What Drives Violations of the Independence Axiom? The Role of Decision Confidence*

Aldo Lucia†

February 24, 2023

Abstract

Recent theoretical work implicates decision confidence as a central component of decision-making under uncertainty, attributing failures of Expected Utility (EU) to a lack of confidence. We design an experiment testing EU's central independence axiom and contemporaneously eliciting measures of decision confidence. We find that choices characterized by high self-reported levels of decision confidence and low response times are more likely to comply with the independence axiom. Contrary to the common certainty effect rationale for independence violations, we show that subjects predominantly violate EU by choosing risky lotteries over certain amounts when they are unconfident in their choices.

Keywords: decision confidence; independence axiom; certainty effect

JEL Classification: C91, D81

*I am grateful to Marina Agranov, Alena Buinskaya, Federico Echenique, Kirby Nielsen, Luciano Pomatto and Charles D. Sprenger for useful comments and suggestions. This study was reviewed and granted exemption by the Institutional Review Boards at Caltech (IR21-1140) and funded by the center for theoretical and experimental social sciences (CTESS) at Caltech.

†Division of the Humanities and Social Sciences, Caltech, allucia@caltech.edu.

1 Introduction

A central research question in economics is how individuals make decisions in the presence of uncertainty. Research has demonstrated critical shortcomings in the neoclassical formulation of Expected Utility (EU) and its central assumption, the independence axiom. In the most famous counterexample of EU proposed by Maurice Allais in 1952, individuals typically violate the independence axiom by showing higher risk aversion when one certain option is available than when all the available options are uncertain. This tendency, known as the "certainty effect" (Kahneman and Tversky, 1979), is a centerpiece of theoretical alternatives to EU—most notably Cumulative Prospect Theory (CPT) introduced by Tversky and Kahneman (1992).

Recent research has challenged the empirical regularities related to the certainty effect and the theoretical explanations proposed to rationalize them. In particular, Blavatskyy et al. (2022a, 2022b) and Jain and Nielsen (2022) show that the certainty effect is a fragile empirical finding whose emergence is systematically affected by features of the experimental design. Moreover, Bernheim and Sprenger (2020) find no evidence of the rank-dependence assumption through which CPT rationalizes the certainty effect as well as other behaviors incompatible with EU. Motivated by these findings, we design an experiment to study the relevance of the certainty effect and to investigate an alternative mechanism for violations of EU's independence axiom: lack of confidence when choosing between different lotteries.

Our paper provides the first experimental investigation of the independence axiom through the lens of the EU core, which captures the largest subrelation of a preference relation that satisfies the independence axiom (Cerreia-Vioglio, 2009). When a decision-maker violates the independence axiom, his EU core is a partial order that is typically interpreted as the subset of his uncontroversial rankings. This interpretation suggests that violations of the independence axiom should only arise in choice problems that are "hard" enough so that the decision-maker feels unconfident about them. Recent non-EU models further appeal to the lack of decision confidence as the driving force for violations of the independence axiom (Cerreia-Vioglio et al. 2015 2020, 2022). However, little is known empirically about the relationship between decision confidence and violations of the independence axiom, which is what our experiment explores.[1]

In our experiment we ask subjects to make incentivized choices between lotteries. After each choice, subjects report on a scale from zero to 100 how confident they are about their choices. We also collect response times as an additional proxy for confidence. To construct the pairs of lotteries in the experiment, we start with an initial set of *unmixed comparisons*, each consisting of a certain prize and a risky lottery. Next, to test the independence axiom, we create *mixed comparisons* by mixing each of the lotteries in an unmixed comparison with a series of third common lotteries. Evaluating independence by assessing behavior in multiple mixed comparisons allows us to assess the independence axiom—and hence adherence to the EU core—for each choice.

We conducted our experiment on the online platform Prolific.co with 300 subjects. Each subject made binary choices over lotteries in 74 comparisons. We assess adherence to the independence axiom for each choice and relate that adherence to measures of decision confidence. A central result of the paper is that behavior is more likely to comply with the independence axiom when subjects report high confidence and make their decisions quickly. In contrast, behavior systematically deviates from EU when subjects report low confidence and take more time to make their choices.

In addition, we study the relationship between risk aversion and decision confidence. It may seem plausible to expect individuals to display higher risk aversion when they are not confident. This prediction is consistent with an interpretation of the negative certainty independence (NCI) axiom introduced by Dillenberger (2010) to rationalize the certainty effect. Cerreia-Vioglio et al. (2015) show that NCI can be interpreted as a completion rule for incomplete but otherwise EU preferences according to which "when in doubt, the decision-maker (DM) chooses a risk-free lottery." Somewhat surprisingly, we find evidence against NCI and its interpretation: individuals are more likely to violate the independence axiom by choosing risky lotteries over certain prizes in situations of low confidence.

[1] A few works on preference imprecision examine the association between the lack of decision confidence and EU violations obtaining mixed evidence (Butler and Loomes, 2011; Cubitt et al., 2015).

Finally, we study whether the mere presence of a certain prize in a choice problem makes independence violations more likely. Recent work finds evidence consistent with the idea that people value certain and uncertain outcomes differently (Halevy, 2008; Andreoni and Harbaugh, 2009; Andreoni and Sprenger, 2010, 2011, 2012). Our experiment produces more data to detect EU violations in choices with a certain alternative than in choices where all alternatives are risky. In the benchmark scenario of an EU decision-maker who makes mistakes, more data to test independence translates into higher expected independence violations. We then adopt the error model of Harless and Camerer (1994) to control for this asymmetry and find that certainty only plays a role when we build mixed comparisons using a fixed third common lottery as it is common for testing the independence axiom.

Our results suggest that decision confidence is a key variable in deepening our understanding of why subjects deviate from EU. In the experimental literature on preferences for randomization, Agranov and Ortoleva (2017) show that individuals strictly prefer to randomize in "hard" comparisons and consequently violate the independence axiom. Arts et al. (2022) design an experiment to elicit both decision confidence and randomization probabilities in choices under risk. In line with the interpretation of the hard questions in Agranov and Ortoleva (2017), they show that subjects tend to choose randomization probabilities that are close to uniform when they report low confidence measures. In our experiment, we obtain the same relationship between decision confidence and violations of the independence axiom.

More broadly, our paper relates to the literature in psychology, neuroscience and economics documenting how behavioral anomalies are often associated with low levels of decision confidence.[2] Enke and Graeber (2019) show that cognitive uncertainty—measured as subjective uncertainty over the ex-ante utility-maximizing decision—is associated with an attenuated relationship between decisions and problem parameters. This attenuated relationship may lead to the emergence of well-known behavioral patterns, such as the fourfold pattern of risk attitudes. In a related work, Oprea (2022) shows apparent evidence of probability weighting and loss aver-

[2]Examples from psychology and neuroscience include De Martino et al. (2013), Aitchison et al. (2015), Meyniel et al. (2015), Folke et al. (2016), Meyniel and Dehaene (2017), Desender et al. (2018), Boldt et al. (2019), Rollwage et al. (2020), da Silva Castanheira et al. (2021), among many others.

sion in problems without risk or loss, suggesting that these behavioral anomalies may be partly driven by the complexity of the lotteries rather than by risk preferences.

We contribute to this literature by focusing on the choice implications of low decision confidence in the risk domain. In particular, we show that subjects tend to adopt a form of "incaution" when reporting low confidence levels by choosing the riskiest available lottery. This finding is consistent with decision models that rely on the positive certainty independence (PCI) axiom (Cerreia-Vioglio et al., 2020). The PCI axiom rationalizes the behavioral pattern opposite to the certainty effect and predicts the relationship between decision confidence and preference for risky alternatives we observe in the data.

Moreover, our paper highlights that features of the experimental design may affect the overall amount of independence violations and the conclusions about their sources. Most experimental research on the independence axiom focuses on variations of the Allais paradox, in which we know that the certainty effect constitutes the modal behavioral pattern.[3] However, our results show that the predominance of the certainty effect may not be robust to richer environments where subjects face a wider variety of unmixed and mixed comparisons.[4]

Another stylized fact in choices under risk that we put under scrutiny is that violations of the independence axiom are less frequent in comparisons with nondegenerate lotteries over common prizes. Camerer (1992) writes that "much as Newtonian mechanics is an adequate working theory at low velocities, EU seems to be an adequate working theory for gambles inside the triangle".[5] Our analysis challenges the generality of this conclusion showing that its validity may depend on the mechanism through which mixed comparisons are constructed.

The paper proceeds as follows. Section 2 describes the theoretical framework in the context of our experimental design. Section 3 illustrates our experimental design. Section 4 presents our main findings and Section 5 concludes.

[3] We refer to Blavatskyy (2010) for a review of this literature.

[4] Jain and Nielsen (2022) also find that the certainty effect is not the most common behavioral pattern that violates the independence axiom.

[5] See also Starmer (2000).

2 Theoretical Framework

We describe the theoretical framework in the context of our experimental design. All questions in the experiment involve lotteries over the set of monetary prizes $X = \{\$1, \$7, \$20\}$. We denote the set of lotteries with prizes in X by $\triangle(X)$. We refer to generic prizes in X by x and denote generic lotteries in $\triangle(X)$ by p, q, r and s. We represent the three-outcome lottery, q, giving \$1 with probability $q(1)$, \$7 with probability $q(7)$ and \$20 with probability $q(20)$ as $(\$1, q(1); \$7, q(7); \$20, q(20))$. We write the lottery that gives \$$x$ for sure as δ_x and we refer to generic pairs of lotteries $(s,r) \in \triangle(X)^2$ as comparisons. Moreover, we denote by N be the set of all the subjects in the experiment, and by $\succsim_i$ and $\succ_i$ the weak and strict preference relations of a subject $i \in N$ over $\triangle(X)$.

The preference $\succsim_i$ satisfies the independence axiom if for all lotteries $q, s, r \in \triangle(X)$ and for all $\lambda \in (0,1]$,

$$s \succsim_i r \Rightarrow \lambda s + (1-\lambda)q \succsim_i \lambda r + (1-\lambda)q.$$

The EU core of $\succsim_i$ is the subrelation $\succsim_i^*$ such that for all lotteries $q, s, r \in \triangle(X)$ and for all $\lambda \in (0,1]$,[6]

$$s \succsim_i^* r \Leftrightarrow \lambda s + (1-\lambda)q \succsim_i \lambda r + (1-\lambda)q.$$

That is, $s \succsim_i^* r$ whenever subject i prefers s to r and mixing both lotteries s and r with a third common lottery q does not affect the relative preferences of i between s and r.[7] Cerreia-Vioglio (2009) proves that $\succsim_i^*$ is the greatest subrelation of $\succsim_i$ that satisfies the independence axiom.[8] Therefore, to study how the independence axiom fails, we test for each subject i separately or for all subjects $i \in N$ at the aggregate

[6] $\succsim_i^*$ is a subrelation of $\succsim_i$ if for all lotteries s and r, $s \succsim_i^* r$ implies $s \succsim_i r$.

[7] In our experiment, we study the EU core by considering only "one-stage" lottery mixtures, rather than two-stage compound lotteries. In other words, we focus on mixture independence, rather than compound independence, as defined in Segal (1990).

[8] That is, if $\succsim_i^{**}$ is another subrelation of $\succsim_i$ that satisfies the independence axiom, then $\succsim_i^{**}$ is a subrelation of $\succsim_i^*$.

level the following hypothesis:

$$s \succsim_i^* r \text{ or } r \succsim_i^* s. \tag{EU-CORE}$$

Throughout the paper, we say that hypothesis EU-CORE holds if we find no evidence against it, while we say that it fails otherwise. By correlating the results of hypothesis EU-CORE with the measures of decision confidence that we collect, we test the interpretation of $s \succsim_i^* r$ as individual i being confident that lottery s is better than lottery r. Moreover, we examine whether individuals are more likely to choose the safer or the riskier lottery when hypothesis EU-CORE does not hold and when they declare to be unconfident. If either lottery s or lottery r is degenerate, this analysis will allow us to shed light on the relevance of the certainty effect. Finally, we study whether hypothesis EU-CORE is more likely to hold in comparisons where the two lotteries are risky and have the same support.

3 Experimental Design

The rationale behind the experimental design is to create a rich dataset to study for what comparisons individuals are more likely to violate hypothesis EU-CORE and test whether the lack of decision confidence can explain failures of hypothesis EU-CORE. This section first illustrates the comparisons that we consider in the experiment. Next, we describe the questions that subjects answer about each comparison. Finally, we discuss the recruitment procedures and the experimental payments.[9]

3.1 Comparisons

There are three treatments: "Worst", "Bad" and "WorstBest." In all treatments, subjects face the same 17 comparisons that we call *unmixed*, each involving the degenerate lottery δ_7 and a risky lottery. The treatments differ in how we construct the additional comparisons to test the independence axiom. In what follows, we use the Marschak-Machina (MM) triangle to describe the lotteries in the experiment

[9] We registered the experimental design and the analysis plan at the AEA RCT Registry as AEARCTR-0008615 (Lucia, 2022).

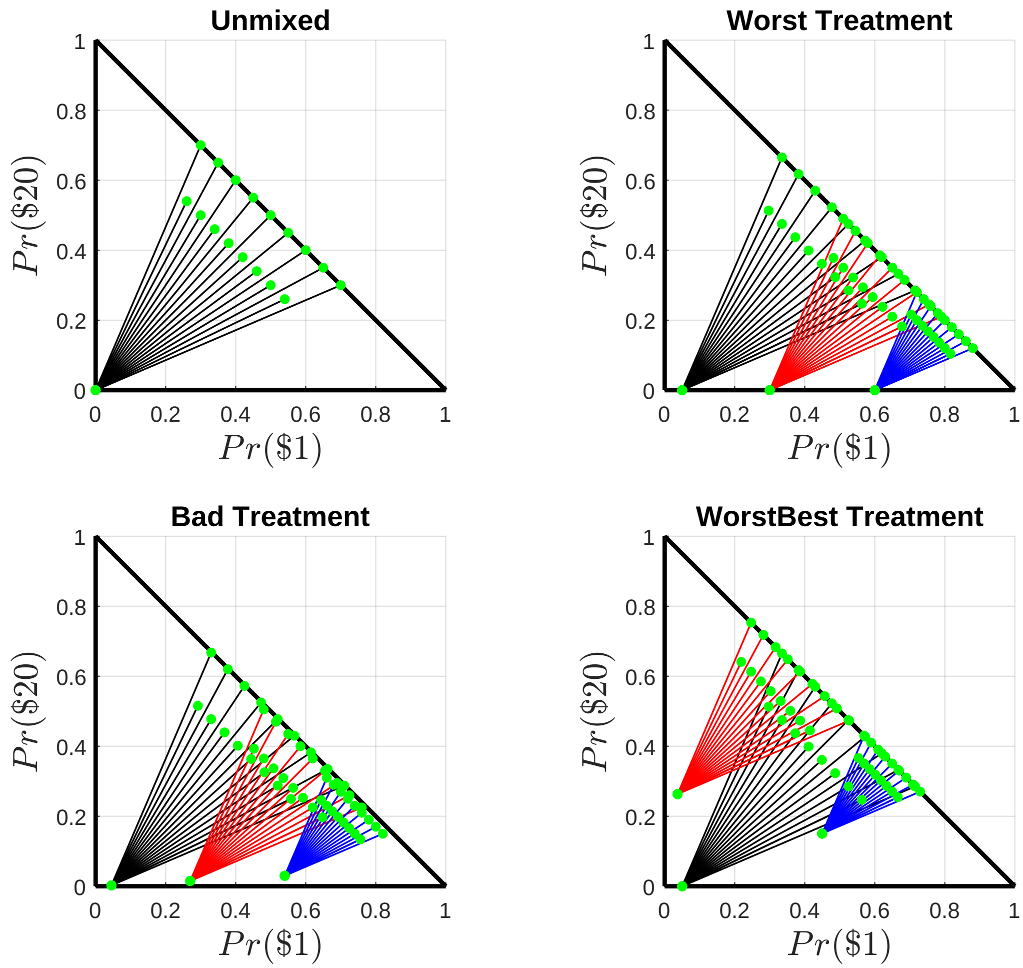


Figure 1: Unmixed and mixed comparisons in the three treatments.

(Marschak, 1950; Machina, 1982). The top-left graph in Figure 1 shows the unmixed comparisons in the MM triangle. In all the graphs of Figure 1, the probability of receiving \$20 is on the vertical axis, and the probability of receiving \$1 is on the horizontal axis. Therefore, the generic point (x, y) in the MM triangle represents the lottery $(\$1, x; \$7, 1-x-y; \$20, y)$. Each segment connecting the degenerate lottery δ_7 with a risky lottery represents an unmixed comparison between these two lotteries.

In order to test hypothesis EU-CORE for an unmixed comparison (δ_7, r), we need at least another comparison (p,q) of the following form:

$$p = \lambda\delta_7 + (1-\lambda)z \text{ and } q = \lambda r + (1-\lambda)z,$$

where both lotteries p and q are constructed by mixing lotteries δ_7 and r with a third common lottery z, using a fixed probability weight $\lambda \in (0,1)$. We call comparisons that satisfy this property *mixed*. In particular, we omit the dependence on the third common lottery and refer to $(\lambda\delta_7, \lambda r)$ as a λ*-mixed comparison*.

Each treatment includes an equal number of 0.95-mixed, 0.7-mixed, and 0.4-mixed comparisons. Overall, there are 51 mixed comparisons in each treatment. The 0.95-mixed comparisons involve one almost degenerate lottery, i.e., $0.95\delta_7 + 0.05z$, while lotteries in the remaining mixed comparisons are all "far" from being degenerate. Therefore, we can study whether the presence of a degenerate or almost degenerate lottery is the main driver for the violations of hypothesis EU-CORE.

In the Worst treatment, we build λ-mixed comparisons by mixing the lotteries in each of the 17 unmixed comparisons with the "worst" lottery $(\$1,1)$ using the probability weight $\lambda \in \{0.95, 0.7, 0.4\}$. To construct mixed comparisons in the Bad treatment, we repeat the same procedure except for replacing lottery $(\$1,1)$ with lottery $(\$1, 0.9; \$7, 0.05; \$20, 0.05)$, which is inside the MM triangle. The bottom-left graph of Figure 1 represents the mixed comparisons in the Bad treatment. Unlike the Worst treatment, mixed comparisons in the Bad treatment have lotteries with the same support. Therefore, we can study the relevance of this feature by comparing failures of hypothesis EU-CORE in the Worst and the Bad treatments.

Mixed comparisons in the Worst and the Bad treatments cluster in the southeast region of the MM triangle. This concentration may preclude us from detecting violations of hypothesis EU-CORE.[10] To account for this potential concern, we consider an additional treatment that we call WorstBest. The WorstBest treatment shares the same 0.95-mixed comparisons of the Worst treatment. The 0.7-mixed comparisons are constructed by mixing the lotteries in each of the 0.95-mixed comparisons with the "best" lottery $(\$20,1)$ using $0.7/0.95$ as probability weight. Finally, for

[10] For instance, this is the case if subjects' preferences are consistent with the fanning-out hypothesis (Machina, 1982, 1987).

Table 1: Experimental design.

	Worst	Bad	WorstBest
# Unmixed Comparisons	17	17	17
# Mixed Comparisons	51	51	51
# Dominance Comparisons	6	6	6
Third Common Lottery	Fixed	Fixed	Alternate
Probability Weights	0.95,0.7,0.4	0.95,0.7,0.4	0.95,0.7,0.4
Sample Size	100	100	100

0.4-mixed comparisons, we mix the lotteries in each of the 0.7-mixed comparisons with lottery $(\$1,1)$ using $0.4/0.7$ as probability weight.[11] The bottom-right graph of Figure 1 describes the mixed comparisons in the WorstBest treatment.

Table 1 summarizes our experimental design. The large number and the diversity of the comparisons in the experiment enable us to test hypothesis EU-CORE throughout the MM triangle and ensure that systematic and persistent violations of hypothesis EU-CORE are not just a reflection of indifference.[12] To further evaluate the reliability of our data, we also include in each treatment six comparisons involving stochastically dominated lotteries.[13] When presenting our results, we exclude all subjects that chose the stochastically dominated lottery more than once.[14]

3.2 Questions

We first asked subjects to indicate the lottery they preferred for each comparison. Next, we asked them to report their confidence level on a scale from zero (not confident at all) to 100 (completely confident). We also collected response times in these answers as an indirect measure of decision confidence.[15] Subjects answered

[11] This approach ensures that differences in expected values between lotteries in unmixed and mixed comparisons are constant across the three treatments. That is, in all treatments, λ-mixed comparisons can be created by mixing the lotteries in unmixed comparisons with some third common lottery z using λ as probability weight.

[12] The "indifference" argument is a common critique for experiments that document preference reversals (Blavatskyy, 2010).

[13] We report these six comparisons in Appendix A.

[14] Overall, 32 out of 300 subjects chose the stochastically dominated lottery more than once.

[15] We report our analysis on response times in Appendix E.

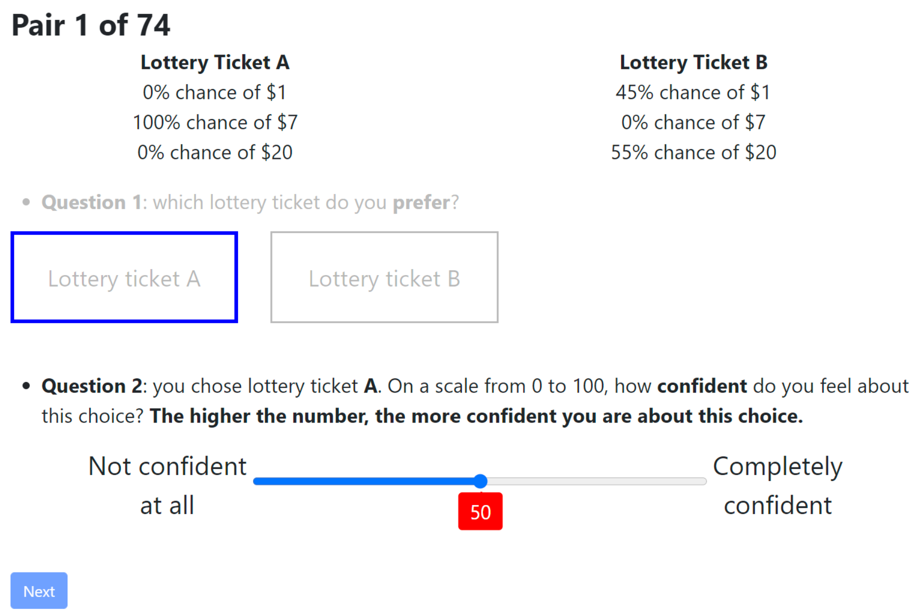


Figure 2: Decision screen from the experiment.

one question at a time. Once subjects selected an answer, they could not modify it. Figure 2 shows a decision screen from the experiment. In this example, a subject who declared to prefer lottery ticket A over lottery ticket B is asked to report how confident he feels about this choice. The slider always started at 50. In order to proceed to the next question, subjects needed to click on the slider at least once.

3.3 Recruitment and Experimental Payments

We recruited 300 subjects through the online platform Prolific.co to run the experiment. A total of nine sessions were conducted between February 22 and February 26 of 2022. Overall, we recruited 100 subjects for each treatment. Our sample consisted of United States citizens between the ages of 18-30 with at least a high-school education. We focused on this sample because most previous experiments involving common ratio questions have been conducted on undergraduate samples. Moreover, given that there were three times more women than men within the population of possible participants that met these criteria, we asked the platform to recruit

Table 2: Hypothesis EU-CORE.

	Worst	Bad	WorstBest
% hp. EU-CORE holds	57.03%	57.58%	52.08%
	(0.71)	(0.76)	(0.73)
N. of subjects	95	82	91
N. of observations	4845	4182	4641

Notes: Percentage of comparisons consistent with hypothesis EU-CORE at the individual level in the three treatments. Standard errors in parenthesis.

an equal number of men and women.[16]

Each subject received a fixed payment of \$4.75 and had a one out of ten chance of receiving a bonus payment. The software randomly selected one of the 74 comparisons for subjects that received a bonus payment. Subjects were paid the realization from whichever lottery they had chosen in the randomly selected comparison.[17]

4 Results

Table 2 summarizes the fraction of comparisons consistent with hypothesis EU-CORE in the three treatments of the experiment. In the Worst and the Bad treatments, approximately 57% of the observations are consistent with hypothesis EU-CORE. The main difference between mixed comparisons in these two treatments is that lotteries have different support in the Worst treatment while sharing the same support in the Bad treatment. However, given the statistically-indistinguishable percentage of comparisons consistent with hypothesis EU-CORE, we find that this difference is inessential. In the WorstBest treatment, the percentage of consistent comparisons goes down by approximately five percentage points. The novel approach used to construct mixed comparisons in the WorstBest treatment allows us to detect more failures of hypothesis EU-CORE. Most importantly, it will enable us in Section 4.3 to shed new light on the role of certainty as a driver for such failures.

The analysis that follows aims to uncover the relevance of decision confidence

[16]Table 7 in Appendix B summarizes the demographic information of the participants.

[17]The complete instructions with screenshots from the experiment are presented in Appendix F.

and the availability of certain alternatives as potential explanations for the violations of hypothesis EU-CORE that we observe. We begin by presenting the estimates from the two linear probability models reported in Table 3. In both regressions, the dependent variable is equal to one if the observation is consistent with hypothesis EU-CORE, zero otherwise. The two regressions differ in the variable used to measure decision confidence. Regression (1) uses the collected measures of decision confidence, while regression (2) uses response times. We find a strong positive correlation between decision confidence and the likelihood of being consistent with hypothesis EU-CORE. Moreover, subjects who spend more time choosing the preferred lottery are more likely to violate hypothesis EU-CORE.[18]

Table 3 also provides new insights into how subjects are violating the independence axiom and what is the role of certainty in driving such violations. Subjects who choose riskier over safer lotteries are approximately a 30% more likely to violate hypothesis EU-CORE. Section 4.1 will further document this observation for unmixed comparisons, concluding that the certainty effect is not the most relevant violation of hypothesis EU-CORE. The analysis of decision confidence in Section 4.2 will also provide a rationale for why this happens, showing that subjects are more likely to prefer the risky lottery over the certain prize when they are less confident.

Moreover, subjects are more likely to violate hypothesis EU-CORE in unmixed comparisons than in mixed comparisons. Taken at face value, this result is consistent with the idea that the availability of certain alternatives plays a role in driving violations of hypothesis EU-CORE. Nevertheless, in Section 4.3, we will show that this conclusion is not entirely robust to a more sophisticated analysis that allows us to control for the different stringency of the requirements that hypothesis EU-CORE imposes on different types of comparisons.

4.1 The Prevalence of the Reverse Certainty Effect

Subjects violate hypothesis EU-CORE in line with the certainty effect if they choose the certain alternative in an unmixed comparison and the riskier lottery in one of the associated mixed comparisons. The reverse certainty effect refers instead to

[18]In Appendix E, we show that response times negatively correlate with decision confidence, confirming the intuition that subjects spend more time when not confident about their choices.

Table 3: Regression results.

	Linear Probability Model					
	Dep. Variable: one if hypothesis EU-CORE holds, zero otherwise					
	(1): Measure of Confidence: Self-Report			(2): Measure of Confidence: Response Time		
	Coefficient	Robust Standard Error	P-value	Coefficient	Robust Standard Error	P-value
$Conf$	0.3611	0.0526	0.000			
$RespTime$				-0.0570	0.0123	0.000
$Risky$	-0.2893	0.0295	0.000	-0.3187	0.0294	0.000
$Treatment$						
Bad	-0.0284	0.0265	0.286	-0.0288	0.0275	0.295
WorstBest	-0.0888	0.0299	0.003	-0.0791	0.0306	0.010
$Type$						
0.95-mixed	0.0870	0.0080	0.000	0.0769	0.0076	0.000
0.7-mixed	0.2620	0.0107	0.000	0.2471	0.0109	0.000
$DistToIndiff$	0.3022	0.0552	0.000	0.3279	0.0555	0.000
Constant	0.2980	0.0440	0.000	0.6667	0.0356	0.000

Notes: Linear probability models predicting consistency with hypothesis EU-CORE at the individual level. The dependent variable is one if hypothesis EU-CORE holds, zero otherwise. $Treatment$ is equal to zero if the observation belongs to the Worst treatment, one if it belongs to the Bad treatment, and two if it belongs to the WorstBest treatment. $Type$ is equal to zero if the comparison is unmixed, one if it is 0.95-mixed, and two if it is 0.7-mixed. $Risky$ is equal to one if the subject chooses the riskier alternative, zero otherwise. $Conf$ is the reported decision confidence divided by 100. $RespTime$ is the logarithm of response times measured in seconds. $DistToIndiff$ is the absolute value difference between the fraction of safer choices in a comparison and 0.5. There are $13,668$ observations. Standard errors are clustered at the individual level (268 clusters).

Table 4: Certainty effect violations

	Worst	Bad	WorstBest
# Total Violations of Hp. EU-CORE	907	762	838
% Certainty Effect Violations	35.94%	29.92%	45.94%

Notes: Percentage of certainty effect violations (unmixed comparisons only).

the opposite behavioral pattern: subjects choosing the risky lottery in an unmixed comparison and then switching to the safer lottery in one of the associated mixed comparisons. Any failures of hypothesis EU-CORE in unmixed comparisons can be then classified as a certainty effect or a reverse certainty effect violation. In this section, we will study which of the two behavioral patterns is more frequent in our experiment.

Table 4 documents that in all treatments of the experiment, certainty effect violations are less frequent than reverse certainty effect violations. However, an important aspect that is worth considering to compare the relevance of these two types of violations is that their emergence in an experiment may be inflated or deflated depending on the relative attractiveness of risky lotteries. For instance, if most subjects prefer risky lotteries over certain prizes in unmixed comparisons, we will have more data to test the certainty effect. In the benchmark scenario of an EU subject who makes mistakes, more data to test the certainty effect automatically translates into higher expected certainty effect violations.

For this reason, we examine the emergence of the certainty effect controlling for the fraction of subjects in each unmixed comparison that prefer the certain prize over the risky lottery. Figure 3 shows for each unmixed comparison the percentage of subjects choosing the certain prize on the x-axis and the percentage of reverse certainty effect violations over all violations of hypothesis EU-CORE on the y-axis. The blue circles represent unmixed comparisons in the Worst treatment, the red circles in the Bad treatment, and the green circles in the WorstBest treatment. The size of the circles informs about the overall percentages of violations of hypothesis EU-CORE, with bigger circles corresponding to higher ranges of percentages.

The fact that the relevance of the certainty effect increases as the fraction of subjects choosing the certain prize in unmixed comparisons increases suggests that one

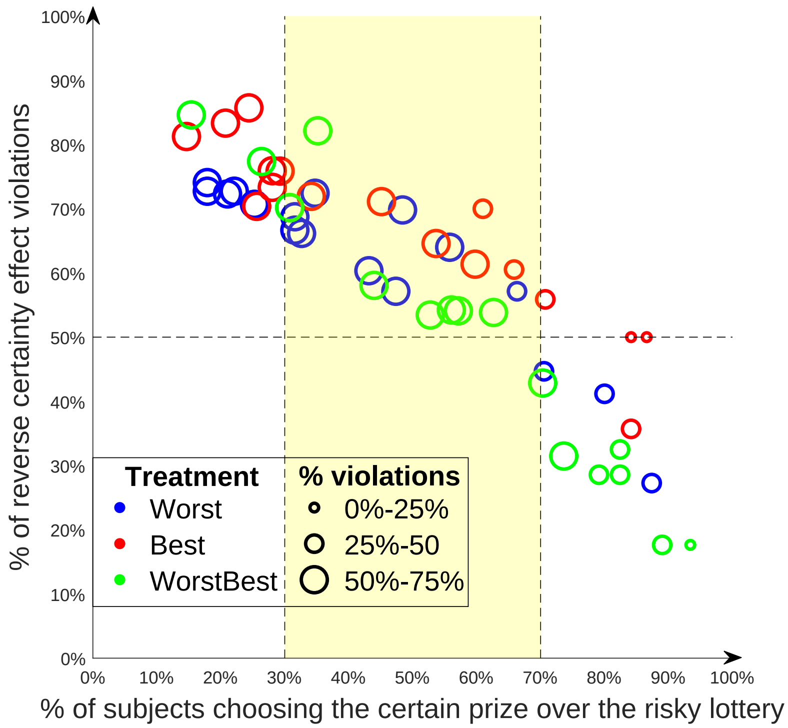


Figure 3: Percentages of reverse certainty effect violations for all comparisons.

has to take seriously the potential bias arising from the unbalance in the available data discussed above. Nevertheless, "fair" comparisons can be made by looking at unmixed comparisons in which both lotteries are chosen by a non-negligible fraction of subjects. For instance, the shaded yellow region in Figure 3 includes all the unmixed comparisons for which the percentage of subjects choosing the certain prize is between 30% and 70%. For comparisons in this region, reverse certainty effect violations are always more frequent than certainty effect violations.

The size of the circles in Figure 3 provides further evidence against the relevance of the certainty effect. As the fraction of subjects choosing the certain prize increases, the size of the circles tends to be smaller. In other words, the more data we have to observe certainty effect violations of hypothesis EU-CORE, the smaller the total number of hypothesis EU-CORE failures that we observe.[19]

[19]In Appendix D, we provide additional evidence of the relevance of the reverse certainty effect by studying the implications of the NCI and PCI axioms on risk attitude.

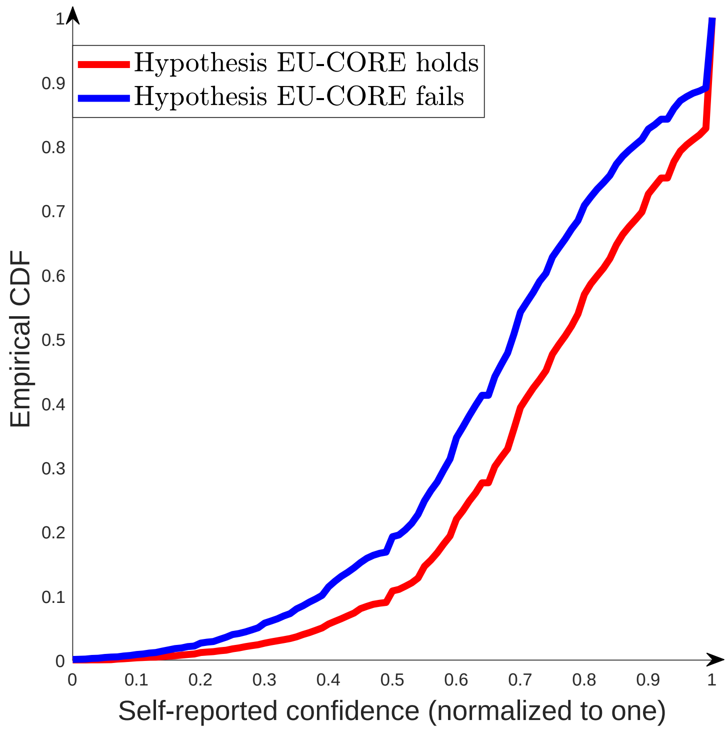


Figure 4: Hypothesis EU-CORE and confidence in all treatments.

4.2 A Possible Mechanism: Decision Confidence

The estimates of the linear probability models in Table 3 indicate that hypothesis EU-CORE is more likely to hold when subjects declare to be confident about their decisions. Figure 4 provides a graphical representation of this result, describing the empirical cumulative distribution functions of decision confidence for observations in which hypothesis EU-CORE holds (red distribution) and does not hold (blue distribution). The red distribution in Figure 4 stochastically dominates the blue one, indicating that conditional on being consistent with hypothesis EU-CORE, subjects in our experiment tend to report higher confidence levels.[20]

The positive correlation between decisions with low confidence and failures of

[20]In Appendix C we further explore the relationship between hypothesis EU-CORE and decision confidence using the notion of "indecisiveness" introduced by Cerreia-Vioglio et al. (2015). An individual is more indecisive than another if his EU core is smaller in the sense of set inclusion. Our analysis provides an empirical justification for the use of the term "indecisive," showing that more indecisive individuals tend to report lower levels of decision confidence.

hypothesis EU-CORE summarized in Figure 4 on the one hand, and the prevalence of the reverse certainty effect documented in Section 4.1 on the other hand, jointly suggest that subjects tend to prefer risky lotteries over certain prizes when reporting low confidence levels. The PCI axiom theorizes this form of incaution in a way that is easily understandable using the notion of EU core. In its canonical formulation, a subject i satisfies the PCI axiom if for all lotteries $p \in \triangle(X)$ and prizes $x \in X$,

$$\delta_x \succsim_i p \Rightarrow \delta_x \succsim_i^* p.$$

In words, the PCI axiom precludes subject i from violating the independence axiom in line with the certainty effect but does not impose any constraint on the reverse certainty effect. An equivalent and insightful way to express the PCI axiom is: for all lotteries $p \in \triangle(X)$ and prizes $x \in X$,

$$\neg\, [\delta_x \succsim^* p] \Rightarrow p \succ \delta_x.$$

Building on the interpretation of the EU core as the subset of uncontroversial comparisons that our analysis supports, the PCI axiom has the following interpretation: when a certain prize is not confidently better than a risky lottery, the risky lottery should be strictly preferred. Motivated by this interpretation, we test whether subjects are more likely to choose the risky lottery or the certain prize in unmixed comparisons when reporting low confidence levels.

Figure 5 shows the likelihood of choosing risky lotteries in unmixed comparisons as a function of the reported level of decision confidence. As decision confidence decreases, subjects are more likely to choose risky lotteries over certain prizes. This finding is consistent with the idea hinted by the PCI axiom of individuals being incautious rather than cautious when reporting low levels of decision confidence. Moreover, it provides a rationale for the prevalence of the reverse certainty effect documented in Section 4.1.

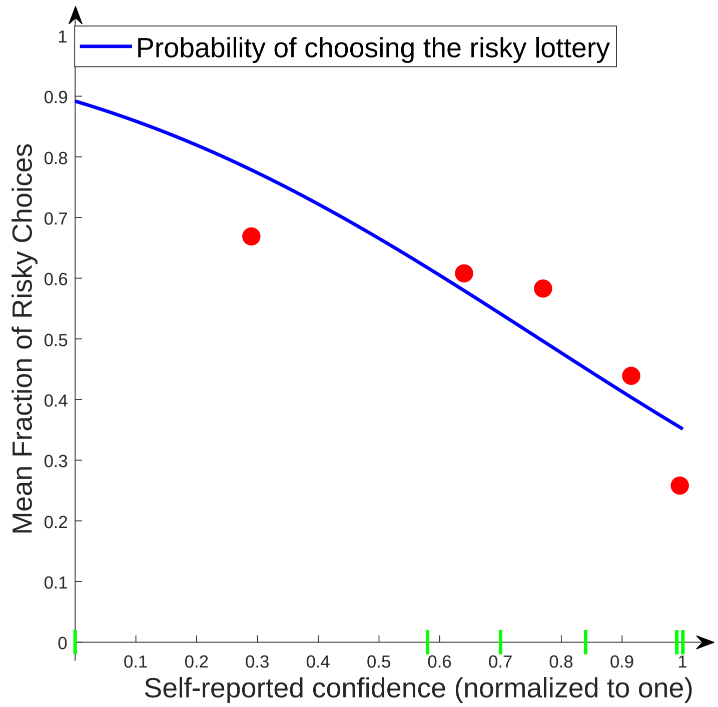


Figure 5: Confidence and preference for risky lotteries in unmixed comparisons.

Notes: We partition observations into five categories of equal size based on decision confidence. The green markers on the x-axis denote the thresholds for each category. The red dots placed at the average confidence values in each category represent the fractions of observations in which the risky lottery is preferred over the certain prize. The predicted probabilities of choosing the risky lottery are computed using a probit model.

4.3 Certainty Does Not Always Matter

Our analysis of behavior thus far highlights the reverse certainty effect as the most relevant violation of hypothesis EU-CORE in unmixed comparisons. In this section, we take a step back and study whether the mere presence of a certain alternative in a comparison is predictive of more violations of hypothesis EU-CORE. The estimation results in Table 3 provide a first positive answer to this question, showing that hypothesis EU-CORE is more likely to fail in unmixed than in mixed comparisons. However, using the same logic adopted to compare certainty effect and reverse certainty effect violations, we now show that this result does not account for the amount of information we have to disprove hypothesis EU-CORE for different categories

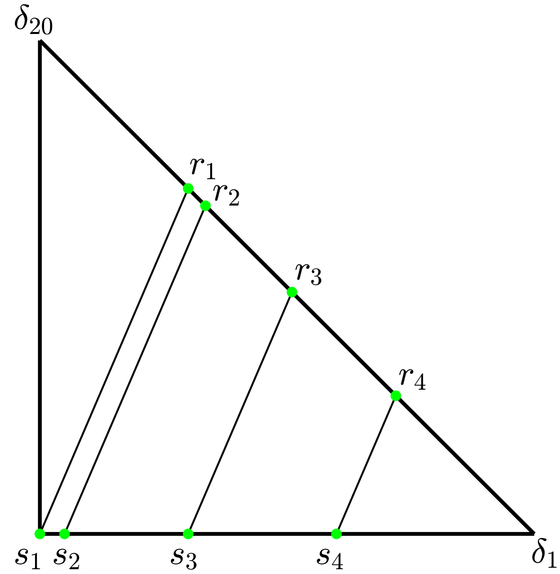


Figure 6: Four comparisons from the Worst treatment.

of comparisons.

Let us consider the four comparisons from the Worst treatment represented in Figure 6 and imagine that a subject declared to prefer lottery s_1 over lottery r_1. Disproving hypothesis EU-CORE for the unmixed comparison (s_1, r_1) amounts to observing a preference for the riskier lottery (r_2, r_3 or r_4) in any of the three associated mixed comparisons. On the contrary, hypothesis EU-CORE for mixed comparisons does not impose any constraint on the preferences expressed in the unmixed comparison. For instance, let us imagine that a subject preferred lottery s_2 over lottery r_2 in the 0.95-mixed comparison (s_2, r_2). Disproving hypothesis EU-CORE for this comparison amounts to observing a preference for the riskier lottery (r_3 or r_4) in any of the remaining mixed comparisons.

Consequently, we have more data to disprove hypothesis EU-CORE for unmixed comparisons than we have for mixed comparisons. To see why this asymmetry can lead to overestimating the role of certainty, let us consider an individual that satisfies EU but with probability 0.1 makes independent mistakes in each of the four comparisons in Figure 6 by choosing the least preferred option. In this case, the probability that this individual satisfies hypothesis EU-CORE in the unmixed comparison (s_1, r_1) is 0.6562, in the 0.95-mixed comparison (s_2, r_2) is 0.73 while in the 0.7-mixed comparison (s_3, r_3) is 0.82. We now describe how we account for this asymmetry by exploiting the error model proposed by Harless and Camerer (1994).

We summarize subjects' choices over the four comparisons in Figure 6 by strings of chosen lotteries. For instance, choosing lottery s_i over lottery r_i for every index $i \in \{1,2,3,4\}$ corresponds to the string $s_1s_2s_3s_4$. In the error model of Harless and Camerer (1994), subjects have strict preferences over lotteries but can make mistakes choosing the least preferred lottery. In each comparison, mistakes happen with probability $\varepsilon \in (0,1)$ and are independent across choices. For instance, a subject with true preferences $s_1s_2s_3s_4$ with probability $\varepsilon(1-\varepsilon)^3$ makes one error and report $r_1s_2s_3s_4$, $s_1r_2s_3s_4$, $s_1s_2r_3s_4$ or $s_1s_2s_3r_4$. We denote by $x_1x_2x_3x_4$ a generic string of chosen lotteries and define by $p(x_1x_2x_3x_4)$ the fraction of subjects in the experiment for which in the absence of mistakes we would observe $x_1x_2x_3x_4$, where $x_i \in \{s_i, r_i\}$ and $i \in \{1,2,3,4\}$. For instance, $p(s_1s_2s_3s_4)$ is the fraction of subjects preferring lottery s_i over lottery r_i for every index $i \in \{1,2,3,4\}$.

Within this framework, consistency with hypothesis EU-CORE in all comparisons amounts to assuming EU:

$$p(s_1s_2s_3s_4) + p(r_1r_2r_3r_4) = 1. \tag{EU}$$

If, instead, we allow hypothesis EU-CORE to fail in unmixed comparisons, the relaxed model becomes:

$$\sum_{x_1 \in \{s_1, r_1\}} p(x_1s_2s_3s_4) + p(x_1r_2r_3r_4) = 1. \tag{CC}$$

Therefore, no matter how "close to certainty" one of the two lotteries in a comparison is, model CC requires consistency with hypothesis EU-CORE. Finally, allowing for failures of hypothesis EU-CORE in unmixed and 0.95-mixed comparisons leads us to the following model:

$$\sum_{x_2 \in \{s_2, r_2\}} \sum_{x_1 \in \{s_1, r_1\}} p(x_1x_2s_3s_4) + p(x_1x_2r_3r_4) = 1. \tag{AC}$$

Table 5: Likelihood ratio tests.

	Worst	Bad	WorstBest
CC-AC	7	9	0
EU-AC	10	3	1
CC-EU	0	0	0
EU-EU	0	5	16

Notes: Each treatment has 17 patterns of comparisons. This table classifies each pattern of comparisons into four possible categories. We denote by i-j the category of all patterns in which model $i \in \{\mathrm{EU}, \mathrm{CC}\}$ prevails in a likelihood ratio test between model EU and model CC, while model $j \in \{\mathrm{EU}, \mathrm{AC}\}$ in a likelihood ratio test between model EU and model AC.

In other words, model AC requires consistency with hypothesis EU-CORE only for comparisons in which both lotteries are "away from certainty".

In each of the three model specifications, the fractions of true preferences and the error term can be estimated using maximum likelihood estimation.[21] The unit of observation in this analysis is the pattern of choices in an unmixed comparison and the three associated mixed comparisons. Figure 6 shows an example of unmixed and associated mixed comparisons from the Worst treatment. Each treatment has 17 unmixed comparisons with their own associated three mixed comparisons. Therefore, we estimate our three models in each treatment 17 times, one for each of the unmixed and associated mixed comparisons.

To evaluate the relevance of certainty for violations of hypothesis EU-CORE, we perform two likelihood ratio tests. The first test compares model EU with model CC, while the second test model EU with model AC. Table 5 summarizes the results of these likelihood ratio tests. In the Worst treatment, model EU is always rejected against either model CC or model AC. In the Bad treatment, accommodating for failures of hypothesis EU-CORE in unmixed comparisons also allows explaining our data significantly better, with the exception of five patterns of comparisons. However, the results in the WorstBest treatment completely overturn this conclusion. In this latter treatment, for 16 out of 17 patterns of comparisons, model EU is never rejected. In other words, allowing for violations of hypothesis EU-CORE in unmixed or 0.95-mixed comparisons does not help to explain our data better.

[21] We refer to Harless and Camerer (1994) for a detailed description of the likelihood function.

5 Discussion

This study sheds new light on what drives violations of the independence axiom. We conduct an experimental investigation involving choices between risky lotteries. Our main finding is that subjects are more likely to be consistent with the independence axiom when they report high decision confidence levels. In this way, we provide empirical support for the psychological interpretation of the EU core as the subset of the uncontroversial rankings. We believe that exploiting the notion of EU core in experimental works, as we do in our paper, represents a promising direction to expand our understanding of decision-making under risk.

Moreover, we analyze decision-making under low decision confidence. Contrary to the certainty effect rationale for independence violations, subjects are more likely to choose a risky lottery over a certain prize and violate the independence axiom when not confident. Given the extensive evidence on the certainty effect and the impact that this evidence had and still has on new theoretical models, more research is plainly needed to test the robustness of our conclusion. An important insight of our work is that the certainty effect may be less relevant in environments where subjects face a greater variety of lotteries than in the Allais paradox.

Our data also questions the relevance of certainty itself. In the WorstBest treatment, where we construct mixed comparisons alternating the third common lottery, we detect more independence violations. Remarkably, we also find that in this treatment, the presence of certain alternatives does not increase independence violations. To our knowledge, this is the first study that alternates the third common lotteries to build mixed comparisons. Given the different conclusions we obtain in the *WorstBest* treatment, we believe that exploring new ways to construct mixed comparisons constitutes a promising line of research for studying the independence axiom.

References

Agranov, M. and Ortoleva, P. (2017). Stochastic choice and preferences for randomization. *Journal of Political Economy*, 125(1):40–68.

Aitchison, L., Bang, D., Bahrami, B., and Latham, P. E. (2015). Doubly bayesian analysis of confidence in perceptual decision-making. *PLoS computational biology*, 11(10):e1004519.

Andreoni, J. and Harbaugh, W. (2009). Unexpected utility: Experimental tests of five key questions about preferences over risk.

Andreoni, J. and Sprenger, C. (2010). Certain and uncertain utility: The allais paradox and five decision theory phenomena. *Levine's Working Paper Archive*, 926159295.

Andreoni, J. and Sprenger, C. (2011). Uncertainty equivalents: Testing the limits of the independence axiom. Technical report, National Bureau of Economic Research.

Andreoni, J. and Sprenger, C. (2012). Risk preferences are not time preferences. *American Economic Review*, 102(7):3357–76.

Arts, S., Ong, Q., and Qiu, J. (2022). Measuring subjective decision confidence.

Bernheim, B. D. and Sprenger, C. (2020). On the empirical validity of cumulative prospect theory: Experimental evidence of rank-independent probability weighting. *Econometrica*, 88(4):1363–1409.

Blavatskyy, P., Ortmann, A., and Panchenko, V. (2022a). On the experimental robustness of the allais paradox. *American Economic Journal: Microeconomics*, 14(1):143–63.

Blavatskyy, P., Panchenko, V., and Ortmann, A. (2022b). How common is the common-ratio effect? *Experimental Economics*, pages 1–20.

Blavatskyy, P. R. (2010). Reverse common ratio effect. *Journal of Risk and Uncertainty*, 40(3):219–241.

Boldt, A., Blundell, C., and De Martino, B. (2019). Confidence modulates exploration and exploitation in value-based learning. *Neuroscience of consciousness*, 2019(1):niz004.

Butler, D. and Loomes, G. (2011). Imprecision as an account of violations of independence and betweenness. *Journal of Economic Behavior & Organization*, 80(3):511–522.

Camerer, C. F. (1992). Recent tests of generalizations of expected utility theory. In *Utility theories: Measurements and applications*, pages 207–251. Springer.

Cerreia-Vioglio, S. (2009). Maxmin expected utility on a subjective state space: Convex preferences under risk. Technical report, Mimeo, Bocconi University.

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

Cerreia-Vioglio, S., Dillenberger, D., and Ortoleva, P. (2020). An explicit representation for disappointment aversion and other betweenness preferences. *Theoretical Economics*, 15(4):1509–1546.

Cerreia-Vioglio, S., Dillenberger, D., Ortoleva, P., et al. (2022). *Caution and Reference Effects*. Penn Institute for Economic Research, Department of Economics, University of ….

Cubitt, R. P., Navarro-Martinez, D., and Starmer, C. (2015). On preference imprecision. *Journal of Risk and Uncertainty*, 50:1–34.

da Silva Castanheira, K., Fleming, S. M., and Otto, A. R. (2021). Confidence in risky value-based choice. *Psychonomic Bulletin & Review*, 28:1021–1028.

De Martino, B., Fleming, S. M., Garrett, N., and Dolan, R. J. (2013). Confidence in value-based choice. *Nature neuroscience*, 16(1):105–110.

Desender, K., Boldt, A., and Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological science*, 29(5):761–778.

Dillenberger, D. (2010). Preferences for one-shot resolution of uncertainty and allais-type behavior. *Econometrica*, 78(6):1973–2004.

Enke, B. and Graeber, T. (2019). Cognitive uncertainty. Technical report, National Bureau of Economic Research.

Folke, T., Jacobsen, C., Fleming, S. M., and De Martino, B. (2016). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, 1(1):1–8.

Halevy, Y. (2008). Strotz meets allais: Diminishing impatience and the certainty effect. *American Economic Review*, 98(3):1145–62.

Harless, D. W. and Camerer, C. F. (1994). The predictive utility of generalized expected utility theories. *Econometrica: Journal of the Econometric Society*, pages 1251–1289.

Hey, J. D. and Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica: Journal of the Econometric Society*, pages 1291–1326.

Jain, R. and Nielsen, K. (2022). A systematic test of the independence axiom near certainty.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–291.

Machina, M. J. (1982). " expected utility" analysis without the independence axiom. *Econometrica: Journal of the Econometric Society*, pages 277–323.

Machina, M. J. (1987). Choice under uncertainty: Problems solved and unsolved. *Journal of Economic Perspectives*, 1(1):121–154.

Marschak, J. (1950). Rational behavior, uncertain prospects, and measurable utility. *Econometrica: Journal of the Econometric Society*, pages 111–141.

Meyniel, F. and Dehaene, S. (2017). Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proceedings of the National Academy of Sciences*, 114(19):E3859–E3868.

Meyniel, F., Schlunegger, D., and Dehaene, S. (2015). The sense of confidence during probabilistic learning: A normative account. *PLoS computational biology*, 11(6):e1004305.

Oprea, R. (2022). Simplicity equivalents. Technical report, Working Paper.

Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., and Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature communications*, 11(1):2634.

Segal, U. (1990). Two-stage lotteries without the reduction axiom. *Econometrica: Journal of the Econometric Society*, pages 349–377.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of economic literature*, 38(2):332–382.

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

FOR ONLINE PUBLICATION

Appendices

A FOSD Questions

Table 6: Comparisons with dominated lotteries.

	Lottery A			Lottery B		
Questions	Pr($1)	Pr($7)	Pr($20)	Pr($1)	Pr($7)	Pr($20)
1	0.3	0.7	0	0.7	0.3	0
2	0.5	0.2	0.3	0.5	0.1	0.4
3	0	0.8	0.2	0	0.6	0.4
4	0.5	0.4	0.1	0.9	0	0.1
5	0.2	0.6	0.2	0.1	0.5	0.4
6	0.5	0.5	0	0.8	0.2	0

Notes: Table 6 describes the six comparisons in which one lottery first-order stochastically dominates the other included in all treatments of the experiment.

B Demographic Summary

Table 7: Demographics, overall sample.

Age	
18-24	54.3%
25-30	45.7%
Gender	
Male	50%
Female	50%
Prefer not to say	0%
Education completed	
High school diploma	50.3%
Undergraduate degree (BA/BSc/other)	40%
Master degree (MA/MSc/MPhil/other)	7%
Doctorate degree (PhD/other)	1.7%
Prefer not to say	1%
Employment	
Not in paid work	5%
Unemployed (and job seeking)	22.3%
New job within the next month	0.7%
Part-time	11.3%
Full-time	22.3%
Other	9%
Prefer not to say	29.3%
Respondents	300

Notes: Information provided by the online platform Prolific.co.

C Confidence and Indecisiveness

Cerreia-Vioglio et al. (2015) classify one individual as more indecisive than another if his EU core is smaller, in the sense of set inclusion.

Definition 1. *Individual i is more indecisive than individual j if for all lotteries p, q:*

$$p \succsim_i^* q \Rightarrow p \succsim_j^* q.$$

We show that the term "indecisive" presumes a relationship between confidence and independence that we observe in the data: subjects that we classify as more indecisive tend to report lower confidence levels. To this end, Section C.1 details our approach to operationalize the notion of indecisiveness. Section C.2 builds a measure of decision confidence comparable across subjects, while Section C.3 uses this new measure to rank subjects in terms of decision confidence. Finally, Section C.3 shows that more indecisive subjects tend to be less confident.

C.1 Pairwise Analysis: Indecisiveness

We denote by t a generic treatment in the experiment and by C_t the set of comparisons in treatment t. For any comparison (s, r), we define the index $Core_i$ for each individual i to be equal to two if there is no evidence against $s \succsim_i^* r$, one if there is no evidence against $r \succsim_i^* s$, and zero otherwise. Using this index, we propose two different criteria to operationalize the notion of indecisiveness proposed by Cerreia-Vioglio et al. (2015) in the context of our experiment:

1. Criterion IND 1: subject i is more indecisive than subject j if for all comparisons $(s, r) \in C_t$,

   $$Core_i(s, r) > 0 \Rightarrow Core_i(s, r) = Core_j(s, r)$$

   and there exists a comparison $(\bar{s}, \bar{r}) \in C_t$ such that $Core_i(\bar{s}, \bar{r}) \neq Core_j(\bar{s}, \bar{r})$.

2. Criterion IND 2: subject i is more indecisive than subject j if

   $$|\{(s, r) \in C_t \, : \, Core_i(s, r) > 0\}| < |\{(s, r) \in C_t \, : \, Core_j(s, r) > 0\}|.$$

Criterion IND 1 amounts to classifying a subject as more indecisive than another only if there is no evidence against it. Because it is very demanding, we show that it does not allow the classification of most pairs of subjects in each treatment. On the contrary, to classify one subject as more indecisive than another according to criterion IND 2, it is enough to compare the number of comparisons with index *Core* greater than zero. The advantage of criterion IND 2 is that being less demanding, it allows the classification of the vast majority of pairs of subjects. The disadvantage is that it moves away from the original notion of indecisiveness captured in Definition 1. In Section C.4, we show that the qualitative results linking indecisiveness with decision confidence are stable across the two criteria.

C.2 Benchmarked Confidence Self-Reports

An important challenge in interpreting the self-reported confidence measures that we collect is that they are subjective and may have different meanings for different subjects. We now describe the approach that we adopt to convert the confidence reports of different subjects into the same unit of measure. To this end, we exploit subjects' confidence statements in comparisons involving stochastic dominance.

Let $FOSD$ be the set of the six comparisons in which one lottery first-order stochastically dominates the other described in Appendix A. Figure 7 shows the distributions of confidence self-reports for different categories of comparisons.[22] A few things emerge from this figure. First, as expected, subjects tend to report higher confidence levels in $FOSD$ (leftmost box plot in Figure 7) than in any other category of comparisons. However, while the median value of the confidence self-reports for $FOSD$ comparisons is close to 100, the interquartile range is equal to 25. In other words, there is significant heterogeneity in how subjects express high confidence using numbers. In what follows, we exploit this heterogeneity to benchmark the confidence self-reports in all other categories of comparisons.

We denote by $Conf_i(s,r)$ the confidence self-report of subject i in comparison (s,r) divided by 100. Moreover, we indicate by $\bar{c}_i$ the average confidence self-report in $FOSD$ comparisons for which subject i declares to prefer the dominant lottery.

[22]The red plus signs denote outliers. We classify any observation that is more than 1.5 times the interquartile range away from the first quantile or the third quantile as an outlier.

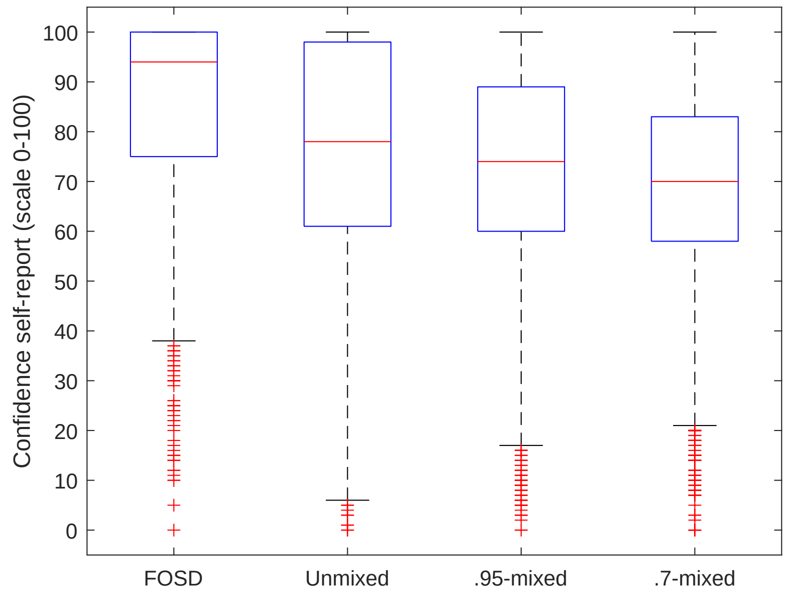


Figure 7: Distribution of confidence self-reports for different categories of comparisons and all subjects in the experiment.

Notes: We classify as outlier any observation that is more than 1.5 times the interquartile range away from the first quantile or the third quantile. The red markers denote outliers.

For all comparisons (s,r) not in $FOSD$ and all subjects i, we construct the following benchmarked index of confidence:

$$AdjConf_i(s,r) = \min\left\{\frac{Conf_i(s,r)}{\bar{c}_i}, 1\right\}.$$

Intuitively, $\bar{c}_i$ is an estimate of what number individual i reports to express extreme confidence. In the extreme case in which $Conf_i(s,r)$ is greater or equal to $\bar{c}_i$, we simply assign to the benchmarked index $AdjConf_i(s,r)$ the value of one.[23]

The underlying assumption behind this benchmarked confidence index is that subjects who use lower numbers to express extreme confidence in FOSD comparisons use lower numbers to express any confidence level. We now test the validity of this assumption. Figure 8 shows the mean confidence levels in unmixed and

[23] This happens overall for the 24.72% of the self-reported confidence measures.

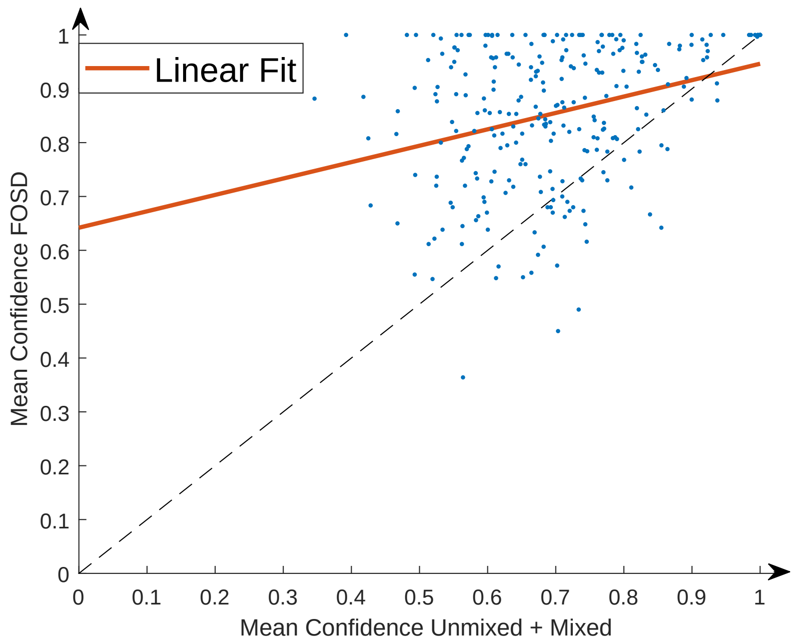


Figure 8: Mean confidence in FOSD comparisons and in all other comparisons for each subject normalized to one.

mixed comparisons (x-axis) and in FOSD comparisons (y-axis) for each subject.[24] The orange line in Figure 8 represents the best linear fit (in a least-squares sense). The positive correlation that we observe between these two quantities justifies our benchmarked index: subjects expressing extreme confidence in FOSD comparisons with lower numbers express any confidence with lower numbers.[25] Figure 8 also allows us to evaluate the reliability of the confidence measures that we collect: for the 83.96% of subjects, the mean confidence in FOSD comparisons is higher than in all other comparisons.

[24]For FOSD comparisons, we exclude confidence self-reports in pairs of lotteries where the dominated lottery was preferred. Given our restriction on the sample, this can happen at most once for each subject.

[25]Correlation coefficient: 0.303.

C.3 Pairwise Analysis: Confidence

In analogy with the pairwise analysis of indecisiveness in Section C.1, we propose two different criteria to rank pairs of subjects in terms of decision confidence using the benchmarked index $AdjConf_i$:

1. Criterion CONF 1: subject i is less confident than subject j if for all comparisons $(s,r) \in C_t$,
$$AdjConf_i(s,r) \leq AdjConf_j(s,r),$$
with strict inequality for some comparison $(\bar{s},\bar{r}) \in C_t$.

2. Criterion CONF 2: subject i is less confident than subject j if for all $x \in [0,1]$,
$$|\{(s,r) \in C_t \, : AdjConf_i(s,r) \leq x\}| \leq |\{(s,r) \in C_t \, : AdjConf_j(s,r) \leq x\}|,$$
with strict inequality for at least one $\bar{x} \in (0,1)$.

Using criterion CONF 1, we classify subject i as less confident than subject j if the benchmarked confidence self-report of i is lower than the benchmarked confidence self-report of j in all comparisons. As with criterion IND 1 for indecisiveness, criterion CONF 1 is very demanding and does not allow the classification of most pairs of subjects. For this reason, we also consider criterion CONF 2, which generalizes the requirement of criterion IND 2 to the continuous index $AdjConf$. According to criterion CONF 2, subject i is less confident than subject j whenever the empirical cumulative distribution function of $AdjConf_j$ first-order stochastically dominates the empirical cumulative distribution function of $AdjConf_i$.

C.4 Results

We now explore the relationship between EU core and decision confidence by using the notion of indecisiveness introduced by Cerreia-Vioglio et al. (2015). In Section C.1 and Section C.3 we propose two criteria to rank subjects in terms of indecisiveness (IND 1 and IND 2) and confidence (CONF 1 and CONF 2). Considering all four possible combinations of these approaches, Table 8 reports the percentage of

Table 8: Confidence and indecisiveness.

Worst			Bad			WorstBest		
	IND 1	IND 2		IND 1	IND 2		IND 1	IND 2
CONF 1	75.00% (4)	66.67% (243)	CONF 1	65.22% (23)	65.27% (334)	CONF 1	54.00% (100)	67.30% (523)
CONF 2	68.85% (61)	59.44% (2,125)	CONF 2	86.49% (111)	60.76% (1,909)	CONF 2	64.22% (218)	56.66% (2,344)

Notes: Percentage of pairs of subjects in which the more indecisive subject is less confident. Number of classifiable pairs in parenthesis.

pairs of subjects, among those that can be classified according to both criteria, for which the more indecisive subject is less confident. The number in parenthesis below each percentage in Table 8 represents the total number of classifiable pairs. For instance, in the Worst treatment, 243 pairs of subjects can be classified according to criterion CONF 1 and criterion IND 2. In the 66.67% of these pairs, the more indecisive subject is less confident.

Overall, for all the possible combinations of criteria in all three treatments, the more indecisive subject is less confident in more than half of the classifiable pairs. We see this result as an empirical justification of the term “indecisive” used in Cerreia-Vioglio et al. (2015) to rank subjects’ EU cores.

Table 9: Indecisiveness and risk aversion.

Worst			Bad			WorstBest		
	IND 1	IND 2		IND 1	IND 2		IND 1	IND 2
RA 1	13.64%	48.85%	RA 1	10.71%	27.33%	RA 1	10.11%	31.03%
	(110)	(4,274)		(168)	(3,154)		(356)	(3,938)
RA 2	12.61%	47.79%	RA 2	10.71%	28.38%	RA 2	12.40%	33.44%
	(111)	(4,340)		(168)	(3,213)		(363)	(4,007)

Notes: Percentage of pairs of subjects in which the more indecisive subject is more risk averse. Number of classifiable pairs in parenthesis.

D Indecisiveness and Risk Aversion

Our analysis shows that the independence axiom is less likely to fail when the safer lottery in a comparison is chosen. Given that the extent to which subjects prefer safer over riskier lotteries positively correlates with their degree of risk aversion, these estimates suggest a negative relationship between indecisiveness and risk aversion. At the same time, Cerreia-Vioglio et al. (2015) show that the opposite relationship holds for preferences that satisfy the NCI axiom. If a subject is more indecisive than another, he must be more risk averse. We now propose a direct test for this prediction.

We adopt two criteria to rank subjects' risk attitudes. The first criterion (RA 1) consists in classifying one subject as more risk averse than another if he chooses the safer over the riskier lottery more often in the experiment. In the second criterion (RA 2), we instead normalize the utility of \$1 to zero, the utility of \$20 to one, and following Hey and Orme (1994), we estimate the utility of \$7 and of the variance of the error term which we assume to be normally distributed with zero mean. The higher the estimated utility, the more risk averse a subject is.

Table 9 reports the percentage of pairs of subjects, among those that can be classified according to both indecisiveness and risk aversion, for which the more indecisive subject is more risk averse. Below each percentage in Table 9 is reported the total number of classifiable pairs. For all the possible combinations of criteria in all treatments, more indecisive subjects are less risk averse. In what follows, we prove that preferences satisfying the PCI axiom may explain the correlation between indecisiveness and risk aversion that we observe in Table 9.

Definition 2. *Let $\succsim$ be a binary relation over the set of lotteries $\triangle(X)$. We say that $\succsim$ satisfies:*[26]

- *Completeness if for each $p,q \in \triangle(X)$, either $p \succsim q$ or $q \succsim p$.*
- *Transitivity if for each $p,q,r \in \triangle(X)$, $p \succsim q$ and $q \succsim r$ imply $p \succsim r$.*
- *Continuity if for each $p \in \triangle(X)$, the sets $\{q \in \triangle(X) : q \succsim p\}$ and $\{q \in \triangle(X) : p \succsim q\}$ are closed.*
- *Strict first-order stochastic dominance if for each $p \in \triangle(X)$, if p first-order stochastically dominates q, then $p \succ q$.*
- *Betweenness if for each $p,q \in \triangle(X)$ and $\lambda \in [0,1]$,*

$$p \sim q \Rightarrow p \sim \lambda p + (1-\lambda)q \sim q.$$

We refer to binary relations that satisfy the five axioms in Definition 2 as betweenness preferences. Next, we introduce the PCI axiom.

Definition 3. *$\succsim$ satisfies the PCI axiom if for each $p,q \in \triangle(X)$, $x \in X$ and $\lambda \in [0,1]$,*

$$\delta_x \succsim p \Rightarrow \lambda \delta_x + (1-\lambda)q \succsim \lambda p + (1-\lambda)q.$$

Let $\succsim_1$ and $\succsim_2$ be the preferences over $\triangle(X)$ for individuals 1 and 2. We use the following standard approach to compare risk attitudes.

Definition 4. *Individual* 1 *is more risk averse than individual* 2 *if for each $p \in \triangle(X)$ and $x \in X$,*

$$p \succsim_1 \delta_x \Rightarrow p \succsim_2 \delta_x.$$

We are now ready to show that betweenness preferences satisfying the PCI axiom rationalize what we observe in our experiment: more indecisive subjects tend

[26] We denote by $\sim$ and $\succ$ the symmetric and the asymmetric parts of $\succsim$. The set X represents any compact set of monetary prizes.

to be less risk averse. The result follows combining the representation result in Cerreia-Vioglio et al. (2020, Remark 1) with the proof technique used in Cerreia-Vioglio et al. (2015, Proposition 2).

Corollary 1. *Suppose that $\succsim_1$ and $\succsim_2$ are betweenness preferences that satisfy the PCI axiom. If individual 1 is more indecisive than individual 2, then individual 1 is less risk averse than individual 2.*

Proof. We denote by $\mathbb{E}[p,v]$ and $c(p,v)$ the EU and the certainty equivalent of lottery p with utility function v, respectively. By Cerreia-Vioglio et al. (2020), for each individual $i \in \{1,2\}$, there exists a set of utility functions $\mathscr{W}_i$ such that

$$p \succsim_i^* q \Leftrightarrow \mathbb{E}[p,v] \geq \mathbb{E}[q,v] \text{ for all } v \in \mathscr{W}_i,$$

and the functional $u_i \colon \triangle(X) \to \mathbb{R}$ defined by

$$u_i(p) = \sup_{v \in \mathscr{W}_i} c(p,v) \text{ for all } p \in \triangle(X),$$

is a continuous utility representation of $\succsim_i$. Given that individual 1 is more indecisive than individual 2,

$$p \succsim_1^* q \Rightarrow p \succsim_2^* q \Rightarrow p \succsim_2 q.$$

By Cerreia-Vioglio (2009, Proposition 22), $\mathscr{W}_2 \subseteq \mathscr{W}_1$. Therefore, for any risky lottery p,

$$u_1(p) = \sup_{v \in \mathscr{W}_1} c(p,v) \geq \sup_{v \in \mathscr{W}_2} c(p,v) = u_2(p).$$

Consequently, individual 1 is less risk averse than individual 2. □

E Response Times

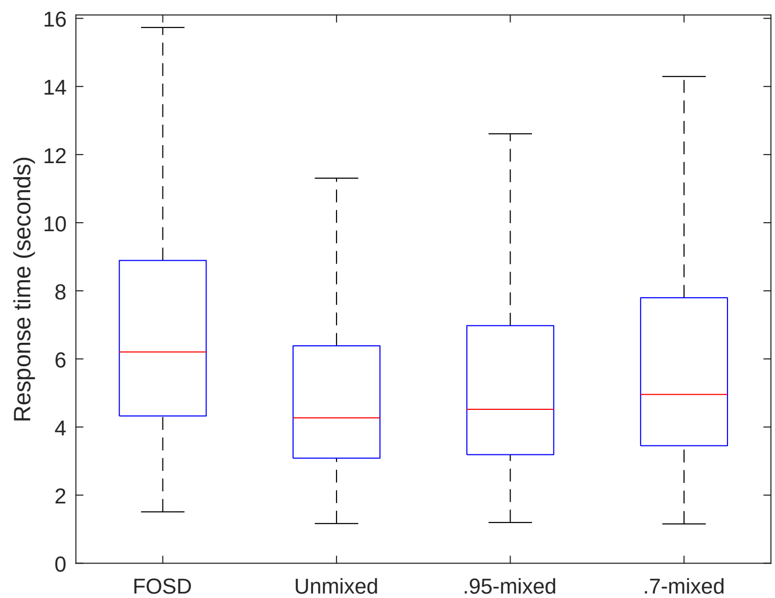


Figure 9: Distribution of response times for all subjects in the experiment.

Notes: We classify as outlier any observation that is more than 1.5 times the interquartile range away from the first quantile or the third quantile. Overall, there are 1,322 outliers. Given that many outliers correspond to particularly high numbers, representing subjects that most likely took a break during the experiment, we do not report them in Figure 9 for clarity.

Figure 9 shows the distribution of these response times for different categories of comparisons. Contrary to the confidence self-reports, response times tend to be lower in unmixed than in 0.95-mixed and 0.7-mixed comparisons. Indeed, we observe a negative correlation between confidence self-reports and response times for these three categories of comparisons.[27] At the same time, analogously to confidence self-reports, response times tend to be higher in FOSD comparison than in any other category of comparisons. We suspect that the high response times observed in FOSD comparisons may result from a "too good to be true" effect. Subjects may lose time in double-checking their understanding of questions in FOSD comparisons, given the lack of trade-offs between the two lotteries.

[27] Correlation coefficient: -0.206. P-value less than 0.001.

F Instructions

General Instructions. The study consists of questions about lottery tickets that pay \$1, \$7 or \$20 with some fixed probabilities. Let us highlight from the start that there are no right or wrong answers. We are only interested in studying your preferences. Here is an example of a pair of lottery tickets:

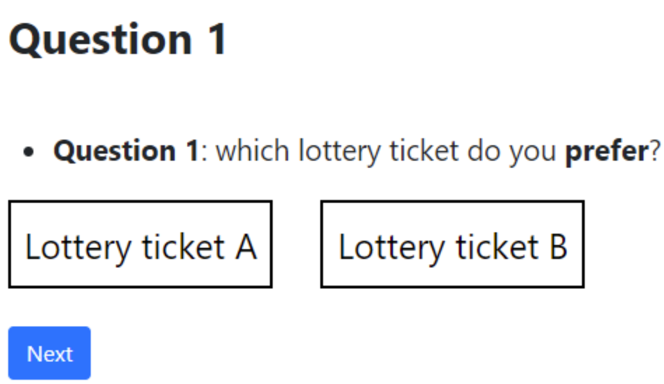
Lottery Ticket A	**Lottery Ticket B**
0% chance of \$1	40% chance of \$1
100% chance of \$7	0% chance of \$7
0% chance of \$20	60% chance of \$20

- Lottery ticket **A** involves no chance at all: it pays \$7 for sure.
- Lottery ticket **B** pays \$20 with probability 60%, or \$1 with probability 40%.

During the experiment, you will encounter 74 pairs of lottery tickets. For each pair, you will answer to two questions.

Notice that after you select an answer to a question and click on Next, you **will not be able to modify it**.

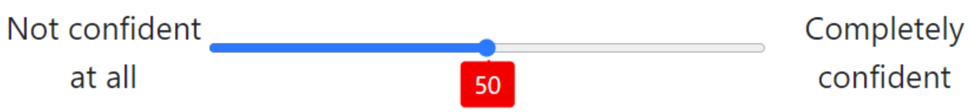
We will now show you these two questions.



Question 2

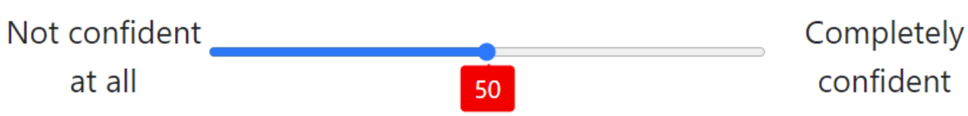
Whenever you select Lottery ticket **A** in Question 1, the **second question** will be:

- **Question 2**: you chose lottery ticket **A**. On a scale from 0 to 100, how **confident** do you feel about this choice? **The higher the number, the more confident you are about this choice.**



Whenever you select Lottery ticket **B** in Question 1, the **second question** will be:

- **Question 2**: you chose lottery ticket **B**. On a scale from 0 to 100, how **confident** do you feel about this choice? **The higher the number, the more confident you are about this choice.**



Next

Training session. To familiarize yourself with the setup of the experiment, you will complete a brief training session. The training session consists of **five** pairs of lottery tickets. For each pair of lottery tickets, you will answer the two questions that we described to you. The answers that you give in this training session **do not affect** your monetary compensation. We will describe the details of your monetary compensation at the end of your training session.

Possible rewards. Now that you have familiarized yourself with the questions of the study, you will learn about the details of your compensation. You will receive a **participation fee** of **$4.75** for completing all the questions in the experiment. Moreover, you may also receive a **bonus payment**:

- At the end of the experiment, the computer will randomly select a number between 1 and 10. Each number has an equal probability (**10%**) of being selected.

- If the randomly selected number is 1, you will receive a bonus payment.
- If you are chosen to receive a bonus payment, the computer will randomly pick one **of the 74 pairs of lotteries** that you encountered in the experiment.
- Then, you will be able to **play the lottery ticket from the selected pair that you declared to prefer in Question 1** (which lottery ticket do you prefer?). That is, the computer will use the probabilities specified in the lottery ticket to select a monetary prize ($1, $7 or $20).
- Your bonus payment will be **the monetary prize that the computer will select**. You will receive your bonus payment together with the participation fee after we review your submission.

Example: Suppose that you are randomly selected for a bonus payment and the randomly picked pair of lottery tickets is:

Lottery Ticket A	**Lottery Ticket B**
0% chance of $1	40% chance of $1
100% chance of $7	0% chance of $7
0% chance of $20	60% chance of $20

- If you chose Lottery ticket **A** in Question 1, you get additional $7 for sure.
- If you chose Lottery ticket **B** in Question 1, you have a 60% chance of getting additional $20 and a 40% chance of getting additional $1.

Begin the experiment. Congratulations, you are now ready to participate in the experiment. If anything is unclear, please let us know through the Prolific anonymized internal messaging service. Otherwise, please click next to begin the experiment.