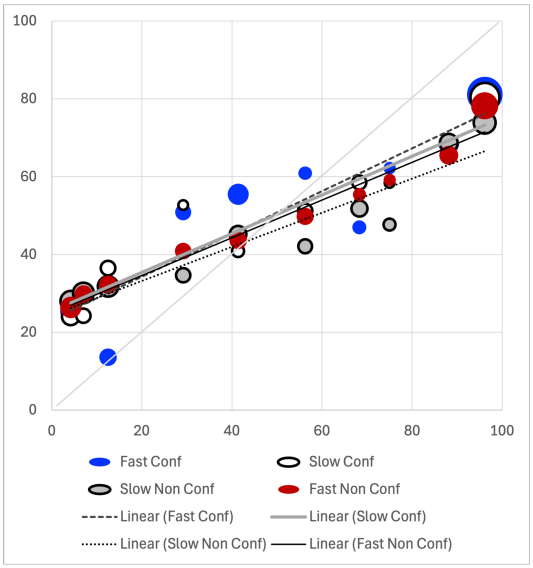
**Cognitive Signatures and Belief Attenuation.** Enke and Graeber (2023) suggest that when people are uncertain about the optimality of their performance, their decisions are heavily attenuated functions of objective probabilities. In terms of belief updating, this attenuation leads to regression to intermediate values. One example is the probability weighting function in belief-updating tasks where posterior probabilities are expected to be greater than Bayesian posteriors for low probability events and lower for high probability events.

Using our data, we can replicate the result of Enke and Graeber (2023) if we drop our combined speed-confidence categorization and instead use the ex-post confidence measure alone (Figure 15 in the Appendix). That is, there is a difference between posteriors reported by subjects with high and low confidence: higher confidence is associated with lower deviations from Bayesian updating.

However, when we take into consideration the full categorization that gives rise to our cognitive signatures (confidence *and* decision speed) we find a different result. Figure 7 presents the estimated linear belief updating functions for each cognitive signature, which depict the relationship between posterior beliefs and the Bayesian benchmark, aggregated

[30] The relationship between cognitive signatures and base-rate neglect is robust to alternative population-based normalizations of contemplation time and confidence, but becomes substantially noisier under the individual-specific normalization (Figure 14 in the Appendix). These robustness results suggest that, while the bias is present, caution is warranted in claiming that it is a universal feature of decision making.

Figure 7: Posterior Beliefs as a function of Bayesian Beliefs, by Assessments




over all belief updating tasks. We observe a pattern familiar in the literature for *all* of the cognitive signatures: people overestimate the probabilities of unlikely events and underestimate the probabilities of very likely events. A test of the slopes and intercepts of the regression lines finds no significant differences between any of these types.

If, as we argued before, confidence alone is not sufficient to capture how people perceive a problem, then one might want to consider the difference in these results to be significant. This would imply that some anomalies in the behavioral literature might actually present for most people regardless of their individual confidence level.

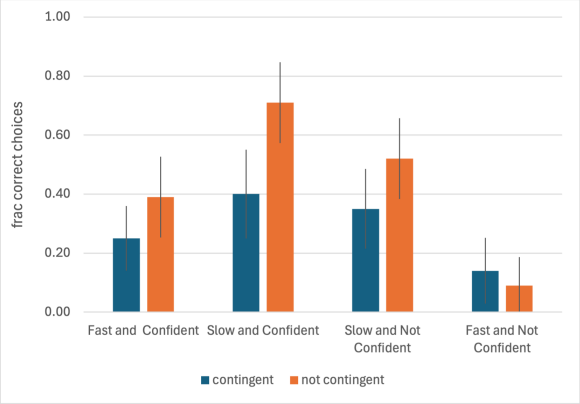
### 5.2.4 Contingent Reasoning

A recent experimental literature has documented the difficulties to think contingently, which is a cornerstone of most economic models (see Esponda and Vespa (2014), Martinez-Marquina et al. (2019), or Ngangoue and Weizsacker (2021)). We use the pivotality task of Esponda and Vespa (2014), with and without the need to think contingently, to study whether this phenomenon is present for all subjects or if it is more likely to arise for certain cognitive signatures (see Section 4 for details about this task). Regardless of the need to think contingently or not, there is an objectively correct answer in these tasks.

Recall from Figure 2 that there is no significant difference in the distributions of cognitive signatures for these two versions of the pivotality task, which is surprising since one task requires subjects to engage in contingent thinking, which could be thought of as ob-

jectively more complex, while the other does not. However, as we can see in Figure 8, the fraction of correct final choices differs across cognitive signatures and, for each category, it is higher for the version of the task where subjects are given enough information so they do not have to engage in contingent reasoning. Likewise, we can observe that spending more time thinking about this problem leads to more correct answers in both versions of the task.

Figure 8: Final Choices in Pivotality Tasks, by Assessment




This result is interesting because it indicates that differences in behavior across problems do not always stem from differences in the distribution of perceptions (Mapping 1), since our subjects did not perceive the contingent and non-contingent problems differently, but rather may stem from differences in behavior conditional on identical perceptions (Mapping 2).

### 5.2.5 Lotteries and Their Deterministic Mirrors

Not all of the problems given to our subjects have clear theoretical predictions. For example, in our lottery experiments, the optimality of choices depends on individual preferences, which are unobserved. In this section we discuss tasks involving binary lottery choices and the valuation of individual lotteries, and we compare cognitive signatures and behavior in each of these decision problems with their deterministic mirrors, which require the same computations to derive valuation as lotteries (multiplying probabilities by values and adding them up) but have no risk (see Oprea (2024b) and Section 4 for details). Even if the valuation of a mirror has an objectively correct answer that corresponds to the expected value of the lottery, we discuss these two types of tasks together to draw comparisons.

**Valuations.** To understand how the four cognitive signatures relate to the way people value lotteries and mirrors, Table 4 presents the certainty equivalents and simplicity equiva-

lents elicited from our subjects for lotteries L1, L3, and L13, whose expected values coincide with the objective value of the corresponding deterministic mirror (15 for L1 and M1 and 15.8 for L3, M3, L13, and M13). In each horizontal block, we first show average valuations for all subjects (Aggregate) and then we disaggregate them by cognitive signature, i.e., according to whether decisions were fast or slow, and confident or not confident.

Table 4: Valuations of Lotteries and their Deterministic Mirrors depending on Individual Assessments

| | Mirror M1 | | Mirror M3 | | Mirror M13 | |
|---|---|---|---|---|---|---|
| | mean (se) | exp value = 15 | mean (se) | exp value = 15.8 | mean (se) | exp value = 15.8 |
| Aggregate | 14.3 (0.29) | $p = 0.02$ | 14.8 (0.36) | $p < 0.01$ | 14.6 (0.31) | $p < 0.01$ |
| Fast and Conf | 13.8 (0.40) | $p < 0.01$ | 13.4 (0.69) | $p < 0.01$ | 14.6 (0.84) | $p = 0.15$ |
| Slow and Conf | 15.4 (0.65) | $p = 0.58$ | 16.2 (0.56) | $p = 0.43$ | 15.3 (0.45) | $p = 0.28$ |
| Slow and not Conf | 15.4 (0.68) | $p = 0.58$ | 15.9 (0.71) | $p = 0.92$ | 15.2 (0.53) | $p = 0.24$ |
| Not Engaged | 13.3 (0.60) | $p < 0.01$ | 13.6 (0.83) | $p = 0.01$ | 13.5 (0.69) | $p < 0.01$ |
| | Lottery L1 | | Lottery L3 | | Lottery L13 | |
| | mean (se) | exp value = 15 | mean (se) | exp value = 15.8 | mean (se) | exp value = 15.8 |
| Aggregate | 13.9 (0.31) | $p < 0.01$ | 12.5 (0.42) | $p < 0.01$ | 14.0 (0.37) | $p < 0.01$ |
| Fast and Conf | 13.5 (0.48) | $p < 0.01$ | 12.4 (0.85) | $p < 0.01$ | 13.6 (0.68) | $p < 0.01$ |
| Slow and Conf | 14.5 (0.73) | $p = 0.47$ | 13.4 (0.99) | $p = 0.02$ | 14.4 (0.76) | $p = 0.08$ |
| Slow and not Conf | 13.7 (0.79) | $p = 0.11$ | 12.1 (0.80) | $p < 0.01$ | 14.5 (0.65) | $p = 0.04$ |
| Fast and not Conf | 14.2 (0.61) | $p = 0.18$ | 12.3 (0.79) | $p < 0.01$ | 13.4 (0.90) | $p = 0.01$ |

Notes: Data from the Non-Binary Treatment. The $p$-values comparing observed average valuations for mirrors and lotteries and comparing each of them with the theoretical values come from regression analysis.

Looking at the average aggregate certainty and simplicity equivalents of all subjects, regardless of their cognitive signature, we replicate the results of Oprea (2024b): On average, the certainty and simplicity equivalents of all lotteries and mirrors are both significantly below the expected value of the lottery, which corresponds to the objective value of the mirror. This pattern, corresponding to deviations in the direction predicted by Prospect Theory, cannot be explained by risk attitudes since mirrors involve no risk. However, our approach allows us to understand whether this result is universal across subjects or if it is related to specific subjective assessments of the complexity of the task. Table 4 suggests that, for all mirror specifications, the low average valuations are mainly driven by people who make fast decisions, that is, who might not take the time to make the mathematical calculations needed to obtain the objective value of the mirror. This is true in all specifications for people who, on top of making fast decisions, are not confident in their final choice, which we can interpret as people who choose not to engage actively in the task. This is also present for people who make fast decisions and are confident in two of the three mirror specifications. On the other hand, people who take their time in this task report simplicity equivalents that are statistically indistinguishable from the correct values of the mirrors, suggesting an important heterogeneity in mirror tasks.

The situation is very different in the valuation of lotteries, where we do not observe a

pattern that suggests that one particular assessment or cognitive signature drives aggregate results. Instead, the valuations across lotteries for the different assessment categories are typically below the risk neutral values, indicating a familiar pattern of risk aversion for small stakes.[31] Our results illustrate once more the importance of disaggregating analysis to incorporate subjective assessments, since results that appear to substantiate a particular conclusion can sometimes be driven by a subclass of people who perceive the problem in a particular way.

**Binary Choices of Lotteries and Mirrors.** We now turn our attention to study the differences in choices that involve two lotteries or two mirrors, as opposed to valuations of individual lotteries and mirrors. Table 5 compares the probability of choosing the riskier lottery (according to the Sharpe ratio), or the corresponding mirror, in each pair of binary choices (ES simp and ES diff).[32] Similar to the results we presented for valuations of lotteries and mirrors, notice that for the choice of L1 vs L2 (M1 vs M2), which is a simple choice according to the index of Enke and Shubatt (2023) and assessed that way (see the top panel of 2), subjects choose among mirrors in the same way that they do among lotteries, for all cognitive signatures. However, for the choice between L3 vs L4 (M3 vs M4), which is considered harder according to Enke and Shubatt (2023), the equivalence in choice between lotteries and mirrors depends on their subjective assessment of the problem. In this case, confidence matters: People who report low confidence in their choices make similar choices in lotteries and mirrors, but people who feel confident in their choices choose differently when the problem involves a lottery or a deterministic mirror.[33]

Our results illustrate the nuanced relationship between lotteries and their deterministic mirrors from two different angles: valuations and binary choices. By studying how the distributions of cognitive signatures translate into final decisions in each task, we observe that, in general, when problems are objectively simpler, both valuations and choices in lotteries and mirrors are indistinguishable from one another. However, for more complicated tasks, the similarity in choices among lotteries and mirrors crucially depends on one of the two components of our cognitive signature measure.

We also examine whether cognitive signatures help explain behavior in the rest of the binary lottery choice tasks in our experiment. Two broad findings emerge (see the Online Appendix for details). First, cognitive signatures have little predictive power in relatively straightforward lottery problems, where subjects with different cognitive signatures make similar choices. In contrast, for more intricate lottery problems, subjects who make fast decisions are substantially more likely to choose the safer lottery, consistent with simplified

[31] The exception is Lottery 1, for which some people report certainty equivalents equal to the expected value of the lottery and others report strictly lower ones.

[32] Section 2 of the Online Appendix presents a comprehensive analysis of risky choices according to the Sharpe ratio for all binary lottery problems subjects encountered in the experiment.

[33] In this particular case, we see more risk averse choices in lotteries with respect to mirrors, but we do not generalize this pattern since we only ask one 'difficult' question.

Table 5: Binary choices of Lotteries and Mirrors

| | Prob of choosing L1 (riskier) or M1 | | | |
|---|---|---|---|---|
| | Fast Conf | Slow Conf | Slow not Conf | Fast not Conf |
| ESsimp lottery task | 0.92 | 0.93 | 0.83 | 0.88 |
| ESsim mirror task | 0.92 | 0.92 | 0.79 | 0.84 |
| Lotteries vs Mirrors | $p = 0.89$ | $p = 0.77$ | $p = 0.55$ | $p = 0.39$ |
| | Prob of choosing L3 (riskier) or M3 | | | |
| | Fast Conf | Slow Conf | Slow not Conf | Fast not Conf |
| ESdiff lottery task | 0.12 | 0.38 | 0.48 | 0.29 |
| ESdiff mirror task | 0.40 | 0.66 | 0.58 | 0.43 |
| Lotteries vs Mirrors | $p < 0.01$ | $p < 0.01$ | $p = 0.14$ | $p = 0.05$ |

Notes: The $p-$values are obtained from the regression analysis comparing the tendency to choose L1 or M1 in the top portion of the table and L3 or M3 in the bottom portion with standard errors clustered at the individual level.

decision making. Second, in the common consequence and common ratio tasks, subjects are substantially less confident in their choices in the second question in each pair compared to the first one. Moreover, violations of Expected Utility in the common ratio problems are significantly less common among subjects who perceive the first question as easy and make fast decisions, whereas no analogous relationship is observed for the common consequence problems.

# 6 Conclusions

This paper provides a new set of tools that allow us to dig deeper into the behavior observed in many familiar and commonly used experiments. Although the tools we use (the Choice Process Protocol, which provides effort measures and an ex-post measure of confidence in choices) are not new, using them in combination is. These tools allow us, for any given experimental task, to define a mapping from that task to the distribution of subjective perceptions (cognitive signatures) of it, and then to observe a second mapping from each subjective perception class to behavior.

In defining subjective assessments of a task, we classify subjects into four categories or classes, depending on the effort the subjects put into solving the problem presented to them and the ex-post confidence they report about the optimality of their final choice. We thus characterize the cognitive signature of a subject in a specific decision task as either Fast and Confident, Slow and Confident, Slow and Not Confident, or Fast and Not Confident. We call these approaches to a problem the subject's cognitive signature.

In the first and last of these assessment classes, subjects make a quick choice that we could interpret as either their intuitive best guess (perhaps based on their informed

understanding of the underlying theory) or the guess of some simple heuristic they employ. They differ in that after making their choice, those who might consider the task as easy are confident they chose correctly, while those who perceive it as too hard, or simply choose not to engage with it, are not confident in their choice. Those subjects who fall into the other two categories, on the other hand, exert a great effort by taking a longer time to make a choice, but come to different conclusions about the optimality of their performance.

Our main point is that how people behave when faced with a decision problem depends on their perception of the problem as indicated by their cognitive signature. However, since these perceptions are heterogeneous, we might expect behavior to be heterogeneous as well, and hence it is important to understand the heterogeneity underlying aggregate choice behavior since that gives us insight into the predictive abilities of the underlying theory.

For example, when the literature implies that people behave in an anomalous manner compared to theory, our techniques allow us to investigate whether this behavior is universal or attributable to one subset of people who have a particular cognitive signature. Similarly, when aggregate behavior suggests support for a theory, our tools allow us to determine whether that behavior is an artifact of the composition of different behaviors across people with different cognitive signatures.

Several of our findings illustrate this point. For example, in belief-updating tasks, using the original parameterization of Kahneman and Tversky (1972), while we find base-rate neglect as they do, we are able to attribute it to those subjects who make Fast and Confident choices. Their mistakes, when aggregated with the other types who update in a Bayesian fashion, drive the Base-rate neglect result.

Contrary to the base-rate neglect problem, recent papers (e.g., Enke and Graeber (2023)) have suggested that the tendency to overweight small probabilities and underweight large ones is attributable to subjects with greater cognitive uncertainty (our not-confident types). However, we find that such behavior exists across all cognitive signatures. These types of results allow us to better understand which phenomena are universal across people and which depend on the way they perceive decision problems.

In other tasks, our approach shows that differences in behavior need not be attributed to different subjective perceptions (what we call Mapping 1) but rather to differences in behavior conditional on a given perception (Mapping 2). This is observed in auctions. Here we find no differences in the distribution of cognitive signatures across the four auction formats with which we present our subjects, yet we do find differences in behavior when we condition on a given signature. For example, subjects who make fast decisions across four auction formats (First-Price, Second-Price, Dutch, and English), tend to use the same simple and salient heuristic of bidding their valuation. In First-Price and Dutch auctions this implies overbidding with respect to the equilibrium prediction ascribed to these formats, the risk-neutral Nash equilibrium. In the Second-Price and English auctions, however, this same heuristic is accounted for as equilibrium play because it coincides with the theoretical prediction. Our results suggest that some of the observed equilibrium

behavior might not reflect the cognitive understanding of the solution to the problem but instead the use of a simple heuristic that is used by those subjects who respond fast and intuitively, regardless of the auction format.

Finally, our results have broad and important implications for different fields of study in microeconomics, such as decision theory and mechanism design. For example, the observation that not all decision biases are universal but may be concentrated among those people who perceive the decision problem they face in certain ways highlights the importance of heterogeneity analysis and suggests that standard economic theory may perform better than we had thought once we take into account who is responsible for the systematic deviations we observe.

Second, in terms of mechanism design, our results also resonate since if mechanisms fail to perform as expected it might be that the failure is attributable to those who perceive the problem as too hard and hence give up. If they are responsible for a large part of the observed behavior, then a simple intervention to better explain the rules of the auction might restore efficiency.

## References


Agranov, M., A. Caplin, and C. Tergiman (2015). Naive play and the process of choice in guessing game. *Journal of Economic Science Association 1(2)*, 146–157.

Agranov, M. and P. Reshidi (2026). Deciphering suboptimal updating: Task difficulty, structure, and sequencing. *Quantitative Economics*.

Ali, S., M. Mihm, S. L., and C. Tergiman (2021). Adverse and advantageous selection in the laboratory. *American Economic Review 111*, 2152–2178.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole americaine. *Econometrica 21(4)*, 503–546.

Allis, L. V. (1994). Searching for solutions in games and artificial intelligence. *Maastricht University*.

Armantier, O. and N. Treich (2016). The rich domain of risk. *Management Science 62*, 1954–1969.

Augenblick, N., E. Lazarus, and M. Thaler (2025). Overinference from weak signals and underinference from strong signals. *Quarterly Journal of Economics Forthcoming*.

Ba, C., A. Bohren, and A. Imas (2023). Over- and underreaction to information. *Working paper*.

Becker, G., M. DeGroot, and J. Marschak (1964). Measuring utility by a singleresponse sequential method. *Behavioral Science 9*, 226–232.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1 2*, 69–186.

Blavatskyy, P., V. Panchenko, and A. Ortmann (2023). How common is the common-ratio effect? *Experimental Economics 26(2)*, 253–272.

Boldt, A., C. Blundell, and B. De Martino (2019). Confidence modulates exploration and exploitation in value-based learning. *Neuroscience of consciousness 1*.

Bordalo, P., J. Conlon, N. Gennaioli, S. Kwon, and A. Shleifer (2025). How people use statistics. *Review of Economic Studies*.

Bordalo, P., N. Gennaioli, and A. Shleifer (2012). Salience theory of choice under risk. *Quarterly Journal of Economics 127(3)*, 1243–1285.

Bordalo, P., N. Gennaioli, and A. Shleifer (2022). Salience. *Annual Review of Economics 14*, 521–544.

Caplin, A., M. Dean, and D. Martin (2011). Search and satisficing. *American Economic Review 1017)*, 2899–2922.

Cerreia-Vioglio, S., D. Dillenberger, and P. Ortoleva (2015). Cautious expected utility and the certainty effect. *Econometrica 83(2)*, 693–728.

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

da Silva Castanheira, K., F. S.M., and A. Otto (2021). Confidence in risky value-based choice. *Psychonomic Bulletin and Review 28*.

Dal Bo, E., D. B. P., and E. Eyster (2018). The demand for bad policy when voters underappreciate equilibrium effects. *The Review of Economic Studies 85*, 964–998.

Danz, D., L. Vesterlund, and A. Wilson (2021). Belief elicitation and behavioral incentive compatibility. *American Economic Review 112(9)*, 2851–2883.

De Martino, B., F. S.M., N. Garrett, and R. Dolan (2012). Confidence in risky value-based choice. *Nature Neuroscience 16(1)*.

Enke, B. and T. Graeber (2023). Cognitive uncertainty. *Quarterly Journal of Economics 138(4)*, 2021–2067.

Enke, B., T. Graeber, and R. Oprea (2025). Complexity and time. *Journal of the European Economic Association Forthcoming*.

Enke, B. and C. Shubatt (2023). Quantifying lottery choice complexity. *Working paper*.

Esponda, I. and E. Vespa (2014). Hypothetical thinking and information extraction in the laboratory. *American Economic Journal: Microeconomics 6*, 180–202.

Esponda, I. and E. Vespa (2023). Contingent thinking and the sure-thing principle: Revisiting classic anomalies in the laboratory. *Review of Economic Studies*.

Esponda, I., E. Vespa, and S. Yuksel (2024). Mental models and learning: The case of base-rate neglect. *American Economic Review 114(3)*, 752–782.

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

Gill, D. and V. Prowse (2023). Strategic complexity and the value of thinking. *The Economic Journal 133(650)*, 761–786.

Gneezy, U., B. Enke, B. Hall, D. Martin, V. Nelidov, T. Offerman, and J. van de Ven (2023). Cognitive biases: Mistakes or missing stakes? *Review of Economics Studies*.

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

Grimaldi, P., H. Lau, and M. Basso (2015). There are things that we know that we know, and there are things that we do not know we do not know: Confidence in decision-making. *Neuroscience Biobehavioral Review 55*.

Gul, F. (1991). A theory of disappointment aversion. *Econometrica 59(3)*, 667–686.

Hu, E. H. (2024). Confidence in inference. *Working Paper*.

Kagel, J. (1995). *Auctions: A Survey of Experimental Research.* The Handbook of Experimental Economics. Princeton: Princeton University Press.

Kahneman, D. and A. Tversky (1972). On prediction and judgement. *ORI Research Monograph 12(4)*.

Kahneman, D. and A. Tversky (1973). On the psychology of prediction. *Psychological review 80(4)*.

Kessler, J., H. Kivimaki, A. Litwin, and M. Niederle (2023). Please take a minute: How prosocial preferences change with deliberation. *Working Paper*.

Ledyard, J. (1995). *Public Goods: A Survey of Experimental Research.* The Handbook of Experimental Economics. Princeton: Princeton University Press.

Loomes, G. and R. Sugden (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal 92(368)*, 805–824.

Luttrell, A., P. Brinol, R. Petty, W. Cunningham, and D. Diaz (2013). Metacognitive confidence: a nauroscience approach. *Revista de Psicología Social 28(3)*.

Martinez-Marquina, A., M. Niederle, and E. Vespa (2019). Failures in contingent reasoning: The role of uncertainty. *American Economic Review 109*, 3437–3474.

McGranaghan, C., T. O'Donoghue, K. Nielsen, J. Somerville, and C. Sprenger (2024a). Connecting common ratio and common consequence preferences. *Working paper*.

McGranaghan, C., T. O'Donoghue, K. Nielsen, J. Somerville, and C. Sprenger (2024b). Distinguishing common ratio preferences from common ratio effects using paired valuation tasks. *American Economic Review 114(2)*, 307–347.

Ngangoue, M. and G. Weizsacker (2021). Learning from unrealized versus realized prices. *American Economic Journal: Microeconomics 13*, 174–201.

Oprea, R. (2020). What makes a rule complex. *American Economic Review 110(12)*, 3913–3951.

Oprea, R. (2024a). Complexity and its measurements. *Working Paper*.

Oprea, R. (2024b). Decisions under risk are decisions under complexity. *American Economic Review 114(12)*, 3789–3811.

Puri, I. (2024). Simplicity and risk. *Journal of Finance forthcoming*.

Rollwage, M., A. Loosen, T. U. Hauser, R. Moran, R. J. Dolan, and S. Fleming (2020). Confidence drives a neural confirmation bias. *Nature communications 11 (1)*.

Rubinstein, A. (2007). Instinctive and cognitive reasoning: A study of response times. *The Economic Journal 117*, 1243–1259.

Spiliopoulos, L. and A. Ortmann (2018). The bcd of response time analysis in experimental economics. *Experimental Economics 21*(2), 383–433.

Tversky, A. and D. Kahneman (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty 5(4)*, 297–323.

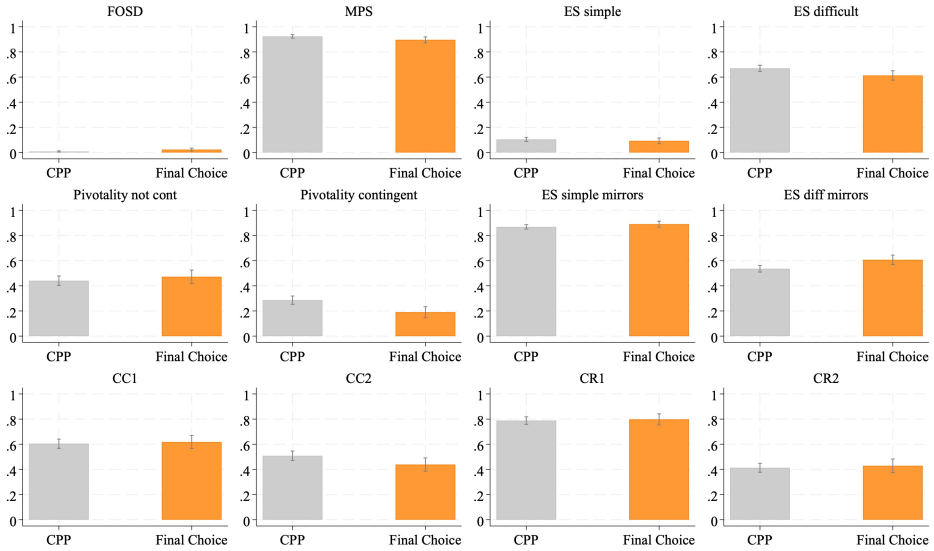
Wilcox, N. (1993). Lottery choice: Incentives complexity and decision time. *Economic Journal 103*(421), 1397–1417.

# A Does the Choice Process Protocol alter behavior?

As we described in Section 3, the CPP protocol provides the complete thinking path showing how participants arrive at a final choice, in addition to the final choice itself. One possible concern about the CPP is that its use affects *how* people think about the problem and, as a result, alters their final choices.

To examine this concern, we conducted an additional set of the Binary treatments, in which only final choices determined payments, so the time series of choices leading to that final choice were irrelevant (not incentivized). We kept all the other experimental details identical to our CPP sessions and only changed the feature that it is the final choice of a subject rather than the choice at a randomly selected second that determined her payment. A total of 173 participants participated in these new Binary sessions: 84 in treatment 1 and 89 in treatment 2.

Figure 9: Comparing Final Choices in CPP and Final Choice sessions.




Notes: For lotteries, we present the fraction of final choices that correspond to the higher Sharpe's ratio. For pivotality tasks and binary mirror choices we present the fraction of optimal (correct) choices.

Figure 9 depicts the final choices of our participants in the CPP sessions and these new sessions, which we refer to as the Final-Choice sessions, separately for each task. It is clear that there are no significant differences in final choices in *any* of the tasks ($p \geq 0.10$ in each task). This is reassuring as it confirms that the set of tools that we propose can be employed broadly across different environments as it does not alter final choices while providing more information about the thinking process.

# B Alternative normalization and robustness of main results

As discussed in the main text, the medians of confidence and contemplation time that we use to normalize the data can be computed in different ways, depending on the available data and the research question. In this section we discuss with more detail the approach we follow on the main text, population-level normalization, and show robustness of our results when we use an individual-level normalizations.

## B.1 Distribution of Cognitive Signatures

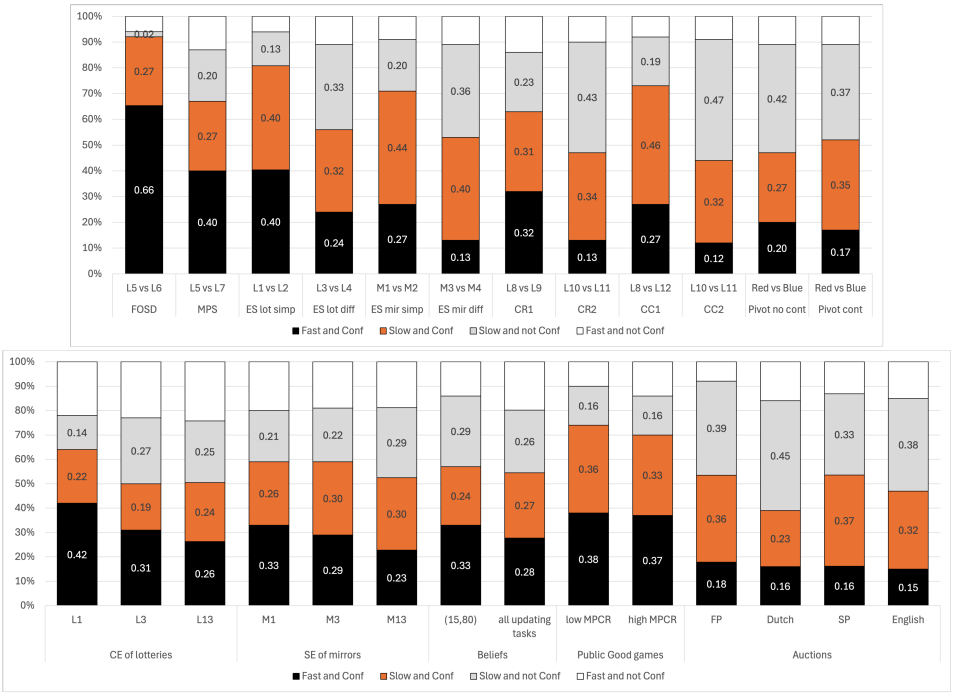
**Individual-level normalization.** When an experiment includes several tasks of a similar kind performed by all subjects, contemplation time and confidence can be normalized at the individual level using each subject's median values across that group of tasks. This normalization classifies how people approach different problems relative to their *own* thinking style and confidence. For example, if a participant completes multiple rounds of binary lottery choices, the median contemplation time of a subject across these rounds serves as a benchmark in determining whether that person took a short or long time to make a decision, relative to her own style. A similar argument applies to confidence.

What types of questions does this normalization allow one to answer? First, we can compare the distribution of assessments across tasks and judge whether some tasks are more likely to induce specific assessments. For instance, one can study which tasks are associated with a larger proportion of fast and confident choices, relative to tasks where we observe a larger proportion of slow and non confident choices. Second, this normalization enables within-subject comparisons of how specific assessments correlate with final choices. For example, one can examine whether a person is more likely to choose a riskier lottery when her decision is fast and confident, and lean toward safer options when decisions are slow and not confident.

This individual normalization can also be applied across all tasks completed by a subject during the session. Such a normalization facilitates comparisons of subjective assessments across tasks, revealing which tasks subjects spend relatively more time contemplating and in which tasks they make decisions with greater or lesser confidence.

Figure 10 illustrates this approach by presenting the distributions of subjective assessments under this alternative normalization. Specifically, confidence and contemplation time are normalized using each subject's median values computed across all tasks completed during the session. The qualitative comparisons and results discussed in the main text remain unchanged, indicating that our results are robust to the choice of normalization. At the same time, this normalization provides additional insights. For example, if we compare how people approach public good games and auctions we can clearly see a larger fraction of fast and confident decisions in the public goods game than in any of the auction formats, suggesting that the auction environments appear to be more cognitively demanding.

Figure 10: Individual Assessments of Tasks and Games, alternative normalization

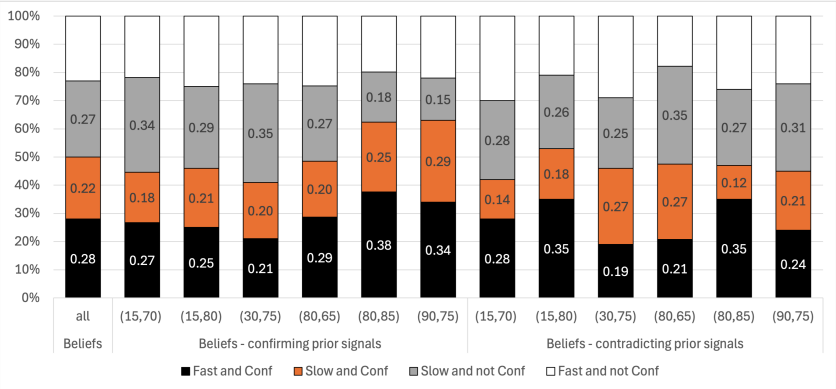



Notes: Assessments are defined relative to the individual medians for all tasks subjects face in the entire experiment. The top figure depicts tasks in the Binary treatment, while the bottom one depicts tasks in the Non-Binary treatment.

**Population-level normalization.** In some experimental designs, participants complete only a single task of some kind, for instance, because of potential spillover concerns. In this case, the individual-level classification described above is not feasible. Instead, one can normalize contemplation time and confidence using the *population* medians. This is what we do in the main text of the paper. Thus, for instance, the category of fast and confident decisions in this case would consist of participants who spend less than the median deliberation time on that particular task (or group of tasks), across all participants who faced that task, and who report above-median confidence.

This population-level normalization enables cross-sectional comparisons across participants within a given task, linking the distribution of final choices to the underlying decision processes. Note that if this normalization is done at the population level separately for each individual task, then, by construction, this approach would yield an approximately uniform distribution of the four decision categories outlined above, which makes the comparison of the distributions of assessments across tasks less informative.

Figure 11: Individual Assessments of Belief-updating Tasks, by parameterization




Notes: Normalization of confidence and speed of decisions is done using the population level medians across all belief updating tasks together (as in the main text of the paper). For belief-updating tasks, we present the distributions separately for confirming and contradicting signals and indicate the prior and the signal precision in the parenthesis in that order. This data is used in the analysis that follows.

If the experiment includes several treatments in which each participant completes only one version of the task, like in our auction treatments, a population-level median normalization across all treatments can be used. This is the normalization we use in the main text of the paper. In this case, the distribution of assessments across treatments (i.e., across different versions of the same task) becomes informative and can reveal which versions elicit more or less confident and more or less rapid decisions, both of which can substantially influence the choices participants make in the task. To complement the distribution of individual assessments presented in the main text of the paper, in Figure 11 we present these distributions for the belief updating tasks where we evaluate separately the cases where subjects observe signals that confirm or contradict the prior, and for each pair of prior and signal accuracy.

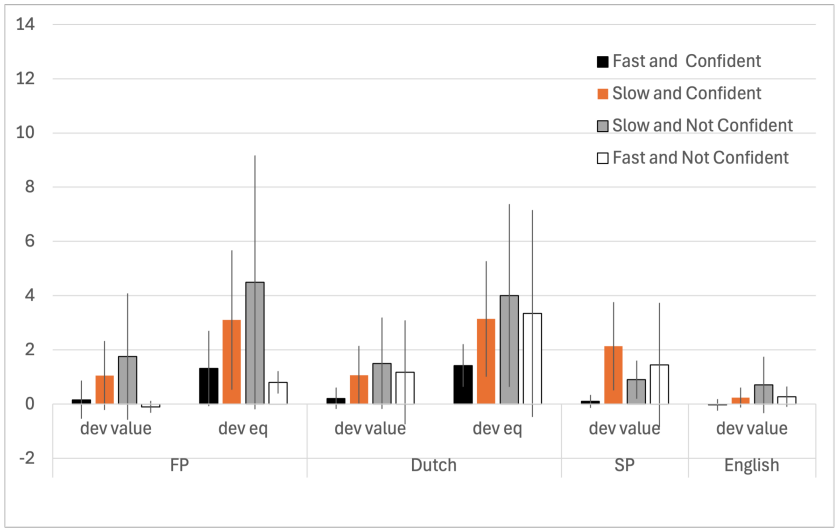
Ultimately, the choice of the normalization procedure depends on the specific research questions one is interested in, and, as a result, on the task or group of tasks that are analyzed. This makes our methodology very portable across domains and experimental designs.

## B.2 Robustness of Main Results

In this section, we replicate the main empirical results reported in the main text for the alternative individual-level normalization for subjective assessments, where confidence and contemplation time are normalized separately for each subject using the individual medians across *all* tasks each subject completes in the experiment.

**Auctions** Figure 12 replicates Figure 3 in the main text under the individual-level normalization. The main patterns discussed in the paper remain intact. Most importantly, across all auction formats, participants who make quick decisions tend to bid close to their valuations, regardless of their ex-post confidence. In addition, overbidding in the second-price auction is primarily driven by participants who make slow and confident decisions, whereas overbidding in the first-price and Dutch auctions (relative to the RNNE) is driven primarily by participants who make slow and non-confident decisions.

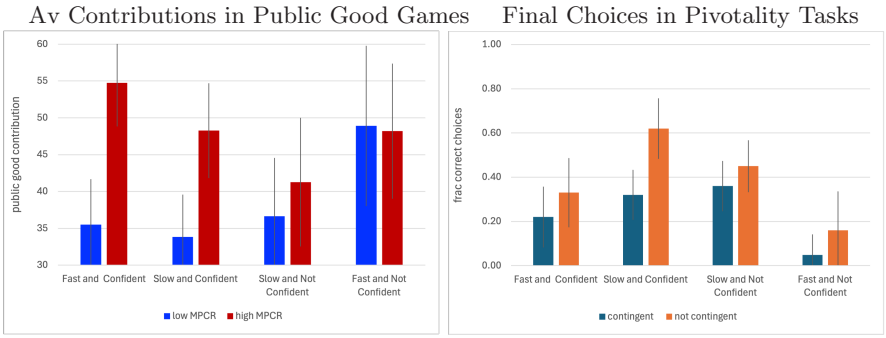
Figure 12: Auctions: Observed Bids relative to RNNE and Own Value for FP and Dutch and relative to Dominant strategy of bidding own value for SP and English, individual-level normalization.




**Public Good games.** The left panel in Figure 13 replicates Figure 4 and shows virtually identical results: subjects who make fast and confident choices are the most sensitive to the focal feature of the environment (the MPCR value), whereas those who deliberate longer and remain confident in their choices are the least sensitive.

**Contingent Reasoning** The right panel in Figure 13 replicates Figure 8 using individual medians to normalize contemplation time and confidence. The results closely mirror those reported in the main text. The fraction of correct final choices remains significantly higher when subjects do not need to engage in contingent reasoning, regardless of their subjective assessment of the task. Furthermore, in both versions of the task, subjects who spend more time contemplating the problem are significantly more likely to make the correct final choice.

Figure 13: Responses in Public Good games and Pivotality Tasks, individual-level normalization



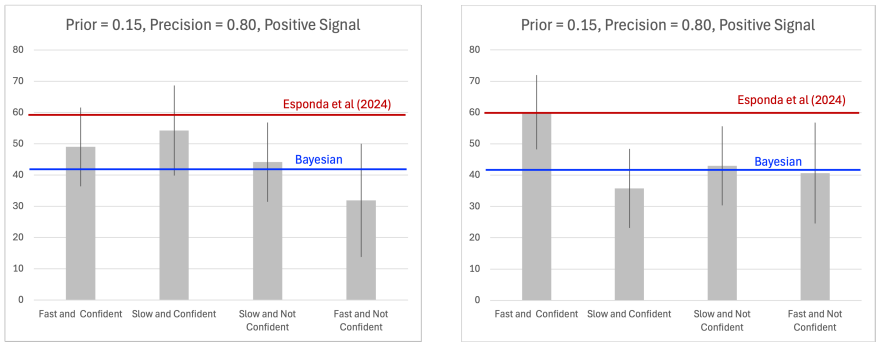

**Base-rate neglect** Figure 14 replicates Figure 6 using two alternative normalizations of contemplation time and confidence. Under the first normalization (shown in the left panel), contemplation time and confidence are normalized relative to each individual's median across all tasks completed during the session. Under the second normalization (shown in the right panel), they are normalized relative to the population medians for the parameterization of the base-rate neglect task used in Kahneman and Tversky (1972) .

Using individual rather than population medians produces considerably noisier results. In particular, the 95% confidence intervals for both groups of confident choices now include both the Bayesian prediction and the level of base-rate neglect reported by Esponda et al. (2024), making the differences across cognitive signatures less pronounced. By contrast, the results remain robust under a population-based normalization that is different to the one used in the main text (separately for this task, instead of jointly across all belief-updating tasks, as in the main text). These findings suggest that the mapping between cognitive signatures and base-rate neglect is more robust to the choice of population-based normalization consistent with our focus on the between-person (rather than within-person) comparisons than to the use of an individual-specific normalization.

**Cognitive signatures and belief attenuation** Figure 15 replicates Figure 3b from Enke and Graeber (2023) using the individual-level normalization. This figure displays coefficients from OLS regressions of reported posterior beliefs on Bayesian posteriors, split by cognitive uncertainty quartiles. Cognitive uncertainty is defined as $\frac{100-\text{confidence}}{100}$. Error bars show 95% confidence intervals. Consistent with the results reported in Enke and Graeber (2023) we observe that reported posteriors are more compressed to intermediate belief when cognitive uncertainty is higher.

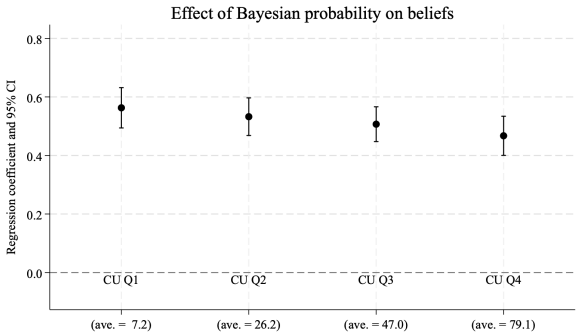
**Valuations of Lotteries and Mirrors** Tables 6 and 7 replicate the analysis in the main text of the paper reported in Tables 4 and 5 but using individual medians to normalize

Figure 14: Posteriors as a function of assessments, base-rate neglect specification, alternative normalizations




Notes: The left panel uses individual medians across all tasks administered in a session, while the right panel uses population medians for this task only (Kahneman and Tversky (1972) parameterization of BRN).

Figure 15: Posteriors as a function of assessments, base-rate neglect specification, individual-level normalization




Notes: We plot coefficients from OLS regressions of reported posterior beliefs on Bayesian posteriors, split by cognitive uncertainty (CU) quartiles. CU is defined as $\frac{100-\text{confidence}}{100}$. Error bars show 95% confidence intervals.

contemplation time and confidence. The results are qualitatively similar to those reported in the main text for both valuations (Table 6) and binary choices (Table 7).

Table 6: Valuations of Lotteries and their Deterministic Mirrors depending on Individual Assessments, individual-level normalization

| | Mirror M1 | | Mirror M3 | | Mirror M13 | |
|---|---|---|---|---|---|---|
| | mean (se) | exp value = 15 | mean (se) | exp value = 15.8 | mean (se) | exp value = 15.8 |
| Aggregate | 14.3 (0.29) | $p = 0.02$ | 14.8 (0.36) | $p < 0.01$ | 14.6 (0.31) | $p < 0.01$ |
| Fast and Conf | 14.1 (0.39) | $p = 0.02$ | 15.3 (0.71) | $p = 0.52$ | 13.6 (0.74) | $p < 0.01$ |
| Slow and Conf | 14.9 (0.53) | $p = 0.78$ | 15.7 (0.61) | $p = 0.87$ | 15.9 (0.43) | $p = 0.84$ |
| Slow and not Conf | 14.9 (0.78) | $p = 0.88$ | 14.7 (0.69) | $p = 0.12$ | 15.0 (0.50) | $p = 0.09$ |
| Not Engaged | 13.3 (0.74) | $p = 0.03$ | 12.6 (0.93) | $p < 0.01$ | 13.3 (0.87) | $p < 0.01$ |
| | Lottery L1 | | Lottery L3 | | Lottery L13 | |
| | mean (se) | exp value = 15 | mean (se) | exp value = 15.8 | mean (se) | exp value = 15.8 |
| Aggregate | 13.9 (0.31) | $p < 0.01$ | 12.5 (0.42) | $p < 0.01$ | 14.0 (0.37) | $p < 0.01$ |
| Fast and Conf | 14.1 (0.41) | $p = 0.02$ | 13.0 (0.92) | $p < 0.01$ | 14.8 (0.81) | $p = 0.20$ |
| Slow and Conf | 13.9 (0.73) | $p = 0.15$ | 13.5 (0.86) | $p < 0.01$ | 13.9 (0.66) | $p < 0.01$ |
| Slow and not Conf | 13.5 (0.91) | $p = 0.12$ | 11.0 (0.66) | $p < 0.01$ | 14.0 (0.66) | $p < 0.01$ |
| Not Engaged | 13.6 (0.74) | $p = 0.07$ | 12.5 (0.93) | $p < 0.01$ | 13.2 (0.83) | $p < 0.01$ |

Notes: Data from the Non-Binary Treatment. The $p$-values comparing observed average valuations for mirrors and lotteries and comparing each of them with the theoretical values come from regression analysis.

Table 7: Binary choices of Lotteries and Mirrors, individual-level normalization

| | Prob of choosing L1 (riskier) or M1 | | | |
|---|---|---|---|---|
| | Fast Conf | Slow Conf | Slow not Conf | Fast not Conf |
| ESsimp lottery task | 0.91 | 0.92 | 0.81 | 0.88 |
| ESsim mirror task | 0.92 | 0.91 | 0.77 | 0.81 |
| Lotteries vs Mirrors | $p = 0.74$ | $p = 0.71$ | $p = 0.52$ | $p = 0.41$ |
| | Prob of choosing L3 (riskier) or M3 | | | |
| | Fast Conf | Slow Conf | Slow not Conf | Fast not Conf |
| ESdiff lottery task | 0.17 | 0.38 | 0.47 | 0.19 |
| ESdiff mirror task | 0.46 | 0.63 | 0.52 | 0.42 |
| Lotteries vs Mirrors | $p < 0.01$ | $p < 0.01$ | $p = 0.41$ | $p = 0.02$ |

Notes: The $p-$values are obtained from the regression analysis comparing the tendency to choose L1 or M1 in the top portion of the table and L3 or M3 in the bottom portion with standard errors clustered at the individual level.