



# The predictive power of risk elicitation tasks

Michele Garagnani<sup>1</sup>

Accepted: 13 March 2023 / Published online: 3 May 2023  
© The Author(s) 2023

## Abstract

This work reports the results of two online experiments with a general-population sample examining the performance of different tasks for the elicitation of risk attitudes. First, I compare the investment task of Gneezy and Potters (1997), the standard choice-list method of Holt and Laury (2002), and the multi-alternative procedure of Eckel and Grossman (2002) and evaluate their performance in terms of the number of correctly-predicted binary decisions in a set of out-of-sample lottery choices. There are limited differences between the tasks in this sense, and performance is modest. Second, I included three additional budget-choice tasks (selection of a lottery from a linear budget set) where optimal decisions should have been corner solutions, and find that a large majority of participants provided interior solutions instead, casting doubts on people's understanding of tasks of this type. Finally, I investigate whether these two results depend on cognitive ability, numerical literacy, and education. While optimal choices in budget-choice tasks are related to numerical literacy and cognitive ability, the predictive performance of the risk-elicitation tasks is unaffected.

**Keywords** Risk preferences · Elicitation methods · Budget sets · Heterogeneity

**JEL Classification** C91 · D81 · C83

## 1 Introduction

The ability to estimate risk preferences at the individual level is of utmost importance for decision analysis and policy evaluation. Accordingly, a number of methods to measure risk attitudes have been proposed, going back at least to Binswanger (1980), see Holt and Laury (2014) and Mata et al. (2018) for recent reviews. A

---

✉ Michele Garagnani  
michele.garagnani@econ.uzh.ch

<sup>1</sup> Department of Economics, University of Zurich, Blümlisalpstrasse 10, CH-8006 Zurich, Switzerland

popular strand of the literature employs elicitation procedures in which subjects make repeated choices between two risky outcomes; the data obtained in this way consist of a finite number of binary choices, which can then be used to partially recover a subject's preference. The most influential among these procedures is the one of Holt and Laury (2002, 2005) (HL), which derives risk parameter intervals from a series of ordered binary lottery choices. However, it has been argued that this method might be too complex and too difficult to understand (Charness & Gneezy, 2010), especially for non-student populations (Yu et al., 2019).<sup>1</sup>

If a method is perceived as being too complex, estimates become inconsistent and reliability might be questionable (Dave et al., 2010; Charness et al., 2018). As a consequence, other methods have tried to simplify the elicitation procedure. Designed as a direct alternative to HL is the procedure by Eckel and Grossman (2002, 2008) (EG) which involves a single choice among 6 gambles, all with 0.5 probability of winning a higher prize. Yet another popular alternative is the Investment Task (INV) of Gneezy and Potters (1997) (see also Charness & Gneezy, 2010; Charness et al., 2013), which simply asks individuals to allocate money between a safe option and a risky one.<sup>2</sup> This latter method belongs to a growing strand employing choice tasks given a fixed budget set, either in the form of an explicit allocation of a monetary budget or as a direct choice from, say, a linear budget set (i.e., Gneezy & Potters, 1997; Choi et al., 2007a, b, 2014; Ahn et al., 2014; Hey & Pace, 2014; Castillo et al., 2017; Halevy et al., 2018; Kurtz-David et al., 2019; Polisson et al., 2020; Daniel et al., 2022, among others). These tasks are also often used to test for the consistency of subjects' choices with the Generalized Axiom of Revealed Preference (GARP) (among many others Drichoutis & Nayga, 2020). In these tasks, which we will refer to as *budget-choice tasks*, subjects choose a preferred option from an effectively-infinite set of alternatives.

However, "risk elicitation is a risky business" (Friedman et al., 2014) and while budget-choice tasks are becoming more popular than binary-choice procedures, their properties remain largely untested. In this work, I tackle two specific questions in this direction. First, it is unclear whether empirical implementations of portfolio-choice tasks have a larger *predictive* validity than binary-choice tasks. By predictive validity, what is meant here is the ability to actually predict risky choices out-of-sample. Second, even though such methods often aim to reduce complexity compared to binary-choice tasks, they involve large choice sets, and hence it is reasonable to ask to what extent do subjects indeed fully understand the involved procedures.

To answer the first question, this work empirically compares the out-of-sample predictive ability of different tasks (HL, INV, and EG), which are the most commonly

<sup>1</sup> In particular, HL assumes a unique switching point as the decision maker works through the list of choices, which is often violated by a significant amount of participants (Andersen et al., 2006). Further, Beauchamp et al. (2019) showed that list-based methods as HL are susceptible to the compromise effect, which might lead to biased results.

<sup>2</sup> Even simpler is the Qualitative Risk Assessment (QRA) of Dohmen et al. (2011) (see also Beauchamp et al., 2017) and Falk et al. (2018), which asks participants to self-report their willingness to take risks on a scale from 0 to 10.

used risk elicitation procedures in economics.<sup>3</sup> To this end, I conducted an experiment including HL, INV, EG and a separate block of 36 lottery choices, hence providing a clear metric to judge the predictive ability of the two methods out-of-sample. Although these two tasks are clearly different, it is an empirical question which one is better at predicting subsequent choices (and hence eliciting risk attitudes). As HL, INV, and EG are often interchangeably used in the literature, it is further important to understand their properties and have an objective criterion which leads the choice of which method to implement in any given experiment. To answer the second question, regarding subjects' understanding of these tasks, the experiment included three further budget-choice tasks which were constructed in such a way that any risk-averse participant should have selected (the same) corner solutions, and hence other choices are indicative of confusion or lack of understanding. Finally, I explore the potential heterogeneity in behavior in these task by eliciting demographic characteristics as well as participants' numerical literacy and cognitive reflection.

The incentivized experiments relied on a general-population sample ( $N = 403$  and  $N = 400$ ). Results show that the out-of-sample predictive ability of HL and INV is undistinguishable, and that overall performance is rather modest. The performance of EG, under some specifications, was worse than that of the other methods, which is not surprising given that it only allows categorization of decision makers into five risk categories, hence it has mechanically less predictive power. Further, I find limited effect of individuals' characteristics on the predictive power of the risk-elicitation tasks. Strikingly, in the additional budget-choice tasks, a large majority of the subjects failed to report the normatively-predicted corner solutions. Moreover, optimal choices in this context seem to depend on numerical literacy as well as cognitive abilities. These results cast doubt on the suitability of general budget-choice tasks for empirical applications in non-student populations or as a general method to test rationality.

The results in this manuscript go beyond the well-known observation that measurements of risk preferences are unreliable (Friedman et al., 2014) and that they often exhibit a limited correlation with real-world behavior (see Charness et al., 2020, for a recent example). First, and in contrast with the literature, I concentrate on (out-of-sample) predictive ability as a well-defined criterion to evaluate measurement methods. Second, this work is part of the more recent but scarce literature investigating why risk preference measurements are unreliable (Crosetto & Filippin, 2016; Holzmeister & Stefan, 2020). Specifically, the results described here suggest that lack of comprehension might be one of the leading explanations. Last, this paper is also related to a different branch of the literature, namely that which extensively uses budget-choice allocation tasks to test the consistency of subjects' choices with GARP (Choi et al., 2007a; Choi et al., 2014; Kurtz-David et al., 2019; Polissou et al., 2020; Drichoutis & Nayga, 2020; Daniel et al., 2022). The results described here should be

<sup>3</sup> In May of 2020 MPL is cited more than six-thousand times and INV more than a thousand. Among other incentivised elicitation methods there are the ordered lottery choice task (Eckel & Grossman, 2008) with less than a thousand citations, which was included in an additional experiment following the request of a referee, and the Bomb Risk Elicitation Task (Crosetto & Filippin, 2013) with less than five hundred.

seen as a caveat on the lack of robustness of tests built around budget-choice allocations. The widespread lack of understanding in the general population for this type of task suggests that systematic attention and comprehension checks should be implemented to increase the reliability of the data in this field.

The paper is structured as follows. Section 2 discusses the experimental design and procedures. Section 3 presents the results on predictive performance and correlation among measures. Section 4 reports the behavior in budget-choice tasks where the optima are corner solutions. Section 5 concludes.

## 2 Experimental design

The two experiments involved 403 and 400 individuals and used Prolific (Palan & Schitter, 2018), an online platform which allows recruiting from the general population.<sup>4</sup> The heterogeneous composition of our sample is confirmed, e.g., by the distribution of age and employment status. Subjects were on average 33 years old (SD 11.546, minimum 18, maximum 82; for the second experiment average age was 38, SD 12.93, minimum 20, maximum 87). Among participants, 49.95% (56.25%) were fully-employed, 19.92% (23.00%) worked part-time, 11.10% (14.00%) were housekeepers, and 9.07% (3.75%) were unemployed. 68% (60%) of our sample was female.

Subjects were paid based on their answers for one randomly-sampled decision. Average earnings were GBP 5.47 (4.45) including 1.25 for completing the experiment (SD = 6.58, min = 1.25, max = 22.25; for the second experiment SD = 5.90, min = 1.25, max = 23.25).

The experiments were programmed in Qualtrics. The first experiment consisted of five parts in the following order: three budget-choice slider tasks, 36 binary lottery choices, an implementation of HL, and implementation of INV, and a repetition of the three budget-choice slider tasks with increased incentives (5×). At the end of the experiment, a (self-reported, non-incentivised) Qualitative Risk Assessment measure (QRA) (Dohmen et al., 2011; Falk et al., 2018) was implemented to investigate its correlation with HL and INV. The second experiment encompassed the first while adding the EG task as well as the Cognitive Reflection Task (Frederick, 2005) and a measure of numeracy ability (Lipkus et al., 2001).

Discussion of the budget-choice tasks is relegated to Section 4 below, which also describes their implementation. Implementation of the other methods was kept as close as possible to the originals, with payoffs scaled to ensure comparability across tasks and guarantee the expected earnings as prescribed by Prolific. HL was implemented using an ordered list of 10 binary choices, such that subjects should start by choosing the safer option (presented on the left) to then indicate a preference for the right option as they proceed along the list of choices, with the switching point indicating their risk attitudes (see Csermely & Rabas, 2016, for an illustration of

<sup>4</sup> The rationale of the sample size followed a power analysis for detecting a small effect size ( $d = 0.2$ ) according to a Wilcoxon Signed-Ranked (WSR) test comparing the performance of the two incentivized elicitation methods.

the different implementations of this task). INV allowed participants to invest part of their endowment in a lottery that paid 2.5 times the amount invested with a 50% chance and that GBP 0 otherwise, while keeping the part of their budget that was not invested. For QRA, subjects were directly asked to state their willingness to take risks on a 0–10 scale. EG was implemented as a choice among six different lotteries following the standard in the literature (Dave et al., 2010). Further details on the implementation of the tasks are given in Appendix A and instructions are presented in Appendix B.

Four of the 36 lottery choices involved a dominance relation. These choices were implemented as a check of participants' attention and comprehension. The remaining 32 lottery choices were used for assessing the out-of-sample predictive performance of the different methods (see Appendix B for the list of lotteries). To ensure an unbiased selection, the set of lotteries used in this phase was constructed following optimal design theory (Silvey, 1980) in the context of non linear (binary) models (Ford et al., 1992; Atkinson, 1996), see also Moffatt (2015) for a detailed explanation of the procedure.

### 3 Comparison of methods

This section presents the results of the experiment. Subsection 3.1 gives an overview of estimated risk attitudes, subsection 3.2 compares the predictive performance of HL, EG, and INV, subsection 3.3 compares HL, EG, and INV with a structural econometric estimation using the block of 32 binary choices, subsection 3.4 explores the potential role of heterogeneity in influencing the predictive power of the different measures. The results of the two experiments are qualitatively identical and they are presented together.

#### 3.1 Descriptive results

Following influential contributions in the estimation of risk attitudes (e.g., Andersen et al., 2008; Wakker, 2008; Dohmen et al., 2011; Gillen et al., 2019), and in agreement with standard analyses of HL and INV, in this paper I adopt the CRRA specification for all incentivized procedures as defined by:

$$U(x) = \begin{cases} x^{(1-r)}, & \text{if } x \geq 0 \\ \ln x, & \text{if } r = 1. \end{cases}$$

In the Appendix A I perform the analyses reported in the main text assuming a different functional form for the utility functions (CARA instead of CRRA). The results are qualitatively unchanged. According to the assumed utility function, the vast majority of subjects are classified as risk averse, as commonly found in the literature (Gneezy & Potters, 1997; Holt & Laury, 2002; Harrison et al., 2007). In particular, according to HL only 27.30% (exp 2: 14.50%) of subjects are classified as risk seeking, while EG classifies 11.25% of participants in this category. INV

does not distinguish between risk-seeking and risk-neutral subjects, since it does not allow for negative values of the relative risk attitude coefficient.

The average estimated risk attitude using HL is 0.309 (SD 0.605; exp 2: 0.598, SD 0.656), for EG is 1.397 (SD 1.204) while with INV is 6.343 (SD 34.996; exp 2: 2.978, SD 22.36). This very large difference is striking. Examination of the data shows that the discrepancy is due to the fact that, in this sample, almost 34.49% of participants (38.00%) gave “focal-point answers” investing amounts of exactly 0% (with an implied  $r \leq 0$ ), exactly 100% (with an implied  $r \geq 223.1$ ), or exactly 50% (with an implied  $r \simeq 0.65$ ). This observation already suggests that budget-choice tasks might be mechanically biased due to subjects’ lack of comprehension or attention. Excluding the 40 (43) subjects who report corner solutions (0% or 100%), the average estimated risk attitude using INV is 0.840 (SD 1.939; exp 2: 1.444, SD 11.925). Excluding all 139 (152) subjects reporting 0%, 100%, or 50%, the average is 0.913 (SD 2.270; exp 2: 1.796, SD 14.302). In the subsequent analysis, no subjects are excluded, but results are qualitatively unchanged when restricting the sample to those subjects not reporting corner solutions in the INV task.

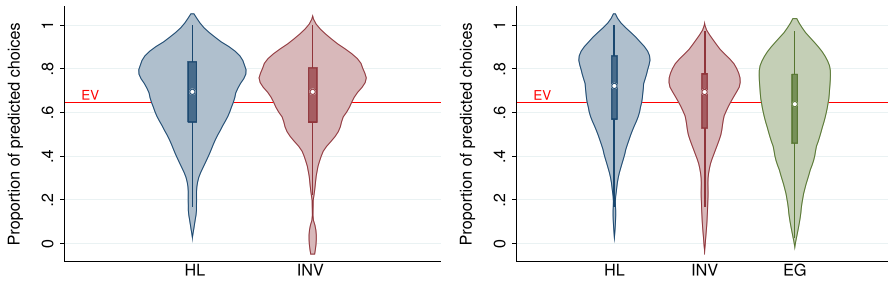
In HL, 31.51% (28.00) of subjects switched from the left to the right option and vice versa multiple times. Following the literature, instead of excluding these subjects, subsequent analyses consider the total number of “safe” (left) choices as an indicator of risk aversion (Holt & Laury, 2002; Holt & Laury, 2005). This yields a comparable sample size for the different measures of risk attitudes. However, the results are not affected if I use “consistent” subjects only (those that switched only once).

The literature has typically found that risk attitudes estimated through different elicitation methods are often uncorrelated (Friedman et al., 2014; Charness et al., 2020).<sup>5</sup> This is also partially true in the datasets at hand: HL and INV are not significantly correlated (Pearson’s  $r = 0.01$ ,  $p = 0.877$ ; exp 2:  $r = -0.05$ ,  $p = 0.283$ ), while EG is correlated with HL but not with INV ( $r = 0.184$ ,  $p = 0.001$ ;  $r = -0.080$ ,  $p = 0.111$ ). However, the three measures are not significantly correlated when restricting to subjects who behaved consistently in the HL task (HL with INV  $r = -0.030$ ,  $p = 0.771$ ; exp 2:  $r = -0.074$ ,  $p = 0.435$ ; HL with EG  $r = 0.145$ ,  $p = 0.125$ ; INV with EG  $r = -0.175$ ,  $p = 0.065$ ;) or to those reporting interior solutions in the INV task (HL with INV  $r = 0.010$   $p = 0.911$ ; exp 2:  $r = -0.068$ ,  $p = 0.406$ ; HL with EG  $r = 0.154$ ,  $p = 0.058$ ; INV with EG  $r = -0.047$ ,  $p = 0.565$ ;) There is a positive, although small, correlation between the self-reported, non-incentivised willingness to take risks (QRA) and HL ( $r = 0.106$ ,  $p = 0.034$ ; exp 2:  $r = 0.231$ ,  $p = 0.001$ ), as well as between QRA and EG ( $r = 0.217$ ,  $p = 0.001$ ).

### 3.2 Predictive performance

Figure 1 presents the out-of-sample predictive performance of INV, HL, and EG in the two experiments. In particular, I report the distribution (violin plot) of the

<sup>5</sup> However, Gillen et al. (2019) show that commonly-used measures of risk attitudes are more correlated than previously thought once measurement errors are accounted for.



**Fig. 1** Distribution of the average proportions of out-of-sample predicted choices between elicitation methods in the two experiments (first in the left, second on the right). Violin plots show the median, the interquartile range, and the 95% confidence intervals as well as rotated kernel density plots on each side. The red-horizontal line indicates the predicted behavior of an expected value maximiser

proportions of correctly predicted choices at the individual level using the individual estimated risk attitudes.<sup>6</sup>

There is no significant difference in the number of correctly predicted choices between INV and HL (INV: 66.51%, HL: 67.18%; Wilcoxon Signed-Ranked (WSR) test,  $N = 403$ ,  $z = 0.485$ ,  $p = 0.628$ ; exp 2: INV: 69.88%, HL: 68.67%; WSR test,  $N = 400$ ,  $z = 1.145$ ,  $p = 0.252$ ), which is modest. However, both measure perform better than EG (60.90%; INV, WSR test,  $N = 400$ ,  $z = 5.863$ ,  $p < 0.001$ ; HL WSR test,  $N = 400$ ,  $z = 5.461$ ,  $p < 0.001$ ).<sup>7</sup> These levels of performance for HL and INV are only slightly above the predictive ability of simply assuming expected-value maximization (i.e., risk neutrality), which is 64.74% (65.07%), while EG does perform statistically significantly worst (for INV: WSR test,  $N = 403$ ,  $z = 2.691$ ,  $p = 0.007$ ; exp 2:  $N = 400$ ,  $z = 3.739$ ,  $p < 0.001$ ; for HL: WSR test,  $N = 403$ ,  $z = 4.451$ ,  $p < 0.001$ ; exp 2:  $N = 403$ ,  $z = 4.932$ ,  $p < 0.001$ ; EG: WSR test,  $N = 400$ ,  $z = -3.319$ ,  $p = 0.001$ ).

At the individual level, INV significantly predicts out-of-sample choices better than HL for 23.57% (32.79%) of individuals.<sup>8</sup> Conversely, HL significantly predicts better than INV for 23.08% (33.53%) of individuals. Therefore, there is no clear difference in the predictive power between these two methods.

Part of the subjects (17.82%, 11.00%) chose a dominated option at least once. The comparison of out-of-sample predictive performance is qualitatively unchanged if those potentially-confused subjects are excluded from the analysis. In particular, there are no significant differences in the percentage of correctly-predicted choices

<sup>6</sup> QRA does not provide a risk parameter, hence it cannot be used to predict choices. However, it can be correlated with choice frequency (Dohmen et al., 2011). In my data, QRA is negatively correlated with the individual proportions of safe choices (Spearman's  $\rho = -0.225$ ,  $N = 403$ ,  $p < 0.0001$ ;  $\rho = -0.225$ ,  $N = 403$ ,  $p < 0.0001$ ).

<sup>7</sup> This is not a surprising result as Dave et al. (2010) already showed that EG has lower accuracy than HL. However, in Appendix A I show that this result is not robust to the use of a different utility function.

<sup>8</sup> The threshold for significance is set at  $p < 0.05$  for a test of proportions, conducted separately for each subject.

(INV 68.97%; HL 69.10%; WSR test,  $N = 332$ ,  $z = 0.757$ ,  $p = 0.449$ ; exp 2: INV 70.73%; HL 69.69%; WSR test,  $N = 356$ ,  $z = 1.008$ ,  $p = 0.313$ ). Moreover, both still perform better than EG (61.42%, WSR test,  $N = 356$ ,  $z = 5.317$ ,  $p < 0.001$ ;  $N = 356$ ,  $z = 5.713$ ,  $p < 0.001$ ). Since not all subjects behaved consistently in the HL task (68.49%; 72.00%), one could argue that the predictive performance of this latter method should be evaluated only on the sub-sample that displayed a unique switching point. However, there are also no differences when focusing on this subset of participants (HL 69.19%; INV 67.75%; WSR test,  $N = 276$ ,  $z = 0.017$ ,  $p = 0.987$ ; exp 2: HL 69.47%; INV 70.59%; WSR test,  $N = 288$ ,  $z = -0.905$ ,  $p = 0.366$ ). Results are also unchanged when restricting the analysis to those participants reporting interior solutions in the INV task (HL 66.75%, INV 67.69%; WSR test  $N = 363$ ,  $z = -1.222$ ,  $p = 0.222$ ; exp 2: HL 68.86%, INV 70.10%; WSR test  $N = 357$ ,  $z = -1.331$ ,  $p = 0.183$ ).

### 3.3 Comparison with structurally-estimated risk attitudes

As an alternative way to compare the predictive performance of the two elicitation methods, I used a maximum likelihood procedure (ML) to estimate each subject's risk attitudes (e.g., Harrison et al., 2005, 2007; Harrison & Rutström, 2008; Harrison et al., 2019) from their decisions in the set of 32 lottery choices. The procedure followed the approach described in Moffatt (2015, Chapter 13) and implemented well-established techniques as used in many recent contributions (Gaudecker et al., 2011; Conte et al., 2011; Moffatt, 2015; Alós-Ferrer & Garagnani, 2022; Alós-Ferrer & Garagnani, 2022; Alós-Ferrer & Garagnani, 2021). I estimated an additive random utility model, which considers a given utility function plus an additive noise component (e.g., Thurstone, 1927; Luce, 1959; McFadden 2001). Specifically, I assumed CRRA utility and normally-distributed errors. To account for individual heterogeneity, I further assumed that the risk parameter is normally distributed over the population and estimated the parameters of this distribution, deriving individual risk attitudes by updating from the so obtained population level prior (e.g., see Harless & Camerer, 1994; Moffatt, 2005; Harrison & Rutström, 2008; Bellemare et al., 2008; Gaudecker et al., 2011; Conte et al., 2011; Moffatt, 2015).

The average estimated risk attitude following this method is 0.418 (SD 0.321; exp 2: 0.397, SD 0.342). I then compared the results to the estimated risk parameters from HL, INV, and EG. There is a positive correlation between HL and ML ( $r = 0.225$ ,  $p < 0.001$ ;  $r = 0.272$ ,  $p < 0.001$ ), and between EG and ML ( $r = 0.166$ ,  $p = 0.001$ ) but there is no significant correlation between INV and ML ( $r = 0.070$ ,  $p = 0.160$ ;  $r = 0.075$ ,  $p = 0.132$ ).

### 3.4 Heterogeneity and predictive performance

This subsection reports an explorative analysis of whether the predictive performance of the different risk-elicitation tasks depends on participants' characteristics.



With this aim in the second experiment I collected several demographic informations as well as answers to the Cognitive Reflection Task (CRT) (Frederick, 2005) and Numeracy Scale (Lipkus et al., 2001).

One possibility is that the predictive performance of the risk-elicitation tasks is poor because the sample is different from the standard (well-educated) student population of laboratory experiments. In the selected sample 2.50% of the participants reported having a doctorate degree while 73.50% had at least an undergraduate degree. However, there are no systematic differences in the predictive performance of the tasks with respect to the highest educational level achieved. In particular, for none of the tasks those with doctoral degrees have more correctly predicted choices than everybody else (INV, 69.91% vs. 68.61%, Wilcoxon rank-sum test (WRS)  $N = 400, z = -0.667, p = 0.504$ ; HL 68.51% vs. 75.00%, WRS,  $N = 400, z = -1.242, p = 0.214$ ; EG 60.86% vs. 62.50%, WRS,  $N = 400, z = -0.410, p = 0.682$ ). There are also no significant differences between the predictive performance of those who have at least an undergraduate degree compared to who do not. There is hence no evidence of an effect of education of the predictive performance of the different risk-elicitation tasks.

Another argument involves the numeracy skills of subjects. Because all the risk-elicitation tasks involve numbers and probabilities, it is reasonable to expect a differential effect of their performance based on the numeracy skill of people, especially for those tasks which have been argued to be more complex, e.g., HL (Dave et al., 2010; Charness et al., 2018). This might be true even in educated individuals as there is evidence that even they often are unable to convert a percentage to a frequency (Lipkus et al., 2001). I measure the participants' numeracy skills with a well-established procedure, the Numeracy Scale by Lipkus et al. (2001), which has been shown to correlate with real-life outcomes such as wealth (Estrada-Mejia et al., 2016). Of the three standard questions no participant answered all three correctly. The average number of correct answers is 1.623 with a median of 2 and 7.50% of subjects answered correctly to no question. Out of the three risk-elicitation tasks, there is a significant difference in the predictive performance only for INV, with participants answering correctly more than one question in the numeracy scale presenting a higher proportion of correctly predicted choices (71.03%) compared to the others (67.24%; WRS,  $N = 400, z = 2.433, p = 0.015$ ). For the other two tasks there are no statistically significant effects (HL 69.69% vs. 66.32%; WRS,  $N = 400, z = 1.276, p = 0.202$ ; EG 62.16% vs. 58.01%; WRS,  $N = 400, z = 1.646, p = 0.1000$ ).

Lastly, I implement the CRT of Frederick (2005) in the version proposed by Alós-Ferrer et al. (2016) in order to avoid recognition effects, as the CRT in its classical form often appears in the popular press. Alós-Ferrer and Hügelschäfer (2016) and Haita-Falah (2017) find that higher test scores in the CRT are correlated with lower incidences of certain economic biases, e.g., the conjunction fallacy, conservatism, and sunk-cost fallacy. Moreover, as argued by Toplak et al. (2011), low CRT scores might indicate a tendency to act on impulse and give an intuitive response. Of the three standard questions 30.75% of participants

answered all of them correctly. The average number of correct answers is 1.615 with a median of 2 and 22.75% of subjects answered correctly to no question. For INV and HL there is a significant difference in the predictive performance, with participants answering correctly more than one question of the CRT presenting a higher proportion of correctly predicted choices compared to the others (INV: 71.99% vs. 64.86%; WRS,  $N = 400$ ,  $z = 3.857$ ,  $p < 0.001$ , HL: 71.15% vs. 68.43%; WRS,  $N = 400$ ,  $z = 2.006$ ,  $p = 0.045$ ). For EG there are no statistically significant effects (61.37% vs. 60.36%; WRS,  $N = 400$ ,  $z = 0.397$ ,  $p = 0.692$ ).

These offer limited insight on potential heterogeneity regarding the predictive performance of different risk-elicitation tasks. Even in a general, but educated, population sample the different tasks seem to not to be difficult to comprehend, as exemplified by the lack of significant differences in their predictive power between differently educated or (numerically) literate groups of individuals. However, there seem to be some indicative results pointing to a relation between impulsivity, as measured by cognitive reflection, and the reliability of most risk-elicitation tasks. Their results should however be taken with a grain of salt and more research on this should be done.

#### 4 Slider budget-choice tasks

In each of the three additional budget choice tasks, participants had to select their preferred lottery from a linear set. The tasks were implemented in the form of sliders as often done in the literature (i.e., Gneezy & Potters, 1997; Kurtz-David et al., 2019; Gillen et al., 2019, among others). In particular, participants were asked to indicate which option they preferred by moving a slider, with values changing in real-time.

The possible values of the sliders were constrained. To avoid losses, all monetary outcomes were positive (larger than or equal to one penny). Further, probabilities belonged to the interval [0.05, 0.95], to avoid confounds due to focal points or heuristics (e.g., the certainty effect). Moreover, in order to show that the results do not depend on the particular probabilities or outcomes chosen by the experimenter, the range of possible values for the second and third slider depended on the subjects' choices in the first slider task. Hence, they potentially assumed different values for each participant. However, the sliders were designed such that, independently of individual differences between subjects, the optimum of the underlying maximization problems was the same corner solution for every (risk-averse) participant.

The set of sliders was presented twice, with different levels of incentives. Specifically, a first version was presented at the beginning of the experiment, and a version with incentives multiplied by five was presented at the end. This aimed to test whether stake sizes influence the stability of risk preferences (Bandyopadhyay et al., 2021).

### 4.1 Design of the sliders

The first slider described the set of lotteries  $\{[p, q; 1 - p, 0] \mid pq = K\}$ . That is, all lotteries in the slider have the same expected value of  $K$ , with  $K = 4$  (GBP) for the first three sliders and  $K = 20$  for the high-incentives version. The slider changed  $p$  and  $q$  simultaneously preserving the expected value. Thus, the underlying maximization problem was

$$\begin{aligned} \max_{p,q} & p \cdot u(q) \\ \text{s.t.} & pq = K, p \in [0.05, 0.95], \end{aligned}$$

or, equivalently,

$$\max_{p \in [0.05, 0.95]} p \cdot u\left(\frac{K}{p}\right).$$

A simple computation (see Appendix C for details) shows that the objective function in the last problem is strictly increasing for any twice-differentiable utility function with  $u''(\cdot) < 0$ . Hence, in normative terms risk-averse participants should report the corner solution  $\hat{p} = 0.95$ .

Let  $p^*$  be the participant’s actual answer to the first slider, and let  $q^* = K/p^*$ . The next two sliders depend on the chosen values  $p^*$  and  $q^*$ .

The second slider described the set of lotteries  $\{[p^*, q + z; 1 - p^*, z] \mid p^*q + z = K\}$ . That is, again all lotteries in the slider have the same expected value  $K$ . The slider moves  $q$  and  $z$  simultaneously preserving the expected value. The maximization problem is

$$\begin{aligned} \max_{q,z} & p^* \cdot u(q + z) + (1 - p^*) \cdot u(z) \\ \text{s.t.} & p^*q + z = K, q \geq 0.01, z \geq 0, \end{aligned}$$

or, equivalently,

$$\max_{q \in [0.01, q^*]} p^* \cdot u(K + (1 - p^*)q) + (1 - p^*) \cdot u(K - p^*q).$$

A direct computation shows that the objective function in this problem is strictly decreasing whenever  $u''(\cdot) < 0$ . Hence, risk-averse participants should report the corner solution  $\hat{q} = 0.01$ .

The third slider described the set of lotteries  $\{[p, q^* + z; 1 - p, z] \mid pq^* + z = K\}$ . As in the previous cases, all lotteries in this slider have the same expected value  $K$ . The slider moves  $p$  and  $z$  simultaneously preserving the expected value. The maximization problem is

$$\begin{aligned} \max_{p,z} & p \cdot u(q^* + z) + (1 - p) \cdot u(z) \\ \text{s.t.} & pq^* + z = K, p \geq 0.05, z \geq 0, \end{aligned}$$

or, equivalently,

$$\max_{p \in [0.05, p^*]} p \cdot u(K + (1 - p)q^*) + (1 - p) \cdot u(K - pq^*).$$

Even replacing  $u(\cdot)$  with a CRRA functional form, the objective function in this problem is not analytically tractable. Numerical results, however, show that the problem has a corner solution at the lower extreme ( $p = 0.05$ ) for all subjects with moderate risk aversion ( $0 < r < 1$ ). Specifically, 247 (exp 2: 249) participants were classified as moderately risk averse according to HL, resulting in four different possible values of  $r$  (but different ranges of  $p$  depending on  $p^*$ ). The numerical solution of the optimization problems using a CRRA utility function with those possible risk parameters has a corner solution at  $p = 0.05$  for all 247 (249) moderately risk-averse participants. An additional 46 (46) subjects were classified with  $r > 1$  according to HL. Of those, 4 (2) had a corner solution at  $p = 0.05$ , 16 (0) had a corner solution at  $p = 0.95$ , and 90 (44) had interior optima. Among the 110 (110) subjects were classified with  $r < 0$  according to HL. Of those, 2 (4) had a corner solution at  $p = 0.05$  and 44 (90) had interior optima.<sup>9</sup> Therefore, all subjects with moderate risk aversion  $0 < r < 1$  had optima at the corner solution  $p = 0.05$  in the third slider in both experiments.

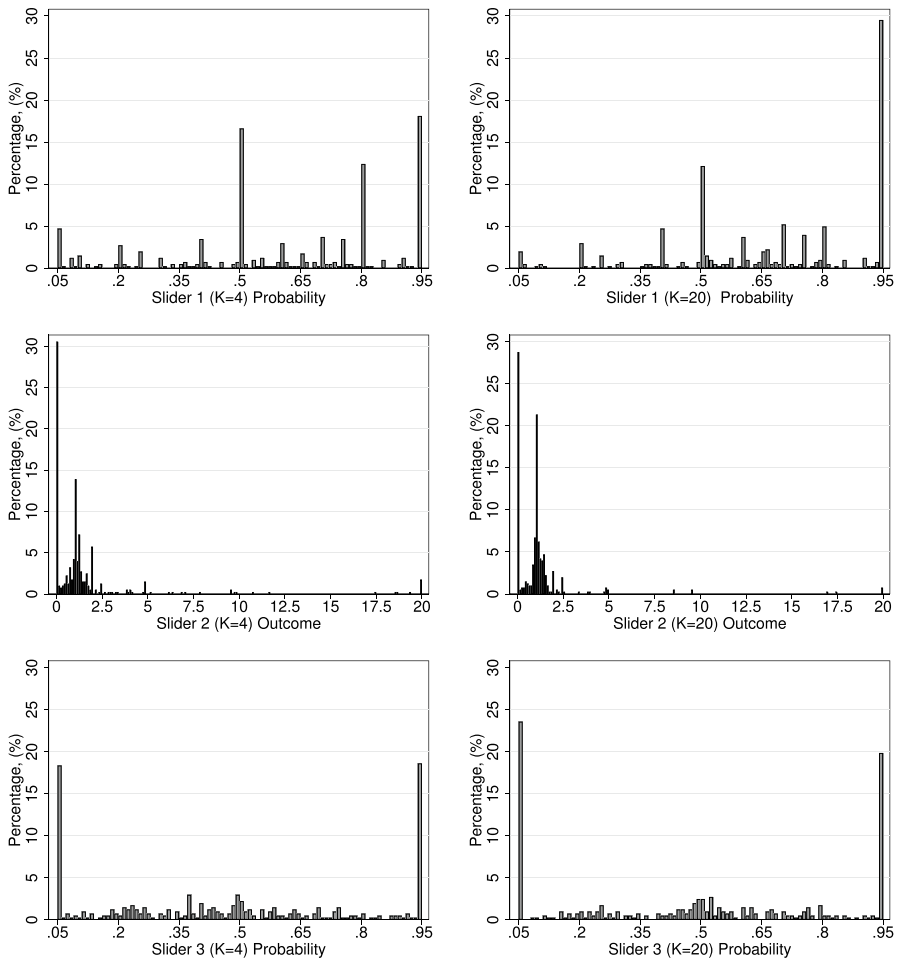
## 4.2 Behavior in budget-choice tasks

To account for imprecision and noise in the use of the interface, and since participants could only select increments of 0.01, I conservatively define a choice to be a corner solution when the chosen value is within the lower (higher) 10% of possible values. For the first slider, this means that an answer was classified as the upper corner solution if it was larger than or equal to  $0.95 - 0.09 = 0.86$ . For the second and third sliders, the range of possible values depended on subjects' previous choice.

Behavior and results relative to the slider tasks are very similar in the two experiments. I report the result of the first experiment in the main text and the second in Appendix E. Figure 2 shows that behavior in the additional budget-choice tasks was far from the normative optima. Choices for the low-incentive ( $K = 4$ ) slider tasks are displayed on the left-hand side panels. Only 4.47% (18) subjects reported all correct corner solutions (of which 13 were risk averse according to HL and 5 had  $r < 0$ ), and only 49.13% (198) reported at least one of the three correct corner solutions (of which 147 were risk averse according to HL).

The extraordinarily high levels of suboptimal behavior in the slider tasks are at odds with what would be expected due to lack of comprehension or attention in other choice tasks. As a comparison, and as reported above, only 17.82% ( $N = 72$ ) of subjects made one or more dominated choices in the binary choice task. This further rules out that the poor performance in the budget-choice tasks was due to general inattention to the experiment.

<sup>9</sup> I also solved the problem numerically for all participants using CRRA with the risk parameter  $r$  derived from the structural estimation ML described in Section 3.3. All 362 subjects with  $0 < r < 1$  according to ML had a corner solution at  $p = 0.05$  in both experiments. Only four (four) were classified as highly risk averse ( $r > 1$ ; Moffatt, 2015), and their optima were interior. The 37 (37) participants classified as risk seeking according to ML had also interior optima.



**Fig. 2** Distribution of answers in the first (upper), second (middle), and third slider (bottom) in the first repetition (left figures) and with higher incentives (right figures). The equivalent pictures for the second experiment are reported in Appendix E

For the first slider, only 80 (19.85%) of the 403 subjects reported the correct (upper) corner solution. Of the 293 participants classified as risk-averse according to HL, only 61 (20.82%) reported that corner solution (recall that INV cannot classify participants as risk-seeking). For the second slider, only 131 (32.51%) of the 403 participants reported the correct (in this case lower) corner solution. This includes only 100 (34.13%) of the 293 participants classified as risk-averse according to HL. In the third slider, the proportion of subjects reporting the correct (upper) corner solution was 20.35% (82 of 403). This includes only 80 (32.39%) of the 247 participants classified as moderately risk-averse ( $0 < r < 1$ ) according to HL.

Needless to say, and given that the sliders should have elicited corner solutions for risk-averse individuals, these numbers are very low, suggesting low levels of understanding in the budget-choice tasks. It is hence natural to ask whether understanding would increase with higher incentives. At the end of the experiment, the sliders were presented again, but with incentives multiplied by 5 ( $K = 20$  instead of  $K = 4$ ), which also changed all involved outcomes. The right-hand side panels in Figure 2 display the choice histograms in these versions of the sliders and illustrate that there is only mixed evidence that more subjects choose the right corner solutions more frequently under increased incentives.

For the first slider, only 129 (32.01%) of the 403 subjects reported the correct (upper) corner solution (99 or 33.79% of the 293 risk-averse ones according to HL). This is a significant increase with respect to the 19.85% under low incentives (test of proportion  $N = 403$ ,  $z = 3.938$ ,  $p < 0.001$ ). For the second slider, only 120 (29.78%) of the 403 participants reported the correct (lower) corner solution (92 or 31.40% of the 293 risk-averse ones according to HL). This is not significantly different from the 32.51% under low incentives (test of proportions  $N = 403$ ,  $z = 0.837$ ,  $p = 0.403$ ). In the third slider, only 97 (24.07%) of the 403 subjects reported the correct (upper) corner solution (59 or 23.89% of the 247 participants classified as moderately risk-averse according to HL). Again, this is not significantly different from the 20.35% under low incentives (test of proportions  $N = 403$ ,  $z = -1.271$ ,  $p = 0.204$ ).

### 4.3 Heterogeneity in the budget-choice tasks

This subsection reports an explorative analysis of whether behavior in the budget-choice tasks depends on participants' characteristics such as their literacy, education, and cognitive reflection.

A similar argument which was made above regarding a potential link between the highest level of educational obtained by the participants and the predictive power of the risk-elicitation tasks can be made also for the behavior in budget-choice tasks. The intuition is that, higher optimal behavior could be relative to more educated subjects as they might have understood the tasks better. However, there are no systematic differences in this respect. Out of the nine sliders, only for one, the second slider for the high incentive, reaches a statistically significant difference in the proportion of optimal behavior between subjects with at least an undergraduate degree compared to all the others (26.42% vs. 36.05%, test of proportion  $z = -1.803$ ,  $p = 0.036$ ).<sup>10</sup>

Compared to education numerical literacy, as measured by the numeracy scale, is able to capture systematic heterogeneity in subjects' behavior in budget-choice tasks. For 5 out of the 6 sliders there are significant differences, with higher numerical literacy corresponding to more optimal behavior in the budget choice tasks (slider

<sup>10</sup> The result is not robust to multiple-tests correction. Similar (non-significant) results are obtained using a division based on those people who got a doctorate compared to everyone else.

1: 31.54% vs. 24.79%, test of proportion,  $z = 1.359, p = 0.087$ ; slider 2: 34.05% vs. 19.83%,  $z = 2.857, p = 0.002$ ; slider 3: 24.37% vs. 14.05%,  $z = 2.318, p = 0.010$ ; slider 4: 40.86% vs. 24.79%,  $z = 3.075, p = 0.001$ ; slider 5: 40.86% vs. 16.53%,  $z = 4.736, p < 0.001$ ; slider 6: 28.67% vs. 19.83%,  $z = 1.851, p = 0.032$ ). This result indicate that there is indeed a relation between numeracy and optimal behavior when using this type of tasks.

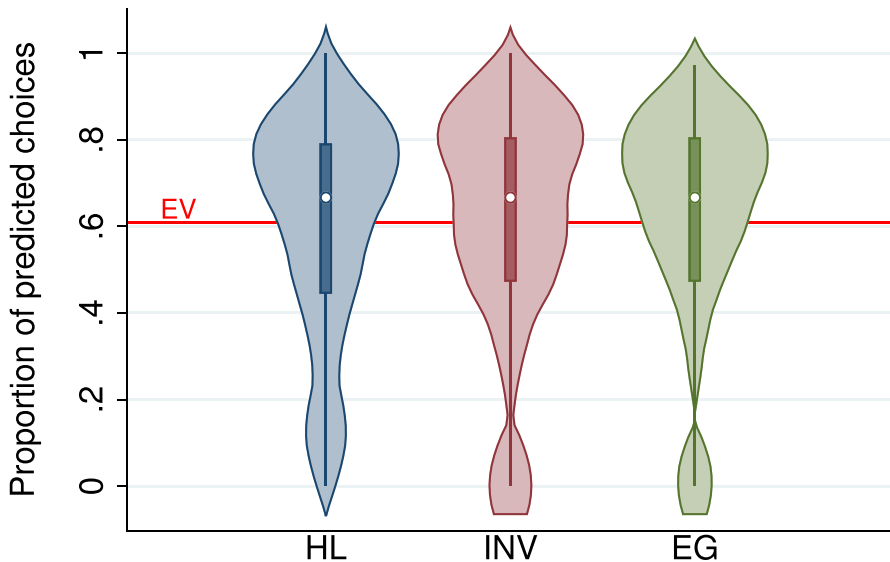
Lastly, I investigate the relation between cognitive reflection as measured by the CRT and the proportion of optimal choices in the budget-choice tasks. The results of the cognitive reflection resemble those of the numeracy skills with 4 out of the 6 sliders presenting significant differences. In particular, higher cognitive reflection corresponds to more optimal behavior in the budget choice tasks (slider 1: 37.85% vs. 19.89%, test of proportion,  $z = 3.928, p < 0.001$ ; slider 2: 31.78% vs. 27.42%,  $z = 0.951, p = 0.171$ ; slider 3: 22.90% vs. 19.35%,  $z = 0.864, p = 0.194$ ; slider 4: 41.59% vs. 29.57%,  $z = 2.498, p = 0.006$ ; slider 5: 42.52% vs. 23.12%,  $z = 4.101, p < 0.001$ ; slider 6: 32.24% vs. 18.82%,  $z = 3.053, p = 0.001$ ).

The results of the exploration of heterogeneity in optimal behavior in budget-choice tasks suggest that even in a general and educated population sample having higher numeracy skills and cognitive reflection correlates positively with optimal behavior in this type of task. This suggests again that budget-choice tasks are hard to comprehend for the general population.

## 5 Discussion and conclusion

In two experiments with a general population sample, I evaluate the predictive validity of three of the most common risk-attitude elicitation procedures, the choice-list procedure of Holt and Laury (2002), the investment task of Gneezy and Potters (1997), and the multi-alternative procedure of Eckel and Grossman (2002). Two of these tasks are undistinguishable in their ability to predict out-of-sample choices, and their performance is moderate at best. That is, it is only slightly better than that obtained by ignoring all individual information and predicting on the basis of expected value maximisation only. However, the procedure introduced by Eckel and Grossman (2002) is outperformed by the other two and by a simple expected value maximization.

The experiment also included budget-choice tasks where participants selected their preferred lotteries out of linear budget sets, and which were such that risk averse participants should have selected corner solutions. On the contrary, the vast majority of participants failed to do so. This strongly suggests that decision makers in non-student samples might have a limited understanding of budget-choice tasks, and hence data collected using such tasks might not adequately reflect attitudes toward risk in the general population. This is a relevant observation, since such methods are widespread.



**Fig. 3** Distribution of the average proportions of out-of-sample predicted choices between elicitation methods assuming CARA utility function. Violin plots show the median, the interquartile range, and the 95% confidence intervals as well as rotated kernel density plots on each side. Red-horizontal line indicate the predicted behavior of an expected value maximiser

Finally, I investigate potential heterogeneous factors influencing comprehension and performance in these tasks. While education, numerical literacy, and cognitive reflection have limited predictive power for the risk-elicitation tasks, they confirm that budget-choice tasks are hard to understand for the general population. In particular, those participants with higher cognitive skills and numerical literacy perform significantly better in this type of task.

The results of this paper are in alignment with the limited external validity of measures of risk attitude found in other studies (Dohmen et al., 2011; Charness et al., 2020). In particular, in light of these results, it should not be expected that these laboratory measures exhibit high levels of correlation with behavior in the field. In particular, measures based on budget-choice or portfolio-choice tasks should not be expected to deliver robust findings.

Furthermore, the results also have implications for other contexts where tasks similar to budget-choice allocations are implemented. For example, these methods are often used to test the consistency of subjects' choices with GARP (i.e., Choi et al., 2007a, 2014; Kurtz-David et al., 2019; Polisson et al., 2020; Drichoutis & Nayga, 2020). However, if a majority of decision makers have a limited understanding of these tasks, the test cannot be expected to be reliable. Therefore, the results speak in favor of implementing systematic understanding checks to increase the reliability of the data and the robustness of the results



when using budget allocation tasks. One possibility is including additional tasks designed in such a way that any risk-averse participant should give the same answer, as in the slider tasks reported in this work.

## Appendix

### A. Robustness analysis

In this section I report the result of the analyses presented in the main text assuming different utility function (constant absolute risk aversion instead of CRRA).<sup>11</sup> Specifically, I assume a constant absolute risk aversion (CARA) function,

$$u(x | r) = \begin{cases} \frac{1-e^{-rx}}{r}, & \text{if } r \neq 0 \\ \frac{x}{x_{\max}}, & \text{if } r = 0, \end{cases} \quad (1)$$

where  $x_{\max} = \max\{x_1, \dots, x_T\}$  is the maximum outcome across all  $T$  trials.

As Figure 3 shows, the results under these different assumptions are qualitatively unchanged from the main text. In particular, there is no significant difference in the number of correctly predicted choices between INV and HL (INV: 59.95%, HL: 58.33%; WSR test,  $N = 403$ ,  $z = 0.132$ ,  $p = 0.895$ ; exp 2: INV: 59.51%, HL: 59.63%; WSR test,  $N = 400$ ,  $z = 1.259$ ,  $p = 0.208$ ), which is still modest. However, both measures do not perform significantly different from EG (59.44%; INV, WSR test,  $N = 400$ ,  $z = 1.874$ ,  $p = 0.061$ ; HL WSR test,  $N = 400$ ,  $z = 0.495$ ,  $p = 0.621$ ).

### B. Risk elicitation methods: Implementation

The investment task (INV) was implemented as closely to the original as possible (Gneezy & Potters, 1997; see also Charness & Gneezy, 2010; Charness et al., 2013). Wording was adapted from Gillen et al. (2019), but I rescaled payoffs to match the intended expected earnings of the experiment. Subjects received an endowment of GBP 4. They were offered to invest in a lottery that paid 2.5 the amount invested with a 50% chance and GBP 0 otherwise. For practical reasons, investment had to be expressed in multiples of 0.01, i.e., no fractions of pennies were allowed. The fraction not invested was kept. Formally, subjects chose an investment  $k \in [0, 4]$  with  $(100 \cdot k) \in \mathbb{N}$  and were paid according to the lottery  $[0.5, 4 - k; 0.5, 4 + 2.5 \cdot k]$ . The expected earnings were thus increasing with the investment. Risk-neutral and risk-seeking subjects should invest their whole endowment, and investment should decrease as risk aversion

<sup>11</sup> The analysis of the behavior in the budget-choice tasks does not depend on the assumed utility function. Moreover, both analyses do not depend on the assumed shape of the noise, e.g., random utility vs. random parameter model. The latter only influences the structurally-estimated risk attitudes.

**Table 1** HL list of lotteries

| A           |           |           | B           |           |           |
|-------------|-----------|-----------|-------------|-----------|-----------|
| Probability | Outcome 1 | Outcome 2 | Probability | Outcome 1 | Outcome 2 |
| 0.1         | 4.00      | 3.20      | 0.1         | 7.70      | 0.20      |
| 0.2         | 4.00      | 3.20      | 0.2         | 7.70      | 0.20      |
| 0.3         | 4.00      | 3.20      | 0.3         | 7.70      | 0.20      |
| 0.4         | 4.00      | 3.20      | 0.4         | 7.70      | 0.20      |
| 0.5         | 4.00      | 3.20      | 0.5         | 7.70      | 0.20      |
| 0.6         | 4.00      | 3.20      | 0.6         | 7.70      | 0.20      |
| 0.7         | 4.00      | 3.20      | 0.7         | 7.70      | 0.20      |
| 0.8         | 4.00      | 3.20      | 0.8         | 7.70      | 0.20      |
| 0.9         | 4.00      | 3.20      | 0.9         | 7.70      | 0.20      |
| 1           | 4.00      | 3.20      | 1           | 7.70      | 0.20      |

increases. In agreement with the literature (Charness et al., 2020), to increase accuracy, for estimation purposes actual decisions were translated into the interval formed by the two closest 5-penny multiples, and the estimated risk attitude was the one that would make a subject indifferent between those two.

For the ordered binary-choice task (Holt & Laury, HL) wording and structure of the lotteries were kept as close as possible to the original (Holt & Laury, 2002). Table 1 presents the list of lotteries. I rescaled payoffs to match the intended expected earnings of the experiment. Ten ordered choices between two lotteries denoted A or B were presented to subjects. Lottery A always paid either GBP 4 or GBP 3.2, while Lottery B paid GBP 7.7 or GBP 0.2. The list is designed so that subjects should switch from choosing A to B according to their risk attitudes, with (at most) one crossing from choosing A to choosing B, and with the last choice involving a dominance relation.

EG was implemented as a choice among six different lotteries closely following the standard in the literature (Dave et al., 2010). I rescaled payoffs to match the intended expected earnings of the experiment. The six lotteries have all the same (50%, 50%) probabilities and the following outcomes (2.8, 2.8), (2.4, 3.6), (2.0, 4.4), (1.6, 5.2), (1.2, 6.0), (0.2, 7.0).

In the QRA, subjects are directly asked how willing they are to take risks. Wording was adapted from the English version of Gillen et al. (2019), who followed the original implementation of Dohmen et al. (2011). Subjects ranked their willingness to take risks on a 0 (lowest) to 10 (highest) scale. In contrast to the other procedures, this mechanism is not incentivized and is based on self-reported rather than revealed preferences. It is thus impossible to estimate risk-attitude parameters based on this question.

### C. Lotteries used for the out-of-sample predictions

See Table 2.

**Table 2** List of lotteries for the out-of-sample predictions

| Trial | Probability 1 | Outcome 1 | Probability 2 | Outcome 2 | Dominated |
|-------|---------------|-----------|---------------|-----------|-----------|
| 1     | 0.6           | 6         | 0.35          | 11        | 0         |
| 2     | 0.52          | 8         | 0.58          | 10        | 1         |
| 3     | 0.6           | 5         | 0.3           | 22        | 0         |
| 4     | 0.15          | 18        | 0.65          | 3         | 0         |
| 5     | 0.8           | 5         | 0.75          | 15        | 0         |
| 6     | 0.2           | 22        | 0.8           | 5         | 0         |
| 7     | 0.7           | 4         | 0.1           | 16        | 0         |
| 8     | 0.55          | 6         | 0.6           | 4         | 0         |
| 9     | 0.6           | 3         | 0.5           | 13        | 0         |
| 10    | 0.8           | 3         | 0.4           | 17        | 0         |
| 11    | 0.5           | 20        | 0.7           | 5         | 0         |
| 12    | 0.7           | 4         | 0.35          | 17        | 0         |
| 13    | 0.4           | 14        | 0.8           | 3         | 0         |
| 14    | 0.7           | 11        | 0.8           | 6         | 0         |
| 15    | 0.65          | 6         | 0.4           | 14        | 0         |
| 16    | 0.4           | 15        | 0.75          | 6         | 0         |
| 17    | 0.5           | 13        | 0.6           | 8         | 0         |
| 18    | 0.7           | 7         | 0.5           | 11        | 0         |
| 19    | 0.42          | 13        | 0.36          | 13        | 1         |
| 20    | 0.2           | 15        | 0.55          | 4         | 0         |
| 21    | 0.55          | 5         | 0.35          | 18        | 0         |
| 22    | 0.75          | 6         | 0.25          | 17        | 0         |
| 23    | 0.85          | 5         | 0.7           | 18        | 0         |
| 24    | 0.55          | 4         | 0.4           | 15        | 0         |
| 25    | 0.55          | 4         | 0.45          | 21        | 0         |
| 26    | 0.6           | 8         | 0.35          | 20        | 0         |
| 27    | 0.7           | 7         | 0.65          | 2         | 1         |
| 28    | 0.75          | 7         | 0.65          | 17        | 0         |
| 29    | 0.65          | 7         | 0.5           | 15        | 0         |
| 30    | 0.4           | 12        | 0.7           | 6         | 0         |
| 31    | 0.3           | 15        | 0.75          | 6         | 0         |
| 32    | 0.75          | 4         | 0.35          | 12        | 0         |
| 33    | 0.4           | 9         | 0.4           | 11        | 1         |
| 34    | 0.7           | 4         | 0.6           | 14        | 0         |
| 35    | 0.6           | 20        | 0.7           | 7         | 0         |
| 36    | 0.05          | 12        | 0.8           | 3         | 0         |

### D. Slider budget-choice tasks

The first slider is equivalent to the one-variable problem

$$\max_{p \in [0.05, 0.95]} p \cdot u\left(\frac{K}{p}\right).$$

The derivative of the objective function is

$$u\left(\frac{K}{p}\right) + p \cdot \left(-\frac{K}{p^2}\right) u'\left(\frac{K}{p}\right) = u\left(\frac{K}{p}\right) - \frac{K}{p} u'\left(\frac{K}{p}\right).$$

Consider a Taylor expansion of  $u$  around  $\frac{K}{p}$  and evaluate it at  $x = 0$ ,

$$0 = u(0) = u\left(\frac{K}{p}\right) - \frac{K}{p} u'\left(\frac{K}{p}\right) + \frac{1}{2} \left(\frac{K}{p}\right)^2 u''(\xi),$$

for some  $\xi \in \left[0, \frac{K}{p}\right]$ . Since  $u''(\cdot) < 0$ , it follows that

$$u\left(\frac{K}{p}\right) - \frac{K}{p} u'\left(\frac{K}{p}\right) > 0,$$

that is, the objective function of the maximization problem above is strictly increasing. Thus the solution to the problem is always the upper corner solution, in this case  $p = 0.95$ .

Let  $p^*$  be the participant’s answer to the first slider. The second slider is equivalent to the one-variable problem

$$\max_{q \in [0.01, K/p^*]} p^* \cdot u(K + (1 - p^*)q) + (1 - p^*) \cdot u(K - p^*q).$$

The derivative of the objective function is

$$p^* (1 - p^*) [u'(K + (1 - p^*)q) - u'(K - p^*q)].$$

Since  $u''(\cdot) < 0$ ,  $u'$  is strictly decreasing, hence  $u'(K + (1 - p^*)q) < u'(K - p^*q)$ . Thus the expression above is strictly negative, implying that the objective function is strictly decreasing. Hence, the solution to the problem is always the lower corner solution, in this case  $q = 0.01$ .

Let  $q^*$  be the participant’s outcome answer to the first slider, i.e.,  $q^* = K/p^*$ . The third slider is equivalent to the one-variable problem

$$\max_{p \in [0.05, p^*]} p \cdot u(K + (1 - p)q^*) + (1 - p) \cdot u(K - pq^*).$$

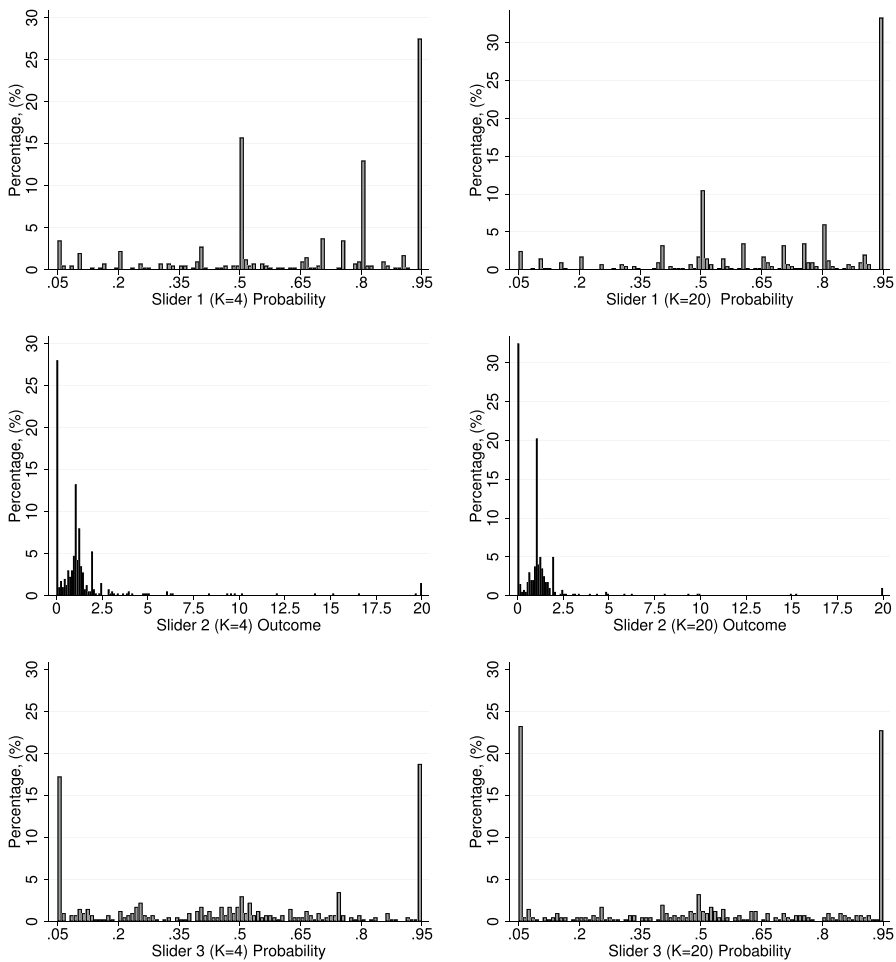
Assume a CRRA ( $u(x) = x^{1-r}$ ) and for the sake of notation let  $u(x) = x^\alpha$  where  $\alpha = 1 - r$ . Substituting, the problem is

$$\max_{p \in [0.05, p^*]} p \cdot (K + (1 - p)q^*)^\alpha + (1 - p) \cdot (K - pq^*)^\alpha.$$

This expression was used to numerically calculate the optimum for each subject.

### E. Behavior in the slider task for the second experiment

Figure 4 is the equivalent of Figure 2 in the main text and it shows that behavior in the additional budget-choice tasks was far from the normative optima also for the second experiment. Choices for the low-incentive ( $K = 4$ ) slider tasks are displayed on the left-hand side panels. Only 5.50% (22) subjects reported all correct corner solutions (3 them had  $r < 0$  according to HL), and only 46.25% (185) reported at least one of the three correct corner solutions.



**Fig. 4** Second experiment: Distribution of answers in the first (upper), second (middle), and third slider (bottom) in the first repetition (left figures) and with higher incentives (right figures)

For the first slider, only 118 (29.50%) of the 400 subjects reported the correct (upper) corner solution. Of the 307 participants classified as risk-averse according to HL, only 98 (31.92%) reported that corner solution. For the second slider, only 119 (29.75%) of the 400 participants reported the correct (in this case lower) corner solution. This includes only 101 (32.90%) of the 307 participants classified as risk-averse according to HL. In the third slider, the proportion of subjects reporting the correct (upper) corner solution was 21.25% (85 of 400). This includes only 54 (21.69%) of the 249 participants classified as moderately risk-averse ( $0 < r < 1$ ) according to HL.

At the end of the experiment, the sliders were presented again, but with incentives multiplied by 5 ( $K = 20$  instead of  $K = 4$ ), which also changed all involved outcomes. The right-hand side panels in Figure 4 display the choice histograms in these versions of the sliders and illustrate that there is only mixed evidence that more subjects choose the right corner solutions more frequently under increased incentives also in the second experiment.

For the first slider, only 144 (36.00%) of the 400 subjects reported the correct (upper) corner solution. This is a significant increase with respect to the 29.50% under low incentives (test of proportion  $N = 400$ ,  $z = 1.959$ ,  $p = 0.025$ ). For the second slider, only 134 (33.50%) of the 400 participants reported the correct (lower) corner solution. This is not significantly different from the 29.75% under low incentives (test of proportions  $N = 400$ ,  $z = 1.1400$ ,  $p = 0.127$ ). In the third slider, only 104 (26.00%) of the 400 subjects reported the correct (upper) corner solution. Again, this is not significantly different from the 21.25% under low incentives (test of proportions  $N = 403$ ,  $z = -1.581$ ,  $p = 0.057$ ).

## F. Experimental instructions

These are the instructions for each part of the experiment, which were presented separately on screen in Prolific. Text in brackets [...] was not displayed to subjects. In all questions an answer was required before participants were able to proceed. A reminder to provide an answer was prompted in case participants had not stated a choice when clicking to proceed with the experiment.

### [General instructions]

This study investigates risky decision-making in five parts and a questionnaire. On top of your fixed earnings of 1.25 GBP, you will earn a bonus payment which will depend on your decisions in the study. Please read all questions carefully. Answer honestly and take care to avoid mistakes. Completing the survey will take about 15 minutes.

### [Explanation lotteries and attention check]

Your bonus payment today depends on the decisions you are about to make and chance. This is because all decisions in the study involve choices between lotteries. A lottery pays one of two potential monetary outcomes each occurring with a given probability.

Here is an example of a lottery:

With 20% probability you get 2 GBP, with 80% probability you get 1 GBP.

This lottery pays 2 GBP with 20% probability or 1 GBP with 80% probability.

After the study the computer will randomly select one among all decisions, and check which lottery you chose. This lottery will be played out and you will be paid according to the resulting outcome.

Each decision could be the one that counts for your bonus. It is therefore in your best interest to consider all your answers carefully.

Before you proceed, please answer the sports test. The test is simple, when asked for your favorite sport you must enter the word *clear* in the text box below.

Based on the text you read above, what favorite sport have you been asked to enter in the text box below?

*[Subjects needed to enter the word “clear” in order to proceed. Fully capitalized, non-capitalized, or capitalized version on the word were accepted. If subjects failed the attention check, the experiment ended.]*

### [Explanation of the three budget-choice tasks]

Part 1:

In this part of the experiment you will be asked to answer 3 questions. Your task is to select your most preferred alternative using a slider. Feel free to explore the possibilities in order to be sure you choose your most preferred alternative.

### [Budget-choice task]

Select the option you prefer by moving the slider (Fig. 5).

*[The three sliders used the same graphical representation.]*

Select the option you prefer by moving the slider.

With 38% probability you get 10.53 GBP, otherwise nothing.

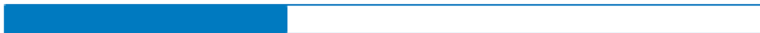
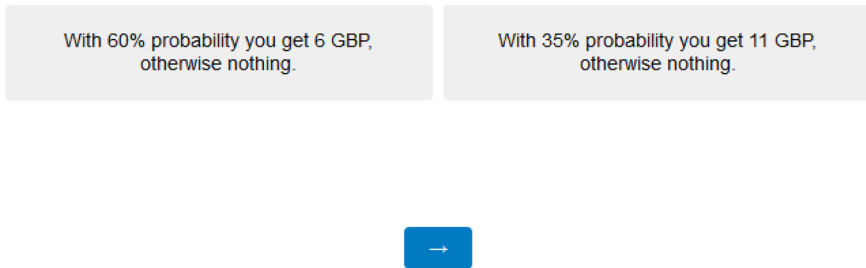


Fig. 5 Example of a slider task

### Question 1 of 36.

Choose one of the two options.



**Fig. 6** Example of a binary choice trial

### [Explanation of the lottery choices]

Part 2:

In this part you will be asked to answer 36 simple questions. Your task is to choose one of the two options.

In this part of the study the second outcome of each lottery is always 0 (zero).

Here is an example of a lottery for this part of the experiment:

With 20% probability you get 2 GBP, otherwise nothing.

This lottery pays 2 GBP with 20% probability or 0 GBP with 80% probability.

### [Lottery choices]

*[The trial number was visible. Participants needed to click over a lottery to choose it and then confirm their preference. The confirmation button was positioned between the two lotteries to avoid biasing answers based on proximity (Fig. 6)]*

### [Multiple price list]

Part 3:

In this part you will be asked to answer 10 simple questions. Your task is to choose one of the two options, A (on the left) or B (on the right).

Here is an example of a lottery for this part of the experiment:

(20% of 2.00 GBP, 80% of 1 GBP) This lottery pays 2 GBP with 20% probability or 1 GBP with 80% probability.

Please choose between Option A and Option B in each line.

*[The 10 lotteries were presented in a ordered-sequential format. See Table 1 for the list of lotteries for this part of the experiment, which follows the actual presentation. Two radio buttons allowed participants to choose between options A and B on*



*each line. Consistency was not enforced, that is, participants could switch back and forth between options A and B. A choice on each line was required and participants were reminded to make a choice on each line in case any was missing.]*

### [Investment task]

Part 4:

In this part you are endowed with 400 Pennies. Your task is to decide which portion of this amount (between 0 and 400 Pennies) you wish to invest in a risky option. The amount of money that you decide not to invest is yours to keep.

The risky option has the following characteristics:

There is a 50% probability that the investment will fail and a 50% probability that it will succeed.

If the investment fails you lose the amount you invested.

If the investment succeeds you receive 2.5 (two and one-half) times the amount invested.

*[A slider with a range between 0 and 400 and precision of 1 unit (one penny) was implemented. The value of the chosen investment was displayed in real time. Subjects needed to confirm their choice in order to proceed.]*

### [EG task]

*Please select from among six different gambles the one gamble you would like to play. The six different gambles are listed below.*

*[Gambles were displayed as a list where only one option could be selected. Participants then needed to confirm their choices].*

- *You must select one and only one of these gambles.*
- *Each gamble has two possible outcomes (ROLL LOW or ROLL HIGH) with the indicated probabilities of occurring. If this choice is randomly selected to determine your compensation, it will be determined by:
 
  - *which of the six gambles you select; and*
  - *which of the two possible payoffs occur.**

*For example, if you select Gamble 4 and ROLL HIGH occurs, you will be paid GBP 5.2. If ROLL LOW occurs, you will be paid GBP 1.6. For every gamble, each ROLL has a 50% chance of occurring.*

### [High-incentive version of the budget-choice tasks]

*[The three budget-choice tasks were implemented in the same way as above, but with incentives increased by a factor of 5.]*

### [Qualitative risk assessment]

Please answer the following question.

How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid risks?

Please indicate an option on the scale, where the value 1 means: not willing to take risks and value 10 means: very willing to take risks.

[10 radio buttons arranged horizontally were implemented for this question. Labels were provided for the lowest outcome (1) “Not willing to take risks” and highest 10 “very willing to take risks.”]

**Acknowledgements** The author thanks Carlos Alós-Ferrer, Björn Bartling, Antonio Filippin, Michel Maréchal, and Roberto Weber for helpful comments and suggestions.

**Funding** Open access funding provided by University of Zurich.

**Data availability** The code and data of the experiment will be freely accessible upon publication. Instructions are enclosed in the Appendix.

## Declarations

**Conflicts of interest** The author declares no conflicts of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Ahn, D., Choi, S., Gale, D., & Kariv, S. (2014). Estimating ambiguity aversion in a portfolio choice experiment. *Quantitative Economics*, *5*, 195–223.
- Alós-Ferrer, C., Fehr, E., & Garagnani, M. (2022). Identifying Nontransitive Preferences. Working Paper, University of Zurich.
- Alós-Ferrer, C., & Garagnani, M. (2021). Choice consistency and strength of preference. *Economics Letters*, *198*, 109672.
- Alós-Ferrer, C., & Garagnani, M. (2022). Strength of preference and decisions under risk. *Journal of Risk and Uncertainty*, *64*, 309–329.
- Alós-Ferrer, C., Garagnani, M., & Hügelschäfer, S. (2016). Cognitive reflection, decision biases, and response times. *Frontiers in Psychology*, *7*, 1–21.
- Alós-Ferrer, C., & Hügelschäfer, S. (2016). Faith in intuition and cognitive reflection. *Journal of Behavioral and Experimental Economics*, *64*, 61–70.
- Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. (2008). Eliciting risk and time preferences. *Econometrica*, *76*, 583–618.
- Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. (2006). Elicitation using multiple price list formats. *Experimental Economics*, *9*, 383–405.
- Atkinson, A. C. (1996). The usefulness of optimum experimental designs. *Journal of the Royal Statistical Society*, *51*, 59–76.
- Bandyopadhyay, A., Begum, L., & Grossman, P. J. (2021). Gender differences in the stability of risk attitudes. *Journal of Risk and Uncertainty*, *63*, 169–201.

- Beauchamp, J. P., Daniel, J. B., David, I. L., & Chabris, C. F. (2019). Measuring and controlling for the compromise effect when estimating risk preference parameters. *Experimental Economics*, 23, 1–31.
- Beauchamp, J. P., Cesarini, D., & Johannesson, M. (2017). The psychometric and empirical properties of measures of risk preferences. *Journal of Risk and Uncertainty*, 54, 203–237.
- Bellemare, C., Kröger, S., & van Soest, A. (2008). Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. *Econometrica*, 76, 815–839.
- Binswanger, H. P. (1980). Attitudes toward risk: Experimental measurement in Rural India. *American Journal of Agricultural Economics*, 62, 395–407.
- Castillo, M., Dickinson, D. L., & Petrie, R. (2017). Sleepiness, choice consistency, and risk preferences. *Theory and Decision*, 82, 41–73.
- Charness, G., Eckel, C., Gneezy, U., & Kajackaite, A. (2018). Complexity in risk elicitation may affect the conclusions: A demonstration using gender differences. *Journal of Risk and Uncertainty*, 56, 1–17.
- Charness, G., Garcia, T., Offerman, T., & Villeval, M. C. (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, 99–123.
- Charness, G., & Gneezy, U. (2010). Portfolio choice and risk attitudes: An experiment. *Economic Inquiry*, 48, 133–146.
- Charness, G., Gneezy, U., & Imas, A. (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior and Organization*, 87, 43–51.
- Choi, S., Fisman, R., Gale, D., & Kariv, S. (2007). Consistency and heterogeneity of individual behavior under uncertainty. *American Economic Review*, 97, 1921–1938.
- Choi, S., Fisman, R., Gale, D., & Kariv, S. (2007). Revealing preferences graphically: An old method gets a new tool kit. *American Economic Review*, 97, 153–158.
- Choi, S., Kariv, S., Müller, W., & Silverman, D. (2014). Who is (more) rational? *American Economic Review*, 104, 1518–1550.
- Conte, A., Hey, J. D., & Moffatt, P. G. (2011). Mixture models of choice under risk. *Journal of Econometrics*, 162, 79–88.
- Crosetto, P., & Filippin, A. (2013). The “bomb” risk elicitation task. *Journal of Risk and Uncertainty*, 47, 31–65.
- Crosetto, P., & Filippin, A. (2016). A theoretical and experimental appraisal of four risk elicitation methods. *Experimental Economics*, 19, 613–641.
- Csermely, T., & Rabas, A. (2016). How to reveal people’s preferences: Comparing time consistency and predictive power of multiple price list risk elicitation methods. *Journal of Risk and Uncertainty*, 53, 107–136.
- Daniel, F., Habib, S., James, D., & Crockett, S. (2022). Varieties of risk preference elicitation. *Games and Economic Behavior*, 133, 58–76.
- Dave, C., Eckel, C. C., Johnson, C. A., & Rojas, C. (2010). Eliciting risk preferences: When is simple better? *Journal of Risk and Uncertainty*, 41, 219–243.
- Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9, 522–550.
- Drichoutis, A. C., & Nayga, R. (2020). Economic rationality under cognitive load. *Economic Journal*, 130, 2382–2409.
- Eckel, C. C., & Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, 23, 281–295.
- Eckel, C. C., & Grossman, P. J. (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization*, 68, 1–17.
- Estrada-Mejia, C., de Vries, M., & Zeelenberg, M. (2016). Numeracy and wealth. *Journal of Economic Psychology*, 54, 53–63.
- Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. *Quarterly Journal of Economics*, 133, 1645–1692.
- Ford, I., Torsney, B., & Wu, C. J. (1992). The use of a canonical form in the construction of locally optimal designs for non-linear problems. *Journal of the Royal Statistical Society*, 54, 569–583.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19, 25–42.
- Friedman, D., Isaac, R. M., James, D., & Sunder, S. (2014). *Risky Curves: On the Empirical Failure of Expected Utility* (1st ed.). London, UK: Routledge.

- Gillen, B., Snowberg, E., & Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the Caltech cohort study. *Journal of Political Economy*, *127*, 1826–1863.
- Gneezy, U., & Potters, Jan. (1997). An experiment on risk taking and evaluation periods. *Quarterly Journal of Economics*, *112*, 631–645.
- Haita-Falah, C. (2017). Sunk-cost fallacy and cognitive ability in individual decision-making. *Journal of Economic Psychology*, *58*, 44–59.
- Halevy, Y., Persitz, D., & Zrill, L. (2018). Parametric recoverability of preferences. *Journal of Political Economy*, *126*, 1558–1593.
- Harless, D. W., & Camerer, C. F. (1994). The predictive utility of generalized expected utility theories. *Econometrica*, *62*, 1251–1289.
- Harrison, G. W., Johnson, E., McInnes, M. M., & Rutström, E. (2005). Temporal stability of estimates of risk aversion. *Applied Financial Economics Letters*, *1*, 31–35.
- Harrison, G. W., Lau, M. I., & Rutström, E. (2007). Estimating risk attitudes in denmark: A field experiment. *Scandinavian Journal of Economics*, *109*, 341–368.
- Harrison, G. W., Lau, M. I., & Yoo, H. I. (2019). Risk attitudes, sample selection, and attrition in a longitudinal field experiment. *Review of Economics and Statistics*, *102*, 1–17.
- Harrison, G. W. & Rutström, E. (2008). Experimental evidence on the existence of hypothetical bias in value elicitation methods. In C. R. Plott & V. L. Smith (Eds.), *Handbook of Experimental Economics Results, Vol. 1* (1st ed., pp. 752–767). Elsevier: North-Holland.
- Hey, J. D., & Pace, N. (2014). The explanatory and predictive power of non two-stage-probability theories of decision making under ambiguity. *Journal of Risk and Uncertainty*, *49*, 1–29.
- Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*, 1644–1655.
- Holt, C. A., & Laury, S. K. (2005). Risk aversion and incentive effects: New data without order effects. *American Economic Review*, *95*, 902–904.
- Holt, C. A., & Laury, S. K. (2014). Assessment and estimation of risk preferences. In M. J. Machina & W. K. Viscusi (Eds.), *Handbook of the Economics of Risk and Uncertainty, Vol. 1*. Elsevier: North Holland.
- Holzmeister, F., & Stefan, M. (2020). The risk elicitation puzzle revisited: Across-methods (in)consistency? *Experimental Economics*, forthcoming.
- Kurtz-David, V., Persitz, D., Webb, R., & Levy, D. J. (2019). The neural computation of inconsistent choice behavior. *Nature Communications*, *10*, 1–14.
- Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, *21*, 37–44.
- Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.
- Mata, R., Frey, R., Richter, D., Schupp, J., & Hertwig, R. (2018). Risk preference: A view from psychology. *Journal of Economic Perspectives*, *32*, 155–172.
- McFadden, D. L. (2001). Economic choices. *American Economic Review*, *91*, 351–378.
- Moffatt, P. G. (2005). Stochastic choice and the allocation of cognitive effort. *Experimental Economics*, *8*, 369–388.
- Moffatt, P. G. (2015). *Experimentics: Econometrics for Experimental Economics*. Palgrave Macmillan: London.
- Palan, S., & Schitter, C. (2018). Prolific. A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.
- Polisson, M., Quah, J. K. H., & Renou, L. (2020). Revealed preferences over risk and uncertainty. *American Economic Review*, *110*, 1782–1820.
- Silvey, S. D. (1980). *Optimal Design: An Introduction to the Theory for Parameter Estimation, Vol. 1*. Chapman and Hall: New York.
- Thurstone, L. L. (1927). Psychophysical analysis. *The American Journal of Psychology*, *38*, 368–389.
- Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*, 1275–1289.
- Von Gaudecker, H. M., Van Soest, A., & Wengström, E. (2011). Heterogeneity in risky choice behavior in a broad population. *American Economic Review*, *101*, 664–694.
- Wakker, P. P. (2008). Explaining the characteristics of the power (CRRA) utility family. *Health Economics*, *17*, 1329–1344.
- Yu, C. W., Zhang, Y. J., & Zuo, S. X. (2019). Multiple switching and data quality in the multiple price list. *Review of Economics and Statistics*, *103*, 1–45.