ONLINE APPENDIX

Complex for Whom? An Experimental Approach to Subjective Complexity

Marina Agranov Andrew Schotter Isabel Trevino

This online appendix presents additional analysis of the data and is organized as follows:

- Section 1 discusses additional predictions of the theoretical model of Goncalves (2024) using tasks with objectively correct answers.
- Section 2 presents additional analyses of the Binary Lottery tasks.
- Section 3 investigates the Common Consequence and Common Ratio effects.
- Section 4 tests the predictions of Agranov and Reshidi (2026) in the belief-updating tasks.

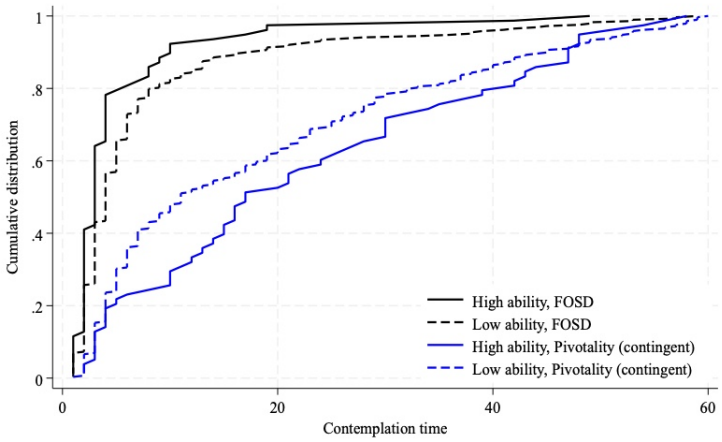
1 Tasks with Objectively Correct Answers

In this section we return to the five tasks in the Binary treatment, in which there is a correct answer. These tasks are analyzed in Section 3.1, in which we showed that neither confidence nor contemplation time alone is enough to track how subjects assess these tasks and approach them. Our empirical results reported in Section 3.1 confirm theoretical results derived by Goncalves (2024). There is an additional prediction in Goncalves (2024) that concerns the relationship between the cognitive ability of a decision-maker and the effort she exerts on a task, i.e., that contemplation 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'. 21% of our subjects are high ability according to this definition. Figure 1 depicts the CDFs of the contemplation 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 1: CDFs of contemplation time in FOSD and Pivotality (contingent) tasks



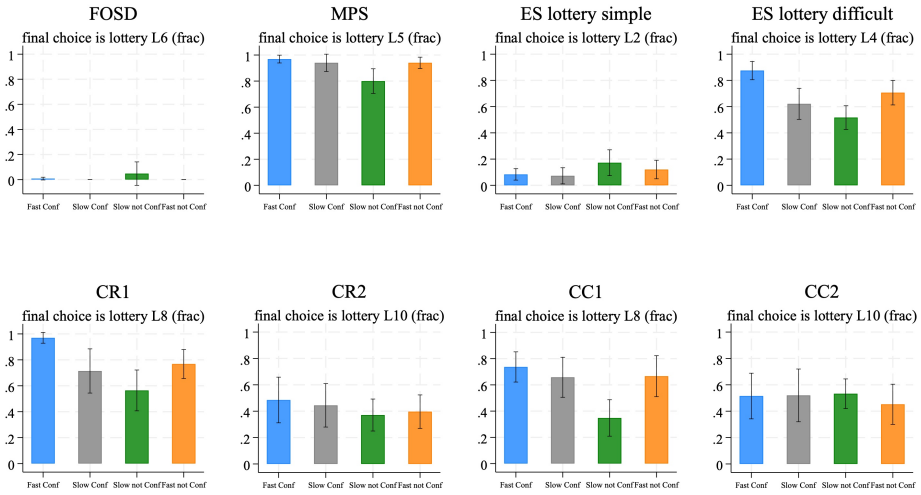
2 Binary Lottery Choices

Figure 2 reports the fraction of subjects choosing the lottery with higher Sharpe's ratio in each binary lottery choice task, by cognitive signature. In our case, lotteries with higher Sharpe's ratio are also those with lower volatility, lending a natural interpretation of choosing higher Sharpe's ratio corresponding to 'safer' or less risky choices.

Three tasks—FOSD, MPS, and ESsimp—do not show a strong relationship between cognitive signatures and final choices, despite substantial heterogeneity in how subjects approach these problems. In the FOSD task, the safer lottery is first-order stochastically dominated, and almost no subject chooses it, regardless of cognitive signature. In the MPS task, nearly all subjects choose the safer lottery, while in ESsimp the vast majority choose the riskier lottery because it offers more attractive prizes. Thus, although subjective assessments differ across these tasks, they have little effect on final choices.

The pattern changes in the more intricate lottery problems. In ESdiff, as well as CR1 and CC1, safer choices are more common among subjects who make fast decisions, whereas choosing the riskier lottery is associated with a longer deliberation time, particularly among slow and non-confident subjects. This finding complements the objective complexity measure of Enke and Shubatt (2023), who show that greater complexity reduces sensitivity to

Figure 2: Choosing Lottery with Higher Sharpe's Ratio in Binary Lottery Tasks as a Function of Individual Assessments



expected-value differences. Our results suggest that, when problems are objectively more complex, subjects who make faster choices are more likely to simplify their decisions by favoring the safer lottery. In the ESdiff task, for example, 88% of fast-and-confident subjects and 71% of fast-but-not-confident subjects choose the safer lottery, compared with only 62% and 52% among subjects who deliberate longer.

This relationship is absent in ESsimp, indicating that cognitive signatures matter primarily when objective complexity is high. Finally, for CR2 and CC2—which present the same lottery pair (L10 versus L11)—approximately half of the subjects choose the riskier lottery in every cognitive-signature category, suggesting that Mapping 2, from cognitive signatures to choices, does not explain behavior in these problems.

3 Common Consequence and Common Ratio Effects

In our design, subjects face two pairs of binary lottery choice problems, CC1–CC2 and CR1–CR2, corresponding to the common consequence and common ratio problems that generate the Allais paradox. Although these problems have no objectively correct answers, Expected Utility theory requires consistency between the two choices within each pair. Thus, inconsistent choices constitute violations of Expected Utility.

An interesting pattern emerges from subjects' cognitive signatures. The first question in each pair is substantially more likely to elicit fast and confident decisions than the second.

Specifically, 35% (32%) of subjects approach CR1 (CC1) in a Fast and Confident manner, compared with only 17% (19%) for CR2 (CC2). Likewise, the fraction of subjects who are not confident in their final choice rises from 50% (47%) in CR1 (CC1) to 64% (67%) in CR2 (CC2).

These differences suggest that subjects might perceive the second problem in each pair as more difficult than the first. Although our data cannot establish this interpretation conclusively, it raises an important possibility: some violations of Expected Utility may reflect differences in subjective assessments about the complexity of the tasks rather than stable preferences. If so, tests of the Allais paradox should ideally compare problems with similar perceived complexity.[1]

To understand whether inconsistent choices can be traced to particular cognitive signatures, Table 1 reports regression coefficients relating subjective perceptions to inconsistency in the Common Ratio questions. Subjects who make fast and confident choices in CR1 are significantly less likely to violate Expected Utility by exhibiting the standard Common Ratio pattern. Figure 3 further shows that inconsistency is least common among subjects who make fast and confident choices in CR1 and fast choices in CR2. Thus, consistency in the Common Ratio problems is most common among subjects who spend relatively little time deliberating, whereas violations are associated with longer deliberation. For the Common Consequence problems, we do not find a clear relationship between individual assessments of the tasks and the likelihood of inconsistencies across two questions (Table 2).

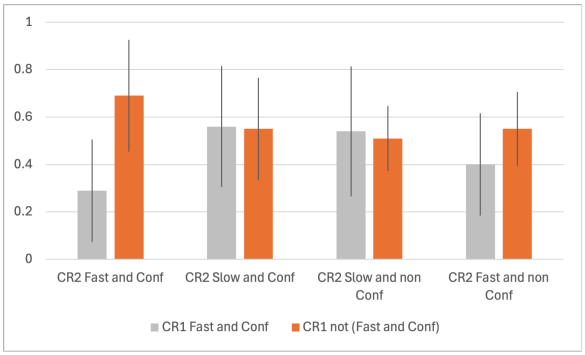
Table 1: Consistency and Perceptions in Common Ratio Questions

	Inconsistent Choices in CR1 and CR2	
	reg (1)	reg(2)
CR1 Fast-Conf	19** (0.08)	
CR2 Fast-Conf	0.01 (0.09)	
CR1 not (Fast-Conf) and CR2 Fast-Conf		0.19 (0.13)
CR1 Fast-Conf and CR2 not (Fast-Conf)		-0.11 (0.09)
CR1 Fast-Conf and CR2 Fast-Conf		-0.30** (0.13)
CR1 safe choice	0.30** (0.09)	0.30** (0.09)
Constant	0.37** (0.09)	0.35** (0.09)
# obs	189	189
adj R-squared	0.05	0.07

Notes: The dependent variable is an indicator of inconsistent choices in CR1 and CR2, i.e., choosing L8 in CR1 and L11 in CR2, or choosing L9 in CR1 and L10 in CR2. We control for the order in which CR1 and CR2 questions appear. CR1 (CR2) Fast-Conf is an indicator that a subject made a fast and confident choice in the CR1 (CR2) question. CR1 (CR2) not (Fast-Conf) is the remaining category. CR1 safe choice indicates that a subject chose L8 in CR1 (getting $12 for sure). The constant in reg (2) represents the frequency of inconsistent choices for subjects who made choices in both CR1 and CR2 in any way but fast and confident.

[1] Designing such a test may be difficult because CC1 and CR1 compare a degenerate and a non-degenerate lottery, whereas CC2 and CR2 compare two non-degenerate lotteries.

Figure 3: Fraction of Subjects Displaying Common Ratio Type Inconsistency



Notes: Inconsistent behavior is choosing L8 in CR1 and L11 in CR2 or choosing L9 in CR1 and L10 in CR2. We plot the fraction of inconsistent subjects depending on their subjective perceptions in both questions.

Table 2: Consistency and Perceptions in Common Consequence Questions

	Inconsistent Choices in CC1 and CC2	
	reg (1)	reg(2)
CC1 Fast-Conf	-0.06 (0.09)	
CC2 Fast-Conf	0.05 (0.11)	
CC1 not (Fast-Conf) and CC2 Fast-Conf		0.11 (0.18)
CC1 Fast-Conf and CC2 not (Fast-Conf)		-0.04 (0.10)
CC1 Fast-Conf and CC2 Fast-Conf		-0.03 (0.11)
CC1 safe choice	0.02 (0.08)	0.02 (0.08)
Constant	0.47** (0.08)	0.46** (0.08)
# obs	177	177
adj R-squared	<0.01	<0.01

Notes: The dependent variable is an indicator of inconsistent choices in CC1 and CC2, i.e., choosing L8 in CC1 and L11 in CC2, or choosing L12 in CC1 and L10 in CC2. We control for the order in which CC1 and CC2 questions appear. CC1 (CC2) Fast-Conf is an indicator that a subject made a fast and confident choice in the CC1 (CC2) question. CC1 (CC2) not (Fast-Conf) is the remaining category. The constant in reg (2) represents the frequency of inconsistent choices for subjects who made choices in both CC1 and CC2 in any way but fast and confident.

4 Belief-updating Tasks

A recent paper by Agranov and Reshidi (2026) discusses the difficulties in updating beliefs when people observe signals that contradict their prior. 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 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 (2026) 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.[2] Table 3 reports the levels and gaps for all parameterizations used in our experiment. Equipped with these two features, Agranov and Reshidi (2026) 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. Table 4 reports average mistakes across parametrizations, which replicate these results in the aggregate.

Table 3: 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 (2026) to define the level and the gap for each belief-updating task with contradictory signals.

We use the data reported in Figure 11 in the Appendix of the main text to see whether cognitive signatures of participants are in line with the non-linearity predictions described above and whether these assessments translate into mistakes in reported posteriors. For a fixed level of 0.70 (parameterizations (0.15,0.70) and (0.30,0.75)), an increase in the gap does not lead to a significant change in assessments, although it leads to larger mistakes, as we documented in Table 4. Similarly, for the higher gap of 0.15, the increase in level does not seem to have a clear pattern in terms of changes in the distributions of assessments (these are parameterizations (0.80,0.65), (0.15,0.70), and (0.90,0.75) in the order of increase in level). Finally, when the gap is fixed at 0.05 and the level increases from 0.70 to 0.80

[2]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. See Agranov and Reshidi (2026) for more details.

Table 4: 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 the 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).

(parameterizations (0.30,0.75), (0.15,0.80), and (0.80,0.85)), more subjects deliver fast and confident choices in the task that leads to more mistakes (level 0.80). In other words, individual assessments of the different belief-updating tasks seem to be at odds with the objective measure of complexity proposed by Agranov and Reshidi (2026) 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. Indeed, Table 5 documents that (a) subjects who deliver fast and confident decisions make larger mistakes and (b) subjects make larger mistakes in belief-updating tasks with higher levels and fixed gap, where mistakes are assessed relative to the Bayesian posteriors.

Table 5: The Effect of Subjective Assessments on Choices in Belief-Updating Tasks

	Dep. Variable: Mistakes
Indicator for Fast and Conf	3.11** (1.31)
Level	0.50** (0.24)
Gap	-0.64 (2.10)
Level × Gap	0.01 (0.03)
Const	-13.97 (17.73)
Nb obs	569
Nb subjects	262
R-squared	0.05

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 3.

References

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

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

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