University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 362

The Predictive Power of Risk Elicitation Tasks

Michele Garagnani

September 2020

The Predictive Power of Risk Elicitation Tasks

Michele Garagnani[1]

[1]Department of Economics, University of Zurich. Blümlisalpstrasse 10, CH-8006 Zurich, Switzerland

Abstract

This work reports an online experiment with a general-population sample examining the performance of budget-choice tasks for elicitation of risk attitudes. First, I compare the investment task of Gneezy and Potters (1997) with the standard choice-list method of Holt and Laury (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 no significant 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 subjects' understanding of tasks of this type.

JEL Classification: C91 · D81 · C83

Keywords: Risk Preferences · Elicitation Methods · Budget Sets · Portfolio Choices

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 Mata et al., 2018, for a recent survey). A 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 and Gneezy, 2010), especially for non-student populations (Yu, Zhang, and Zuo, 2019).[1]

[1]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.

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. A popular alternative is the Investment Task (INV) of Gneezy and Potters (1997) (see also Charness and Gneezy, 2010; Charness, Gneezy, and Imas, 2013), which simply asks individuals to allocate money between a safe option and a risky one.[2] 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 and Potters, 1997; Choi et al., 2007a,b, 2014; Ahn et al., 2014; Hey and Pace, 2014; Castillo, Dickinson, and Petrie, 2017; Halevy, Persitz, and Zrill, 2018; Kurtz-David et al., 2019; Polisson, Quah, and Renou, 2020, 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 and 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 HL and INV, which are two of the most commonly used risk elicitation tasks in economics. To this end, I conducted an experiment including HL, INV, 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. To answer the second question, 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.

The incentivized experiment relied on a general-population sample ($N = 403$). Results show that the out-of-sample predictive ability of HL and INV is undistinguishable, and that overall performance is rather modest. Strikingly, in the additional budget-choice tasks, a large majority of the subjects failed to report the normatively-predicted corner solutions. These results cast doubt on the suitability of general budget-choice tasks for empirical applications in non-student populations.

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

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 and Filippin, 2016; Holzmeister and Stefan, 2019). 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, 2014; Kurtz-David et al., 2019; Polisson, Quah, and Renou, 2020; Drichoutis and Nayga, 2020). The results described here should be 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 tasks 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 experiment involved 403 individuals using Prolific (Palan and Schitter, 2018), an online platform which allows to recruit from the general population.[3] 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). Among participants, 49.95% were fully-employed, 19.92% worked part-time, 11.10% were housekeepers, and 9.07% were unemployed. 68% of our sample was female.

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

The experiment was programmed in Qualtrics and it 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 3 budget-choice slider tasks with increased incentives (5×). At the end of the experiment, a

[3]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.

(self-reported, non-incentivised) qualitative risk assessment measure (Dohmen et al., 2011; Falk et al., 2018) was implemented to investigate its correlation with HL and INV.

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 and Rabas, 2016, for an illustration of 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. Further details on the implementation of the tasks are given in Appendix A and instructions are presented in Appendix D.

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, Torsney, and Wu, 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 and INV, and subsection 3.3 compares HL and INV with a structural econometric estimation using the block of 32 binary choices.

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, Snowberg, and Yariv, 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:

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

According to the assumed utility function, the vast majority of subjects are classified as risk averse, as commonly found in the literature (Gneezy and Potters, 1997; Holt and Laury, 2002; Harrison, Lau, and Rutström, 2007). In particular, according to HL only 27.30% (110) of subjects are classified as risk seeking. 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), while with INV is 6.343 (SD 34.996). 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 (139) gave "focal-point answers" investing amounts of exactly 0% (10 participants, with an implied $r \leq 0$), exactly 100% (30 participants, with an implied $r \geq 223.1$), or exactly 50% (99 participants, 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 subjects who report corner solutions (0% or 100%), the average estimated risk attitude using INV is 0.840 (SD 1.939). Excluding all 139 subjects reporting 0%, 100%, or 50%, the average is 0.913 (SD 2.270). 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% (127) 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 and Laury, 2002, 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).[4] This is also true in the dataset at hand: HL and INV are not significantly correlated (Pearson's $r = 0.01$, $p = 0.877$). This does not change when restricting to subjects who behaved consistently in the HL task ($r = -0.03$, $p = 0.771$) or to those reporting interior solutions in the INV task ($r = 0.010$ $p = 0.911$). However, there is a positive, although small, correlation between the self-reported, non-incentivised willingness to take risks (QRA) and HL ($r = 0.106$, $p = 0.034$), as well as between QRA and INV ($r = 0.134$, $p = 0.007$).

[4]However, Gillen, Snowberg, and Yariv (2019) show that commonly-used measures of risk attitudes are more correlated than previously thought once measurement errors are accounted for.

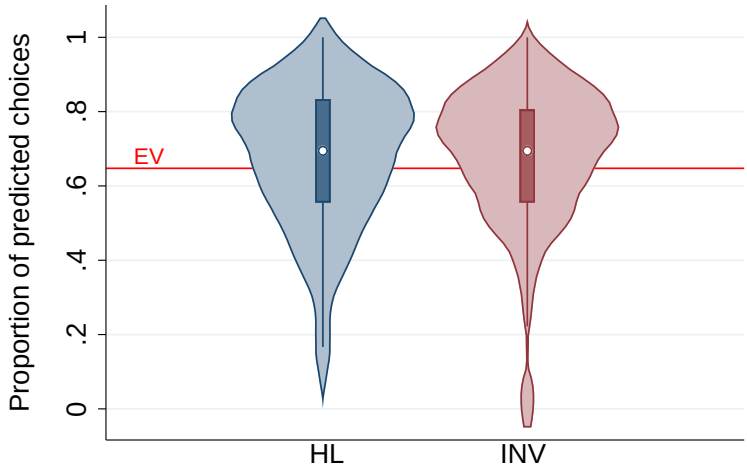

Figure 1: Distribution of the average proportions of out-of-sample predicted choices between elicitation methods. Violin plots show the median, the interquartile range, and the 95% confidence intervals as well as rotated kernel density plots on each side. Red-horizontal line indicate the predicted behavior of an expected value maximiser.

3.2 Predictive Performance

Figure 1 presents the out-of-sample predictive performance of INV compared to HL. In particular, I report the distribution (violin plot) of the proportions of correctly predicted choices at the individual level using the individual estimated risk attitudes.[5]

There is no significant difference in the number of correctly predicted choices between the two methods (INV: 66.51%, HL: 67.18%; WSR test, $N = 403$, $z = 0.485$, $p = 0.628$), which is modest. These levels of performance are only slightly above the predictive ability of simply assuming expected-value maximization (i.e., risk neutrality), which is 64.74% (for INV: WSR test, $N = 403$, $z = 2.691$, $p = 0.007$; for HL: WSR test, $N = 403$, $z = 4.451$, $p < 0.001$).

At the individual level, INV significantly predicts out-of-sample choices better than HL for 23.57% (95) of individuals.[6] Conversely, HL significantly predicts better than INV for 23.08% (93) of individuals. Therefore, there is no clear difference in the predictive power between the two methods.

Part of the subjects (17.82%, $N = 72$) 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 (INV 68.97%;

[5]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 $\rho = -0.225, N = 403, p < 0.0001$).

[6]The threshold for significance is set at $p < 0.05$ for a test of proportions, conducted separately for each subject.

HL 69.10%; WSR test, $N = 332$, $z = 0.757$, $p = 0.449$). Since not all subjects behaved consistently in the HL task (68.49%), 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, $N = 276$, $z = 0.017$, $p = 0.987$). 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 $N = 363$, $z = -1.222$, $p = 0.222$).

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; Harrison, Lau, and Rutström, 2007; Harrison and Rutström, 2008; Harrison, Lau, and Yoo, 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 (Von Gaudecker, Van Soest, and Wengström, 2011; Conte, Hey, and Moffatt, 2011; Moffatt, 2015; Alós-Ferrer and Garagnani, 2018). I estimated an additive random utility model (RUM), 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 and Camerer, 1994; Moffatt, 2005; Harrison and Rutström, 2008; Bellemare, Kröger, and van Soest, 2008; Von Gaudecker, Van Soest, and Wengström, 2011; Conte, Hey, and Moffatt, 2011; Moffatt, 2015).

The average estimated risk attitude following this method is 0.418 (SD 0.321). I then compared the results to the estimated risk parameters from HL and INV. There is a positive correlation between HL and ML ($r = 0.225$, $p < 0.001$), but there is no significant correlation between INV and ML ($r = 0.070$, $p = 0.160$). Since both INV and HL correlated with QRA (but not with each other), and only HL correlated with ML, this suggests that INV might be closer to a measure of self-evaluated willingness to take risks, with HL capturing aspects both of the latter and a more normatively-defined risk attitude.

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 and Potters, 1997; Kurtz-David et al., 2019; Gillen, Snowberg, and Yariv, 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.

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

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

or, equivalenty,

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

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

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

or, equivalently,

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

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

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

or, equivalently,

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

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 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 moderately risk-averse participants. An additional 46 subjects were classified with $r > 1$ according to HL. Of those, 4 had a corner solution at $p = 0.05$, 16 had a corner solution at $p = 0.95$, and 90 had interior optima. Among the 110 subjects were classified with $r < 0$ according to HL. Of those, 2 had a corner solution at $p = 0.05$ and 44 had interior optima.[7] Therefore, all subjects with moderate risk aversion $0 < r < 1$ had optima at the corner solution $p = 0.05$ in the third slider.

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

[7] 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 $p = 0.05$. Only four were classified as highly risk averse ($r > 1$; Moffatt, 2015), and their optima were interior. The 37 participants (9.18%) classified as risk seeking according to ML had also interior optima.

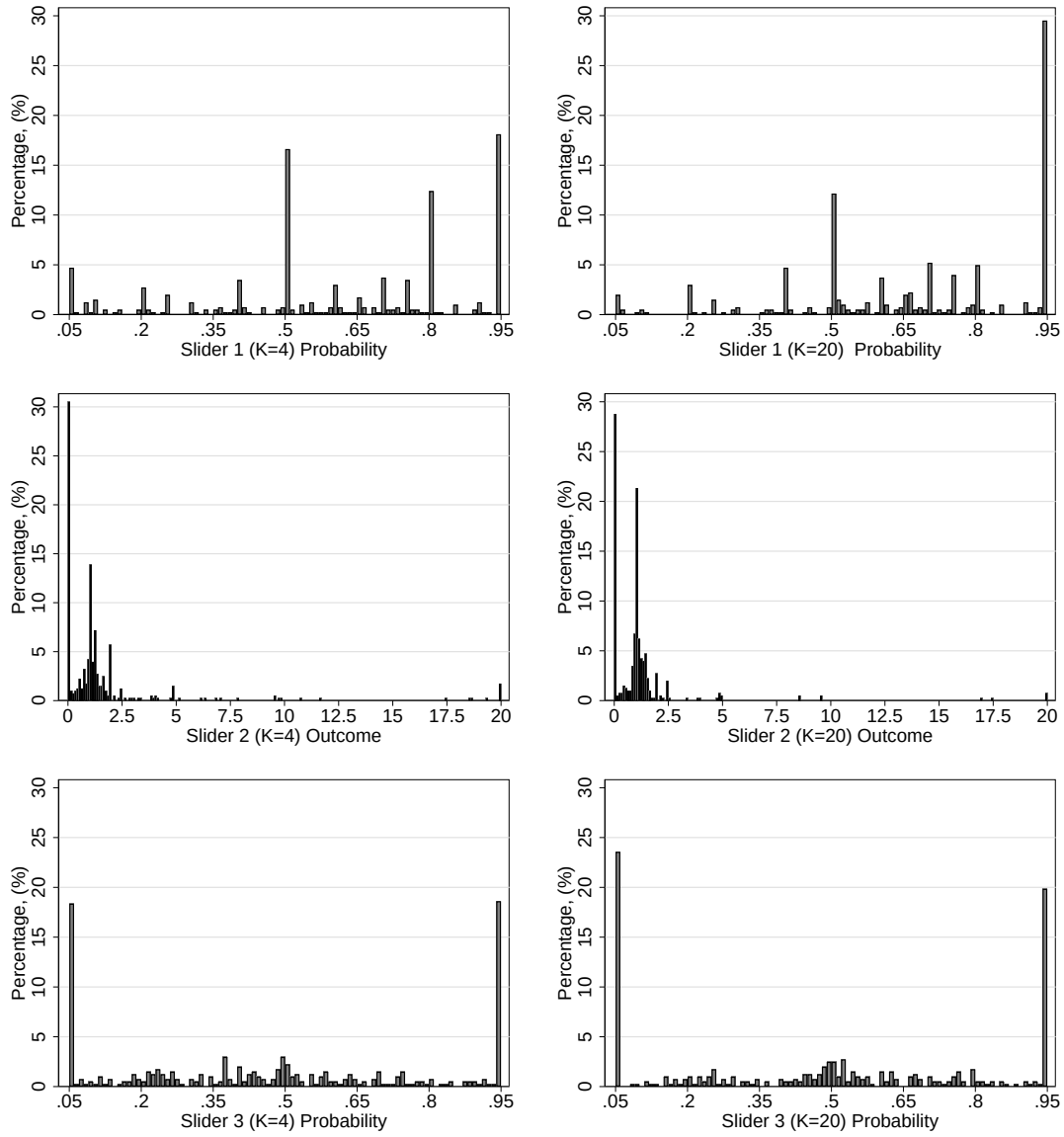

Figure 2: Distribution of answers in the first (upper), second (middle), and third slider (bottom) in the first repetition (left figures) and with higher incentives (right figures).

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.

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.

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 $N = 403$, $z = -1.271$, $p = 0.204$).

5 Discussion and Conclusion

In an experiment with a general population sample, I evaluate the predictive validity of two of the most common risk-attitude elicitation procedures, the choice-list procedure of Holt and Laury (2002) and the investment task of Gneezy and Potters (1997). The tasks are undistinguishable in their ability to predict out-of-sample choices, and this ability is moderate at best. Strikingly, performance in this sense is only slightly better than that

obtained by ignoring all individual information and predicting on the basis of expected values only.

Additionally, the experiment included choice-budget 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, an 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 currently becoming widespread.

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, Quah, and Renou, 2020; Drichoutis and 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.

References

Ahn, David, Syngjoo Choi, Douglas Gale, and Shachar Kariv. 2014. "Estimating Ambiguity Aversion in a Portfolio Choice Experiment." *Quantitative Economics* 5 (2):195–223.

Alós-Ferrer, Carlos and Michele Garagnani. 2018. "Strength of Preference and Decisions Under Risk." Working Paper, University of Zurich.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström. 2008. "Eliciting Risk and Time Preferences." *Econometrica* 76 (3):583–618.

Andersen, Steffen, Glenn W. Harrison, Morten Igel Lau, and E. Elisabet Rutström. 2006. "Elicitation Using Multiple Price List Formats." *Experimental Economics* 9 (4):383–405.

Atkinson, Anthony C. 1996. "The Usefulness of Optimum Experimental Designs." *Journal of the Royal Statistical Society* 51 (1):59–76.

Beauchamp, Jonathan P., Daniel J. Benjamin, David I. Laibson, and Christopher F. Chabris. 2019. "Measuring and Controlling for the Compromise Effect when Estimating Risk Preference Parameters." *Experimental Economics* :1–31.

Bellemare, Charles, Sabine Kröger, and Arthur van Soest. 2008. "Measuring Inequity Aversion in a Heterogeneous Population Using Experimental Decisions and Subjective Probabilities." *Econometrica* 76 (4):815–839.

Binswanger, Hans P. 1980. "Attitudes Toward Risk: Experimental Measurement in Rural India." *American Journal of Agricultural Economics* 62 (3):395–407.

Castillo, Marco, David L Dickinson, and Ragan Petrie. 2017. "Sleepiness, Choice Consistency, and Risk Preferences." *Theory and Decision* 82 (1):41–73.

Charness, Gary, Catherine Eckel, Uri Gneezy, and Agne Kajackaite. 2018. "Complexity in Risk Elicitation may Affect the Conclusions: A Demonstration Using Gender Differences." *Journal of Risk and Uncertainty* 56 (1):1–17.

Charness, Gary, Thomas Garcia, Theo Offerman, and Marie Claire 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* forthcoming.

Charness, Gary and Uri Gneezy. 2010. "Portfolio Choice and Risk Attitudes: An Experiment." *Economic Inquiry* 48 (1):133–146.

Charness, Gary, Uri Gneezy, and Alex Imas. 2013. "Experimental Methods: Eliciting Risk Preferences." *Journal of Economic Behavior and Organization* 87:43–51.

Choi, Syngjoo, Raymond Fisman, Douglas Gale, and Shachar Kariv. 2007a. "Consistency and Heterogeneity of Individual Behavior under Uncertainty." *American Economic Review* 97 (5):1921–1938.

Choi, Syngjoo, Raymond Fisman, Douglas M. Gale, and Shachar Kariv. 2007b. "Revealing Preferences Graphically: An Old Method Gets a New Tool Kit." *American Economic Review* 97 (2):153–158.

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman. 2014. "Who is (More) Rational?" *American Economic Review* 104 (6):1518–50.

Conte, Anna, John D. Hey, and Peter G. Moffatt. 2011. "Mixture Models of Choice Under Risk." *Journal of Econometrics* 162 (1):79–88.

Crosetto, Paolo and Antonio Filippin. 2016. "A Theoretical and Experimental Appraisal of Four Risk Elicitation Methods." *Experimental Economics* 19 (3):613–641.

Csermely, Tamás and Alexander Rabas. 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 (2-3):107–136.

Dave, Chetan, Catherine C. Eckel, Cathleen A. Johnson, and Christian Rojas. 2010. "Eliciting Risk Preferences: When is Simple Better?" *Journal of Risk and Uncertainty* 41 (3):219–243.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." *Journal of the European Economic Association* 9 (3):522–550.

Drichoutis, Andreas C. and Rodolfo Nayga. 2020. "Economic Rationality Under Cognitive Load." *Economic Journal* forthcoming.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. "Global Evidence on Economic Preferences." *Quarterly Journal of Economics* 133 (4):1645–1692.

Ford, Ian, Bernard Torsney, and C.F. Jeff Wu. 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 (2):569–583.

Friedman, Daniel, R. Mark Isaac, Duncan James, and Shyam Sunder. 2014. *Risky Curves: On the Empirical Failure of Expected Utility*. New York, NY: Routledge.

Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2019. "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study." *Journal of Political Economy* 127 (4):1826–1863.

Gneezy, Uri and Jan Potters. 1997. "An Experiment on Risk Taking and Evaluation Periods." *Quarterly Journal of Economics* 112 (2):631–645.

Halevy, Yoram, Dotan Persitz, and Lanny Zrill. 2018. "Parametric Recoverability of Preferences." *Journal of Political Economy* 126 (4):1558–1593.

Harless, David W. and Colin F. Camerer. 1994. "The Predictive Utility of Generalized Expected Utility Theories." *Econometrica* 62 (6):1251–1289.

Harrison, Glenn W., Eric Johnson, Melayne M. McInnes, and E. Elisabet Rutström. 2005. "Temporal Stability of Estimates of Risk Aversion." *Applied Financial Economics Letters* 1 (1):31–35.

Harrison, Glenn W., Morten I. Lau, and E. Elisabet Rutström. 2007. "Estimating Risk Attitudes in Denmark: A Field Experiment." *Scandinavian Journal of Economics* 109 (2):341–368.

Harrison, Glenn W., Morten I. Lau, and Hong I.l. Yoo. 2019. "Risk Attitudes, Sample Selection, and Attrition in a Longitudinal Field Experiment." *Review of Economics and Statistics* 102 (3):1–17.

Harrison, Glenn W. and E. Elisabet Rutström. 2008. "Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods." In *Handbook of Experimental Economics Results*, vol. 1, Part 5, edited by Charles R. Plott and Vernon L. Smith, chap. 81. Elsevier, 1 ed., 752–767.

Hey, John D. and Noemi Pace. 2014. "The Explanatory and Predictive Power of Non Two-Stage-Probability Theories of Decision Making Under Ambiguity." *Journal of Risk and Uncertainty* 49 (1):1–29.

Holt, Charles A. and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review* 92 (5):1644–1655.

———. 2005. "Risk Aversion and Incentive Effects: New Data Without Order Effects." *American Economic Review* 95 (3):902–904.

Holzmeister, Felix and Matthias Stefan. 2019. "The Risk Elicitation Puzzle Revisited: Across-Methods (in)Consistency?" *SSRN 3471852 Working Paper* .

Kurtz-David, Vered, Dotan Persitz, Ryan Webb, and Dino J. Levy. 2019. "The Neural Computation of Inconsistent Choice Behavior." *Nature Communications* 10 (1):1–14.

Luce, R. Duncan. 1959. *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

Mata, Rui, Renato Frey, David Richter, Jürgen Schupp, and Ralph Hertwig. 2018. "Risk Preference: A View from Psychology." *Journal of Economic Perspectives* 32 (2):155–72.

McFadden, Daniel L. 2001. "Economic Choices." *American Economic Review* 91 (3):351–378.

Moffatt, Peter G. 2005. "Stochastic Choice and the Allocation of Cognitive Effort." *Experimental Economics* 8 (4):369–388.

———. 2015. *Experimetrics: Econometrics for Experimental Economics*. London: Palgrave Macmillan.

Palan, Stefan and Christian Schitter. 2018. "Prolific.ac – A Subject Pool for Online Experiments." *Journal of Behavioral and Experimental Finance* 17:22–27.

Polisson, Matthew, John K.H. Quah, and Ludovic Renou. 2020. "Revealed Preferences over Risk and Uncertainty." *American Economic Review* 110 (6):1782–1820.

Silvey, Samuel David. 1980. *Optimal Design: An Introduction to the Theory for Parameter Estimation*, vol. 1. New York: Chapman and Hall.

Thurstone, L. 1927. "A Law of Comparative Judgement." *Psychological Review* 34:273–286.

Von Gaudecker, Hans-Martin, Arthur Van Soest, and Erik Wengström. 2011. "Heterogeneity in Risky Choice Behavior in a Broad Population." *American Economic Review* 101 (2):664–694.

Wakker, Peter P. 2008. "Explaining the Characteristics of the Power (CRRA) Utility Family." *Health Economics* 17 (12):1329–1344.

Yu, Chi Wai, Y Jane Zhang, and Sharon Xuejing Zuo. 2019. "Multiple Switching and Data Quality in the Multiple Price List." *Review of Economics and Statistics* :1–45.

Table A.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

APPENDICES – For Online Publication Only

A Risk Elicitation Methods: Implementation

The investment task (INV) was implemented as closely to the original as possible (Gneezy and Potters, 1997; see also Charness and Gneezy, 2010; Charness, Gneezy, and Imas, 2013). Wording was adapted from Gillen, Snowberg, and Yariv (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 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 and Laury, HL) wording and structure of the lotteries were kept as close as possible to the original (Holt and Laury, 2002). Table A.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.

In the Qualitative Risk Assessment (QRA), subjects are directly asked how willing they are to take risks. Wording was adapted from the English version of Gillen, Snowberg, and Yariv (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.

B Lotteries Used for the Out-of-Sample Predictions

Table B.1: 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

C 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

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

The derivative of the objective function is

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

Since $u''(\cdot) < 0$, u' is strictly decreasing, hence $u'\left(K + (1-p^*)q\right) < u'\left(K - p^* q\right)$. 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

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

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

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

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

D 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.

[The three sliders used the same graphical representation.]

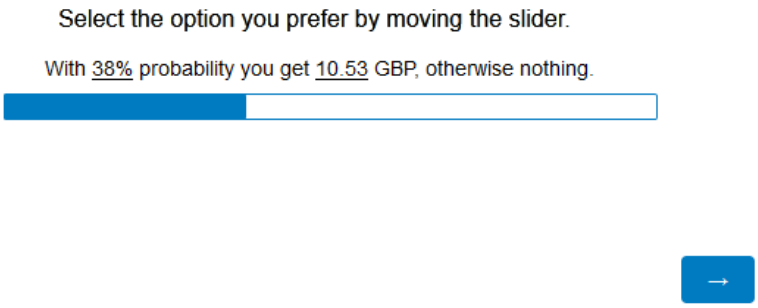

Figure D.1: Example of a slider task.

[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]

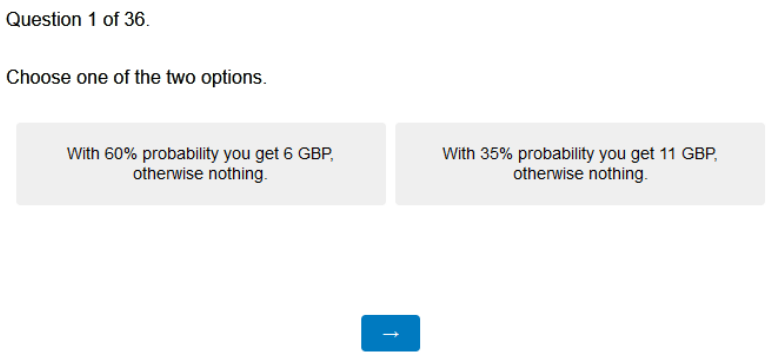

Figure D.2: Example of a binary choice trial.

[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.]

[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 A.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.]

[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."]