Complex for Whom?
An Experimental Approach to Subjective Complexity

Marina Agranov* Andrew Schotter† Isabel Trevino‡

February 27, 2025

Abstract

We present a set of tools to elicit subjective perceptions of the complexity of a variety of choice problems that have been thoroughly studied in economics. Our object of interest is the mapping from the description of a problem to the distribution of subjective perceptions of its complexity, and then the mapping of such perceptions to choices. We find that, in general, different problems are perceived differently by the people who engage in them, which, in turn, induces different distributions over behavior. Depending on the task, we find that aggregate results, established in the literature, might reflect the behavior of only a subset of these perception classes and, therefore, the specific distribution over perception classes has a strong effect on what average or aggregate behavior is found to be. As a result, our findings shed a new light on the predictive power of economic theory, since they suggest that observed deviations might be driven by a subset of people who perceive problems in a particular way and not universally by all economic agents.

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, but that explanation fails to consider the heterogeneity of perceptions of it. Because the distribution of subjective perceptions of the complexity of a problem is unobserved, just focusing on the aggregate behavior of agents is somewhat uninformative since we do not know which subjective perceptions are behind the observed behavior. The object of interest, therefore, is the mapping from the description of a problem to the distribution of

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

subjective perceptions of complexity it induces in the population, and then the mapping of such perceptions to choices. If changing the problem induces a change in the distribution of perceptions, then as we sweep across different problems, we also sweep across different perception distributions, and hence different assessments of complexity. Put differently, the question is not whether a problem is complex, but rather for whom it is complex and how people respond to this complexity.

In this paper, we present a set of tools that allows us to make the heretofore unobserved distribution of subjective perceptions of a problem observable and, as a result, gives us a more in-depth understanding of its complexity. Our approach differs from most recent attempts to measure complexity, which typically ignore the heterogeneity of subjective perceptions and apply a one-size-fits-all approach that declares one problem more complex than another without recognizing a subjective component of complexity that varies across people.[1] Our approach to thinking about complexity in terms of the distribution of perceptions it induces is similar to the approach of Bordalo et al. (2025) to modeling the emergence of biases in decision making as a two-step process, where people's responses to the salient features of a problem lead to a distribution of individual representations of it, which ultimately affects choices.

To define the subjective perception of the complexity of a problem we combine the Choice Process Protocol (CPP) of Caplin et al. (2011) with the 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 effort a subject puts into solving a problem, while confidence measures how certain she is about her solution.[2] On the basis of these two tools, we classify the perception of the complexity of a problem into four categories. In our classification, a problem is perceived as EASY if the subject exerts low effort when solving it and, at the same time, is very confident that she made the right decision (she need not actually choose correctly, it is enough that she feels she has). A problem is perceived as DOABLE if the person puts in relatively high effort and afterward is confident in her final choice. A problem is perceived as HARD if the person puts relatively high effort into trying to find the best decision and at the end of the process is still unsure of her decision (low confidence). Finally, we say that a problem is perceived as TOO HARD if the person gives up and does not engage in the problem by putting in low effort and expressing low confidence in her final decision.[3] It is the distribution across these four categories or perception classes that will serve as our main analytical tool.

[1] One notable exception is Enke and Graeber (2023) who link some behavioral patterns in valuations of risky lotteries and belief-updating tasks to the confidence in choices, elicited at the individual level.

[2] 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.

[3] Calling the last category TOO HARD implies that when facing the problem, the subject perceives it as too hard to attempt to solve it conscientiously. This does not necessarily mean that the problem requires a high cognitive ability to be solved, since subjects might give up on a problem, for example, if they are tired or disengaged.

Defining an objective measure of complexity is difficult because complexity is ultimately in the eye of the beholder. This does not imply that theoretical notions of objective complexity cannot be predictive of the distribution of subjective perceptions we discuss here. As we show, some of them may capture features that correlate with how people perceive them. However, the burden of proof of how relevant an objective measure is falls on the objective side, since behavior is ultimately determined by subjective perceptions.

In this experiment, we study behavior (the process of choosing, final choices, and confidence) in a series of individual choice tasks and games that have been thoroughly studied in behavioral and experimental economics: binary lottery choices, certainty equivalents of lotteries, simplicity equivalents of the deterministic mirrors of lotteries, tasks that require contingent reasoning, public good games, auctions, and belief updating tasks. In some of these tasks, objective measures of complexity reflect subjective perceptions, but in others they do not. For example, in binary lottery choices, the distribution of subjective perceptions changes between pairwise choices in the direction predicted by the objective complexity index of Enke and Shubatt (2023), suggesting that their objective measure, based on the excess dissimilarity between lotteries, captures the cognitive process driving the subjective perceptions we generate. However, in other cases, changes in objective notions of complexity do not affect subjective perceptions. For example, in the pivotality task of Esponda and Vespa (2014), the distribution of subjective perceptions is not different in the version of the task that requires subjects to engage in contingent reasoning and in the one where this is not necessary. However, within a task, different perceptions of complexity give rise to different choices.

In general, our results across this large set of tasks confirm our a priori intuition: Different problems lead to different distributions over perceptions of complexity, which in turn induce different distributions over behavior. This richness, based in heterogeneity, is lost when we only focus on aggregate behavior. Aggregate results sometimes reflect the behavior of only a subset of these perception classes and, therefore, the specific distribution over perception classes has a strong effect on what average or aggregate behavior is found to be. For example, in belief updating tasks, we document that the distributions of perceptions of complexity respond to changes in the priors, signal precisions, and whether the signal observed confirms or contradicts the prior. In terms of how these differences in perception affect behavior, we find that regardless of their subjective perception, people overestimate the probabilities of unlikely events and underestimate the probabilities of very likely events, a pattern that has been extensively documented in the literature. But the heterogeneity in subjective perceptions uncovers important differences that would not be observed otherwise. Focusing on the parametrization of Kahneman and Tversky (1972) that originally identified base rate neglect, we find that the 43% of subjects who perceive this task as EASY are the ones who drive the mistakes in updating in the direction of base rate neglect. This result is consistent with recent evidence by Esponda et al. (2023), who show that the base-rate neglect phenomenon is persistent, despite experience and extensive feedback, for those people with incorrect mental models, which would be consistent with

the interpretation of a subject who perceives this task as EASY and does not exert effort to find a solution.

A similar result, where one specific perception class is responsible for an anomaly, arises when we compare the certainty equivalents of lotteries and the simplicity equivalents of their deterministic mirrors (see Oprea (2024b)). Our results indicate that, as in Oprea (2024b), aggregating across all subjects, the average valuations of lotteries and their deterministic mirrors are similar and lower than the valuation of a risk neutral agent, and this phenomenon cannot be explained by risk attitudes since no risk is present in the mirrors. However, our methodology reveals that the low average valuations for mirrors are driven by subjects who perceive the task as TOO HARD, that is, those who do not exert effort in the task and know they have performed poorly. In contrast, the valuations of mirrors for the other three categories (EASY, DOABLE, and HARD) are statistically indistinguishable from the objectively correct value of the mirrors (the expected value of the corresponding lotteries).

Our methodology also allows us to uncover underlying behavioral mechanisms that might change the way we think about equilibrium behavior and overbidding in auctions. We show that for the four classic auction formats mostly studied in the literature (First-Price, Second-Price, Dutch, and English independent private value auctions), subjects who exert low effort and make intuitive choices (classified as perceiving the auctions as EASY or TOO HARD) tend to follow the simple and salient heuristic of bidding their own valuation, regardless of the auction format. In the Second-Price and English auctions this heuristic coincides with theoretical predictions, but in the First-Price and Dutch auctions it implies departures from equilibrium in the direction of overbidding. These findings raise the question of whether some of the observed equilibrium behavior in Second-Price and English auctions reflects the cognitive understanding of the solution to the problem or if it is just the result of a salient heuristic that coincides with theoretical predictions. Likewise, they suggest that some of the overbidding that has been widely documented in First-Price and Dutch auctions could, in part, be due to the behavior of these subjects. The rest of the subjects in FP and Dutch auctions, who exert higher effort (classified as perceiving the auctions as DOABLE or HARD), are more likely to set even higher bids than their valuations, suggesting that higher effort in these formats leads to choices that are further away from the equilibrium than simple heuristics.

In sum, our novel approach, which is highly portable across decision domains, highlights the importance of heterogeneity analysis and allows us to provide nuance to well-known experimental results and to revisit the predictive power of theoretical models.

In the remainder of the introduction, we review the relevant literature related to the measurement of task complexity. In Section 2 we present our measure based on subjective perceptions of complexity and discuss its interpretation and relation with existing measures. Section 3 describes our experimental design. Section 4 presents our results for each family of tasks, focusing first on the distribution of subjective perception classes induced by a task and then on the behavior associated to each of these classes. Section 5 contains robustness

checks and a discussion. Finally, Section 6 presents our conclusions.

Related Literature. 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.[4] This distribution of representations, via salience, leads to a distribution of final choices across people. Biases arise when relevant features are neglected due to not being salient to some.

2 Subjective Perceptions of Complexity

Before we present our experimental design, let us pause to introduce our measure of subjective perception of complexity and discuss why it is necessary. As alluded to in the introduction, the method we use to classify subjective perceptions is based on the combination of two experimental tools, each providing a non-choice measure of task complexity: the Choice Process Protocol of Caplin et al. (2011) and reported confidence in individual choices.[5] Let us describe these tools one at a time.

Choice Process Protocol (CPP). The use of the CPP is motivated by the observation that when people recognize difficult tasks, this is reflected in the cognitive effort they exert. If this is true, then effort measures could be used to track the complexity of the task. If a decision-maker exhibits high effort by either taking a long time to make a decision or, during her deliberations, changes her mind many times, we might be inclined to take those behaviors as evidence that the decision-maker found the problem complex.

The CPP provides us with these and other effort measures in the following way. Each round of the experiment starts with instructions about the rules of the task. Immediately afterward, participants observe a screen with a number of buttons, each representing a possible choice in this task. When a button is clicked, it stays selected until a participant clicks on another button, and there are no restrictions on the number of switches a subject can make throughout the round, which has a fixed time length that is known to participants (for details, see Section 3). Figures 13 and 14 in Appendix A present the screenshots of a binary lottery task for illustration.

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

[5] The CPP has been used also in Agranov et al. (2015) and Kessler et al. (2023).

To incentivize subjects, the payment in the CPP protocol is based on the choice made at a *random* second within the specified deliberation time. The exact second that matters for payment is drawn by the computer at the end of a task, and thus is not known in advance to participants. If a participant has not yet clicked on any button at the randomly selected second, then she would receive zero payment for this round. Since it is no longer only the final choice that is potentially rewarded, this payment scheme incentivizes participants to make what they perceive as the best decision at each time point. In particular, they are incentivized to make a quick first decision to avoid zero payment, and, whenever further thinking about the task causes people to revise their choices, they are incentivized to immediately implement this change, that is, click on a different button, to reduce the likelihood that their previous choice, which they came to realize is inferior, is chosen for payment.[6] Therefore, the CPP provides us with the whole thinking path of subjects for each decision problem, in addition to their final choices.

The CPP provides several effort measures: incentivized response time (total contemplation time, i.e., time to last click), the period of active consideration (difference between time to last click and first click), and the number of choice revisions made by a subject in a task (number of switches). The analysis in the paper uses response time as the CPP measure of effort for simplicity, but in the Online Appendix we show that the qualitative results are robust to using either of these three measures of effort and we discuss the connections between them. In addition, in Appendix B.1, we show that the CPP does not alter final choices in our tasks.[7]

Confidence. To measure how confident subjects are in their final choices, we ask people at the end of a round how certain they are, on a scale of 0 to 100, that the choice they made was the correct one for them. This measure is similar to cognitive uncertainty measure of (Enke and Graeber, 2023) and confidence measures used in Psychology.[8]

Measuring Subjective Perceptions of Complexity. The combination of these two measures, effort via the CPP and confidence self-reports, provides us with a way to classify how different subjects perceive different problems by recognizing heterogeneities in their choice process (how fast they reach a decision and how many times they change their mind) and the way they reflect on their performance in a task. We normalize effort and confidence at the individual level by taking the *average* effort and confidence of each subject

[6] See complete instructions in the Online Appendix.

[7] To make this point, we conduct an additional set of experiments in which participants are paid based on their final choices rather than their choices at a random second. We compare the distributions of final choices in these additional sessions with the CPP sessions and observe no significant differences across the two in any task we administered (see Figure 15).

[8] For instance, in De Martino et al. (2012), participants answer the question "How confident are you that the choice you made was the right one for you?" on a continuous sliding scale between 1 (low confidence) and 6 (high confidence) after every choice.

across all decision tasks that she encounters. This normalization is necessary because we are interested in classifying how people approach different problems relative to their own thinking style and confidence. So, for each subject, we define her personal average effort and say that she exhibits high (low) effort in a task if her effort is higher (lower) than her average effort across all tasks. Similarly, high (low) confidence in a task means that a subject's confidence is higher (lower) than her average confidence.

Equipped with these individual levels of effort and confidence, we define the four categories for subjects' perceptions of task complexity. We say that a task is perceived as:

1. EASY, when a participant exerts low effort (less than her average effort across tasks) and, at the same time, is confident that she made the right decision (higher reported confidence than her average across tasks).

2. DOABLE, when a participant exhibits lower effort than her own average and is confident in her final choice.

3. HARD, when a participant exerts high effort and reports low confidence in her decision.

4. TOO HARD, when a participant exerts low effort and shows low confidence in her final decision, i.e., when she gives up.

The labels of these four perception categories reflect our interpretation of how the decision process might influence the perception of a task's complexity. We consider a decision as a two-stage process. In the first stage, the decision maker observes the task and decides whether to actively think about it (exert effort) or not. Low effort can arise from two very different subjective perceptions: either the task is viewed as easy and hence requires little thought, or the task appears so difficult that the decision maker opts out (gives up) and exerts little to no effort thinking about it. For other tasks, the decision maker thinks the problem is tractable but not obvious and decides to exert effort 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. In the second stage, 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 evaluation provides us with a way to distinguish two different perceptions of a task, for a given effort exerted. As mentioned above, exerting low effort and making a quick choice can be due to perceiving the task as EASY, in which case the decision maker should be confident about this intuitive choice, or TOO HARD, in which case the decision maker gives up and reports a low confidence in her quick choice. Likewise, when subjects choose to exert high effort and engage in a process of thoughtful consideration, we can distinguish between tasks perceived as what we call DOABLE, when they feel confident about their

choice, and HARD, when they express low confidence in their choice. In a sense, exerting low effort in a task and choosing quickly (as in tasks perceived as EASY and TOO HARD) reflects an ex ante evaluation of the complexity of the task and might lead to more intuitive choices, while exerting high effort in a task and deciding when to stop (as in tasks perceived as DOABLE and HARD) implies an interim evaluation of the task's complexity, since the process of choosing also informs this perception.

To sum up, our categories incorporate elements of ex ante judgment of subjective task complexity (whether to engage thoughtfully with the problem or not), the endogenous evaluation of when to stop thinking, and the ex post evaluation of final choices. This is clearly just one possible interpretation of our categorization. Regardless of the interpretation, our two-by-two classification of subjective perceptions of task complexity is rich yet manageable, providing a new tool to understand the heterogeneity of individual perceptions of a task and how these perceptions affect choices in a variety of decision problems.

Why we need our measure. It is important to establish why we need both effort and confidence measures to understand subjective perceptions of complexity, rather than just one. Intuitively, effort alone might not distinguish between tasks that are easy and tasks that are so hard that people give up on them. 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 problems that are decided quickly by subjects may indicate that the subject finds that problem easy and hence solves it correctly very fast, or so hard that they do not attempt to solve it accurately. Similarly, decision confidence might not be enough to track individual perceptions of complexity. 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 (Grimaldi et al., 2015).

We illustrate the limitations associated with inferring task complexity using only effort or confidence in Table 1. We focus on five tasks in our experiment that have an objectively correct answer. We define choice accuracy as the proportion of subjects who arrived at the correct final choice at the end of a round and use this as an aggregate measure of the difficulty to solve the task. We then present, for each task, the choice accuracy broken down by quartiles for effort (top panel) or confidence (bottom panel). Clearly, choice accuracy is not monotonic with respect to either effort or confidence across tasks. In other words, exerting more effort does not necessarily lead to better choices and expressing more confidence in a choice does not necessarily reflect better performance. In addition, in Appendix B.2, we present a detailed analysis of tasks with objectively correct answers to characterize and further establish the non-monotonicity of both effort and confidence as individual measures of accuracy across these tasks.[9]

[9]In Appendix B.2, we show that the accuracy of choices increases over time for these tasks and establish an endogenous ordering of tasks to show that both confidence and effort levels across tasks are non-monotonic in choice accuracy. We also test a more nuanced prediction of Goncalves (2024) and show

Table 1: Accuracy of Choices in Tasks with Correct Answers, by Effort and Confidence

	FOSD	ESsimp mirrors	ESdiff mirrors	Pivotality (non-cont)	Pivotality (cont)
Effort					
Q1	1.00	0.95	0.39	0.33	0.13
Q2	0.99	0.82	0.54	0.28	0.28
Q3	0.98	0.88	0.64	0.56	0.37
Q4	0.99	0.82	0.58	0.59	0.38
Confidence					
Q1	0.99	0.78	0.52	0.43	0.31
Q2	0.99	0.86	0.50	0.27	0.21
Q3		0.92	0.56	0.47	0.32
Q4				0.59	

Notes: We report accuracy of final choices measured separately for 4 effort or confidence quartiles, computed separately for each task. Missing values for confidence indicate that there is a mass of people with the same confidence levels so it is not possible to distinguish between, say, Q2 and Q3 or Q4 in FOSD task since more than 50% of people reported a confidence level of 100. Effort is measured as total thinking time.

Despite the fact that our effort and confidence measures in isolation are unsatisfactory in capturing a task's complexity, combining them offers a way to observe how the task is perceived. This is achieved by taking advantage of the relevant information that each measure provides and combining them to overcome their individual limitations.[10] As mentioned above, we do so by categorizing people's perception of the decision problems into four distinct classes that are based on their own relative effort and confidence across the tasks they face: EASY, DOABLE, HARD, and TOO HARD.

This classification will be the workhorse for our analysis. For each problem that subjects encounter, we are interested in understanding how subjective perceptions are distributed across our four complexity categories. This distribution is important because it illustrates that complexity is necessarily subjective. Ultimately, our goal with this classification of subjective perceptions is to characterize how different problems lead to different distributions over perceptions of complexity 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 perceptions of the complexity of the problem, which is reflected in the distribution over our four categories.

that higher ability participants exert less effort in high-accuracy tasks but more effort in low-accuracy tasks, compared to those with lower ability.

[10] For example, ex-post confidence can help distinguish when a person exerted low effort in a task because they found it easy and when they gave up because it was too hard. Similarly, effort allows us to distinguish between quick and thoughtful responses, for a given level of confidence.

3 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 perceptions of complexity shape established results in the literature, e.g., do people free ride in a public goods game because they exert effort to make that choice and think it is the best choice for them or is it a quick and thoughtless decision? We also study tasks that have gained attention in recent years to study notions of complexity with the objective of understanding how their a priori notions and findings correlate to ours.

Subject Pool. We conducted our experiment on Prolific with a total of 976 participants between the ages of 18 and 65, who were living in the United States, were fluent in English, and had a high approval rating on 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).[11] 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 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.[12] 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.

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

[12] 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.

- **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 2. 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, L8 vs L12 and L10 vs L11, (CC1 and CC2 hereafter).

Table 2: 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 first-order stochastic dominance choice is simple and is included, in part, to make sure that the instructions and presentation of the lotteries were clear to the subjects. The other lottery choices are more 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 to the excess dissimilarity of the cumulative distribution functions of the lotteries in the choice set. According to this measure, if the lotteries 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 choice 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).[13]

[13] 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.

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. These two effects were originally proposed by Allais (1953) and are known in decision theory as primary deviations from expected utility.[14] 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[15]

- **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, it features disaggregated prizes but, unlike lotteries, involves no risk at all. As an example, imagine a set of 100 boxes, 50 of which contain \$20 each and the other 50 contain \$10 each. Say that this collection of boxes was offered to a person and the person was told that they would keep the average amount of the money contained in the boxes. There is no uncertainty, so all that needs to be done to determine the value of the boxes is to calculate its worth. In other words, this collection of boxes has an objective value of \$15, which is the same as the expected value of lottery L1 depicted in Table 2, 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 perceptions of complexity across different parametrizations of mirrors and, for a given parametrization, across a lottery and its deterministic mirror.

- **Contingent Reasoning.** Anticipating the consequences of one's actions is an integral part of the economic analysis of decision making. There is a growing experimental literature documenting the difficulty in performing contingent reasoning and how this tendency translates into mistakes in 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 con-

[14] 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., 2012a).

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

tingently 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 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.[16]

- **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 to obtain certainty equivalents of these lotteries to facilitate comparisons with the index of complexity of individual lotteries developed by Enke and Shubatt (2023) and the recent work by Puri (2024).[17] 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.[18] On the other hand, Puri (2024) suggests that decision makers may 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

[16] 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.

[17] 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.

[18] Indeed, L1 complexity index is 2.57, while L3 has 4.168 and L13 has 4.176. Higher numbers indicate higher complexity according to Enke and Shubatt (2023).

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 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., 2023; Augenblick et al., 2025; Ba et al., 2023; Agranov and Reshidi, 2024). In each task there are 100 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 3 contains the exact parameters we use. Participants were incentivized to reveal their posteriors honestly using a standard incentive-compatible BDM mechanism[19] 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., 2023; 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 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

[19]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).

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 α 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 α-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 3 displays all tasks and games performed in each treatment, which we call Binary 1, Binary 2, Non-Binary 1 and Non-Binary 2. Different subjects participated in each 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.[20]

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. According to the CPP, the choice selected in the chosen round at a 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.

4 Results

Approach to Data Analysis. We investigate the relationship between the distribution of perceived complexity and behavior in both our binary and non-binary tasks. The dia-

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

Table 3: 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

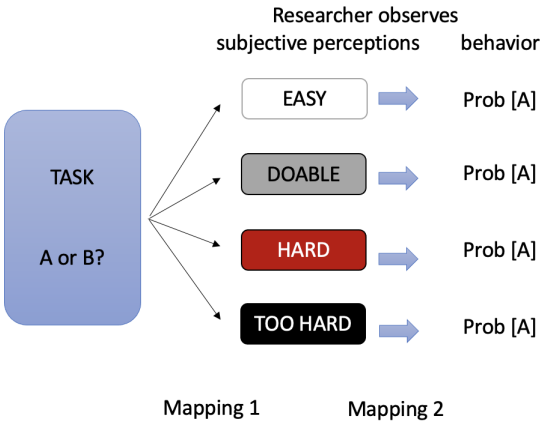
Notes: The exact parameters of lotteries are described in Table 2. For the belief-updating tasks, the pair (x, y) represents the parameters where x depicts the prior and y depicts the precision of a binary signal.

gram in Figure 1 presents the schematic way in which we approach the data analysis, using a binary task as an example. The toolbox described in Section 2 allows us to: (i) observe the otherwise unobservable subjective perceptions of task complexity and characterize the distribution over perception classes associated to each task, and (ii) study how behavior differs across these perception classes. Therefore, there are two mappings: one from the task to the distribution of subjective perceptions it induces (Mapping 1) and one from each perception class to choices in the task (Mapping 2). Whenever possible, we will also study how the subjective perceptions that we elicit relate to existing objective measures of the complexity of a particular task.

In what follows, each subsection presents the results of a different task in our experiment, and for each task we discuss these two mappings. That is, for each task, we will concentrate on how that task is perceived by subjects and how these perceptions affect behavior. Whenever such comparisons involve the same subjects completing both tasks, we use regression analysis and cluster standard errors at the individual level to account for the interdependencies of observations that come from the same subject. Otherwise, we use the Test of Proportions. We report $p-$values that indicate whether there is a significant difference in the two proportions.[21]

[21] There are a few participants who click uncontrollably in most of the tasks. We exclude them from the

Figure 1: Approach to data analysis (binary case)



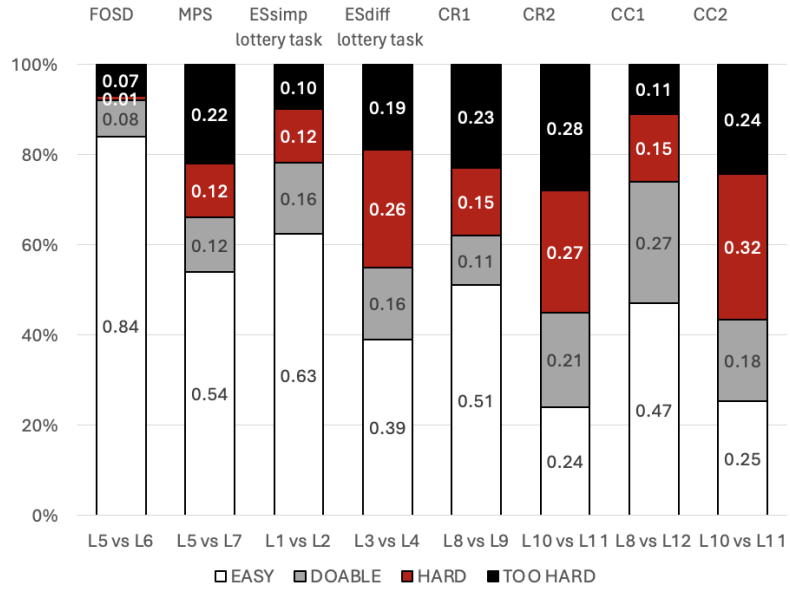
4.1 Binary Lottery Tasks

Subjective Perceptions of Binary Lottery Choices. To get an overview of how subjective perceptions vary across lottery choice problems, Figure 2 presents the distribution of perceived complexity by our subjects in all binary lottery choices. Notice how different binary lottery tasks induce very different distributions of perceived complexity. For instance, while over 80% of people perceive the first-order stochastic dominance task as EASY, less than 60% of people think that the mean preserving spread task is EASY ($p < 0.01$).[22] Similarly, 63% of subjects perceive the ESsimp lottery choice as EASY and 12% as HARD, compared to 39% who perceive the ESdiff lottery task as EASY (63% vs 39%, $p < 0.01$), and 26% who find it HARD (12% vs 26%, $p < 0.01$). The distributions of subjective perception of complexity that we elicit are thus confirmatory of the a priori index of objective complexity of Enke and Shubatt (2023) that relates the complexity of the choice problem to the excess dissimilarity of the cumulative distribution functions of the lotteries in the choice set.

analysis to minimize outliers. There are 14 participants like this in the Binary treatment (4% of our Binary sample). These subjects switch more than 15 times on average in each task, while 95% of subjects in the Binary tasks have less than 2 switches, on average. In the Non-Binary treatment, there are 32 participants who switch on average more than 30 times in each task (5% of our Non-Binary sample), while 95% of participants in Non-Binary tasks switch less than 10 times, on average. After excluding these participants, we are left with a total of 930 participants: 366 in the Binary and 564 in the Non-Binary treatments. Figure 6 in the Online Appendix depicts the distributions of perceptions of complexity in several tasks when we include these subjects, showing that they do not change our qualitative results. Table 2 in the Online Appendix presents the summary statistics of different markers of behavior of these outliers and compares them to the rest of the sample.

[22]The responses in the FOSD choice provide a first pass sanity check: over 99% of subjects make the 'correct' choice in this question and do this fast.

Figure 2: Distribution of Perceptions of Subjective Complexity in Binary Lottery Choices



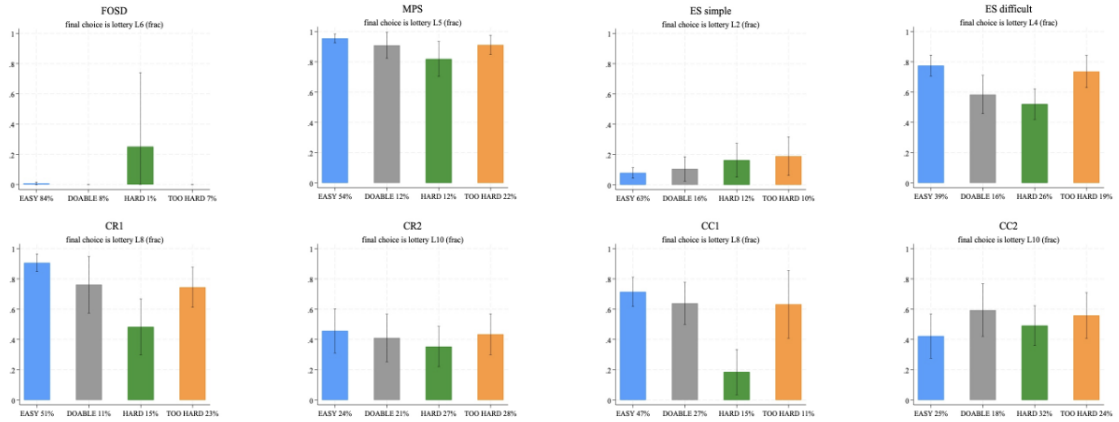
For the common consequence and the common ratio problems, subjects made two choices in random order, either CR1 and CR2, or CC1 and CC2. In both cases, the second question (CR2 and CC2) is perceived as more difficult than the first one (CR1 and CC1): almost 60% of participants perceive the second questions as either HARD or TOO HARD compared with less than 40% for the first questions ($p < 0.01$ in both comparisons).

In short, our measure of subjective complexity seems to capture differences in the way subjects perceive the binary lottery choices assigned to them.

Risk and Complexity in Binary Lottery Choices. Figure 3 presents, for each subjective perception within a task, the fraction of subjects who made the risk-averse (safer) choice between the two lotteries presented to them in each task. We use the Sharpe ratio to determine which lottery in a pair is less risky.

Some interesting results are shown in Figure 3. First, for the three choice problems that have the highest percentage of people who perceive them as EASY (FOSD, MPS, and ESsimp lottery task, see Figure 2), while perceived complexity can vary greatly between subjects (Mapping 1), behavior appears to be almost invariant to subjective perceptions of complexity (Mapping 2). In particular, in the lottery choice with FOSD, the safe choice is the dominated lottery, and almost no subject makes this choice, for all subjective perceptions. In the choice with a MPS, almost all people choose the safer lottery over the more dispersed one. The same is true in the ESsimp lottery task, where the vast majority of peo-

Figure 3: Low Risk Final Choices as a Function of Perceptions in Binary Lottery Tasks



Notes: For each task, we plot the fraction of final choices with a higher Sharpe's ratio (higher Sharpe's ratio indicates safer choices). The frequencies of perceptions are in the horizontal axes.

ple, regardless of their perception, choose the riskier lottery, which is the more attractive in terms of prizes. Although there are noticeable differences in the way subjects perceive the complexity of FOSD, MPS, and ESsimp lottery tasks, within each choice problem, the responses to these perceptions are largely the same. Therefore, we can conclude that for these simpler choice problems, behavior is governed more by Mapping 1 (task perception) than Mapping 2 (task behavior conditional on perception).

The situation changes for more intricate lottery choices. In the ESdiff lottery task, as well as CR1 and CC1, safe choices seem to be related to intuitive, less thoughtful choice processes where subjects exert low effort.[23] On the flip side, choosing the risky option seems to require a more deliberate decision where subjects exert high effort (those who perceive the problem as DOABLE or HARD). This effect is particularly strong for those who perceive the tasks as HARD, i.e., subjects who, on top of exerting high effort, do not feel confident about their final choice.

Our results suggest that the objective complexity measure of Enke and Shubatt (2023), together with our heterogeneity analysis, might be complementary in explaining behavior in binary lottery choices. Enke and Shubatt (2023) document a reduced sensitivity to expected value differences in the lotteries in the choice set as the complexity of the problem increases. This could imply that, in more complex environments, people might simplify their decision making by favoring safer (less risky) options, as these are easier to justify or

[23] Our results raise an interesting question of how people behave when they admit that a problem is too difficult for them to solve. While in some problems the heuristic they use leads them to take risks, in other contexts it leads them to play it safe.

understand.

The behavior of our subjects is consistent with this interpretation. Focusing on the ESdiff lottery task, there is clearly a higher propensity to choose the safer lottery than the riskier one, especially among subjects who exert low effort (77% for those who perceive the problem as EASY and 74% for those who perceive it as TOO HARD).[24] Intuitively, these are the types that would be more likely to simplify their decisions along the lines of Enke and Shubatt (2023). Subjects who perceive the task as DOABLE or HARD, on the other hand, exert more effort in their deliberations and, as a result, are more balanced in their choices, reflecting preferences (58% and 52% of safer choices, respectively). This result, however, is not present in the ESsimp lottery task, suggesting an interesting relationship between our measure based on subjective perceptions of complexity and the measure of objective complexity of Enke and Shubatt (2023): For decision problems that are objectively more complex, subjective perceptions of complexity affect the propensity to simplify the decision process and favor safer choices.

For the remaining lottery problems, CR2 and CC2, which are, in fact, the same question (L10 vs L11), we find that, for all perception classes, about half of the subjects choose the riskier lottery, thus suggesting that Mapping 2, from perceptions to choice, is not responsible for observed behavior.

In common ratio and common consequence problems, subjects fall prey to the Allais paradox if they make inconsistent choices in two lottery choice problems (CR1 and CR2 or CC1 and CC2).[25] In our analysis, the question arises as to who is responsible for these inconsistent choices, i.e., can they be attributed to subjects with different perceptions of the problem's complexity.

Table 3 in the Online Appendix (Section 3) reports the results of a series of regressions that explore the relationship between choice inconsistency in the Common Ratio questions and subjective perceptions of complexity, controlling for having made the safer choice in CR1, which is highly correlated with inconsistencies across the two questions (see McGranaghan et al. (2024a)). Our results indicate that the perception of CR1 is an important determinant of choice inconsistency: People who perceive CR1 as EASY are less likely to exhibit inconsistencies that imply violations of expected utility than those who perceive it in any other way, and, in particular, they are less likely to exhibit standard CRE-type behavior. Moreover, Figure 7 in the Online Appendix (Section 3) suggests that this is particularly true for subjects who perceive CR1 as EASY and CR2 as either EASY or TOO HARD, in other words, those who do not exert effort in the second question. This suggests that consistency in the Common Ratio questions is more likely to arise when people perceive the problem as EASY or TOO HARD and hence make more intuitive (faster) choices, and that violations might be related to more effort (over thinking). This result suggests that the inconsistent behavior so often observed when discussing the Allais

[24] Note that in this task it appears that both Mapping 1 and Mapping 2 are responsible for our results.

[25] Choices in CR1 and CR2 are consistent with expected utility if the responses to both questions are either for the safer lottery or for the riskier one. The same is true for CC1 and CC2.

Paradox may not be a general phenomenon but rather one determined by those subjects who perceive the problem in a particular way (as either EASY or TOO HARD). For the Common Consequence effect, we observe no such relationship (Table 4 in Section 3 of the Online Appendix).

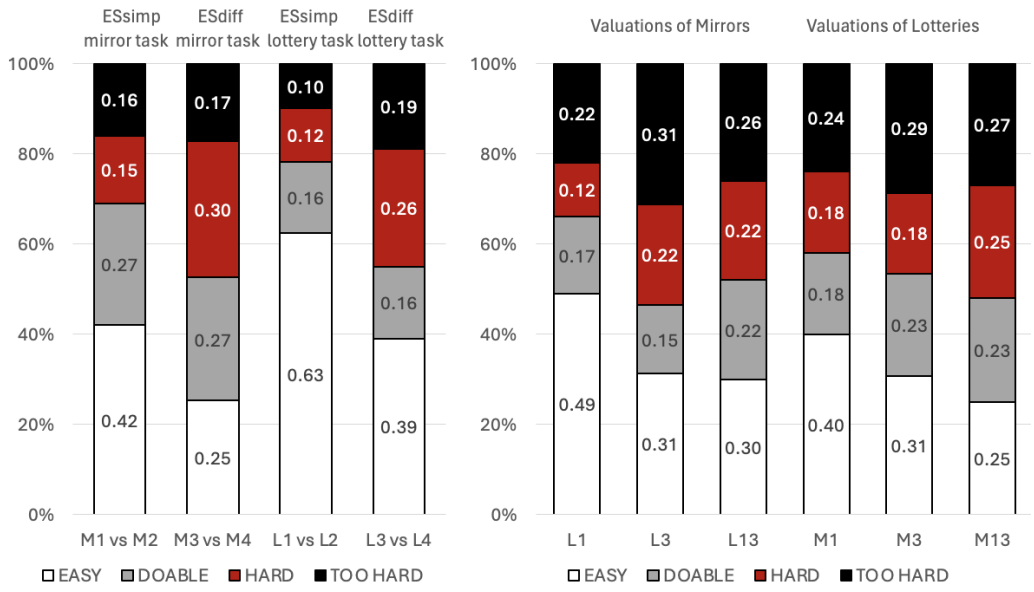
4.2 Lotteries and their Deterministic Mirrors

Studying the behavior of our subjects when they choose between pairs of lotteries or their deterministic mirrors and when they value these objects individually with the certainty and simplicity equivalents provides a perfect opportunity to illustrate how our two mappings, from task to perception and perception to choice, enrich our understanding of choice behavior. We first discuss the distributions of subjective perceptions of complexity of the different binary choices and valuation tasks (Mapping 1) and then proceed to discuss how the heterogeneity of these perceptions affects choices (Mapping 2).

4.2.1 Distributions of Perceived Complexity (Mapping 1)

Figure 4 presents the distributions of subjective perceptions of complexity for binary choices and valuation tasks for lotteries and their deterministic mirrors.

Figure 4: Distribution of Perceived Complexity: Lotteries vs Mirrors



Subjective Perceptions of Binary Lottery Choice tasks. The left-hand panel in Figure 4, which presents the distributions of subjective perceptions of the ESsimp and ESdiff lottery and mirror binary choice tasks, suggests two clear patterns. First, the objective complexity index of Enke and Shubatt (2023) is predictive of the distribution of subjective perceptions of mirror choices. Just as in the case of lotteries, significantly more people perceive the ESsimp mirror task as EASY and less people perceive it as HARD than the ESdiff mirror task (42% vs 25%, $p < 0.001$ for EASY, and 15% vs 30%, $p < 0.001$ for HARD, respectively). Second, for a given parametrization (ESsimp or ESdiff) we find a stark difference in the perception of lottery and mirror binary choices. In particular, more people perceive lottery choices as EASY than mirror choices (63% vs 42%, $p < 0.001$ for the ESsimp parameterization, and 39% vs 25% $p < 0.001$ for the ESdiff parametrization). There are slightly less people who perceive lottery choices as HARD, compared to the equivalent mirror choices, but the differences are not statistically significant (12% vs 15% for the ESsimp and 26% vs 33% for the ESdiff parameterizations, respectively). In other words, people are more likely to make intuitive choices in lotteries (faster and more confident) than in the corresponding mirrors. This could be thought as surprising, since lottery choices involve risk, while mirror choices are deterministic.[26]

Subjective Perceptions of Valuation Tasks. Comparing the distributions of subjective perceptions of the valuation task of individual lotteries and mirrors (right panel of Figure 4), we see similar, but less stark patterns. 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]

Just as in binary choices, our measure is consistent with existing objective measures of complexity for both lotteries and mirrors, i.e., more people perceive the valuation of L1

[26] This evidence is at odds with the following decision-making process: when comparing two lotteries, 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 complex than the evaluation of mirrors because of the additional dimension of risk, but this is not what we find. In contrast, people seem to find lottery comparisons easier than those of mirrors. However, this evidence is consistent with the idea that in the mirror tasks there is one correct answer, while in the lottery task there is not, which could be driving differences in perceived complexity across lotteries and mirrors.

[27] While stating a valuation of either a lottery or a mirror forces subjects to reduce the object to a single number, choosing between lotteries or mirrors might imply taking an alternative approach, as suggested by Kahneman and Tversky (1973) and Rubinstein (2006), who posit that decision-makers compare lotteries by comparing their attributes or dimensions one by one and looking for similarities and differences (see also Bordalo et al. (2012b), which does not require any aggregation, unless the attributes that the decision-maker compares are expected values or other moments of the distribution of lotteries.

as EASY compared to L3 (49% vs. 31%, $p < 0.001$) and the same is true for M1 and M3 (40% vs 31%, $p = 0.01$).[28] Increasing the support of the lottery to 3 possible prizes as is the case in L13 leads to fewer people finding it EASY than when there are 2 prizes as is true for lottery L1 (as Puri (2024) suggests). However, perceptions are similar between L3 and L13, consistent with Enke and Shubatt (2023).[29] The same is true for the valuation of mirrors.

4.2.2 Perceived Complexity and Choices (Mapping 2)

We now look at how the different distributions of subjective perceptions of complexity affect behavior. We start with the valuations of lotteries and mirrors (certainty and simplicity equivalents) from our Non-Binary treatment and then discuss choices between lotteries and mirrors in our Binary treatment.

Valuations of Lotteries and their Deterministic Mirrors. Table 4 presents the certainty equivalents and simplicity equivalents elicited from our subjects for three lotteries and their deterministic mirrors, which have identical expected values (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 perception class, i.e., according to whether they perceived the task as EASY, DOABLE, HARD or TOO HARD.

What we find is very revealing and captures the importance of heterogeneity analysis, which is one of the main contributions of our approach to subjective complexity. First, when we look at the average certainty and simplicity equivalents of all subjects, regardless of their subjective perception (Aggregate row, in the first two horizontal blocks of Table 4), we replicate the results of Oprea (2024b): On average, the certainty and simplicity equivalents of 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 there is no risk in the mirrors. However, our approach allows us to dig deeper into the observed behavior by exploring whether this result is universal across subjects or if it is related to specific perceptions of the complexity of the task. Table 4 suggests that, for all mirror specifications, the low average valuations are mainly driven by people who perceive the task as TOO HARD, i.e., who give up and do not put effort into thinking about the task and know their final choice is not correct. At the same time,

[28] 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 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 3).

[29] 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 4: Valuations of Lotteries and their Deterministic Mirrors depending on Perceived Complexity

	Mirror M1		Mirror M3		Mirror M13	
	mean (se)	exp value = 15	mean (se)	exp value = 15.8	mean (se)	exp value = 15.8
Aggregate	14.4 (0.28)	$p = 0.03$	14.9 (0.35)	$p = 0.01$	14.7 (0.30)	$p < 0.01$
EASY	14.4 (0.38)	$p = 0.14$	14.7 (0.64)	$p = 0.10$	14.5 (0.65)	$p = 0.05$
DOABLE	15.4 (0.66)	$p = 0.56$	17.1 (0.48)	$p = 0.01$	16.0 (0.47)	$p = 0.65$
HARD	14.6 (0.82)	$p = 0.64$	15.3 (0.93)	$p = 0.56$	15.8 (0.51)	$p = 0.99$
TOO HARD	13.4 (0.59)	$p = 0.01$	13.1 (0.71)	$p < 0.01$	12.8 (0.67)	$p < 0.01$
	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)	$p < 0.01$	12.6 (0.42)	$p < 0.01$	14.1 (0.36)	$p < 0.01$
EASY	13.9 (0.40)	$p = 0.01$	12.7 (0.86)	$p < 0.01$	14.6 (0.69)	$p = 0.08$
DOABLE	14.4 (0.85)	$p = 0.52$	14.2 (1.08)	$p = 0.13$	14.0 (0.82)	$p = 0.03$
HARD	13.3 (1.01)	$p = 0.10$	11.1 (0.73)	$p < 0.01$	13.4 (0.69)	$p < 0.01$
TOO HARD	13.9 (0.67)	$p = 0.11$	12.7 (0.71)	$p < 0.01$	14.4 (0.70)	$p = 0.05$
	L1 vs M1		L3 vs M3		L13 vs M13	
Aggregate	$p = 0.20$		$p < 0.01$		$p = 0.07$	
EASY	$p = 0.29$		$p = 0.05$		$p = 0.96$	
DOABLE	$p = 0.36$		$p = 0.01$		$p = 0.03$	
HARD	$p = 0.29$		$p < 0.01$		$p = 0.01$	
TOO HARD	$p = 0.47$		$p = 0.64$		$p = 0.09$	

Notes: Data from the Non-Binary Treatment 1. The p-values comparing observed average valuations for mirrors and lotteries and comparing each of them with the theoretical values come from regression analysis.

the other three categories of participants in general report simplicity equivalents that are statistically indistinguishable from the correct values theoretically predicted.

The situation is very different in the valuation of lotteries. Here, we do not observe such a pattern that suggests that one particular perception class drives aggregate results. Instead, we see that the valuations of most perception classes across lotteries are below the risk neutral values, indicating a familiar pattern of risk aversion for small stakes.[30]

Finally, in the third horizontal block of Table 4, we compare the valuations of lotteries and mirrors, for each parametrization, to further assess their equivalence. We find two main results. First, for the L1/M1 parametrization, deemed less complex by the index of Enke and Shubatt (2023) (perceived as EASY more frequently than the other two parametrizations, see Figure 4), the valuations of this lottery and its corresponding mirror are indistinguishable from each other across all perception classes. However, for more difficult parameterizations (L3/M3 according to Enke and Shubatt (2023) and L13/M13 according to Puri (2024)), mirrors and lotteries are, on average, valued differently in the aggregate. Looking deeper into our perception classes, we see a clear pattern that suggests that people that exert high effort in this task (who perceive it as DOABLE or HARD)

[30]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.

value mirrors differently than lotteries. In other words, subjective perceptions of complexity reveal that people value lotteries and their corresponding mirrors differently when they choose to exert significant effort in their evaluation *and* when the object at hand is objectively not trivial to evaluate (either because it has a large variance or a larger support).

Binary Choices of Lotteries and Mirrors. We now turn 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 lottery with the highest risk, or the corresponding mirror, in each pair of binary choices (ESsimp and ESdiff). Similar to the results we presented for valuations of lotteries and mirrors, notice that when the binary choice is simple (according to the index of Enke and Shubatt (2023) and perceived that way, see left panel of Figure 4), subjects choose among mirrors in the same way that they do among lotteries, for all subjective perceptions. However, when the task is perceived as harder (ESdiff), the equivalence in choices between lotteries and mirrors depends on their subjective perception of the complexity of the problem. In this case, it is via confidence: People who report low confidence in their choices (who perceive the problem as HARD or TOO HARD) make similar choices in lotteries and mirrors, but people who feel confident in their choices (who perceive the problem as EASY or DOABLE) choose differently when the problem involves a lottery or a deterministic mirror.[31]

Table 5: Binary choices of Lotteries and Mirrors

	Prob of choosing L1 (riskier) or M1			
	EASY	DOABLE	HARD	TOO HARD
ESsimp lottery task	0.92	0.89	0.84	0.81
ESsim mirror task	0.92	0.89	0.75	0.81
Lotteries vs Mirrors	$p = 0.97$	$p = 0.91$	$p = 0.28$	$p = 0.96$
	Prob of choosing L3 (riskier) or M3			
	EASY	DOABLE	HARD	TOO HARD
ESdiff lottery task	0.23	0.42	0.48	0.26
ESdiff mirror task	0.53	0.65	0.53	0.38
Lotteries vs Mirrors	$p < 0.01$	$p < 0.01$	$p = 0.49$	$p = 0.16$

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.

Our results illustrate the nuanced relationship between lotteries and their deterministic mirrors from two different angles, valuations and binary choices. By studying how the

[31] 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.

distributions of subjective perceptions of complexity 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 subjective complexity measure. People who exert a high effort in their deliberation value lotteries differently to mirrors. In binary choice problems, people who report high confidence in their choices tend to choose differently when the problem involves lotteries or mirrors.

4.3 Contingent Reasoning

To study contingent reasoning, we use the pivotality task of Esponda and Vespa (2014) where one member of a committee votes along with two independent computers whose decision rules they are informed of. In one treatment, the decision maker needs to engage in contingent reasoning because the choices of the two computers are not known, in the other treatment these choices are observed so there is no need to engage in contingent reasoning. We refer to these as the "contingent" and "non-contingent" treatments, respectively. In both tasks, there is one objectively correct answer (see Section 3 for details). Our purpose is to understand how these two versions of the same decision problem are perceived by subjects and how these perceptions influence their ability to arrive at the correct decision.

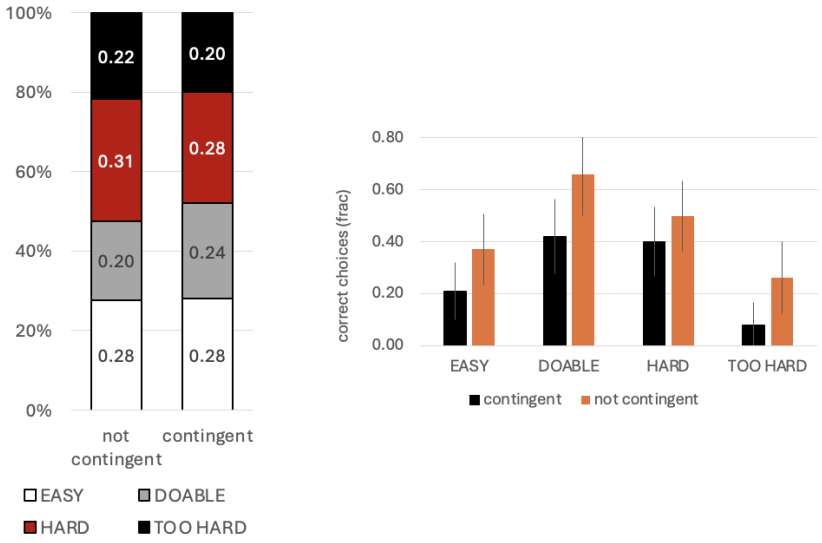
In Figure 5 we present the distribution of perceptions of complexity in the left panel and the fraction of correct final choices for subjects in each perception category in the two versions of the pivotality task.

First, notice in the left panel of Figure 5 that the distribution of perceived complexity does not seem to vary across its contingent and non-contingent versions. While one task requires subjects to engage in contingent thinking, which could be thought of as objectively more complex, and the other does not, this fact seems lost on our subjects. Hence, if subjects behave differently across these two treatments, we cannot ascribe it to their different perceptions of the problem (Mapping 1).

Second, despite their similar perceptions, as was discovered in Esponda and Vespa (2014), people are more likely to make the right decision in the task that does not require contingent reasoning. We show that this is true for all perceived complexity classifications in the right panel of Figure 5. That is, for a given perception of complexity, the frequency of correct choices is higher in the non-contingent version of the task than in the contingent version ($p < 0.01$ for all perception classes).

When we look at the proportion of correct choices across perception classes, it becomes clear that exerting effort leads to better choices in both versions of the task, regardless of their opinion about their performance. That is, for a given level of confidence, those who put more effort into thinking about the task perform better than those who do not, whether the task requires contingent reasoning or not.

Figure 5: Perceived Complexity and Final Choices in Pivotality Tasks



4.4 Public Good Games

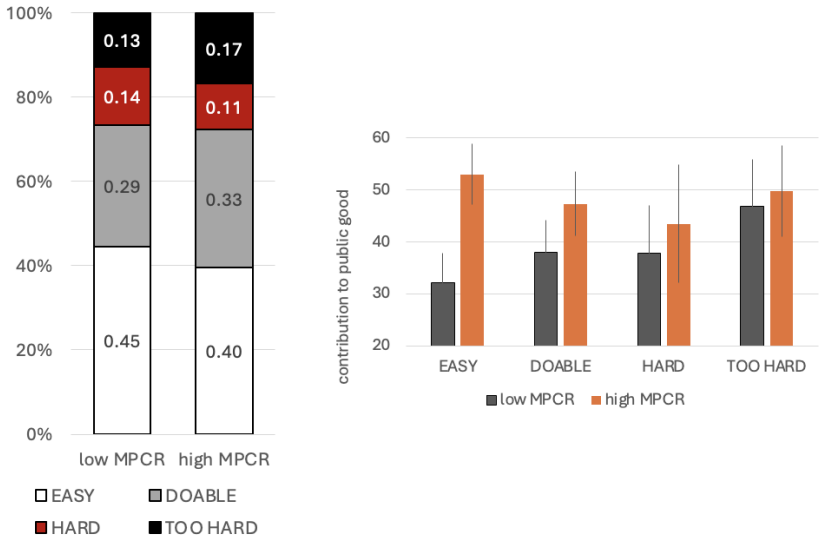
To study behavior in a public good game through the lens of subjective perceptions of complexity, Figure 6 presents, on the left panel, the distribution of perceived complexity (Mapping 1) in the two versions of the game (low and high MPCR), and, on the right panel, the final contributions to the public good in both treatments, by perception class (Mapping 2).

The first thing to notice in Figure 6 is that the two versions of the public good game presented to our subjects led to similar perceptions of complexity. This can be rationalized since the two problems are strategically equivalent and involve the same dominant strategy equilibrium. However, despite the similarity of the perception distribution, these two games generated very different behavior. For example, subjects in the high MPCR game contributed significantly more of their endowment to the public good, on average, and also within perception classes.

The pattern of contributions across these games is especially interesting. People who perceive the games as EASY are much more responsive to the game primitives and salient features of the environment. These lead participants to contribute almost twice as much in the game with the high MPCR than the low one.[32] The difference in contributions across

[32]Figure 18 in the Appendix presents the evolution over time of contributions to the public good, by perceived complexity, for low and high MPCR. Subjects who perceive the game as EASY consistently choose lower contributions than all other types across the consideration period for the low MPCR treatment, while the opposite is true for the high MPCR treatment, which might reflect different intuitive responses to the

Figure 6: Perceived Complexity and Average Contributions in Public Good Games



Notes: The left figure depicts the distribution of perceived complexity in two versions of the public good game. The right figure depicts the final choices by perceived complexity, with 95% CI.

the two treatments decreases when subjects exert more effort, i.e., when they perceive the game as DOABLE or HARD. Notice that those subjects who perceive the game as TOO HARD and thus do not engage actively with the task, are not sensitive to the parameters of the game at all and display the same behavior across games, contributing roughly half of their endowment to the public good, regardless of the MPCR.

4.5 Auctions

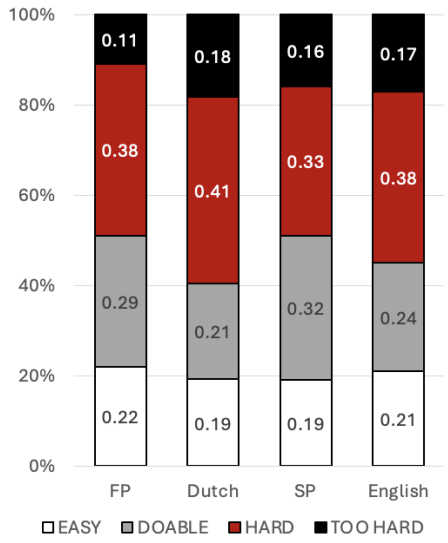
Along with pubic goods, the experimental investigation of auction mechanisms has attracted a lot of attention in the past few decades (Kagel, 1995). In this literature, two types of question have dominated. One is whether the predictions of the Nash Equilibrium theory are valid for the four major types of auctions typically studied: the Risk-Neutral Nash Equilibrium (RNNE) for the First-Price (FP) and Dutch auctions, and the Nash Equilibrium in dominant strategies for the Second-Price (SP) and English auctions. The second question pertains to the revenue equivalence of these auction formats, since the FP and Dutch auctions are isomorphic and the SP and English auctions are isomorphic, and the four auction formats are revenue equivalent in theory.

parameters: A low MPCR tilts the trade-off between opportunistic and prosocial behavior towards smaller contributions, while the opposite is true for a high MPCR.

We find evidence of the common pattern found in the literature: subjects tend to bid higher than the RNNE predictions in FP and Dutch , and Second-Price auctions but close to that prediction in the English auction. Figure 19 in the Appendix displays the final bids in the two pairs of auctions. As can be seen, subjects often bid higher than the RNNE in the FP and Dutch auctions and overbid in the SP auction, with less overbidding in the English auction than in the SP auction.[33]

Our methodology allows us to study whether these results can be attributed to the way these different auctions are perceived by our subjects. We present the distributions of subjective perceptions of complexity for each auction format in Figure 7. First, notice that the four auction formats give rise to very similar distributions of perceptions of their complexity. Therefore, differences in behavior are unlikely to be a result of Mapping 1. Notice also that, for all auctions, the vast majority of people find the auctions either DOABLE or HARD, that is, no matter the auction format most people exert a high effort in thinking about them.

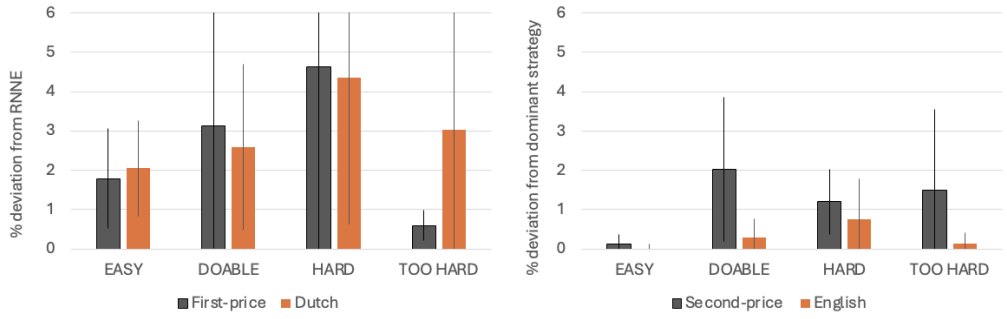
Figure 7: Distribution of Perceived Complexity



To understand how differences in perception affect final bids (Mapping 2), Figure 8 presents deviations of the average bids made by subjects in each perception class from

[33]Note that in our experiment the only difference between the FP (SP) and the Dutch (English) auction is the way it is framed. The Dutch and English auctions were not conducted as oral auctions in which bidders see the behavior of others and respond to it in real-time. Instead, all auctions were conducted as sealed-bid with the only difference being the way we described the rules (see the Online Appendix for instructions)

Figure 8: Final Bids by Perceived Complexity



Notes: We report average percentage deviation of observed bids from equilibrium predictions. Positive values correspond to over-bidding and negative values correspond to under-bidding.

the RNNE bid for FP and Dutch auctions and from the dominant strategy of bidding one's valuation in SP and English auctions. Positive numbers indicate that subjects were bidding, on average, above the equilibrium bids, while negative values indicate the opposite. Because overbidding is such a common phenomenon, there are no negative entries. Zero indicates exact equilibrium bids.

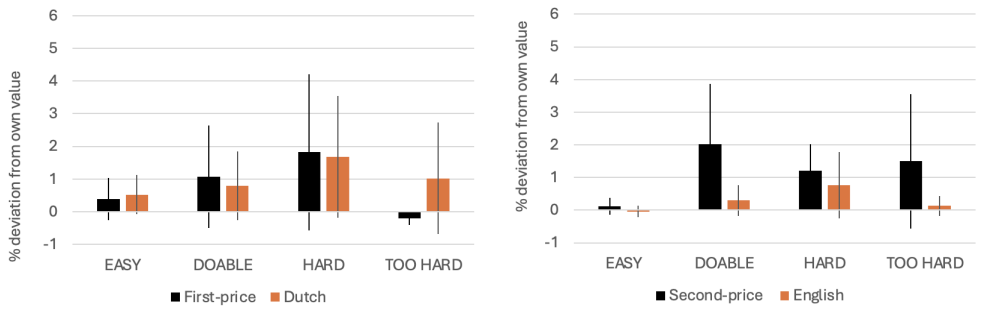
The first thing to notice in Figure 8 is the difference between the two formats that have dominant-strategy equilibria (the SP and the English auctions) and those that do not (the FP and Dutch auctions). Despite perceiving the complexity of all these auctions similarly, perception-type by perception-type, bids are closer to the equilibrium bid in the SP and English auctions than in the FP and Dutch auctions.

In addition, those subjects who exert low effort in the English Auction (classified as EASY or TOO HARD) bid, on average, almost identically, with both types bidding their values. This illustrates an interesting aspect of the relationship between perception and choice. For those who find the English auction EASY, we might think that it appears obvious that bidding one's value is optimal. Those who find it TOO HARD and give up, however, need to fall back on some heuristic that seems salient in order to bid, and, in this case, bidding one's valuation seems to be the default. However, this raises the question of whether bidding one's valuation might be a salient heuristic for those subjects who make quick, intuitive choices, across auction formats.

To investigate whether subjects who exert low effort (classified as EASY or TOO HARD) follow a simple heuristic of bidding their valuation, Figure 9 presents the average deviations of final bids to own valuations, for all auction formats. Remarkably, subjects who perceive all auction formats as EASY bid according to their valuation. This is also the case for subjects who perceive the auctions as TOO HARD, although with more noise for the Dutch and SP auctions. These results illustrate not only that fast, intuitive choices

might follow the natural heuristic of bidding one's valuation and that subjects feel confident about them, but they also that bidding behavior alone cannot identify the cognitive understanding of the problem. In the case of SP and English auctions, this heuristic corresponds to the equilibrium prediction, so the conclusion of a high proportion of equilibrium bidding in these auction formats might actually hide the fact that the natural heuristic (used by the intuitive subjects across auction formats) coincides with equilibrium. In other words, when playing the SP or English auctions, these subjects behaved optimally by luck. For the FP and Dutch auctions, however, this heuristic implies higher bids than the RNNE, which can explain, at least in part, the overbidding behavior documented extensively in the literature (Kagel (1995)).

Figure 9: Final Bids Relative to Own Value, by Perceived Complexity



Notes: We report average percentage deviation of observed bids from individual valuations.

To further understand overbidding in FP and Dutch auctions, if we look at Figures 8 and 9 together, they seem to suggest that high effort (subjects who perceive the auctions as DOABLE and HARD) is related to more pronounced overbidding than low-effort subjects who overbid by choosing their own valuations. For these formats, this suggests that exerting high effort not only does not lead to more equilibrium behavior, but it leads to choices that are further away from the equilibrium choice than simple heuristics.

In summary, our results suggest that subjects who exert low effort and make intuitive choices (classified as perceiving the auctions as EASY or TOO HARD) tend to follow the simple and salient heuristic of bidding their own valuation. In some auction formats (FP and Dutch), these heuristics imply departures from equilibrium predictions in the direction of overbidding, but in auction formats where these heuristics coincide with equilibrium predictions, some of the observed equilibrium behavior might not reflect the cognitive understanding of the solution to the problem. In FP and Dutch auctions, those who exert higher effort overbid even more.

4.6 Belief-Updating Tasks and Base-Rate Neglect

Bayesian updating is the canonical procedure prescribed for updating beliefs. Despite being the rational way to update beliefs, it is well documented that people have difficulties updating beliefs as Bayes suggested (see Benjamin (2019) for a survey of this literature). What is not well documented is which type of decision maker is most vulnerable to the common mistakes involved in belief updating (for instance, base-rate neglect). Our interest in studying belief updating tasks relies on understanding how the perception of a task influences the type of updating that the subject engages in. Put differently, we are interested in understanding if deviations from Bayesian updating are more likely to arise for subjects with specific perceptions of the task's complexity.

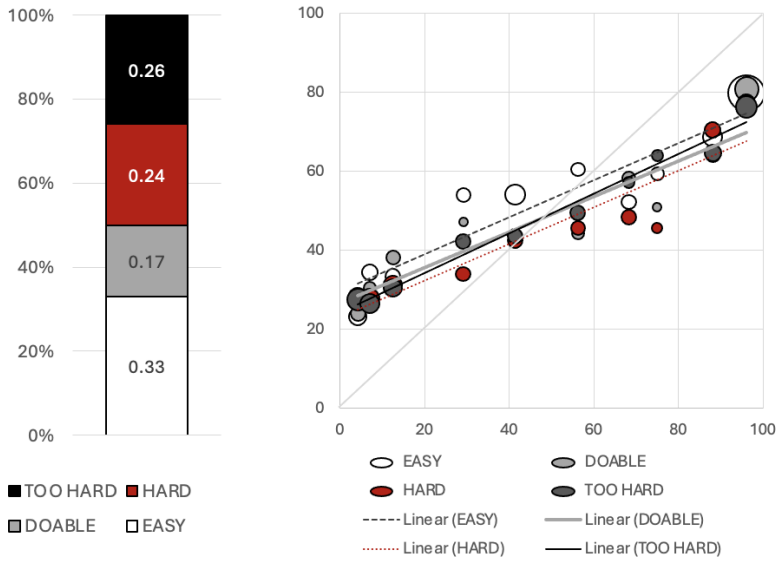
In our experiment each belief updating task can be indexed by a triple (p_0, q, s) and a binary state space $\omega \in \{0, 1\}$, where $p_0 = \Pr[\omega = 1]$ represents the prior, i.e., the probability that the state is positive (indexed by 1), $q = \Pr[s = \omega|\omega]$ represents the signal precision, and $s \in \{0, 1\}$ depicts the signal realization. We used six different belief updating tasks each with a different parametrization with two possible signal realizations 0,1, for a total of 12 cases.[34]

Overestimating small probabilities and underestimating large probabilities. We first present Figure 10 which, in the left panel, presents the distribution of subjects over our four perception categories aggregated over the 12 updating tasks, and in the right panel depicts the relationship between the posterior beliefs of these subjects and the Bayesian benchmark, again aggregated over all of our subjects and parametrizations. Looking at the left panel first, we observe a clear heterogeneity in the perception of our tasks. Despite these diverse perceptions, when it comes to updating, we see a pattern, familiar in the literature, in the way all perception types update their beliefs, as can be seen by the estimated linear updating functions in the right panel: regardless of their subjective perception, 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 in the function used to update across any of these types.[35] We confirm that the pattern found in the literature is robust and not dependent on subjective perceptions of the task's complexity.

[34] The different parameters were chosen to relate our findings to the literature on base rate neglect (see Kahneman and Tversky (1972) and, more recently, Esponda et al. (2023)) and to existing measures of objective complexity proposed by Agranov and Reshidi (2024).

[35] Enke and Graeber (2023) do a similar exercise but differentiating subjects according to their reported confidence and find significant differences across high and low confidence, indicating that higher confidence is associated with less departures from Bayesian updating. Figure 20 in the Appendix replicates their result, focusing only on confidence. Our measure of subjective complexity allows us to differentiate even further by focusing also on effort, therefore disentangling differences in perception that arise even when subjects report similar confidence.

Figure 10: Belief-updating tasks



Notes: The left panel presents the distribution of perceived complexity across all updating tasks. The right panel depicts final posteriors as a function of Bayesian posteriors for each subjective perception.

Base-rate neglect. Base-rate neglect is one of the most well-documented biases in the belief-updating literature. This phenomenon is characterized by under weighting information contained in the prior when encountering new information.[36]

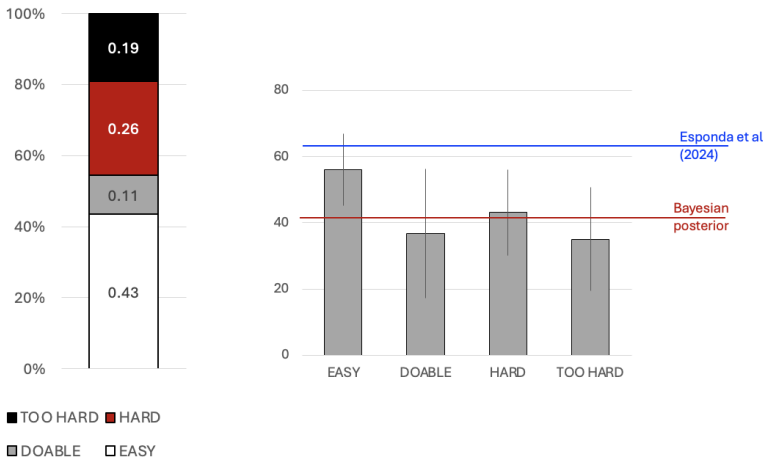
To illustrate this bias, we focus on the parameterization used originally in Kahneman and Tversky (1972) and in many follow-up papers, included in our experiment. The classic example is to imagine a person who is tested for a disease which has a prevalence of 15% in the general population and the test used for diagnosis has an accuracy of 80%. With these primitives, the chance that the person is sick, conditional on a positive test result, is 41%, but a very robust finding is that many subjects (and doctors!) incorrectly consider this chance to be much higher. Focusing on learning dynamics, Esponda et al. (2023) show that the average reported posterior is around 60% the first time people encounter this problem and it decreases slightly, but not much, after abundant feedback.

Figure 11 shows the distribution of perceived complexity for this specific parameterization and the average reported posterior for each perception class.

As we can see, there is a clear heterogeneity of perceptions for this belief updating task, with 43 percent of subjects finding it EASY while 45 percent find it either HARD

[36] For a survey of this literature see Benjamin (2019) and for recent papers on the topic, see Esponda et al. (2023) and Gneezy et al. (2023).

Figure 11: Posteriors as a function of perceived complexity, base-rate neglect specification



Notes: The left panel reports the distribution of perceived complexity when $p_0 = 0.15, q = 0.80, s = 1$. The right panel reports final reported posteriors, conditional on perceived complexity.

or TOO HARD. Interestingly, it is the 43 percent who find the problem EASY that make the biggest mistakes in updating and report posteriors that are significantly higher (in the direction of Base-Rate Neglect) than other perception types and greater than the Bayesian prediction. This observation is consistent with the results of Esponda et al. (2023) which suggest that mistakes are more likely to persist among people with incorrect mental models, who therefore miss or misrepresent important aspects of the environment. One possible interpretation is that such models might induce confidence in a subject's instinctive response to the problem or make the updating task seem easy, which prevents them from learning over time as new information arrives. This is consistent with how we categorize EASY perceptions, i.e., subjects who exert low effort but report high confidence in their performance.

Mistakes in Updating A recent paper by Agranov and Reshidi (2024) discusses the difficulties in updating beliefs when people observe signals that contradict their prior.[37] The authors show that subjects have trouble understanding the relative informativeness of the signal and the prior in this case and this difficulty increases when Bayesian posteriors change significantly with small changes in primitives' values. Regions in which these non-linearities are more pronounced are precisely the regions in which people make larger

[37] For example, when the prior is $p_0 = 0.15$, the signal is $s = 1$ contradicts the prior which indicates that the state $\omega = 0$ is more likely than state $\omega = 1$. Similarly, when the prior is $p_0 = 0.80$, the signal $s = 0$ contradicts the prior.

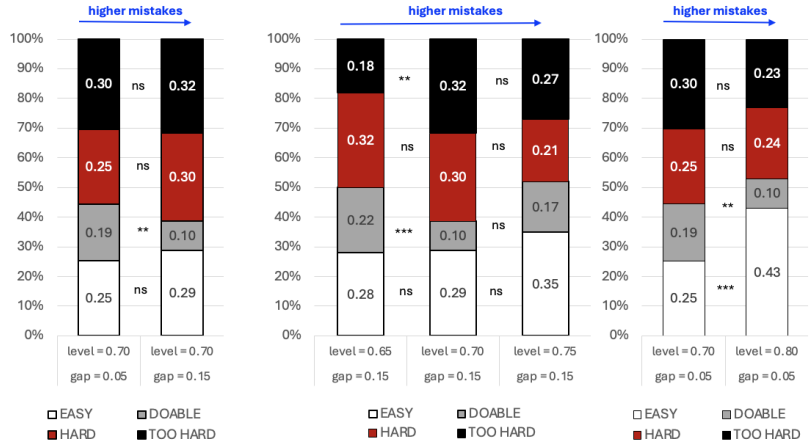
mistakes, unable to fully incorporate the extent to which a signal is more or less informative than the prior. To operationalize this idea and derive testable implications, Agranov and Reshidi (2024) define two features of the belief-updating task, the *level* and the *gap* of relative signal informativeness, both of which are calculated based on the task primitives (the prior and the signal precision). The gap corresponds to the difference in the probability with which the two signals give rise to a contradictory signal, and the level corresponds to the lowest of these two numbers.[38] Table 8 in the Appendix reports the levels and gaps for all parameterizations used in our experiment. Equipped with these two features, Agranov and Reshidi (2024) report two regularities which track non-linearity in Bayesian updating: (1) for a fixed gap, a higher level leads to larger mistakes and (2) for a fixed level, a higher gap leads to larger mistakes. In this context, mistakes refers to departures from the Bayesian predictions. In Table 9 in the Appendix we report average mistakes across parametrizations, which replicate these results in the aggregate.

Figure 12 depicts the distributions of subjective perceptions of the complexity of our belief updating tasks, classified by their levels and gaps according to Table 8 in the Appendix. We explore how the perception of the task's complexity relates to the extent of mistakes in the updating task. The left panel shows that, for a fixed level of 0.70, an increase in the gap does not lead to a significant change in perceptions, although it leads to larger mistakes, as we documented in Table 9. Similarly, for the higher gap of 0.15 (central panel) the increase in level does not seem to have a clear pattern in terms of changes in the distributions of perceptions. The right panel, on the other hand, shows that when the gap is fixed at 0.05 and level increases from 0.70 to 0.80, more subjects perceive the task that leads to more mistakes (level 80) as EASY than the task with a lower level. In other words, subjective perceptions seem to be at odds with the objective measure of complexity proposed by Agranov and Reshidi (2024) and captured by the magnitudes of mistakes. The overall evidence across these different parameterizations suggests that people do not perceive tasks with higher non-linearities as more complex, despite the fact that they make larger mistakes in them.

However, the mapping from perceptions to actions shows that perceptions matter in terms of mistakes. Table 10 in the Appendix shows the coefficients from a regression that controls for the level and gap and their interaction and shows that subjects who perceive the task as EASY make significantly larger mistakes than subjects in other perception classes, suggesting that subjective perceptions matter for choices beyond the effect of the objective measure of Agranov and Reshidi (2024) based on non-linearity.

[38] For example, consider the case in which the prior is 0.15, the signal precision is 0.80, and the signal realization is 1. This situation is mathematically equivalent to another problem in which the prior is 0.50 and the decision-maker receives two signals: the first one is 0 and comes from the original prior of 0.15, which we transform to an information source with precision 0.85 (1 - 0.15, since our focus is on contradictory signals) and the second is 1 and comes from the source with precision 0.80 (the signal in the original formulation). This alternative formulation is handy to define the two features we discussed above: the gap, which is the difference in signals' precisions, i.e., 0.05 in this case, and the level, which is the lowest precision among the two signals and is 0.80 in this case. We refer the reader to Agranov and Reshidi (2024) for more details.

Figure 12: Subjective Perceptions of Belief-Updating Tasks



Notes: The significance of the Test of Proportions that compare size of perception categories across tasks follow the notation: 'ns' for not significant and *** (**) indicates significance at 1% level (5% level). The blue line on the top of each panel depicts the direction of larger mistakes from Table 9.

4.7 Summary of Results

We have discussed two avenues through which behavior in a task can be affected by subjective perceptions of its complexity: the distribution of subjective perceptions induced by the task (Mapping 1) and the distribution of observed behavior, conditional on those perceptions (Mapping 2).

In an attempt to organize our findings about the the role that Mappings 1 and 2 play in determining behavior across tasks, we present two tables. Table 6 compares the distributions of subjective perceptions of pairs of tasks that are comparable and categorizes these pairs of tasks according to whether they lead to statistically different distributions of subjective perceptions or not. That is, we categorize those tasks for which Mapping 1 has a strong effect. We use the chi-squared independence test to compare the distributions of perceived complexity between pairs of tasks.

Table 6 shows that, in general, pairwise comparisons of related binary tasks are more likely to lead to different distributions of perceptions than non-binary tasks.

The second table, Table 7, focuses on our Mapping 2 from subjective perceptions to final choices and distinguishes tasks for which different perceptions lead to different choices and those tasks for which different perceptions lead to similar choices.

Table 6: Differences in Distributions of Subjective Perceptions across Tasks (Mapping 1)

Distributions of subjective perceptions are **different**		Distributions of subjective perceptions are **similar**	
Pairs of Tasks	Chi-sq (p-value)	Pairs of Tasks	Chi-sq (p-value)
BINARY treatments			
FOSD vs MPS	88.75 ($p < 0.01$)	Pivotality cont vs non-cont	1.04 ($p = 0.79$)
ESsimp lotteries vs ESdiff lotteries	49.84 ($p < 0.01$)		
ESsimp mirrors vs ESdiff mirrors	34.14 ($p < 0.01$)		
ESsimp lotteries vs ESsimp mirrors	31.29 ($p < 0.01$)		
ESdiff lotteries vs ESdiff mirrors	21.36 ($p < 0.01$)		
Common Consequence: CC1 vs CC2	34.64 ($p < 0.01$)		
Common Ratio: CR1 vs CR2	30.10 ($p < 0.01$)		
NON-BINARY treatments			
ESsimp lottery vs ESdiff lottery	23.84 ($p < 0.01$)	Public Goods: low vs high MPCR	3.86 ($p = 0.28$)
ESsimp lottery vs Puri lottery	23.63 ($p < 0.01$)	First-price vs Dutch	4.37 ($p = 0.22$)
		First-price vs Second-price	2.08 ($p = 0.56$)
		Second-price vs English	2.04 ($p = 0.56$)
		ESdiff lottery vs Puri lottery	4.45 ($p = 0.22$)
		ESsimp lottery vs ESsimp mirror	6.09 ($p = 0.11$)
		ESdiff lottery vs ESdiff mirror	5.47 ($p = 0.14$)
		Puri lottery vs Puri mirror	1.91 ($p = 0.59$)

Table 7: Do Perceptions affect Final Choices in a Task? (Mapping 2)

Choices are **different** across perception classes	Choices are **similar** across perception classes
BINARY treatments	
ESdiff lotteries	ESsimp lotteries
CR1	FOSD
CC1	MPS
	CC2 and CR2
ESdiff mirrors	ESsimp mirrors
Pivotality tasks	
NON-BINARY treatments	
Valuations of mirrors	Valuations of lotteries
Base-rate Neglect	Probability weighting function
Public Good games	
First-price auction	English auction
Dutch auction	
Second-price auction	

5 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 that 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 of it, and then to observe a second mapping from each subjective perception class to behavior. This procedure allows us to sort out what is responsible for the aggregate behavior observed in the experiment: is it the way the problem is perceived or behavior conditional on perception?

In defining subjective perception of a task's complexity, we classify subjects into four categories or classes: those who perceive the problem as EASY, DOABLE, HARD, and TOO HARD. These classes differ by 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. In two of these perception classes, EASY and TOO HARD, subjects exert low effort into solving a task either because they think they can solve it easily or because they perceive it too hard to be worth trying. These subjects make a quick choice that we could interpret as either their intuitive best guess or the guess of some simple heuristic they employ. They differ in that after making their choice, those who consider the task EASY are confident they chose correctly, while those who perceive it as TOO HARD are not confident in their choice. Those subjects who perceive the problem as DOABLE and HARD, on the other hand, exert high effort to solve the task but come to different conclusions about their performance, with those who found it doable thinking that they solved it correctly, while those who found it hard are not confident in their choice.

With this apparatus, we are able to make observations that allow us to provide nuance to well-known experimental results and to revisit the predictive power of theoretical models under the lens of heterogeneity in subjective perceptions. For example, in belief-updating tasks, using the original parameterization of Kahneman and Tversky (1972), we too find base-rate neglect, but attribute it to those subjects who perceive the problem as EASY. Their mistakes, when aggregated with the other types, drive the result. Contrary to the base-rate neglect problem, the often observed tendency to overweight small probabilities and underweight large ones is attributable to subjects in all perception classes. This type 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 the behavior of subjects in a specific perception class is invariant to changes in the decision environment. This is observed in auctions. Subjects who exert low effort across four auction formats (First-Price, Second-Price, Dutch, and English), and thus perceive the problem as EASY or TOO HARD, depending on their reported confidence, 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 also used by those subjects who respond intuitively, regardless of the auction format.

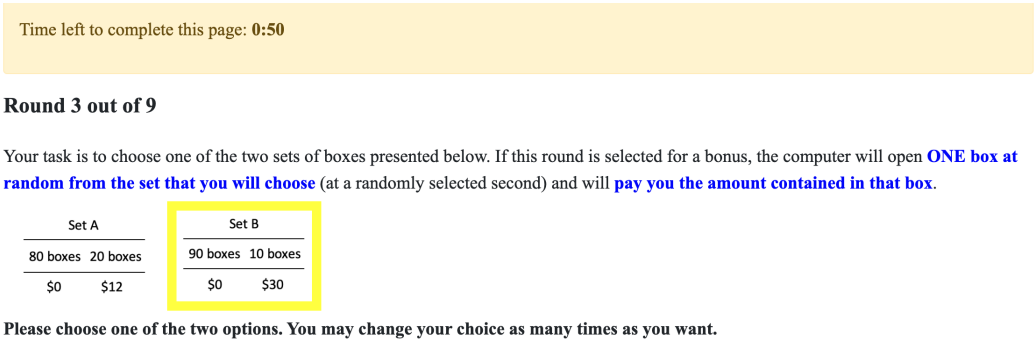
Our methodology allows us, therefore, to identify which subjects are responsible for

different types of behavior behind well-known results in the literature across a variety of tasks. In this sense, the methodology is portable across decision domains.

Finally, our results have broad and important implications for different fields of study in microeconomics, like 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.

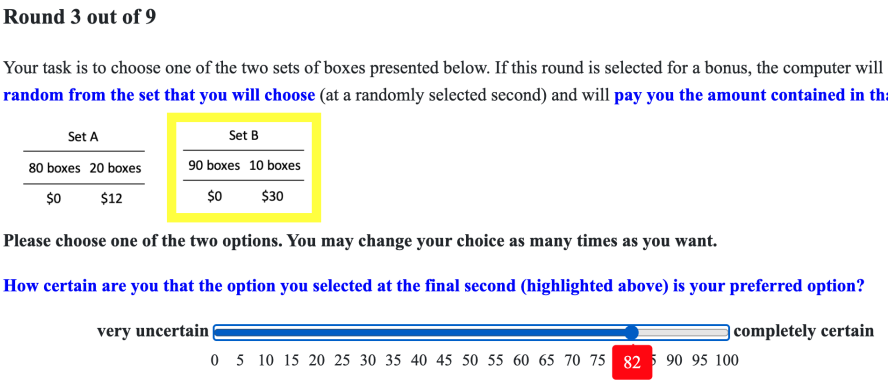
A Screenshots

Figure 13: CPP Interface for Binary Lottery Task



Notes: This is the binary lottery task in which subjects choose between a lottery that pays $12 with probability 20% and a lottery that pays $30 with probability 10%. The time on the top of the screen indicates the number of seconds left in this round. The yellow border indicates the lottery that the participant has currently selected.

Figure 14: Confidence Interface for Binary Lottery Task



Notes: The confidence question appears after the time is up for this round (after 60 seconds). The yellow box indicates the final choice of the participant.

B Additional Analysis

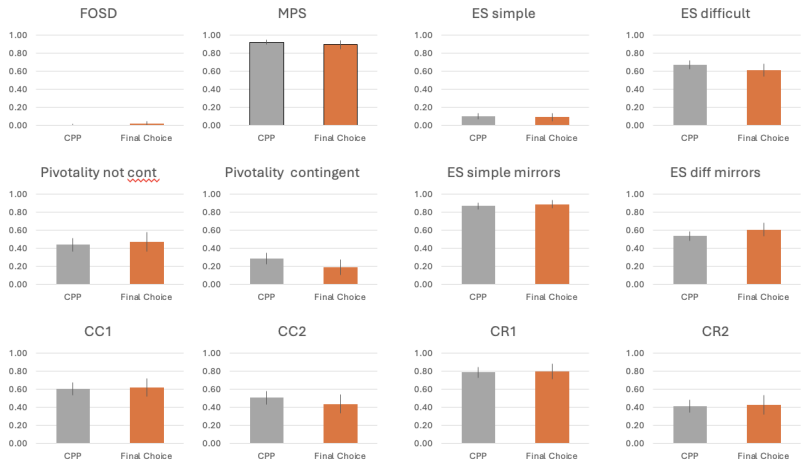
B.1 Does the CPP protocol alter behavior?

As we described in Section 2, 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 Binary treatments, in which only final choices determined payments, so the time series of choices leading to that final choice were irrelevant. 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 15 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 our set of tools for eliciting subjective perceptions of complexity can be employed broadly across different environments as it does not alter final choices while providing more information about the thinking process.

Figure 15: Final Choices in CPP sessions vs Final-Choice sessions



Notes: For binary lottery comparisons, we present the fraction of final choices that correspond to the safe lottery measured by Sharpe's ratio. For pivotality tasks and binary mirrors choices, we present the fraction of correct (optimal) final choices.

B.2 Tasks with Objectively Correct Answers

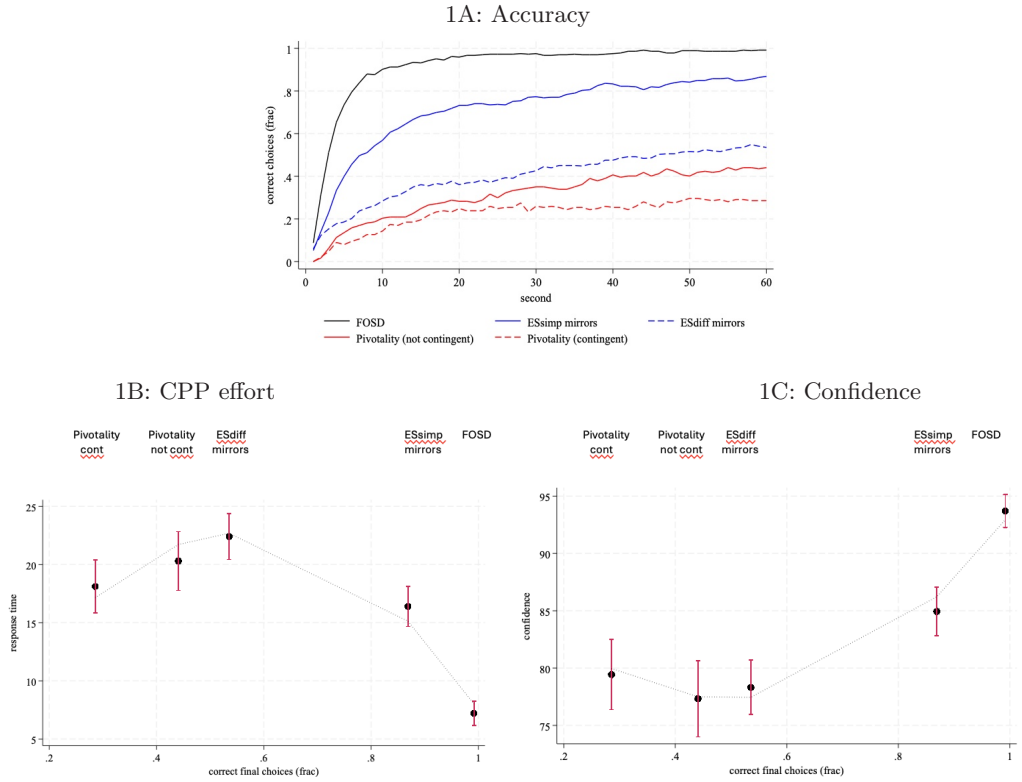
In this section, we focus on the five tasks administered in our Binary treatments that have an objectively correct answer: FOSD, ESsimp mirrors, ESdiff mirrors, Pivotality requiring contingent reasoning, and Pivotality without contingent reasoning. These tasks are special since the optimal choice does not depend on participants' preferences: any person who wants to maximize her expected payoff in these tasks should choose the same action. This analysis complements what we presented in Section 2 where we explore the variation of effort and confidence *within a task* (Table 1). Here, instead, we compare the effort and confidence levels *across tasks*. The ultimate objective of this section is to show that effort and confidence alone do not track task difficulty, measured by accuracy of choices, across tasks. This exercise also allows us to test several theoretical predictions of Goncalves (2024).

Evolution of accuracy and ordering of tasks based on accuracy. We define the accuracy of choice as simply the fraction of people who solve each problem correctly. First, we ask whether thinking more about a problem is helpful, on average. Panel 1A in Figure 16 shows the fraction of people who selected the correct choice at each second in the consideration period. We can clearly see that more time to think about a task leads to better choices, on average (all lines are increasing with time), for all tasks. Panel 1A in Figure 16 also provides us with an endogenous ordering of tasks according to their difficulty (measured by the fraction of subjects who found the correct answer in the last second): FOSD has the highest accuracy, followed by ESsimp mirrors, followed by ESdiff mirrors, followed by Pivotality (non-contingent), followed by Pivotality (contingent).

Does effort or confidence alone track choice accuracy? We answer this question in panels 1B and 1C in Figure 16. In each of these graphs, we order the five tasks by choice accuracy (correct final choices) from the lowest (Pivotality with contingent reasoning) to the highest (FOSD) and plot the average effort level (measured by total response time) in panel 1B and the average level of reported confidence in panel 1C, for each of these five tasks.

Figure 16 illustrates that neither effort nor confidence alone can capture task difficulty, measured by choice accuracy. Consistent with Goncalves (2024), we find a non-monotonic relationship between effort and task difficulty (measured by how accurate final choices are), suggesting that subjects that exert low effort in trying to solve the problem do so either because they find the task easy and solve it accurately, or they find it so hard that they give up and exhibit low accuracy. Similarly, the non-monotonic relation between confidence and task difficulty shows that subjects might report similar ex post confidence in their choices for tasks where the accuracy of their choices is very dissimilar, suggesting that confidence in the accuracy of choices might not correspond to observed accuracy.

Figure 16: Tasks with Correct Answers in Binary treatments



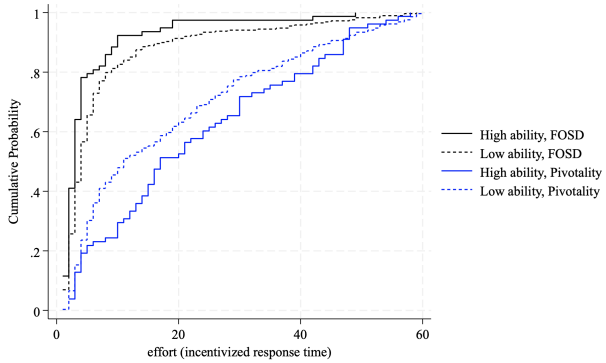
Notes: Data from FOSD, ESsimp mirrors, ESdiff mirrors, and two Pivotality tasks. Panel 1A presents the fraction of correct choices in each task as the round progresses. Panel 1B depicts the average effort (response time) of all participants plotted against the fraction of correct final choices associated to each task. Panel 1C depicts the average reported confidence plotted against the fraction of correct final choices associated to each task.

Between subjects comparison. One of the predictions of Goncalves (2024) concerns the relationship between the cognitive ability of a decision-maker and the effort she exerts on a task. Goncalves (2024) predicts that response time is not a good predictor of ability and suggests that fast responses are indicative of high ability in simple tasks and, on the contrary, are indicative of low ability in complex tasks. This means that high ability subjects should be faster on average in simple tasks and slower in more difficult ones.

Our data provides a natural testing environment for this prediction. We classify subjects as 'high ability' if they correctly solved all the tasks we consider in this section; the

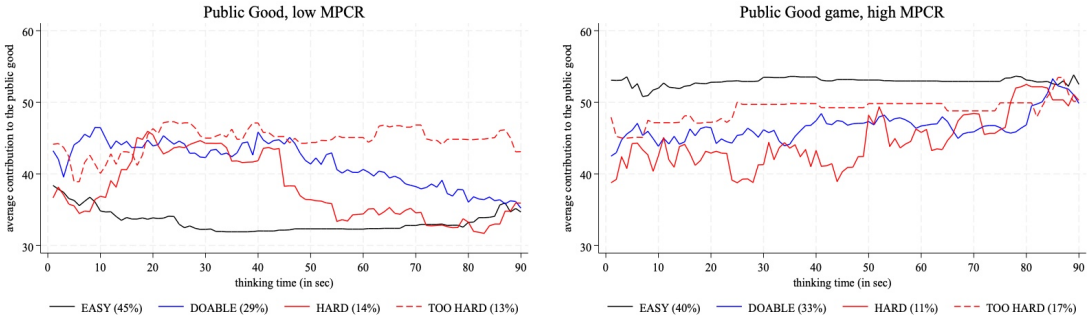
remaining subjects are classified as 'low ability'.[39] Figure 17 depicts the CDFs of the response times of subjects according to this classification in the task with the highest fraction of correct answers, which could be interpreted as the simplest (FOSD), and the task with the lowest fraction of correct answers, interpreted as the most difficult (Pivotality with contingent reasoning). Our data supports this prediction: high-ability subjects complete the simple FOSD task faster than low-ability ones, but the reverse is true in the not so simple Pivotality-contingent task.

Figure 17: CDFs of effort in FOSD and Pivotality-contingent tasks



B.3 Additional Figures and Tables

Figure 18: Evolution of Choices in Public Good Games



Notes: We plot the evolution of contributions to the public good as a function of the thinking time depicted on the horizontal axis. For each subject, the thinking time starts at the time of the first click and progresses onward.

[39] 21% of our subjects are high ability according to this definition.

Figure 19: Final Bids in Auctions

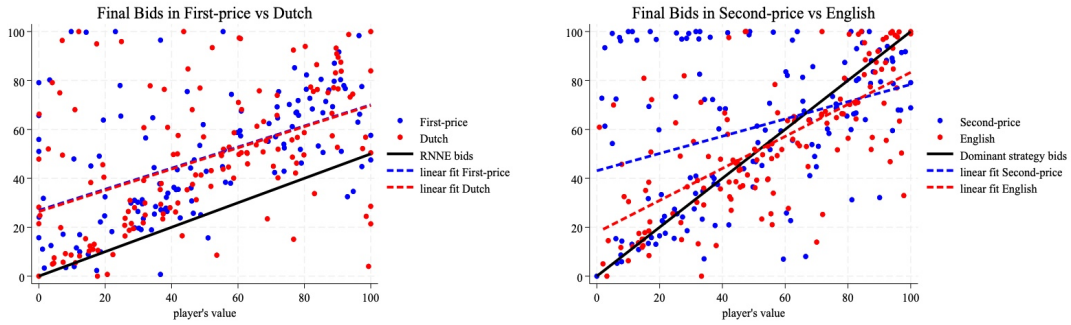
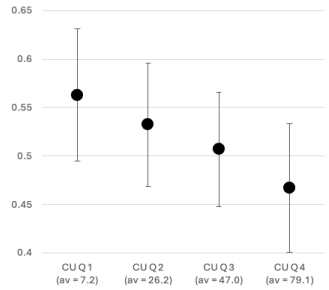


Figure 20: Effect of Bayesian Posteriors on Observed Posteriors



Notes: This figure replicates Figure 3b in Enke and Graeber (2023). It displays coefficients from OLS regressions of observed posteriors on Bayesian posteriors, split by quartiles of cognitive uncertainty (CU), which is the reciprocal of our confidence measure. Standard errors are clustered at the individual level.

Table 8: Gaps and Levels in Belief-Updating Tasks

prior	signal precision	signal realization	level	gap
0.15	0.70	$s = 1$	0.70	0.15
0.15	0.80	$s = 1$	0.80	0.05
0.30	0.75	$s = 1$	0.70	0.05
0.80	0.65	$s = 0$	0.65	0.15
0.80	0.85	$s = 0$	0.80	0.05
0.90	0.75	$s = 0$	0.75	0.15

Notes: We follow Agranov and Reshidi (2024) to define the level and the gap for each belief-updating task with contradictory signals.

Table 9: Mistakes in Observed Posteriors, Aggregate Data

	gap = 0.05	gap = 0.15
level = 0.65		0.213 (0.01)
level = 0.70	0.229 (0.01)	0.268 (0.02)
level = 0.75		0.281 (0.02)
level = 0.80	0.291 (0.01)	

Notes: We report the absolute difference between observed posteriors and Bayesian predictions across all parameterizations focusing on the signals contradicting priors. All pairwise comparisons are statistically significant at 1% level except for one, in which we hold the gap fixed at 0.15 and change the level from 0.70 to 0.75 ($p = 0.68$).

Table 10: The Effect of Subjective Perceptions on Choices in Belief-Updating Tasks

	Dep. Variable: Mistakes
Indicator for EASY	3.42^{**} (1.43)
Level	0.49^{**} (0.24)
Gap	-0.65 (2.09)
Level × Gap	0.013 (0.03)
Const	-13.44 (17.73)
Nb obs	569
Nb subjects	262
R-squared	0.0498

Notes: The dependent variable is the absolute value difference between reported posterior and the Bayesian prediction. We use all parameterizations of belief-updating tasks and focus on signals contradicting priors. The level and the gap in each task are defined in Table 8.

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 (2024). Deciphering suboptimal updating: Task difficulty, structure, and sequencing. *Working paper*.

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.

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 (2012a). Salience theory of choice under risk. *Quarterly Journal of Economics 127(3)*, 1243–1285.

Bordalo, P., N. Gennaioli, and A. Shleifer (2012b). 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 (2023). 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. (2006). Dilemmas of economic theorist. *Econometrica 74(4)*, 865–883.

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.