A Non-Parametric Test of Risk Aversion

Jacob K. Goeree and Bernardo García-Pola[1]

April 27, 2026

Abstract

In economics, risk aversion is modeled via a concave Bernoulli utility within the expected-utility paradigm. We propose a non-parametric test of expected utility and concavity. Data from a large-scale experiment show little support for either. Only 3% of the participants are consistent with expected utility and only 39% of the choices are consistent with a concave utility. Our findings contrast with a large body of work that employs the Holt–Laury task. We demonstrate theoretically and empirically that this task cannot identify risk aversion. The preponderance of "risk averse" choices it produces merely reflects parametric assumptions. Using data from both the Holt–Laury task and our task, we reject these assumptions. A minority of the participants have utilities that are convex (2%) or concave (32%). A majority are risk averse for small prizes and risk loving for large prizes (55%) or vice versa (11%). Preference heterogeneity augmented with decision error provides the best "in sample" and "out of sample" fit of the combined data as well as correct comparative statics predictions.

Keywords: *Risk elicitation, mean-preserving spreads, non-parametric test, multiple-price list, non-identifiability, decision error, preference heterogeneity*

[1]Goeree: AGORA Center for Market Design, UNSW, Sydney, Australia. García-Pola: Department of Economics, Universidad Pública de Navarra and INARBE, Pamplona, Spain. Goeree gratefully acknowledges funding from the Australian Research Council (DP220102893). Bernardo García-Pola acknowledges financial support from the Spanish Ministry of Science and Innovation (PID2022-138774NB-I00, PID2021-127119NB-I00). We thank the Editor, an anonymous referee, Felix Holzmeister, Peter Wakker, and Brett Williams for helpful comments, Filip Fidanoski and Andreas Ortmann for an overview of the many risk-elicitation methods, and Charles Holt for sharing instructions for the Holt–Laury task.

1. Introduction

Risk attitudes play an important role in economic and financial decisionmaking. Risk aversion gives rise to demand for insurance, impacts consumption and saving levels, shapes moral hazard and the design of contracts, and affects bidding in pay-your-bid auctions. In finance, risk attitudes are key when evaluating the trade-off between an asset's risk and return and when composing investment portfolios. Because of the ubiquitous role of risk in decisionmaking, the proper measurement of risk preferences is important for economic analysis, policy, and financial advice.

Financial institutions typically use surveys and self reports to assess clients' risk preferences.[1] The data so obtained are qualitative in nature and do not lend themselves to extrapolation beyond the survey's hypothesized scenarios. Incentivized experiments offer a viable alternative for the quantitative measurement of risk preferences. Holt and Laury (2002) proposed a simple task based on the multiple-price list methodology that has a long tradition in development economics and psychology.[2] Their seminal contribution led to a surge of interest in measuring risk attitudes in the laboratory. In the past two decades, a wide variety of risk elicitation methods have been proposed by experimental economists.[3]

However, results from these elicitation methods can significantly differ from those of surveys, see e.g. Anderson and Mellor (2009). While Falk et al. (2023) report a robust relationship between survey and experimental results, Chapman et al. (2025) caution that the observed correlations may reflect confounding factors rather than a common underlying preference. Regardless, results vary considerably across the different elicitation methods and sometimes contradict each other. There is no consensus why the results vary across elicitation methods nor which elicitation method is most accurate.[4] This lack of consistency, dubbed the *risk elicitation puzzle* by Pedroni et al. (2017), undermines the external validity

[1] Many countries require financial institutions to gauge investors' risk attitudes. For instance, the European Securities and Markets Authority dictates that "firms should specify the general attitude that target clients should have in relation to the risks of investment" (Guidelines on MiFID II, 2018).

[2] See, e.g., Binswanger (1980) and Tversky and Kahneman (1982, p.305-306). The brief description of the multiple-price list methodology in the latter underlines it was considered standard at the time. We thank Peter Wakker (private communication) for pointing out that multiple-price lists have been used in decision theory as long as the field exists.

[3] See, e.g., Holt and Laury (2002, 2005); Eckel and Grossman (2002, 2008); Choi et al. (2007); Crosetto and Filippin (2013); Dohmen et al. (2010); Abdellaoui et al. (2011). This list is far from complete. Fidanoski and Ortmann (private communication) report there are over a hundred different risk-elicitation methods.

[4] See, e.g. Charness et al. (2013); Crosetto and Filippin (2015); Pedroni et al. (2017); Holzmeister and Stefan (2021). Harrison and Rutström (2008) provide a comprehensive summary of the various elicitation methods, see also Belzil and Jagelka (2020) for an update that takes into account rational inattention. Friedman et al. (2022) show that task attributes, e.g. text versus visualization and continuous versus discrete choices, explain some of the (lack of) correlation across various tasks.

of laboratory risk-elicitation methods (Smith, 1989; Friedman et al., 2014).

Despite the large number of risk-elicitation methods that have been proposed, none of them *directly tests* the two key assumptions that underlie the economic model of risk aversion. Expected utility is linear in a lottery's probabilities and the lottery's prizes are evaluated using a concave Bernoulli utility. We introduce a simple task that entails comparing a baseline lottery to two lotteries obtained via mean-preserving spreads. This task allows us to test for concavity of the utility without making parametric assumptions. Moreover, by varying the probabilities of the baseline lottery it allows us to test the linearity assumption.

We report the results of a large-scale experiment in which 370 participants compared fourteen baseline lotteries to mean-preserving spreads of the baseline lotteries. The experiment, main hypotheses, and empirical strategy were preregistered. We imposed several measures to mitigate noisy decision making in our task and to make it comparable to the Holt–Laury task, e.g. by using the same prize values and the same range of probabilities. These measures allowed us to define a "low noise" sub-sample consisting of those participants who scored well on all of them. For both the low-noise sub-sample and the full sample, our results refute both the concavity and linearity assumptions.

Our results contradict a large body of work that finds a preponderance of "risk averse" choices in the Holt and Laury task, arguably the "gold standard" for risk elicitation. We prove that the Holt–Laury task cannot identify risk aversion. The high prevalence of "risk averse" choices it produces merely reflects the assumption of constant relative risk aversion. We demonstrate this assumption is untenable. First, when applied to data from our task, the predictions of the homogeneous constant relative risk aversion model are rejected. Second, using data from both the Holt–Laury task and our task, we non-parametrically estimate individual Bernoulli utilities and reject the heterogeneous constant relative risk aversion model. We find that a majority of the participants have Bernoulli utilities that are neither convex (2%) nor concave (32%). They are risk averse for small prizes and risk loving for large prizes (55%) or vice versa (11%). We demonstrate that preference heterogeneity augmented with decision error provides the best "in sample" and "out of sample" fit of the combined data as well as correct comparative statics predictions.

The next section proposes a simple non-parametric test of expected utility and risk aversion and revisits the Holt–Laury task. Section 3 discusses the experimental design and protocol, the hypotheses, and details the measures we took to mitigate noise in decision making. Section 4 reports the results from a large-scale experiment and Section 5 discusses survey results. Section 6 concludes. The Appendices contain proofs, instructions for the experiment, and additional empirical support.

2. Non-Parametric Versus Parametric Risk Measures

2.1. Mean-Preserving Spreads

A classic result due to Rothschild and Stiglitz (1970) is that *any* risk averse individual prefers a lottery to a mean-preserving spread of itself.[5] Surprisingly, this result has hitherto not been systematically exploited in the large literature on risk elicitation.[6]

Consider lotteries over four prizes $\pi = (€1, €16, €21, €38.5)$ with utilities $u(1)$, $u(16)$, $u(21)$, and $u(38.5)$.[7] We can subtract a constant from these utilities or multiply them by a positive constant and the result would describe the same preferences. So let us normalize the utilities as 0, u_1, u_2, and 1 where $0 \leq u_1 \leq u_2 \leq 1$.[8] This utility-possibility set corresponds to the triangular area above the 45-degree line in Figure 1. The Bernoulli utility is concave if and only if its slope is non-increasing, i.e.

$$\frac{u_1}{15} \geq \frac{u_2 - u_1}{5} \geq \frac{1 - u_2}{17.5} \tag{1}$$

where the numerators are utility differences and the denominators are prize differences.[9] The black lines correspond to utility pairs (u_1, u_2) for which these inequalities hold with equality. The red area shows utility pairs such that both inequalities in (1) hold and the blue area shows utility pairs for which both inequalities are reversed. The yellow (green) area shows utility pairs for which only the first (second) inequality of (1) is reversed.

Suppose lottery $\mathscr{L}_A$ offers the prizes π with probabilities $p = (p_1, p_2, p_3, p_4)$. A mean-preserving spread of $\mathscr{L}_A$ results when moving all mass of the second prize to the first and third prizes. This yields lottery $\mathscr{L}_B$ with probabilities $p = (p_1 + \frac{1}{4}p_2, 0, p_3 + \frac{3}{4}p_2, p_4)$ over the same prizes π. The reason only one-quarter of the mass is moved downward and three-quarters are moved upward is to keep the mean the same. A second mean-preserving spread of $\mathscr{L}_A$ results from moving all mass of the third prize to the second and fourth prizes, which

[5] This result is derived under the assumption of equal means, which is a special case of the second-order stochastic dominance criterion developed by Hadar and Russell (1969) and Hanoch and Levy (1969) around the same time.

[6] See, however, Levy and Levy (2001) who use an elicitation procedure similar to ours, although they do not vary the probabilities of the lotteries. We would like to thank Felix Holzmeister for pointing this paper out to us, which has helped us improve our design.

[7] These prizes were chosen to match those of the Holt–Laury task, see Section 2.2.

[8] This follows by subtracting $u(1)$ from all utilities and dividing the resulting utilities by $u(38.5) - u(1)$ so that $u_1 = (u(16) - u(1))/(u(38.5) - u(1))$ and $u_2 = (u(21) - u(1))/(u(38.5) - u(1))$. Risk neutrality corresponds to $u(x) = x$ in which case $u_1 = \frac{2}{5}$ and $u_2 = \frac{8}{15}$.

[9] In detail, the slope of the utility between the first two prizes, 0 and 16, is $(u_1 - 0)/(16 - 1)$, the slope between the middle prizes, 16 and 21, is $(u_2 - u_1)/(21 - 16)$, and the slope between the last two prizes, 21 and 38.5, is $(1 - u_2)/(38.5 - 21)$. The requirement that the slope is non-increasing yields (1).

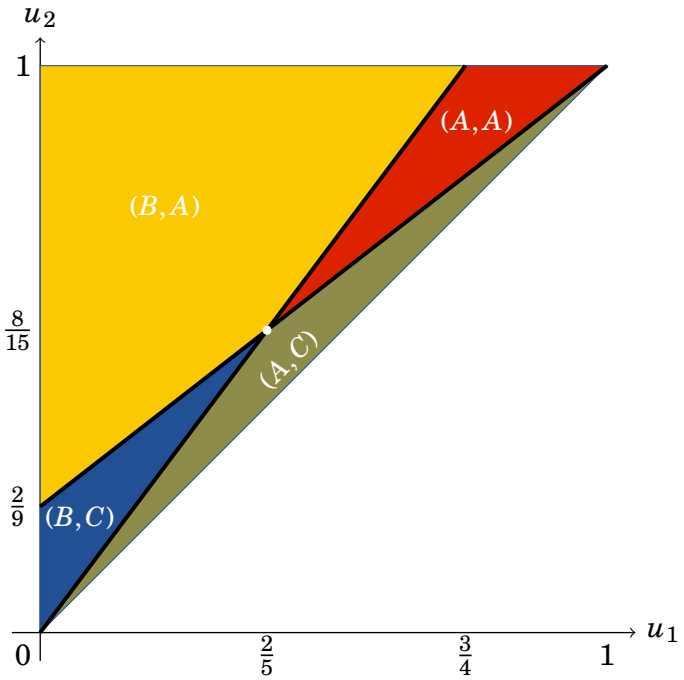


Figure 1: The colored areas represent possible utility pairs (u_1,u_2) for each of the four outcomes when an individual chooses between lotteries $(\mathscr{L}_A,\mathscr{L}_B)$ and between lotteries $(\mathscr{L}_A,\mathscr{L}_C)$. The red area corresponds to concave Bernoulli utilities. Risk neutrality implies $(u_1,u_2)=(\frac{2}{5},\frac{8}{15})$, which is indicated by the white dot at the intersection of all four regions.

yields lottery $\mathscr{L}_C$ with probabilities $p=(p_1,p_2+\frac{7}{9}p_3,0,p_4+\frac{2}{9}p_3)$.[10]

Consider choosing between lotteries $(\mathscr{L}_A,\mathscr{L}_B)$ and between lotteries $(\mathscr{L}_A,\mathscr{L}_C)$. If $\mathscr{L}_A$ is preferred over $\mathscr{L}_B$ then $u_1\geq\frac{3}{4}u_2$, which is equivalent to the left inequality in (1). If $\mathscr{L}_A$ is preferred over $\mathscr{L}_C$ then $u_2\geq\frac{7}{9}u_1+\frac{2}{9}$, which is equivalent to the right inequality. So, if $\mathscr{L}_A$ is chosen twice, the utility pair (u_1,u_2) belongs to the red area and the Bernoulli utility is concave. If $\mathscr{L}_A$ is not chosen at all then the utility pair belongs to the blue area and the Bernoulli utility is convex. If $\mathscr{L}_B$ and $\mathscr{L}_A$ are chosen then the utility pair belongs to the yellow area, which contains "S shaped" Bernoulli utilities that indicate risk lovingness for small prizes and risk aversion for large prizes. Finally, if $\mathscr{L}_A$ and $\mathscr{L}_C$ are chosen then the utility pair belongs to the green area, which contains "inverted S shaped" Bernoulli utilities that reflect risk aversion for small prizes and risk lovingness for large prizes.

The analysis of the previous paragraph applies regardless of the probabilities that define lottery $\mathscr{L}_A$. This is a consequence of the linearity assumption that underlies expected utility. We can test for both linearity and concavity by letting individuals choose between $(\mathscr{L}_A,\mathscr{L}_B)$ and between $(\mathscr{L}_A,\mathscr{L}_C)$ for various values of $p=(p_1,p_2,p_3,p_4)$.

[10]Note that we cannot move mass of the extreme prizes €1 and €38.5 to any of the other prizes while keeping the mean the same.

The next theorem generalizes to lotteries with an arbitrary number of prize values. If there are $K > 2$ prizes then there are $2(K-2)$ *types* of Bernoulli utilities, depending on whether the slope of the utility rises or falls at any of the $K-2$ non-extreme prizes.

Definition 1 *A task can identify risk aversion if it determines the type of Bernoulli utility.*

Theorem 1 (Non-parametric) *Consider any lottery $\mathscr{L}$ over $K > 2$ ascending prizes $\pi = (\pi_1, \ldots, \pi_K)$ that occur with strictly positive probabilities $(p_1, \ldots, p_K)$. Let $\mathscr{L}_k$ for $1 < k < K$ denote the lotteries in which all probability mass of prize k is moved to the prizes $k-1$ and $k+1$ in a mean-preserving manner. Comparing $\mathscr{L}$ to each of the $\mathscr{L}_k$ for $1 < k < K$ determines the type of Bernoulli utility. In particular, an individual is risk averse iff $\mathscr{L}$ is preferred to all $\mathscr{L}_k$ for $1 < k < K$. An individual is risk loving iff all $\mathscr{L}_k$ for $1 < k < K$ are preferred to $\mathscr{L}$.*

The $K-2$ comparisons of Theorem 1 yield necessary and sufficient conditions for concavity of the Bernoulli utility (see Appendix A for the proof).[11] Varying the probabilities that define the lottery $\mathscr{L}$ in Theorem 1 yields a simple test of expected utility.

2.2. Multiple-Price Lists

The Holt–Laury task employs a multiple-price list. Participants choose between a "safe" lottery with prizes of €16 and €21 and a "risky" lottery with prizes of €1 and €38.5. They face a total of ten "safe" versus "risky" lottery comparisons that result by varying the probability, p, of the best outcome (€21 for the safe lottery and €38.5 for the risky lottery) from $p = 0.1$ to $p = 1.0$, see Appendix B.

Multiple-price lists are popular as they are easy to use and interpret. Participants typically choose the "safe" lottery when p is below some threshold and then switch to the "risky" lottery. The threshold is then used to parametrically identify risk aversion by imposing constant relative risk aversion (CRRA), i.e.

$$u_i(x) = \frac{x^{1-r_i}}{1-r_i}$$

where $r_i \in \mathbb{R}$ is the coefficient of constant relative risk aversion.[12]

[11] The few existing non-parametric tests of risk aversion typically exploit an implication of concavity, rather than testing for concavity directly. Baillon and L'Haridon (2021) propose a non-parametric version of the well-known Arrow–Pratt measure. Johnson et al. (2021) measure the utility function using the adaptive trade-off method proposed by Wakker and Deneffe (1996). L'Haridon and Vieider (2019) elicit risk preferences and prospect theory parameters using certainty equivalents. Our task, which entails two choices between two lotteries, is simpler and avoids the certainty effect.

[12] Another common choice is constant absolute risk aversion (CARA), i.e. $u_i(x) = (1 - \exp(-\alpha_i x))/\alpha_i$ where $\alpha_i \in \mathbb{R}$ is the coefficient of constant absolute risk aversion.

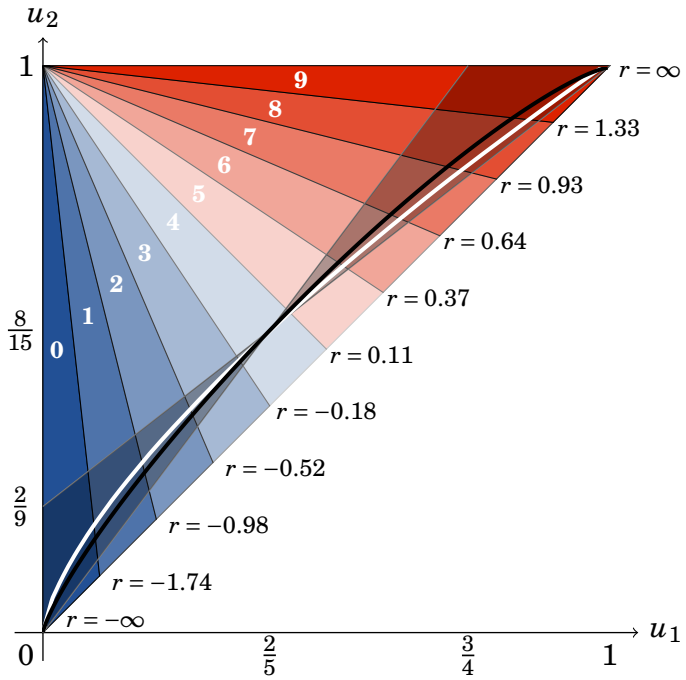


Figure 2: The colored triangles show the possible utilities based on the number of "safe" choices ranging from zero (dark blue triangle) to nine (dark red triangle), as indicated by the white numbers. The upper shaded area corresponds to concave utilities and the lower shaded area to convex utilities, cf. the red and blue areas in Figure 1. The white curve corresponds to CRRA utilities $u(x) = x^{1-r}/(1-r)$ for $r \in \mathbb{R}$ and the black curve to CARA utilities $u(x) = (1-e^{-\alpha x})/\alpha$ for $\alpha \in \mathbb{R}$. The lower corners of each triangle indicate the range of r values for which that triangle applies (assuming CRRA).

To evaluate the multiple-price methodology non-parametrically, let us use the same normalization of the utilities for the four prizes as in Section 2.1. If an individual makes s safe choices and then switches, the utilities satisfy

$$1 - \frac{10-s}{s}\,u_1 \le u_2 \le 1 - \frac{9-s}{s+1}\,u_1 \tag{2}$$

where the first inequality follows since the safe lottery was chosen when $p = \frac{s}{10}$ and the second inequality follows since the risky lottery was chosen when $p = \frac{s+1}{10}$. The inequalities in (2) define the colored triangles in Figure 2. In this figure, the dark blue triangle on the left corresponds to zero safe choices and the dark red triangle at the top corresponds to nine safe choices. The upper shaded area corresponds to concave utilities and the lower shaded area to convex utilities, cf. the red and blue areas in Figure 1.

The white (black) curve that runs through these shaded areas corresponds to the possible CRRA (CARA) utilities. For instance, for participants that made six safe choices the inferred parameter is in the range $r \in [0.37, 0.64]$. The lower number of this range corresponds to the value of r for which the white curve enters the triangle from the left and the higher number

corresponds to the value of r for which the white curve exits the triangle on the right. Figure 1 shows the range of r values for any number of safe choices.

Theorem 2 (Parametric) *Assuming CRRA or CARA utilities, individuals are risk averse if they make more than four "safe" choices in the Holt–Laury task and risk loving if they make fewer than four "safe" choices.*

If CRRA or CARA applies then the Holt–Laury task somewhat underreports risk aversion, e.g. individuals with $0 < r_i < 0.11$ make four safe choices and aren't labeled risk averse. In contrast, in our task they choose lottery $\mathscr{L}_A$ twice and are counted as risk averse.

2.3. Decision Error and (Non) Identification of Risk Aversion

Theorems 1 and 2 assume perfect expected-utility maximization, i.e. there are no mistakes. This yields stark predictions that are typically refuted by the data, see e.g. Goeree et al. (2016). For instance, risk averse individuals who face several choices between a lottery and a mean-preserving spread of it should *always* choose the former. The data may show they almost always did but that they deviated a few times. To accommodate such deviations, we consider a model of decision error in which mistakes can happen but more costly mistakes are less likely. This can be operationalized using an additive random utility model (ARUM) in which the expected utilities are augmented with utility perturbations.

To illustrate, consider the choice between lotteries $\mathscr{L}_A$ and $\mathscr{L}_B$ of Section 2.1. The lotteries' expected utilities are perturbed by additive disturbances $\widetilde{U}_A = U_A + \mu\varepsilon_A$ and $\widetilde{U}_B = U_B + \mu\varepsilon_B$, where ε_A and ε_B are iid random variables with a strictly positive density and μ measures their impact.[13] If individuals maximize perturbed utilities, lottery $\mathscr{L}_A$ is chosen with probability

$$P_A = F\Big(\frac{U_A - U_B}{\mu}\Big)$$

where F denotes the distribution of $\varepsilon_B - \varepsilon_A$. For example, if the utility disturbances are extreme-value distributed then F is the logistic distribution. In any ARUM, behavior becomes random for large μ while perfect maximization results as μ tends to zero.

Recall that $U_A - U_B = u_1 - \frac{3}{4}u_2$ where the normalized utilities of the four prizes are 0, u_1, u_2, and 1 as before. Let $(u_1', u_2') = \lambda(u_1, u_2) + (1-\lambda)(\frac{2}{5}, \frac{8}{15})$ and $\mu' = \lambda\mu$ for λ between 0 and 1 then

$$\frac{u_1' - \frac{3}{4}u_2'}{\mu'} = \frac{u_1 - \frac{3}{4}u_2}{\mu}$$

[13] Utility disturbances are independent across individuals.

In other words, the probability of choosing lottery $\mathscr{L}_A$ is constant on line segments that emanate from the risk-neutral point $(\frac{2}{5}, \frac{8}{15})$ in the triangular region of Figure 1. Importantly, the risk-neutral point itself is not included since $U_A = U_B$ implies $P_A = \frac{1}{2}$.

This invariance result is the statistical consequence of the test's non-parametric nature and means it cannot identify the exact Bernoulli utility. For instance, it may determine that the utility belongs to the red area of Figure 1, but it cannot pinpoint its exact location in the area. In line with Definition 1, the test identifies the *type* of Bernoulli utility and, hence, *can* identify risk aversion. That P_A is constant on line segments is irrelevant since the line segments are contained within one of the colored regions of Figure 1.

We next show that the Holt–Laury task *cannot* identify risk aversion. The expected utility of the "safe" lottery is $U_S = pu_2 + (1-p)u_1$ and that of the "risky" lottery is $U_R = p$, with p the chance of the best outcome. Hence, the probability of a "safe" choice is

$$P_S = F\Big(\frac{p(u_2 - 1) + (1-p)u_1}{\mu}\Big)$$

Let $(u_1', u_2') = \lambda(u_1, u_2) + (1-\lambda)(0,1)$ and $\mu' = \lambda\mu$ for $0 < \lambda \leq \frac{1}{1-u_2+u_1}$ then

$$\frac{p(u_2' - 1) + (1-p)u_1'}{\mu'} = \frac{p(u_2 - 1) + (1-p)u_1}{\mu}$$

which implies that P_S is constant on line segments emanating from the top-left corner of the triangle in Figure 1 and ending on the diagonal boundary of the triangle.

Theorem 3 (Non-Identification) *The Holt–Laury task cannot identify risk aversion.*

Theorem 3 is the statistical consequence of the Holt–Laury task determining triangles that overlap the yellow, green, and red (or blue) areas in Figure 1. It implies that for any distribution of "safe" choices there is a continuum of "S shaped" Bernoulli utilities (i.e. steepest for intermediate prizes) and a continuum of "inverted S shaped" Bernoulli utilities (i.e. least steep for intermediate prizes) that explain the data as well as any of the parametric risk models, such as CRRA or CARA.

Theorem 3 holds irrespective of whether switching back-and-forth between the "safe" and "risky" lotteries is allowed, see the proof in Appendix A. Section 4.2 illustrates non-identification in the Holt–Laury task using experimental data.

3. Experiment

3.1. Experimental Design and Protocol

We conducted eight sessions and recruited a total of 370 participants via ORSEE (Greiner, 2015). The sessions were conducted from January 20 to January 27, 2025, at the Laboratory of Experimental Economics (LEE) of Jaume I University of Castellón. The experimental tasks were administered using z-Tree (Fischbacher, 2007). Participants were randomly selected from the LEE pool without applying any filter. The design and analysis of the experiment was pre-registered at the AEA RCT registry.[14]

Sessions consisted of two parts in which participants executed either the Holt–Laury task or our task. In half the sessions the Holt–Laury task preceded our task (Order A) and in the other half this order was reversed (Order B). Participants did not learn any details about the second part until they had finished the first part. Before each part, instructions were shown using a slide presentation (see Appendix B) that was read aloud. Participants had the opportunity to ask questions at any time during the experiment.

The Holt–Laury task employs the multiple-price list of Appendix B. One difference with the Holt and Laury (2002) study is that in our experiment, participants could switch between safe and risky lotteries only once.

In our task, participants went through fourteen screens, each displaying one of the lottery triples $(\mathscr{L}_A, \mathscr{L}_B, \mathscr{L}_C)$ shown in Table 1.[15] Participants chose one lottery from the pair $(\mathscr{L}_A, \mathscr{L}_B)$ and one lottery from the pair $(\mathscr{L}_A, \mathscr{L}_C)$, resulting in a total of 28 decisions. The $\mathscr{L}_B$ and $\mathscr{L}_C$ lotteries in C1-C6 are constructed from lottery $\mathscr{L}_A$ as per Theorem 1, i.e. probability mass is moved from the two middle prizes to neighbouring prizes.

The $\mathscr{L}_B$ and $\mathscr{L}_C$ lotteries in C7-C12 are spreads of $\mathscr{L}_A$ that involve moving mass to non-neighbouring prizes. As such, they do not fit Theorem 1 and we do not use them to test the linearity assumption underlying expected utility. One reason we included C7-C12 is to check that our results are robust to the number of possible prizes. C7 is special in that the baseline lottery $\mathscr{L}_A$ assigns near-uniform probabilities to all four prize values.[16] In C8-C12, all three lotteries have only two possible prize values. Another special case is FOSD, in which $\mathscr{L}_B$ and $\mathscr{L}_C$ are *not* mean-preserving spreads of $\mathscr{L}_A$. Despite these differences, a risk averse individual prefers $\mathscr{L}_A$ over $\mathscr{L}_B$ and $\mathscr{L}_C$ in all cases of Table 1.

[14] Reference AEARCTR-0015143.

[15] The lotteries of C3 were shown twice, as explained below.

[16] Interestingly, C7 led to the second-highest preference for $\mathscr{L}_A$ over the spreads $\mathscr{L}_B$ and $\mathscr{L}_C$ even though the latter have fewer possible outcomes.

Case	$\mathscr{L}_A$	$\mathscr{L}_B$	$\mathscr{L}_C$
C1	(21, 16, 63, 0)	(25, 0, 75, 0)	(21, 65, 0, 14)
C2	(15, 40, 45, 0)	(25, 0, 75, 0)	(15, 75, 0, 10)
C3	(43, 12, 45, 0)	(46, 0, 54, 0)	(43, 47, 0, 10)
C4	(0, 40, 45, 15)	(10, 0, 75, 15)	(0, 75, 0, 25)
C5	(0, 52, 36, 12)	(13, 0, 75, 12)	(0, 80, 0, 20)
C6	(0, 40, 18, 42)	(10, 0, 48, 42)	(0, 54, 0, 46)
C7	(26, 24, 27, 23)	(32, 35, 0, 33)	(53, 0, 0, 47)
C8	(60, 40, 0, 0)	(70, 0, 30, 0)	(84, 0, 0, 16)
C9	(0, 40, 60, 0)	(10, 0, 90, 0)	(52, 0, 0, 48)
C10	(0, 55, 45, 0)	(0, 90, 0, 10)	(54, 0, 0, 46)
C11	(0, 0, 45, 55)	(0, 35, 0, 65)	(21, 0, 0, 79)
C12	(0, 0, 90, 10)	(0, 70, 0, 30)	(42, 0, 0, 58)
FOSD	(40, 30, 30, 0)	(50, 40, 10, 0)	(60, 30, 10, 0)

Table 1: Probabilities (in percentages) of the lotteries used in our task. In each of the thirteen cases, the prize values were $\pi = (€1, €16, €21, €38.5)$. The probabilities of lotteries $\mathscr{L}_B$ and $\mathscr{L}_C$ follow from those of lottery $\mathscr{L}_A$ by moving all mass of the second and/or third prize. For C1-C6 this involves mean-preserving spreads to neighboring prizes, see Theorem 1. For C8-C12 this may involve mean-preserving spreads to non-neighboring prizes to create three lotteries that have only two possible prize values. Special cases are C7, in which the baseline lottery $\mathscr{L}_A$ has four possible prize values, and FOSD, in which $\mathscr{L}_B$ and $\mathscr{L}_C$ are *not* mean-preserving spreads of $\mathscr{L}_A$.

Another reason we included C7-C12 is that they divide the upper-triangular region $0 \leq u_1 \leq u_2 \leq 1$ differently than in Figure 1. Using the expressions of lotteries as vectors of four probabilities, as in Table 1, the red region follows from the inequalities

$$\begin{aligned} (\mathscr{L}_A - \mathscr{L}_B) \cdot (0, u_1, u_2, 1) &\geq 0 \\ (\mathscr{L}_A - \mathscr{L}_C) \cdot (0, u_1, u_2, 1) &\geq 0 \end{aligned}$$

where $v \cdot w = \sum_i v_i w_i$ denotes the usual inner product. The yellow region follows by reversing only the top inequality and the green region follows by reversing only the bottom inequality. Finally, the blue region follows by reversing both inequalities.

The results of this exercise are shown in Figure 3. The top panels correspond to C7 (left) and C8 (right), the middle panels to C9 (left) and C10 (right), and the bottom panels to C11 (left) and C12 (right). The red region is (much) larger than the one in Figure 1 in all cases. Also, in four cases, the green and yellow regions have swapped position compared to Figure 1. These differences make C7-C12 ideal for out-of-sample testing of models estimated using data from C1-C6.

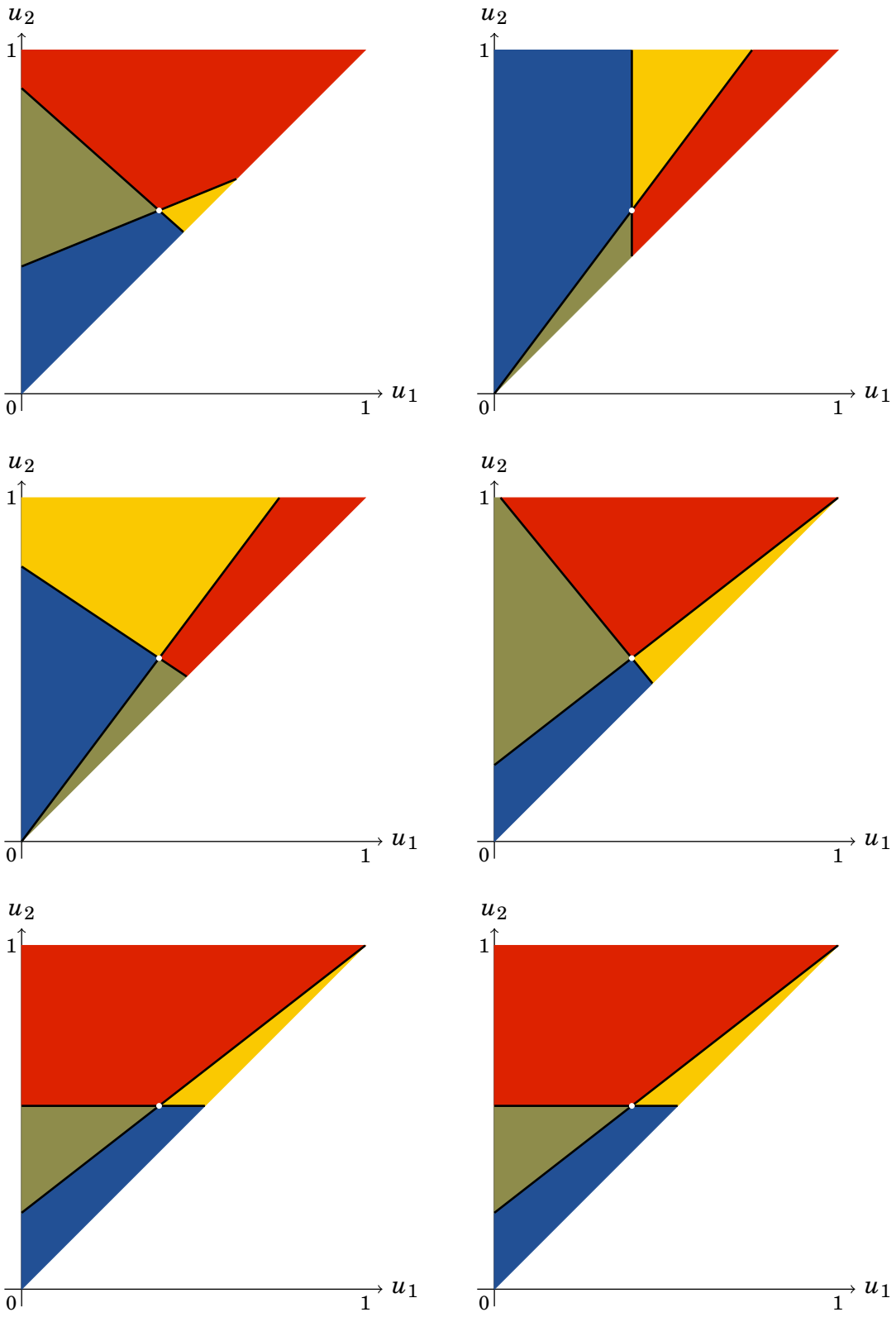


Figure 3: Division of the upper triangle $0 \leq u_1 \leq u_2 \leq 1$ for the cases C7 (upper-left) to C12 (bottom-right). The red area corresponds to (A,A), the yellow area to (B,A), the green area to (A,C) and the blue area to (B,C).

The order of participants' screens was randomized (except for two noise control screens, see Section 3.2), and this order was the same for all participants. Within each screen we randomized three features independently for each participant: (i) the position of the lotteries (and, hence, which one was labeled A, B, or C), (ii) the order in which the two choice tasks were presented, i.e. $(\mathscr{L}_A, \mathscr{L}_B)$ and $(\mathscr{L}_A, \mathscr{L}_C)$, and (iii) the left–right position of the buttons corresponding to the two lotteries within each choice. Each lottery was shown with a textual and a visual representation to facilitate comprehension.

At the end of the experiment, we paid participants in private and in cash. Each participant was paid for two randomly-selected choices, one from each part. This ensured that both parts were incentivized and reduced variance in participants' earnings. On average, participants earned roughly €11.64 for the first part and €9.06 for the second part, plus a €4 show-up fee, for a total of €24.70.

3.2. Noise-Reducing Measures

We took several measures to reduce noise in decisionmaking in our task and make it comparable to the Holt–Laury task.

Simple non-extreme lotteries:

(i) Same prize values. Our task uses the same four prizes as the Holt–Laury task, i.e. €1, €16, €21, and €38.5.

(ii) Few possible outcomes. We restricted the lotteries to have two or three possible outcomes (with the exception of C7). This was done to keep the lotteries simple while avoiding the certainty-effect bias that arises when a lottery has a single outcome.

(iii) Non-extreme probabilities. If a prize value occurs with positive probability then this probability is at least 10%. First, this avoids behavioral distortions that often arise for small (but positive) probabilities. Second, it mirrors the lowest probability of the best outcome in the Holt–Laury task. Hence, not only the prize values, but also the range of probabilities is similar in our task.

(iv) Integer percentages. Probabilities appear as integer percentages so that participants can process each lottery at a glance.

Explicit indifference option: indifference can add noise to choices, especially when participants are forced to make a choice. This is true for our task as well as for the Holt–Laury task. To mitigate the effects of indifferences in our task, we added a button labeled "I don't mind" (see Appendix B). Participants were informed that if they pressed that button, the computer would choose an option for them. This may not completely eliminate the effects of

indifferences, as participants may be indifferent between pressing the button and choosing any lottery, but it should reduce the frequency of randomly clicking on one of the lotteries.

FOSD screen: even with an explicit indifference option, participants may occasionally click without thinking. To assess this possibility, we inserted a case that had a strict first-order stochastically dominant option, see "FOSD" in the bottom row of Table 1. Unlike the other cases, $\mathscr{L}_B$ and $\mathscr{L}_C$ are *not* a mean-preserving spreads of $\mathscr{L}_A$. The display resembles that of spread lotteries, but $\mathscr{L}_A$ first-order stochastically dominates the other two lotteries. We placed this check in the middle of the sequence (position 7) rather than at the start or end.

Repeated-choice screen: we included a within-participant coherence check by repeating one case. Specifically, we showed C3 twice with five other cases in between. The idea is that participants with strict preferences should respond the same when facing the same case.

Low-Noise sub-sample: the above measures allow us to define the *low-noise sub-sample* consisting of participants who (i) never selected the indifference button (88%), (ii) chose the first-order stochastically dominant option (90%), and (iii) made the same choice in the repeated lottery (56%). Applying these three criteria, 123 participants remain (3,444 choices).

3.3. Hypotheses

The lotteries in C1-C6 of Table 1 can be used to test the linearity assumption underlying expected utility. Since $\mathscr{L}_B$ and $\mathscr{L}_C$ are obtained from $\mathscr{L}_A$ via mean-preserving spreads to neighboring prizes, Theorem 1 applies and a necessary condition for expected utility to hold is that participants make the same choices in these cases. In addition, participants should choose lottery $\mathscr{L}_A$ in the FOSD case.

Hypothesis 1 (Expected Utility) *Participants' choices between $\mathscr{L}_A$ and $\mathscr{L}_B$, and between $\mathscr{L}_A$ and $\mathscr{L}_C$, are the same for C1-C6 of Table 1. Participants choose $\mathscr{L}_A$ in the FOSD case.*

The next hypothesis reflects the finding of numerous risk-elicitation experiments that the bulk of participants' choices are risk averse. For our task this means that there should be a preponderance of (A,A) choices, as this is the only choice compatible with risk aversion.

Hypothesis 2 (Risk Aversion) *The vast majority of participants' choices are* (A,A).

These are the two main hypotheses that were preregistered. Both of them are required for the economic model of risk aversion to hold.

In addition, we consider two hypotheses that concern the robustness of our task and how it compares to the Holt–Laury task. As noted in Section 2.2, if the parametric CRRA

assumption holds then the Holt–Laury task somewhat underreports risk aversion as individuals with $0 < r_i < 0.11$ make four "safe" choices and are labeled risk neutral. But in our task they choose (A,A) and are labeled risk averse.

Hypothesis 3 (CRRA) *Under the CRRA assumption, the fraction of* (A,A) *choices in our task is no less than the fraction of four or more "safe" choices in the Holt–Laury task.*

Finally, the creation of a low-noise sub-sample allows us to test whether our task is robust.

Hypothesis 4 (Robustness) *The choice frequancies in our task are the same whether our task comes before or after the Holt–Laury task. Moreover, the choice frequencies of the low-noise sub-sample are the same as those of the full sample.*

The second part of this hypothesis stipulates that choices in our task are not biased by noisy decisionmaking. The choices of more-noisy participants may show greater variability but the average choice frequencies are the same.

4. Results

The experimental design, Hypotheses 1 and 2, and statistical analyses were pre-registered prior to data collection and implemented as planned. In addition to the pre-registered analyses, we report results for Hypotheses 3 and 4, several auxiliary results, as well as structural estimates.

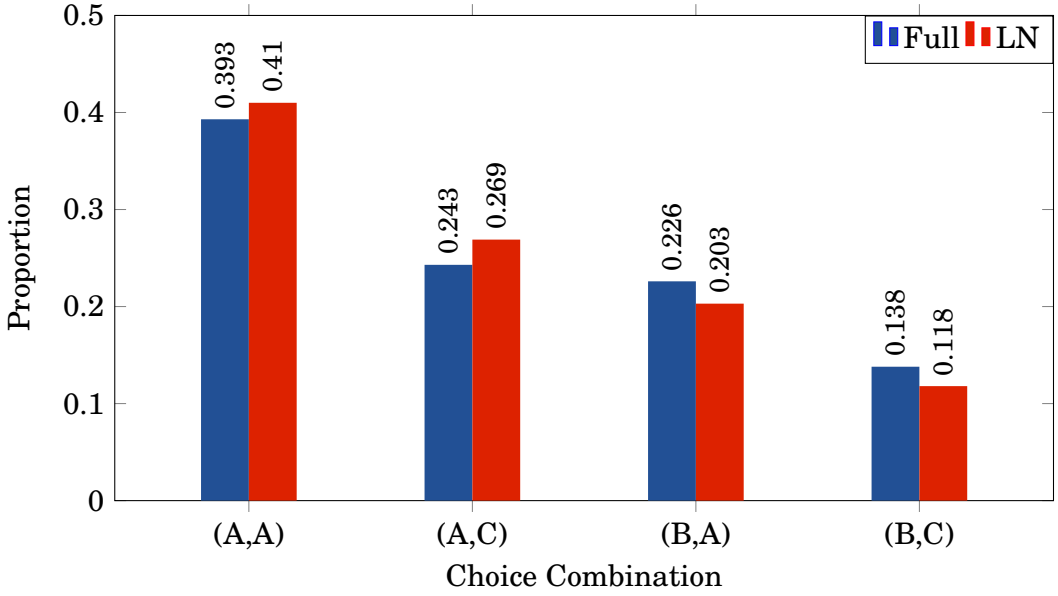
4.1. Descriptive Analysis

An overview of the choice data for our task and for the Holt–Laury task can be found in Figures 4 and 5 respectively. In the top panels the choice data is split up by sample (full or low noise). In the bottom panels the choice data is split up by task order (Order A and Order B). We use these figures to discuss the evidence for the above hypotheses. In Sections 4.2 and 4.3 we follow up with structural estimation.

Result 1 *Hypothesis 1 is rejected: participants' choices between* $\mathscr{L}_A$ *and* $\mathscr{L}_B$*, and between* $\mathscr{L}_A$ *and* $\mathscr{L}_C$*, are not the same for C1-C6 of Table 1.*

Support: Choices differ across C1-C6 at both the individual and population levels. Out of 370 participants, only 9 make consistent choices across C1-C6, including the repetition of C3, mostly choosing (A,C). A minority of participants (103) make the same choice in at

(a) Choice proportions in our task by sample.



(b) Choice proportions in our task by task order.

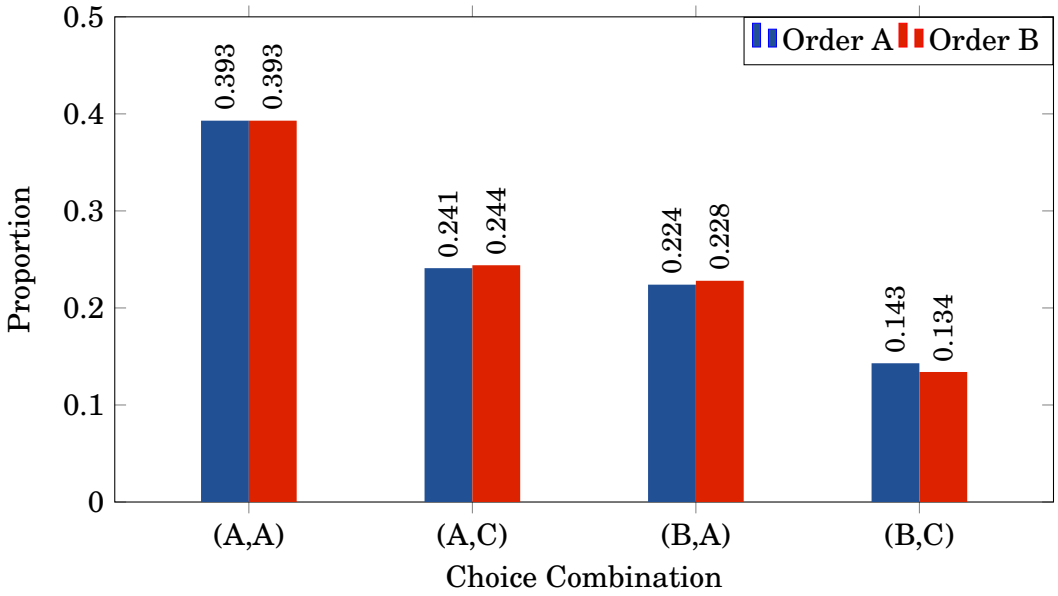
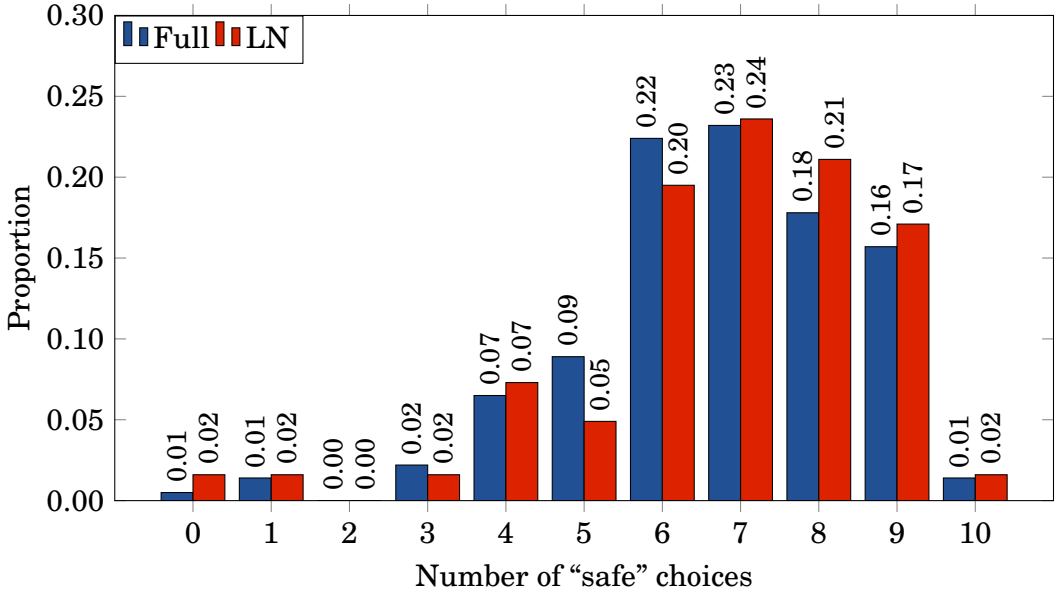


Figure 4: Choice proportions in our task. Panel (a) shows choice proportions by sample (Full vs. LN). Panel (b) shows choice proportions by task order (Order A vs. Order B).

(a) Choice proportions in the Holt–Laury task by sample.



(b) Choice proportions in the Holt–Laury task by order.

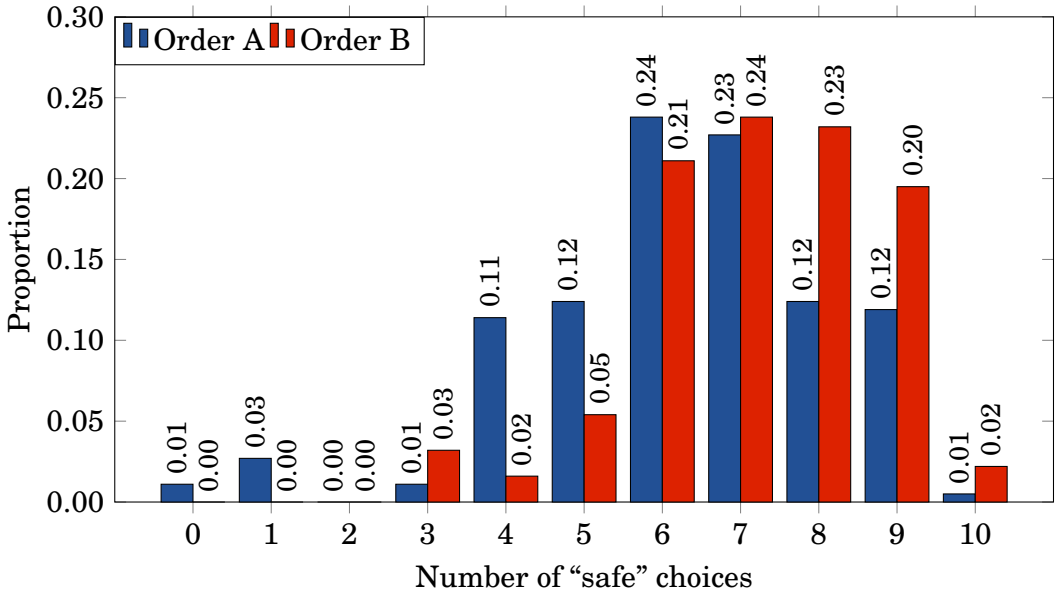


Figure 5: Choice proportions in the Holt–Laury task. Panel (a) reports proportions by sample (Full vs. LN). Panel (b) reports proportions by task order (Order A vs. Order B).

least five out of seven cases. Even if we ignore individual consistency and consider choice distributions, a Pearson's χ^2 test reveals they differ across the six cases ($p < 2.2 * 10^{-16}$).

Result 2 *Hypothesis 2 is rejected: only a minority of participants' choices are* (A,A).

Support: Pooling across C1-C12 of Table 1, with C3 occurring twice, the percentage of (A,A) choices is only 39%. Almost half the choices belong to the yellow and green areas, i.e. they are neither risk averse nor risk loving. Among the low-noise sub-sample the percentage of (A,A) choices is similar (41%).

Result 3 *Hypothesis 3 is rejected: the fraction of* (A,A) *choices in our task is far less than the fraction of four or more "safe" choices in the Holt–Laury task.*

Support: Panel (a) of Figure 5 shows the distributions of "safe" choices for both the full and low noise (LN) samples. For the full sample, the percentage of (A,A) choices in our task, 39%, is significantly lower than the fraction of more-than-four "safe" choices in the Holt–Laury task, 89%. For the low-noise sample these percentages are 41% and 88%.

Result 4 *Hypothesis 4 cannot be rejected: the choice frequencies in our task are the same whether our task comes before or after the Holt–Laury task, and the choice frequencies of the low-noise sub-sample are the same as those of the full sample.*

Support: For our task, results are very stable across orders. The choice proportions are essentially identical and a two-sample proportion tests confirm no significant differences, see the left panel of Table 2. The right panel compares choice proportions of the full sample and the low-noise (LN) sub-sample. Two-sample proportion tests show no significant differences for any of the four choices.

Choice	Order A	Order B	p-value
(A,A)	0.393	0.393	1.000
(A,C)	0.241	0.244	0.788
(B,A)	0.224	0.228	0.730
(B,C)	0.143	0.134	0.404

Choice	Full	LN	p-value
(A,A)	0.393	0.410	0.237
(A,C)	0.243	0.269	0.037
(B,A)	0.225	0.203	0.062
(B,C)	0.138	0.118	0.037

Table 2: Choice proportions under Order A and Order B (left) and for the Full and LN samples (right). The p-values are based on two-sample proportion tests.

We end this section with two auxiliary results that compare results in our task to those in the Holt–Laury task.

Result 5 *The results of the Holt–Laury task are affected by the order of the tasks.*

Support: Panel (b) of Figure 5 shows the distribution of "safe" choices in the Holt–Laury task when it came first (Order A) or second (Order B). The distribution of "safe" choices shifts rightward in Order B, i.e. participants who went through our task first made more "safe" choices. The difference is statistically significant (χ-square test, $p < 0.0001$).

Result 6 *More "safe" choices in the Holt–Laury task do not result in more* (A,A) *choices in our task.*

Support: Figure 13 in Appendix C shows the proportion of (A,A) choices by the number "safe" choices. Logit regressions that test for a linear trend yield estimated slopes that are positive but not statistically significant (full sample: $\hat{\beta} = 0.122$ (0.076), with the standard error in parenthesis, and $p = 0.108$; low-noise sub-sample: $\hat{\beta} = 0.072$ (0.130) and $p = 0.578$).

4.2. Non-Identification in the Holt–Laury Task

In our version of the Holt–Laury task, participants can switch only once from the "safe" to the "risky" lottery, so the expected utility of s "safe" choices is

$$U(s) = \frac{1}{10}\sum_{k=1}^{s} U_S(\tfrac{k}{10}) + \frac{1}{10}\sum_{k=s+1}^{10} U_R(\tfrac{k}{10}) \qquad \text{for } s = 0,\ldots,10, \tag{3}$$

where $U_S(p) = pu_2 + (1-p)u_1$ and $U_R(p) = p$, with p the chance of the best outcome and utilities normalized to 0, u_1, u_2, and 1. The $\frac{1}{10}$ in front of the sums is the chance that any of the ten lotteries of the Holt–Laury task is selected. Expected utilities map to choice probabilities via a logit rule, i.e.

$$P(s) = \frac{e^{U(s)/\mu}}{\sum_{s'=0}^{10} e^{U(s')/\mu}} \qquad \text{for } s = 0,\ldots,10.$$

The loglikelihood is given by

$$\log\mathrm{L}(u_1,u_2,\mu) = \sum_{s=0}^{10} n(s)\log(P(s))$$

with $n(s)$ the observed number of s "safe" choices. Let $\log\mathrm{L}(u_1,u_2)$ denote the loglikelihood that results by maximizing over μ for given (u_1,u_2). The grey surface in the right panel of Figure 6 shows $\log\mathrm{L}(u_1,u_2)$ over the triangle $0 \le u_1 \le u_2 \le 1$. Note that $\log\mathrm{L}(u_1,u_2)$ is constant on line segments emanating from the top-left corner $(u_1,u_2) = (0,1)$. In particular,

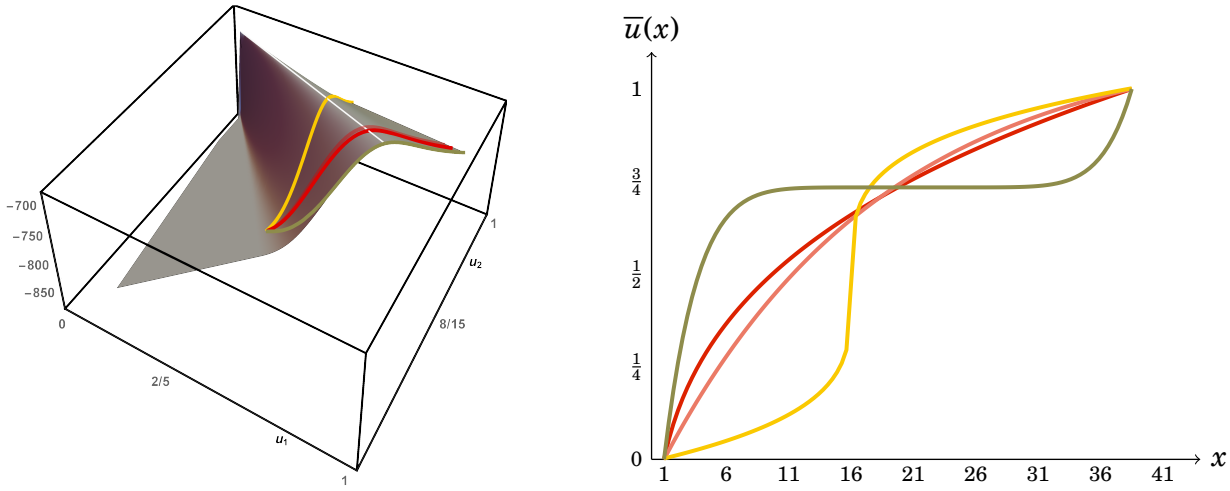


Figure 6: The grey surface in the left panel shows the loglikelihood $\log L(u_1, u_2)$ over the triangle $0 \leq u_1 \leq u_2 \leq 1$. The loglikelihood is constant on line segments emanating from the top-left corner $(u_1, u_2) = (0, 1)$. The white line segment $u_2 = 1 - 0.37u_1$ for $0 < u_1 \leq 0.73$ indicates utility pairs for which the loglikelihood reaches its maximum value. The (light) red curve depicts the loglikelihood of (CARA) CRRA utility pairs for $(\alpha \geq 0)$ $r \geq 0$. The yellow and green curves show the loglikelihoods of the utility in (4) when r varies and $t = 16$ and $t = 21$ respectively. All four curves cross the white line segment. The right panel shows the utility functions that correspond to these crossings. While the yellow "S shaped" and the green "inverted S shaped" utilities differ markedly from the red concave utilities, they fit the data from the Holt–Laury task equally well. Note: in both panels utilities are normalized, i.e. $\overline{u}(x) = (u(x) - u(\underline{\pi}))/(u(\overline{\pi}) - u(\underline{\pi}))$ with $\overline{\pi}$ ($\underline{\pi}$) the highest (lowest) possible prize.

$\log L(u_1, u_2)$ reaches its maximum value of -728.2 when $u_2 = 1 - 0.37u_1$ for $0 < u_1 \leq 0.73$, as indicated by the white line segment. Any (u_1, u_2) utility-pair that belongs to this segment provides the best possible fit of the Holt–Laury data. Since the white line segment runs through the yellow, red, and green areas of Figure 1, the Holt–Laury task cannot identify the type of Bernoulli utility and, hence, cannot identify risk aversion. The left panel of Figure 6 demonstrates this non-identifiability result in a non-parametric manner.

The preponderance of "risk averse" choices typically inferred from the Holt–Laury task merely reflect parametric assumptions. To elucidate, consider the Bernoulli utilities:

$$u(x) = \text{sign}(x - t)\frac{|x - t|^{1-r}}{1 - r} \tag{4}$$

When $t = 0$, this simplifies to a CRRA utility. But a strictly positive t acts as a reference point around which the shape of the utility is determined by r. When $0 < r < 1$ the utility in (4) is "S shaped," i.e. risk loving when $x < t$ and risk averse when $x > t$. When $r < 0$, it is "inverted S shaped," i.e. risk averse when $x < t$ and risk loving when $x > t$.

The yellow curve in the left panel of Figure 6 shows the utility pairs, and the associated loglikelihood, when we fix $t = 16$ in (4) and vary r. The green curve shows the same when we

$\hat{r}$	$\hat{\alpha}$	$\hat{\mu}$	logL	t
0.73 (0.03)		0.032 (0.002)	−728.2	0
0.74 (0.02)		0.023 (0.002)	−728.2	16
−6.6 (0.4)		0.036 (0.003)	−728.2	21
	0.056 (0.003)	0.031 (0.003)	−728.2	

Table 3: Estimates for the parametric model in (4) and CARA (bottom row). Each model yields the same maximum likelihood despite the utilities being starkly different, see the right panel of Figure 6.

fix $t = 21$. These values of t were chosen so that the utility's inflection point coincides with one of the middle prizes. The yellow and green curves cross the white segment. Hence, these alternative utilities describe the Holt–Laury data as well as the CARA and CRRA models that are indicated by the light and dark red curves, which also cross the white segment.

The estimation results of Table 3 provide the parameter values (and standard errors) for which these crossings occur. For each of the models in the first three rows, t is fixed and only the r and μ parameters are estimated.[17] All three models yield the same maximum likelihood, as does CARA in the bottom row. This despite the utilities being starkly different, see the right panel of Figure 6. The estimated CARA and CRRA utilities correspond to the light and dark red concave functions. The yellow "S shaped" and green "inverted S shaped" utilities are based on the estimates in the second and third row of Table 3.

In the left panel of Figure 6, the green curve may appear "close" to the red CRRA curve. In part, this is due to the normalization of utilities to an upper triangle. More importantly, lottery-choice behavior is not determined by utility levels but by the second derivative of the utility function, i.e. by the *type* of Bernoulli utility. As exemplified by the right panel of Figure 6, the "inverted S shaped" utility that corresponds to a point on the green curve in the left panel can be vastly different from the concave utility that corresponds to a point on the red curve. Even when the green and red points seem "close" in the left panel.

We end this section with a discussion of factors that could affect non-identification in the Holt–Laury task. For brevity, we do so point-by-point in a series of remarks.

Remark 1 In parametric models such as CRRA or CARA, a higher degree of risk aversion means that expected utility differences get smaller as all options become less attractive. Hence, the ARUM choice probability of a more risky option may be higher for a more risk averse individual (Wilcox, 2011; Apesteguia and Ballester, 2018).

[17] If also t were estimated then it would not be identified. Table 3 shows that the same maximum loglikelihood occurs for three different values of t. More generally, there is a continuum of t values that produce the same maximum loglikelihood.

A random preference model (RPM) avoids such non-monotonicities by perturbing preference parameters rather than expected utilities, see e.g. Jagelka (2024). The loglikelihood $\log L(u_1, u_2)$ depends on the utilities u_1 and u_2. Adding perturbations to these parameters, i.e. $\tilde{u}_1 = u_1 + \varepsilon_1$ and $\tilde{u}_2 = u_2 + \varepsilon_2$, changes the expected utility of s "safe" choices in (3) as follows

$$\widetilde{U}(s) = U(s) + \frac{s}{200}\big(20\varepsilon_1 + (1+s)(\varepsilon_2 - \varepsilon_1)\big)$$

The second term on the right side is a mean zero perturbation, so the resulting choice probabilities fit the ARUM framework and Theorem 3 applies.

Remark 2 Holt and Laury (2002) apply the Luce rule

$$P(s) = \frac{U(s)^{1/\mu}}{\sum_{s'=0}^{10} U(s')^{1/\mu}} \qquad \text{for } s = 0, \ldots, 10, \tag{5}$$

to determine choice probabilities, which is not an ARUM so Theorem 3 does not apply. However, as detailed in Appendix C, the likelihood is near constant for $u_2 = 1 - 0.37u_1$ and $0 < u_1 < 0.73$. A likelihood-ratio test cannot reject the hypothesis that the model belongs to this line segment. Hence, Theorem 3 applies in a statistical sense.

Remark 3 Holt and Laury (2002) ran multiple versions of their task in which the stakes were varied: in one version the prizes were (\$0.10,\$1.60,\$2,\$3.85) and in three other versions these prizes were multiplied by 20, 50, and 90. This improved identification as Holt and Laury conclude that a "power-expo" model

$$u(x) = \frac{1}{\alpha}\big(1 - \exp(-\alpha x^{1-r})\big) \tag{6}$$

best fits the data from all tasks (with a single task the "power-expo" would produce the same fit as other parametric models). We replicate their results in Appendix C. We further show how to create a wide range of utilities that fit the data equally well. To summarize, non-identification remains an issue even for the Holt–Laury task with varying stakes.

4.3. Estimating Individual Utilities and Out-of-Sample Predictions

A consequence of non-identification is that the Holt–Laury task cannot be used for out-of-sample predictions. There is a continuum of Bernoulli utilities that fit the data from this task equally well, so which one to use for out-of-sample predictions? One could insist on CRRA utilities, but the homogeneous CRRA model fares poorly when applied to our task.

Model	Sample	Parameters	logL
CRRA-Hom	Full	$\hat{r} = 0.54(0.03)$, $\hat{\mu} = 0.056(0.003)$	−6894.1
NP-Hom	Full	$\hat{u}_1 = 0.59(0.01)$, $\hat{u}_2 = 0.66(0.007)$, $\hat{\mu} = 0.050(0.03)$	−6801.7
CRRA-Het	Full	$\overline{r} = 1.5(0.1)$, $\overline{\mu} = 0.052(0.005)$	−5743.2
NP-Het	Full	$\overline{u}_1 = 0.61(0.01)$, $\overline{u}_2 = 0.69(0.01)$, $\overline{\mu} = 0.031(0.002)$	−5203.5
NP-Het	LN	$\overline{u}_1 = 0.61(0.02)$, $\overline{u}_2 = 0.70(0.02)$, $\overline{\mu} = 0.023(0.002)$	−1523.5

Table 4: Parameter estimates and log-likelihoods for the homogeneous and heterogeneous constant relative risk aversion (CRRA) and non-parametric (NP) models. The estimates are based on data from the Holt–Laury task and our task for a total of 29 decisions per participant. The hatted parameters in the top two lines are maximum likelihood estimates and the barred parameters in the bottom three lines are averages of individual estimates.

Result 7 *The homogeneous CRRA model is rejected by the choice data from our task.*

Support: Figure 7 shows the choice distributions for the twelve cases of Table 1. Figure 8 shows the predictions of the homogeneous CRRA model with $\hat{r} = 0.73$ and $\hat{\mu} = 0.032$. For each of the twelve cases, the CRRA model over-predicts (A,A). This is intuitive since without noise, (A,A) would be the only predicted choice. So any probability the estimated CRRA model puts on other choices is merely the result of noise diluting (A,A)-probability mass toward them. As a result, the ordering of choice frequencies in Figure 8 is wrong in eleven out of twelve cases (only C4 is correct). A Pearson's χ^2 test rejects the estimated CRRA model at $p < 0.001$ in all twelve cases except C4 ($p = 0.090$) and C5 ($p = 0.043$).

Of course, a main reason for rejection is the homogeneity assumption. To estimate individual utilities we combine our task and the Holt–Laury task. In our task, choice probabilities are constant along line segments emanating from the risk-neutral point in Figure 1 and in the Holt–Laury task they are constant along line segments emanating from the top-left corner. Combining the choice data from both tasks resolves the indeterminacy as the line segments cross in a single point.

Table 4 shows logit estimation results.[18] The heterogeneous non-parametric model (NP-Het) assumes individual-specific utility and noise parameters (u_1^i, u_2^i, μ^i). It nests the heterogeneous CRRA model (CRRA-Het) that restricts normalized utilities to

$$(u_1^i, u_2^i) = \left(\frac{16^{1-r^i}-1}{38.5^{1-r^i}-1}, \frac{21^{1-r^i}-1}{38.5^{1-r^i}-1}\right)$$

[18] In the third and fourth row of Table 4 we exclude one participant who chose the dominated option in the Holt–Laury task (10 safe choices) and the FOSD case of our task (lotteries B and C), chose the "I don't mind" option twice, and made otherwise inconsistent choices. The estimated noise parameter $\hat{\mu}^i$ for this participant is infinite (whence it could not be included in Table 4), i.e. the participant's behavior is completely random.

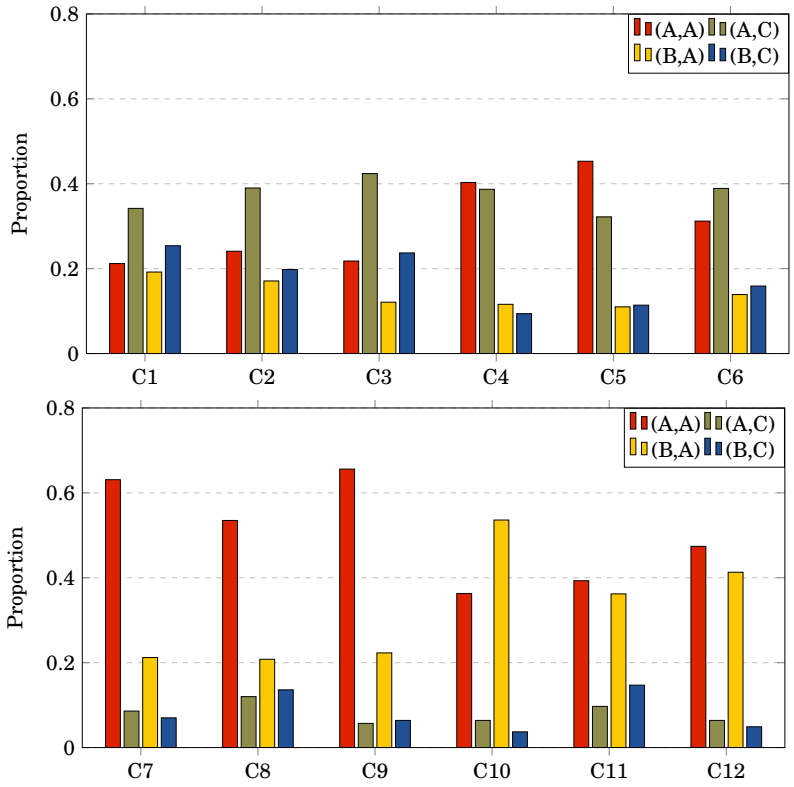


Figure 7: Observed choices for C1–C6 (top) and C7–C12 (bottom).

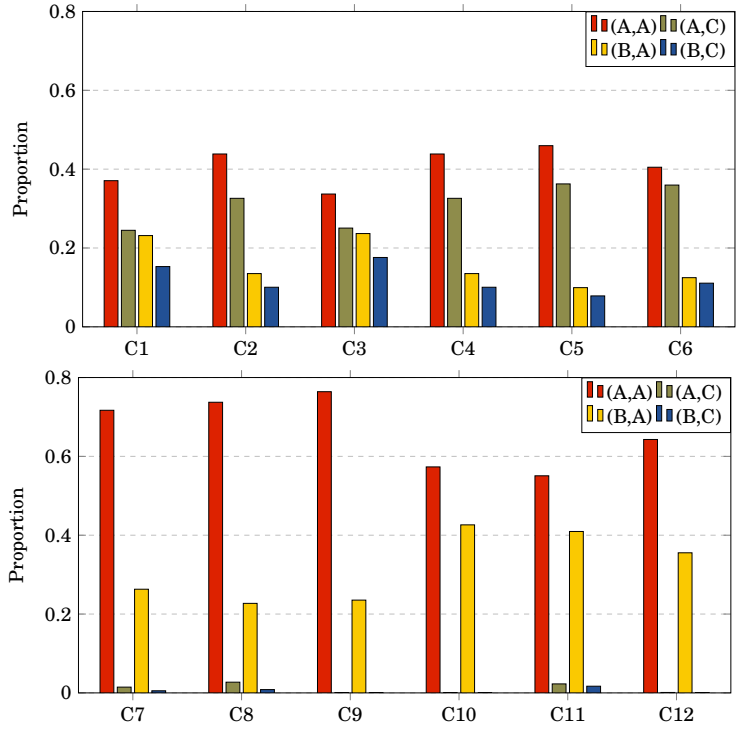


Figure 8: CRRA predictions for C1–C6 (top) and C7–C12 (bottom) when $\hat{r} = 0.73$.

CRRA-Het was estimated under the constraint $-10 \leq r_i \leq 5$, where the boundary values are such that the implied CRRA utilities are very close to $(u_1^i, u_2^i) = (0,0)$ and $(u_1^i, u_2^i) = (1,1)$ respectively. NP-Het also nests the homogeneous non-parametric model (NP-Hom) that restricts parameters to be the same across all participants, i.e. $(u_1^i, u_2^i, \mu^i) = (u_1, u_2, \mu)$ for all i. Finally, the homogeneous CRRA model (CRRA-Hom), which restricts r_i to be the same for all participants, is nested in both CRRA-Het and NP-Hom.

Result 8 *The homogeneous and heterogeneous CRRA models are rejected by the choice data from both tasks.*

Support: Twice the difference in loglikelihoods yields a likelihood-ratio test statistic of 184.8 (one degree of freedom) for the homogeneous case and of 1079.4 (369 degrees of freedom) for the heterogeneous case, see Table 4. Hence, both the homogeneous and heterogeneous CRRA models can be rejected at any conventional level of significance.

Result 9 *The heterogeneous CRRA model yields biased estimates.*

Support: For almost a third of the participants (115 out of 369) the estimated CRRA parameter exceeds 1.33, which is the threshold for at least nine "safe" choices in the Holt–Laury task. This is double the observed fraction, see the top panel of Figure 5. The corresponding non-parametric utility estimates for these 115 participants reveal that 100 of them (86%) have an "inverted S shaped" utility. To illustrate, consider the choices of one of few participants who did not make any mistakes: eight "safe" choices in the Holt–Laury task followed by (A,C) choices in C1–C6 of our task, (A,A) in C7–C9, (B,A) in C10–C12, (A,C) in the repeated C3, and (A,A) in FOSD. Under NP-Het, the estimated noise parameter is zero and so is the loglikelihood as all choices are consistent with an "inverted S shaped" utility. Under the heterogeneous CRRA model, the estimated risk-aversion parameter is the highest possible, $\hat{r}_i = 5$, and the loglikelihood is -8.89. The reason for this high estimate is that it implies utilities $(u_1, u_2) = (1,1)$, creating indifferences between (A,A) and the chosen options. While this is less accurate than predicting the chosen options with certainty, as NP-Het does, it is the best that CRRA-Het can do. To summarize, to account for deviations from (A,A), CRRA-Het produces unrealistically high estimates that lead to indifferences between (A,A) and other options. Even the average CRRA estimate, $\overline{r} = 1.5$, see Table 4, exceeds the threshold for at least nine "safe" choices in the Holt–Laury task.

The bottom line of Table 4 shows estimates of the NP-Het model using data from only the low-noise (LN) sub-sample.

Result 10 *The LN sub-sample effectively captures less noisy participants. Compared to the full sample the noise estimates for the LN sub-sample are significantly lower while the utility estimates are the same. In other words, observed choice frequencies are not biased by noisy decisionmaking (cf. Result 4).*

Support: The average estimated noise level in the LN sub-sample, $\overline{\mu} = 0.023(0.002)$, is significantly lower compared to the full sample, $\overline{\mu} = 0.031(0.002)$. A Wilcoxson rank-sum test (comparing the LN sample to the remaining non-LN participants) yields $p < 0.001$ and a two-sample Welch t-test yields $p = 0.043$. The average estimated utility parameters in the LN sub-sample, $\overline{u}_1 = 0.61(0.02)$ and $\overline{u}_2 = 0.70(0.02)$, are not significantly different from those in the full sample, $\overline{u}_1 = 0.61(0.02)$ and $\overline{u}_2 = 0.69(0.02)$. A Wilcoxson rank-sum test yields $p = 0.18$ and $p = 0.29$ respectively. A two-sample Welch t-test yields $p = 0.36$ and $p = 0.39$ respectively. In the full sample, 32% of the estimated utilities are concave, 2% are convex, 11% are "S shaped," and 55% are "inverted S shaped." The low-noise sub-sample produces virtually the same results: 32% of the estimated utilities are concave, 2% are convex, 7% are "S shaped," and 59% are "inverted S shaped."

NP-Het can be used to "out-of-sample" predict behavior. Specifically, we estimate individual utility and noise parameters (u_1^i, u_2^i, μ^i) using data from the Holt–Laury task and C1–C6 of our task to predict choices in C7–C12. The average estimates are similar to those of Table 4: $\overline{u}_1 = 0.62(0.01)$, $\overline{u}_2 = 0.71(0.01)$, and $\overline{\mu} = 0.024(0.002)$, and the loglikelihood is -2534.4. Figure 9 shows the estimated utilities:[19] 36% utilities are concave (red), 4% are convex (blue), 9% are "S shaped" (yellow), and 51% are "inverted S shaped" (green). These proportions are similar to the ones obtained using data from the Holt–Laury task and all fourteen cases of our task, see the support of Result 10. In other words, preferences are stable across the various cases of our task.

Preference stability allows us to make comparative statics predictions for C7–C12. A majority of the estimated utilities in Figure 9 fall in the upper green region of Figure 1. In C7–C9, this area is red, see Figure 3. So, compared to C1–C6, we should expect (i) an increase in (A,A) choices for these case. In C10–C12, this area is yellow so we should expect (ii) an increase (B,A) choices for these cases. There are few estimates in Figure 9 that fall into the green regions in any of the six panels of Figure 3 so we should expect (iii) a decrease in (A,C) choices in C7–C12.

Result 11 *The comparative statics predictions (i)–(iii) are borne out by the data.*

[19] Recall that even though some of the green estimates may appear "close" to some of the red estimates, the associated Bernoulli utilities can be vastly different. See the discussion in Section 4.2, in particular, Figure 6.

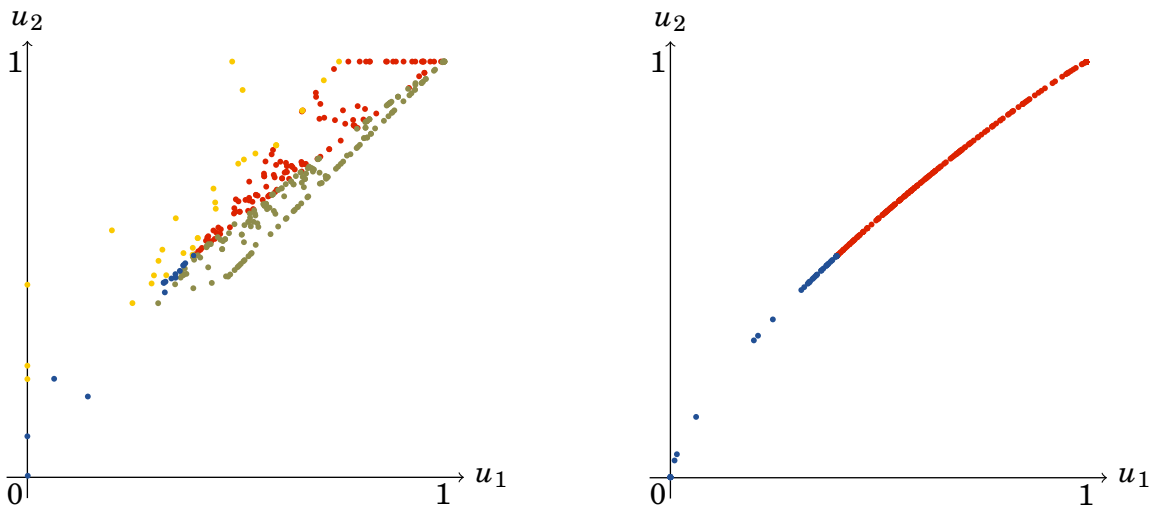


Figure 9: The colored dots show $(\hat{u}_1, \hat{u}_2)$ estimates for the 370 participants using data from C1-C6 of Table 1 and the Holt–Laury task. A red dot corresponds to a concave utility, a blue dot to a convex utility, a yellow dot to an "S shaped" utility, and a green dot to an "inverted S shaped" utility. In the left panel, estimates are unconstrained. In the right panel, estimates are constrained to lie on the CRRA curve. The numbers of (red, green, yellow, blue) dots are $(133, 187, 34, 16)$ in the left panel and $(318, 0, 0, 52)$ in the right panel.

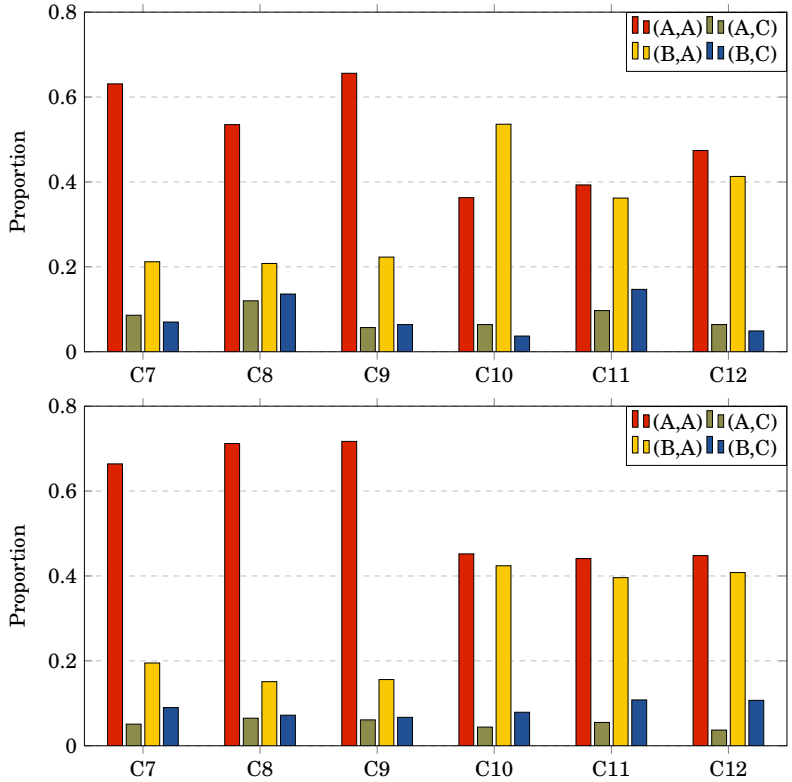
Support: See Figure 10. Panel (a) corresponds to the full sample and panel (b) to the low-noise sub-sample. In both panels, observed choices appear in the upper plot and predicted choices in the lower plot. For both the full sample and the low-noise sub-sample, the estimated model reproduces key features of the C7–C12 choice data. In particular, the increased frequency of red (A, A) choices in C7–C9 and the increased frequency of yellow (B, A) choices in C10–C12. This comes at the expense of the green (A, C) choices, which occur less frequently in C7–C12 compared to C1–C6. Besides reproducing comparative statics, the out-of-sample predictions are close in a quantitative sense. For the low-noise sub-sample, the estimated model cannot be rejected at the 1% level in four out of six cases (a χ^2 test yields $p = 0.817$ in C7, $p = 0.435$ in C9, $p = 0.239$ in C10, and $p = 0.228$ in C11).

Result 11 reinforces that preferences are stable and that a large fraction of the Bernoulli utilities are "inverted S shaped."

5. Survey Results

At the end of the experiment, participants answered questions regarding demographics, self reported risk, and real-life risky behaviors.

(a) Observed (top) and predicted (bottom) choices in C7–C12 (full sample).



(b) Observed (top) and predicted (bottom) choices in C7–C12 (low-noise sample).

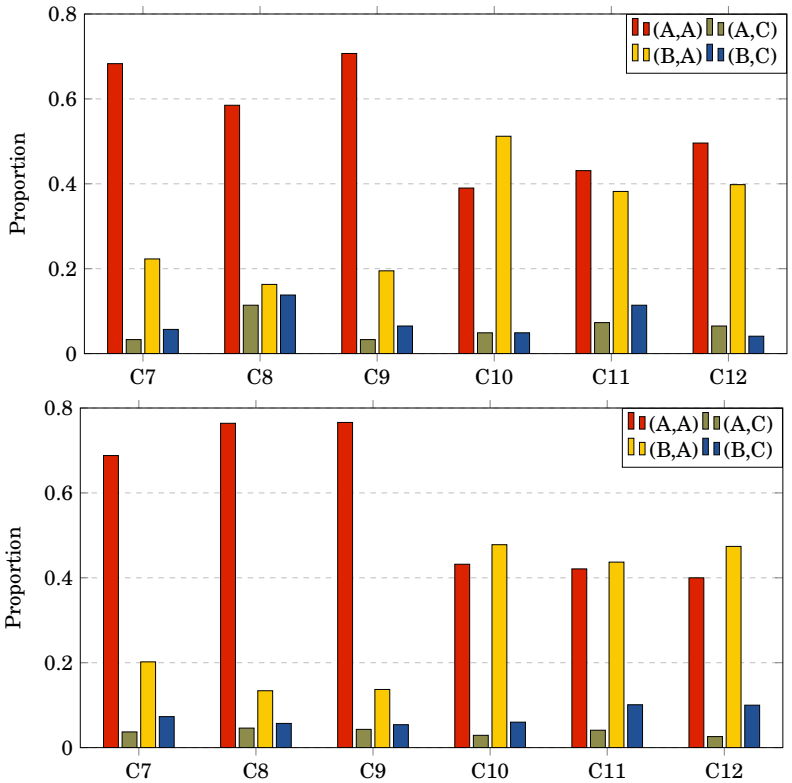


Figure 10: Observed (top) and predicted (bottom) choices in C7–C12. Panel (a) reports results for the full sample. Panel (b) reports results for the low-noise sample.

Task	Risk Measure	Spearman's ρ	p-value
Holt–Laury "safe" choices	General risk	-0.255	6.46×10^{-7}
Holt–Laury "safe" choices	Financial risk	-0.286	2.08×10^{-8}
Our task risk averse choices	General risk	-0.322	2.31×10^{-10}
Our task risk averse choices	Financial risk	-0.336	3.07×10^{-11}
Our task risk averse choices (LN)	General risk	-0.305	5.76×10^{-4}
Our task risk averse choices (LN)	Financial risk	-0.371	2.21×10^{-5}

Table 5: Spearman correlations between self-reported risk attitudes and risk-averse behavior in the Holt–Laury task and our task

5.1. Demographics

We find no statistically significant differences by gender (chi-square test, $p = 0.11$) or age (Spearman's $\rho = 0.072$, $p = 0.167$) in the Holt–Laury task. Figure 11 shows average choice frequencies by gender in our task. Differences across genders are statistically significant only for the risk-averse and risk-loving options, which is in line with the literature (Croson and Gneezy, 2009). None of the choices is significantly correlated with age.

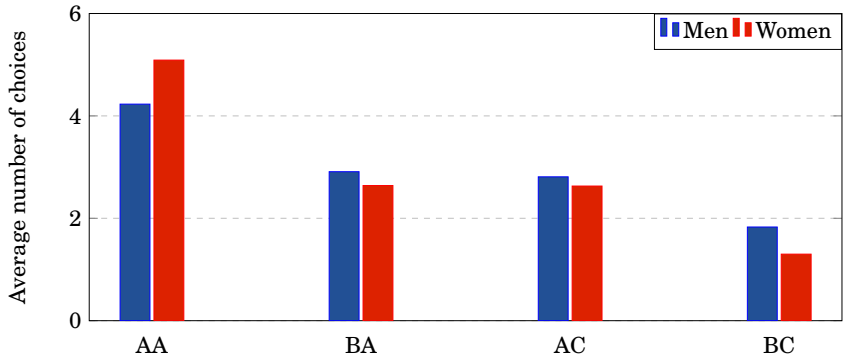


Figure 11: Average number of each decision type by gender in our task.

5.2. Self-Reported Risk

We included two commonly-used self-evaluations: the general willingness-to-take-risk question and its financial variant from the German Socio-Economic Panel, see Dohmen et al. (2011); Beauchamp et al. (2017); Eckel (2019); Charness et al. (2020); Falk et al. (2023). Both are correlated with risk aversion in both tasks in the expected direction.

Table 5 shows that all Spearman correlations are negative and significant. Higher self-reported willingness to take risks is associated with fewer risk-averse choices in our task and fewer "safe" choices in the Holt–Laury task. The financial question is generally more

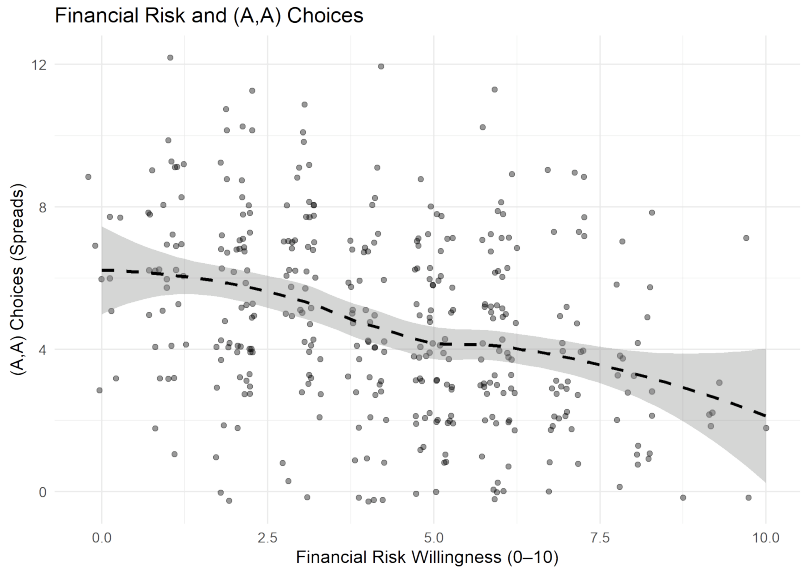


Figure 12: Financial self-reported risk (0–10) vs. number of risk averse choices (0–12) in our task. Points are individual participants (jittered for visibility). The dashed curve is a LOESS smoother (a nonparametric curve from local regressions), and the shaded band is its 95% confidence interval for the conditional mean.

correlated and the strongest relationship is between the financial question and risk aversion in our task. Figure 12 shows the corresponding trend, and Table 8 in Appendix C reports OLS regressions. Recent work by Chapman et al. (2025) cautions against interpreting correlations between qualitative self-assessments and incentivized choice measures as evidence that both capture the same underlying preferences. They argue that such correlations may reflect confounding factors, such as response heuristics or cognitive ability, rather than genuine risk preferences.

6. Conclusions

Following Holt and Laury (2002), and continuing a long tradition in psychology, multiple-price lists have been employed in hundreds of risk-elicitation experiments. They may well be the most commonly-used methodology in experimental economics today. One reason for their frequent use is that risk-elicitation tasks are routinely appended to experiments in which risk aversion is thought to play a role. Another is that multiple-price lists are easy to implement and seemingly easy to interpret. Assuming some parametric risk aversion model, the number of "safe" choices in the Holt–Laury task pins down a lower and upper bound for the risk aversion parameter, see Figure 2. Following this practice, many papers conclude that a vast majority of participants are risk averse.

One contribution of this paper is to elucidate why this conclusion is not justified. The four-prize setup of the Holt–Laury task is simple enough to graph the risk-loving and risk-averse Bernoulli utilities as subsets of the set of all non-decreasing utilities, see the blue and red areas in Figure 1. There is no a priori reason that participants' utilities belong to these areas. Expected utility theory allows for "S shaped" and "inverted S shaped" Bernoulli utilities that belong to the yellow and green areas respectively.

These alternative Bernoulli utilities are not only consistent with expected utility theory, they are also plausible. Participants may value prizes like €16 and €21 similarly, and differently from extreme prizes like €1 and €38.5. Such an "inverted S shaped" utility is least steep in the middle and implies risk seeking if there is a chance of the best prize (€38.5) and risk aversion if there is a chance of the worst prize (€1). Alternatively, participants may be risk seeking over the "loss" of getting €1 and risk averse over the "gain" of getting €38.5, which corresponds to an "S shaped" utility that is most steep in the middle.

Theorem 3 implies there always exist "S shaped" and "inverted S shaped" Bernoulli utilities that can explain any distribution of "safe" choices in the Holt–Laury task as well as any (parametric) risk aversion model can. In other words, the Holt–Laury task cannot identify the type of Bernoulli utility and, hence, cannot identify risk aversion.

A second contribution is that we introduce a task that generates the correct inequalities to test for concavity, i.e. whether utilities belong to the red area in Figure 1. It is based on Rothschild and Stiglitz's (1970) result that *any* risk averse individual prefers a lottery to a mean-preserving spread of itself. By varying the probabilities of the lottery we can further test the linearity assumption that underlies expected-utility theory.

A third contribution is empirical. Our task shows virtually no evidence for expected-utility maximization, see Result 1. Choices exhibit substantial decision error, which raises the question whether our data can be described by risk aversion plus noise. In line with previous experiments, we do find that a vast majority of participants make five or more "safe" choices in the Holt–Laury task. However, more "safe" choices do not imply more risk averse choices in our task, see Result 6. Moreover, the homogeneous CRRA model that is estimated using data from the Holt–Laury task is rejected when applied to data from our task, see Result 7. Finally, structural estimation of individual utilities using data from both tasks allows us to reject the homogeneous and heterogeneous CRRA models, see Result 8.

This failure does not come unexpected as our task provides little evidence of risk aversion, see Result 2. Only 39% of the choices are red (A,A) choices, see Figure 4, and only 32% of the estimated Bernoulli utilities are concave, see Result 10. A majority (55%) of the estimated utilities are "inverted S shaped," 11% is "S shaped," and a small fraction (2%) is

convex. We demonstrate that preference heterogeneity augmented with decision error provides the best "in sample" and "out of sample" fit of the data from the combined tasks, as well as correct comparative statics predictions.

Preference heterogeneity and decision error likely play a role in explaining data from other lottery-choice experiments as well. Which Bernoulli utilities are most prevalent in other experiments will depend on the context. Specifically, the choice of the lotteries' prizes and probabilities will determine the extent to which reference points, choice set effects, and decision error affect behavior.

References

Abdellaoui, M., A. Driouchi, and O. L'Haridon (2011). Risk aversion elicitation: Reconciling tractability. *Theory and Decision 71*, 63–80.

Anderson, L. R. and J. M. Mellor (2008). Predicting health behaviors with an experimental measure of risk preference. *Journal of health economics 27*(5), 1260–1274.

Anderson, L. R. and J. M. Mellor (2009). Are risk preferences stable? comparing an experimental measure with a validated survey-based measure. *Journal of Risk and Uncertainty 39*(2), 137–160.

Apesteguia, J. and M. A. Ballester (2018). Random models for the joint treatment of risk and time preferences. *Journal of Political Economy 126*(1), 74–106.

Baillon, A. and O. L'Haridon (2021). Discrete arrow–pratt indexes for risk and uncertainty. *Economic Theory 72*(4), 1375–1393.

Barsky, R. B., F. T. Juster, M. S. Kimball, and M. D. Shapiro (1997). Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement study. *The quarterly journal of economics 112*(2), 537–579.

Beauchamp, J. P., D. Cesarini, and M. Johannesson (2017). The psychometric and empirical properties of measures of risk preferences. *Journal of Risk and uncertainty 54*(3), 203–237.

Belzil, C. and T. Jagelka (2020). Separating true preferences from noise and endogeneous effort. Technical report.

Binswanger, H. P. (1980). Attitude toward risk: Experimental measurement in rural india. *American Journal of Agricultural Economics 62*, 395–407.

Chapman, J., P. Ortoleva, E. Snowberg, L. Yariv, and C. Camerer (2025). Reassessing qualitative self-assessments and experimental validation. Technical report, National Bureau of Economic Research.

Charness, G., T. Garcia, T. Offerman, and M. C. Villeval (2020). Do measures of risk attitude in the laboratory predict behavior under risk in and outside of the laboratory? *Journal of Risk and Uncertainty 60*(2), 99–123.

Charness, G., U. Gneezy, and A. Imas (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization 87*, 43–51.

Choi, S., R. Fisman, D. Gale, and S. Kariv (2007). Consistency and heterogeneity of individual behavior under uncertainty. *American Economic Review 97*(5), 1921–1938.

Crosetto, P. and A. Filippin (2013). The "bomb" risk elicitation task. journal of risk and uncertainty. *Journal of Risk and Uncertainty 47*, 31–65.

Crosetto, P. and A. Filippin (2015). A theoretical and experimental appraisal of four risk elicitation methods. *Experimental Economics 18*(6), 1–29.

Croson, R. and U. Gneezy (2009). Gender differences in preferences. *Journal of Economic literature 47*(2), 448–474.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2010). Are risk aversion and impatience related to cognitive ability? *American Economic Review 100*(3), 1238–1260.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the european economic association 9*(3), 522–550.

Eckel, C. C. (2019). Measuring individual risk preferences. *IZA World of Labor*.

Eckel, C. C. and P. J. Grossman (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior 23*, 281–295.

Eckel, C. C. and P. J. Grossman (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization 68*, 1–17.

Falk, A., A. Becker, T. Dohmen, D. Huffman, and U. Sunde (2023). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Management Science 69*(4), 1935–1950.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics 10*(2), 171–178.

Friedman, D., S. Habib, D. James, and B. Williams (2022). Varieties of risk elicitation. *Games and Economic Behavior*, forthcoming.

Friedman, D., R. M. Isaac, D. James, and S. Sunder (2014). *Risky curves: On the empirical failure of expected utility*. Routledge.

Goeree, J. K., C. A. Holt, and T. R. Palfrey (2016). *Quantal Response Equilibrium*. Princeton, USA: Princeton University Press.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association 1*(1), 114–125.

Hadar, J. and W. R. Russell (1969). Rules for ordering uncertain prospects. *American Economic Review 59*, 25–34.

Hanoch, G. and H. Levy (1969). The efficiency analysis of choices involving risk. *Review of Economic Studies 36*, 335–346.

Harrison, G. and E. Rutström (2008). Risk aversion in the laboratory. In J. C. Cox and G. W. Harrison (Eds.), *Risk Aversion in Experiments*, pp. 41–196.

Holt, C. A. and S. Laury (2002). Risk aversion and incentive effects. *American Economic Review 92*(5), 1644–1655.

Holt, C. A. and S. Laury (2005). Risk aversion and incentive effects: New data without order effects. *American Economic Review 95*(3), 902–904.

Holzmeister, F. and M. Stefan (2021). The risk elicitation puzzle revisited: Across-methods (in)consistency? *Experimental Economics 24*, 593–616.

Jagelka, T. (2024). Are economists' preferences psychologists' personality traits? a structural approach. *Journal of Political Economy 132*(3), 910–970.

Johnson, C., A. Baillon, H. Bleichrodt, Z. Li, D. Van Dolder, and P. Wakker (2021). Prince: An improved method for measuring incentivized preferences. *Journal of Risk and Uncertainty 62*(1), 1–28.

Levy, M. and H. Levy (2001). Testing for risk aversion: A stochastic dominance approach. *Economics Letters 71*, 233–240.

L'Haridon, O. and F. Vieider (2019). All over the map: A worldwide comparison of risk preferences. *Quantitative Economics 10*, 185–215.

Pedroni, A., R. Frey, A. Bruhin, G. Dutilh, R. Hertwig, and J. Rieskamp (2017). The risk elicitation puzzle. *Nature Human Behavior 1*, 803–809.

Rothschild, M. and J. Stiglitz (1970). Increasing risk i: A definition. *Journal of Economic Theory 2*, 225–243.

Smith, V. L. (1989). Theory, experiment and economics. *Journal of Economic Perspectives 3*(1), 151–169.

Tversky, A. and D. Kahneman (1982). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty 5*(4), 297–323.

Wakker, P. and D. Deneffe (1996). Eliciting von neumann-morgenstern utilities when probabilities are distorted or unknown. *Management Science 42*(8), 1131–1150.

Wilcox, N. T. (2011). 'stochastically more risk averse': A contextual theory of stochastic discrete choice under risk. *Journal of Econometrics 162*(1), 89–104.

A. Proofs

Proof of Theorem 1: The "only if" part of Theorem 1 follows from Rothschild and Stiglitz's (1970) classic result that a risk averse individual prefers any lottery to a mean-preserving spread of itself. For the "if" part we need to show that the inequalities generated by the choices between $(\mathscr{L}, \mathscr{L}_k)$ for $1 < k < K$ are sufficient to establish concavity of the Bernoulli utility u over the prizes $\pi = (\pi_1, \ldots, \pi_K)$. Concavity of u means that its slope is non-increasing, i.e.

$$\frac{u(\pi_k) - u(\pi_{k-1})}{\pi_k - \pi_{k-1}} \geq \frac{u(\pi_{k+1}) - u(\pi_k)}{\pi_{k+1} - \pi_{k-1}} \qquad \text{for } 1 < k < K \tag{A.1}$$

Let p and q denote the probabilities defining lotteries $\mathscr{L}$ and $\mathscr{L}_k$ respectively. The difference $p - q$ has entries

$$(p-q)_\ell = \begin{cases} 0 & \text{if} \quad \ell < k-1 \\ -\frac{\pi_{k+1}-\pi_k}{\pi_{k+1}-\pi_{k-1}}\, p_k & \text{if} \quad \ell = k-1 \\ p_k & \text{if} \quad \ell = k \\ -\frac{\pi_k-\pi_{k-1}}{\pi_{k+1}-\pi_{k-1}}\, p_k & \text{if} \quad \ell = k+1 \\ 0 & \text{if} \quad \ell > k+1 \end{cases}$$

If $\mathscr{L}$ is preferred to $\mathscr{L}_k$ for $1 < k < K$ then

$$u(\pi_k) \geq \frac{\pi_{k+1} - \pi_k}{\pi_{k+1} - \pi_{k-1}}\, u(\pi_{k-1}) + \frac{\pi_k - \pi_{k-1}}{\pi_{k+1} - \pi_{k-1}}\, u(\pi_{k+1}) \qquad \text{for } 1 < k < K \tag{A.2}$$

and it is straightforward to verify that (A.2) is equivalent to (A.1). ■

Proof of Theorem 3: Here we show non-identification for our version of the Holt–Laury task in which participants can switch only once from the "safe" to the "risky" lottery. As in Section 2.1, we normalize the utilities of the four prize values to 0, u_1, u_2, and 1 and let p denote the chance of the best outcome. The expected utility of s "safe" choices is

$$U(s) = \frac{1}{10} \sum_{k=1}^{s} U_S(\tfrac{k}{10}) + \frac{1}{10} \sum_{k=s+1}^{10} U_R(\tfrac{k}{10}) \qquad \text{for } s = 0, \ldots, 10,$$

where $U_S(p) = pu_2 + (1-p)u_1$ and $U_R(p) = p$ and the $\frac{1}{10}$ in front of the sums is the chance that any of the ten lotteries of the Holt–Laury task is selected.

To account for noise in decisionmaking the expected utilities are perturbed by additive shocks $\widetilde{U}(s) = U(s) + \mu\varepsilon_s$. The ε_s are (possibly correlated) mean-zero random variables with a strictly positive density, and $\mu > 0$ is a noise parameter that measures the impact of the shocks. The additive random utility model (ARUM) assumes participants maximize their

perturbed utilities, i.e. the probability of s "safe" choices is

$$P(s) = \text{Prob}\big(\varepsilon_s - \varepsilon_t \geq \tfrac{U(t)-U(s)}{\mu} \ \ \forall\, t\big) \tag{A.3}$$

Estimated utilities $(\hat{u}_1, \hat{u}_2)$ and an estimated noise parameter $\hat{\mu}$ follow from maximizing the loglikelihood

$$\text{logL} = \sum_{s=0}^{10} n(s)\log(P(s))$$

with $n(s)$ the number of participants who made s "safe" choices.

When $u_2 = 1 - \alpha u_1$, the expected utility of s "safe" choices is

$$U(s) = \frac{11}{20} + \frac{u_1}{200}(s+1)(20 - s(1+\alpha))$$

The utility difference between s and t "safe" choices is

$$U(s) - U(t) = \frac{u_1}{200}(s-t)(20 - (1+\alpha)(1+s+t))$$

When $\mu = \beta u_1$ the ratio $(U(s) - U(t))/\mu$ in (A.3) is independent of u_1. Hence, so is $P(s)$ and the loglikelihood logL. In other words, when $\mu = \beta u_1$, the loglikehood is constant on the line segment $u_2 = 1 - \alpha u_1$ for $0 < u_1 \leq \frac{1}{1+\alpha}$. The $\alpha > 0$ and $\beta > 0$ parameters can be estimated from the data. ■

B. Translated Instructions (Order A)

WELCOME

Thank you for participating in today's session.

The entire experiment will be conducted through your computer. Please avoid any distractions during the session.

1

General instructions

- The experiment will last at most one hour and a half.
- During the experiment you can earn money.
- The amount earned may be different for each participant.
- The amount you earn depends on the choices you make and on chance. In addition, you will receive a 4€ show-up fee.

2

General instructions

- The experiment has two parts and a questionnaire.
- At the end of the experiment, the computer will randomly select one choice from each part. You will be paid the average of the results in those two choices.
- I will start explaining Part 1. When it is finished, please wait for further instructions.

3

Part 1

- You will choose between 2 different lotteries: A or B.
- You will choose between A or B ten different times.
- These ten choices are shown in ten different rows.
- The euro amounts in lotteries A and B will always be the same.
- The probabilities in lotteries A and B will change from row to row.

4

Part 1

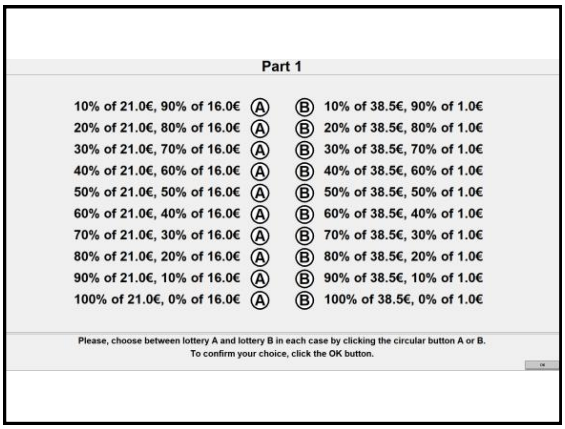
- For both lotteries A and B, the probailitiy of the larger amount starts at 10% and increases to 100%.
- For both lotteries A and B, the probability of the smaller amount starts at 90% and decreases to 0%.
- If you select B in a certain row, then you cannot select A in the rows below it.

5

Part 1: payments

- At the end of the experiment, the computer first randomly selects one of your choices (one of the rows).
- Then, the computer rolls a 100-side die, which determines the payoff for the selected choice.
- For example, option A of row 1 pays 21€ if the roll of the die is 1-10 and 16€ if the roll is 11-100.

6



7

Part 2

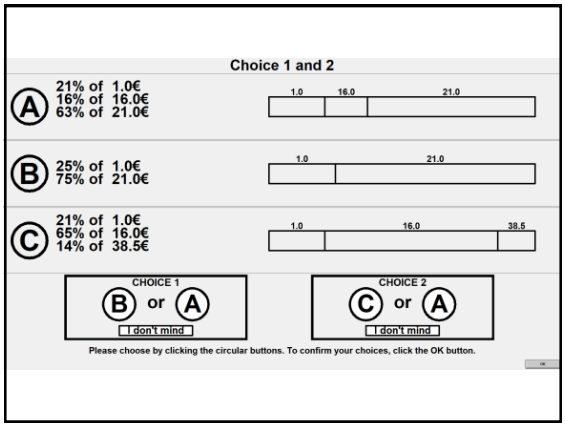
- This part consists of 14 periods.
- Each period has two choices, where you will choose between two different lotteries.
- The lotteries are represented in two ways.
- On the left you see the probabilities and the euro amounts. On the right you see the same information represented by bars.
- The size of the bar represent the probability of obtaining the amount shown on top of the bar.

8

Part 2

- To make your choice, click the circle with corresponding choice “A”, “B” or “C”.
- You can also click in the “I don’t mind” button and the choice will be made for you.

9



10

Part 2: payments

- At the end of the experiment, the computer first randomly selects one of your choices in this part.
- Then, the computer rolls a 100-sided die, which determines the payoff for the selected choice.
- For the example on the previous slide, option A pays 1€ if the roll of the die is 1-21, 16€ if the roll of is 22-37, and 21€ if the roll is 38-100.

11

C. Additional Empirical Support

Result 4: We estimate a multinomial logit with four choice categories and test if group membership (Order A vs Order B; Full sample vs LN) affects the distribution of choices. Specifically, we compare the loglikelihood, LogL, of the model that allows probabilities to vary by group to the loglikelihood, LogL_c, of the model that constrains them to be equal across groups, using the likelihood–ratio statistic $LR = 2(\text{LogL} - \text{LogL}_c)$. Because each participant makes multiple choices, we obtain a cluster-robust p-value by calibrating the LR against its permutation distribution formed by shuffling group labels at the participant level, using 2,000 permutations, which preserves within-participant choice patterns.

Comparison	LR	p-value
Order A vs Order B	0.687	0.943
Full sample vs LN	21.545	0.015

Table 6: Overall permutation LR tests by comparison (permutation p-values).

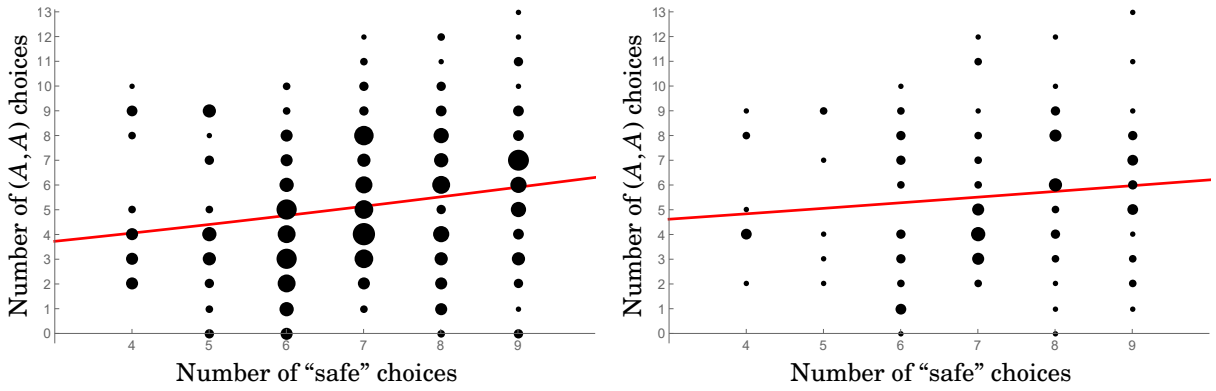


Figure 13: Number of (A,A) choices in our task (ranging from 0 to 13) by the number of "safe" choices (ranging from 4 to 9) for the full sample (left panel) and the low-noise sub-sample (right panel). In both panels, the area of a point is proportional to the number of observations. The red line shows the estimated logit model with a linear trend. The estimated slope is positive but not significant for either the full or low-noise sub-sample.

Result 6: Figure 13 provides an overview of the distributions of (A,A) choices by the number of "safe" choices for those participants who that could be classified as risk averse (i.e. with four or more "safe" choices). The left panel pertains to the full sample and the right panel to the low-noise sub-sample. In both panels, the area of a dot is proportional to the number of observations. Logit regressions that test for a linear trend in the proportion of (A,A) choices by the number "safe" choices yield estimated slopes that are positive but not

statistically significant. Full sample: $\hat{\beta} = 0.122$ (0.076), with the standard error in parenthesis, and $p = 0.108$. Low-noise sub-sample: $\hat{\beta} = 0.072$ (0.130) and $p = 0.578$.

Remark 2: We normalize the utilities of the four prize values to 0, u_1, u_2, and 1, and estimate u_1 and u_2 using the Luce model in (5). Specifically, for $u_1 = 0.01, 0.02, \ldots, 0.73$ we maximize the loglikelihood with respect to u_2 and the noise parameter μ. The resulting (u_1, u_2) pairs are plotted in the four-colored triangular region, see Figure 14.

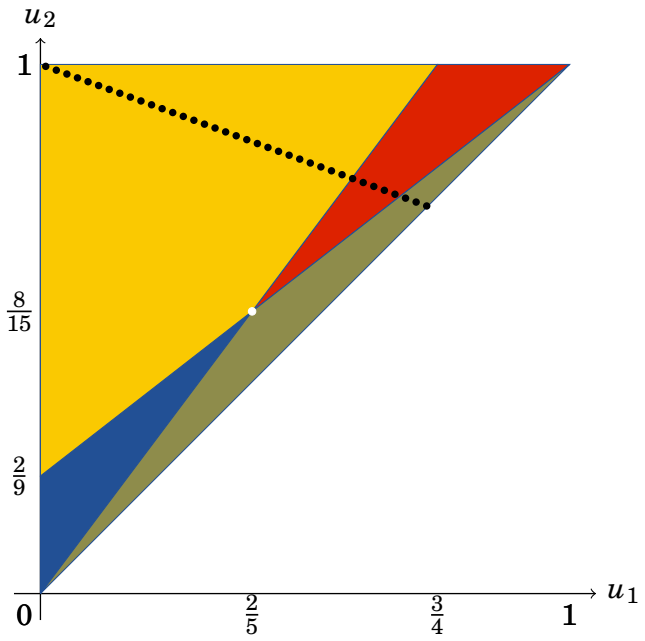


Figure 14: The black dots show (u_1, u_2) estimates using the Luce model in (5). These estimates fall (roughly) on the same line segment as the logit estimates. Also the maximum loglikelihood is very similar although it is non-constant: it falls from -728.8 in the upper-left corner of the yellow area to -731.1 on the diagonal boundary of the green area. A likelihood-ratio test shows that these differences are not statistically significant.

The Luce estimates are very similar to the Logit estimates and (roughly) fall on the line segment $u_2 = 1 - 0.37u_1$. The loglikelihood is not constant on this line segment. It falls from -728.8, for $(u_1, u_2) = (0.01, 0.996)$ in the yellow region, to -731.1, for $(u_1, u_2) = (0.73, 0.73)$ on the diagonal in the green region. But a likelihood-ratio test cannot reject any (u_1, u_2) on the line segment $u_2 = 1 - 0.37u_1$. In contrast with logit, the Luce noise parameter does not grow linearly with u_1. Instead, the best fit is $\mu = 0.069 * u_1^{0.926}$.

Remark 3: Holt and Laury (2002) conducted four treatments. In treatment Low, the prize values of the "safe" lottery were \$1.60 and \$2, and the prize values of the "risky" lottery were \$0.10 and \$3.85. In the other three treatments, labeled $\times 20$, $\times 50$, and $\times 90$, these prize values were multiplied by 20, 50, and 90 respectively.

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Low	1.	0.99	0.98	0.92	0.66	0.4	0.17	0.04	0.01	0.
×20	1.	0.99	0.98	0.94	0.81	0.62	0.39	0.17	0.06	0.
×50	1.	1.	1.	1.	1.	0.84	0.65	0.36	0.16	0.
×90	1.	1.	1.	1.	0.86	0.83	0.66	0.44	0.38	0.

Table 7: The probability of choosing the "safe" lottery when the chance of the high prize p varies from 0.1 to 1.0 in each of the four treatments that Holt and Laury (2002) report.

Table 7 shows the observed frequencies of choosing the "safe" lottery in each treatment as the chance p of the high prize varies from 0.1 to 1.0. We read these observed frequencies off their Figure 6, which likely introduced small measurement error. But the estimates we find are very close to those reported by Holt and Laury (2002). Using the Luce model in (5) and the "power-expo" utility $u(x)$ in (6), we obtain $\hat{r} = 0.263$, $\hat{\alpha} = 0.029$, $\hat{\mu} = 0.11$, and the loglikelihood is -812.4 .

A non-parametric approach follows by assigning a utility to any of the prizes (across all treatments). There are 15 prizes (the \$2 prize occurs in both Low and ×20), which we label $\pi = \pi_1, \ldots, \pi_{15}$.[20] The associated utilities are $u_1, \ldots, u_{15}$, where we normalize the highest utility to one, i.e. $u_{15} = 1$. Concavity of the Bernoulli utility can be imposed by restricting slopes to be non-increasing:

$$\frac{u_k - u_{k-1}}{\pi_k - \pi_{k-1}} \geq \frac{u_{k+1} - u_k}{\pi_{k+1} - \pi_k} \quad \text{for} \;\; 2 \leq k \leq 14 \tag{C.2}$$

Estimating the data from the Holt–Laury paper under these constraints yields a loglikehood of -803.0. This means that there is a continuum of concave Bernoulli utilities that fit the data from the combined treatments as well as (or slightly better than) the "power-expo" model.

In addition, we can estimate non-concave Bernoulli utilities with "inverted S shaped" features, e.g. by imposing that the utilities of the middle prizes in each of the four treatments are the same. Estimating the data from the Holt–Laury paper under the constraints $u_2 = u_3$, $u_7 = u_8$, $u_{10} = u_{11}$, and $u_{12} = u_{13}$, yields a loglikelihood of -799.8. So also these utilities fit the Holt–Laury data as well as (or better than) the "power-expo" model.

Section 5.2: We ran OLS regressions to test whether self-reported willingness to take risks predicts risk averse choices in both tasks. The single coefficient is the expected change in risk-averse choices per one-point increase on the self-report scale. Table 8 shows that the coefficients are negative and significant. The financial measure slightly outperforms the

[20] The list of prizes is $\pi = \{0.1, 1.6, 2, 3.85, 5, 9, 32, 40, 77, 80, 100, 144, 180, 192.5, 346.5\}$.

Task	Predictor	Coef.	p-value	R^2
Holt–Laury "safe" choices	General Risk	-0.237	7.57×10^{-8}	0.076
Holt–Laury "safe" choices	Financial Risk	-0.239	4.49×10^{-9}	0.089
Our task risk averse choices	General Risk	-0.416	1.58×10^{-10}	0.105
Our task risk averse choices	Financial Risk	-0.400	2.83×10^{-11}	0.114
Our task risk averse choices (LN)	General Risk	-0.384	3.84×10^{-4}	0.098
Our task risk averse choices (LN)	Financial Risk	-0.418	1.82×10^{-5}	0.141

Table 8: OLS regressions of the number of safe choices on self-reported willingness to take risks (0–10). Coefficients are changes in the expected number of safe choices per one-point increase in the self-report scale.

general one. Self-reports explain more variation in our task than in the Holt–Laury task, and predictive power increases within the low-noise sub-sample.

Real-Life Risk Behaviors: In the survey, we asked participants about real-life risk proxies adapted from Barsky et al. (1997), with modifications from Anderson and Mellor (2008): risky driving, smoking, drinking, and body-mass index. We find no significant correlations with the Holt–Laury task or our task, for the full sample or the low-noise sub-sample, whether we look at the number of risk–averse (A,A) choices, the number of risk–loving (B,C) choices, or the number of "safe" choices in the Holt–Laury task.