Exp Econ
DOI 10.1007/s10683-015-9457-9

ORIGINAL PAPER

A theoretical and experimental appraisal of four risk elicitation methods

Paolo Crosetto[1] · **Antonio Filippin**[2,3]

Received: 1 April 2014 / Revised: 25 June 2015 / Accepted: 30 June 2015

Abstract The paper performs an in-depth comparison of four incentivised risk elicitation tasks. We show by means of a simulation exercise that part of the often observed heterogeneity of estimates across tasks is due to task-specific measurement error induced by the mere mechanics of the tasks. We run a replication experiment in a homogeneous subject pool using a between subjects one-shot design. Results shows that the task estimates vary over and above what can be explained by the simulations. We investigate the possibility the tasks elicit different types of preferences, rather than simply provide a different measure of the same preferences. In particular, the availability of a riskless alternative plays a prominent role helping to explain part of the differences in the estimated preferences.

Keywords Risk attitudes · Elicitation methods · Experiment

JEL Classification C81 · C91 · D81

Electronic supplementary material The online version of this article (doi:10.1007/s10683-015-9457-9) contains supplementary material, which is available to authorized users.

✉ Paolo Crosetto
paolo.crosetto@gmail.com

Antonio Filippin
antonio.filippin@unimi.it

[1] INRA and Univ. Grenoble Alpes, UMR 1215 GAEL, 38000 Grenoble, France

[2] DEMM, University of Milan, Via Conservatorio 7, 20122 Milano, Italy

[3] Institute for the Study of Labor (IZA), Schaumburg-Lippe-Str. 5-9, 53113 Bonn, Germany

Published online: 28 July 2015

1 Introduction

Since uncertainty is a pervasive phenomenon in economic decisions, properly measuring attitudes toward risk is crucial in drawing conclusions from economic theory. Scholars have proposed a strikingly long list of methods to measure risk preferences usually by making subjects choose among lotteries. This is done in a variety of ways. The task can entail a *single* choice among a set of predetermined prospects presented in an abstract way (Binswanger 1981; Eckel and Grossman 2008a) or be framed as an investment decision (Gneezy and Potters 1997; Charness and Gneezy 2010). Alternatively, subjects might be asked to take *multiple* decisions between pairs or sets of risky lotteries presented in a structured (Holt and Laury 2002; Garcia-Gallego et al. 2012) or random way (Hey and Orme 1994). Lotteries are sometimes presented by means of visual tasks without making explicit reference to probabilities (Slovic 1966; Lejuez et al. 2002; Crosetto and Filippin 2013). Other designs elicit the certainty equivalent of some lotteries (Becker et al. 1964), let the subjects choose among an increasing sure amount and a fixed lottery (Abdellaoui et al. 2011), or ask subjects to input a value for one of the outcomes of a lottery that would make them indifferent with respect to another offered lottery (Wakker and Deneffe 1996). Risk preferences have also been indirectly derived from bids in first price sealed bid auctions (Cox et al. 1982).

All the aforementioned tasks make use of remunerated choices within incentive compatible designs. A different and widely used approach is to ask subjects to directly report their risk preferences. This can be done using a single question such as the one contained in the German Socio-Economic Panel Study (SOEP, Wagner et al. 2007) or asking questions about hypothetical real-life decisions, as done by the Domain-Specific Risk-Taking Scale (DOSPERT, Blais and Weber 2006). Such a florilegium of alternatives can at least in part be explained by different research goals. For instance, different tasks should be used if the researcher wants to investigate risk preferences *per se*, or if the aim is instead to control for risk attitudes while analysing choices in other contexts that nonetheless involve uncertainty. While some characteristics should be common to both goals, e.g., a sound theoretical underpinning, others are more goal-specific. If the target is just to control for risk preferences, the ideal risk elicitation mechanism should also be easy to understand and fast to implement, possibly paying the lowest possible price in terms of loss of precision.

In this paper we focus on a battery of incentivised tasks that are well-suited to elicit risk preferences as controls, i.e., to be used as companion tasks in experimental sessions in which the core treatments deal with other topics involving uncertainty:

- the multiple price list, in its Holt and Laury (2002) incarnation (henceforth, HL);
- an ordered lottery choice task, in the version implemented by Eckel and Grossman (2002, 2008a) (EG);
- the Investment Game by Gneezy and Potters (1997) (GP);
- the Bomb Risk Elicitation Task by Crosetto and Filippin (2013) (BRET).

Moreover, we include two self-reported questionnaire measures, and namely

- the German Socio-Economic Panel Study risk question (SOEP, Wagner et al. 2007), and
- the Domain-Specific Risk-Taking Scale (Blais and Weber 2006, DOSPERT).

While many other risk elicitation mechanisms exist,[1] we focus on the ones mentioned above as they are among the most commonly used, they arguably result in a relatively lower cognitive load for the subjects,[2] and they are fast and easy to implement.

Other scholars have already compared some of these tasks. Deck et al. (2010) compare four common risk elicitation tasks: HL, EG, the Balloon (Lejuez et al. 2002), and a version of the 'Deal or Not Deal' TV show; Deck et al. (2013) also include the DOSPERT questionnaire. Bruner (2009) uses multiple price-lists. Harbaugh et al. (2010) compare the price-based Becker–DeGroot–Marschack (BDM) mechanism with a choice-based procedure. Reynaud and Couture (2012) elicit risk preferences of a random sample of French farmers using four different elicitation methods (HL, EG, the DOSPERT, and the SOEP). Dave et al. (2010) stress the trade-off between the comprehensibility and precision of the task, comparing HL and EG. Charness et al. (2013) survey the literature to discuss merits and weaknesses of HL, GP, EG and the Ballon.

A low correlation, if any, in the observed behaviour across tasks is a recurrent finding; Isaac and James (2000) even find a negative correlation between choices in different tasks. Even abandoning the assumption of a single individual risk attitude and adopting the concept of a rich, domain-specific risk trait does not solve the problem of a low individual correlation, as documented by Deck et al. (2013).

In this paper we perform an in-depth comparison of four of the most used risk elicitation tasks, with the aim of investigating if the tasks themselves might be heterogeneous enough to generate (at least some of) the observed instability of behavior. We do so in three ways.

First, we run a simulation exercise. Risk attitudes are a latent construct that can only be indirectly and imperfectly measured, and the degree of measurement error is possibly influenced by the characteristics of the elicitation methods. The simulations allow us to measure the bias introduced by the *mere mechanics* of the tasks in the estimation of an underlying known distribution of risk preferences, imposing that no behavioral artifacts, like framing effects, enter the picture. We find that the different methods do introduce systematic task-specific measurement errors. When coupled with stochastic preferences and trembles, the estimated preferences diverge considerably and in task-specific directions from the underlying true values.

Second, we run a *between subjects*, *one-shot* replication of the chosen tasks within a homogeneous subject pool. All the aforementioned comparisons opted for a within-

[1] For an extensive review, including other elicitation tasks with respect to those analysed here (e.g., random lottery pairs as in Hey and Orme (1994), the Becker–DeGroot–Marschak mechanism, auctions, and the trade-off method as in Wakker and Deneffe (1996), see Harrison and Rutström (2008), who underline pros and cons and provide different estimation techniques for the risk preference parameter(s) of different theories.

[2] While an absolute measure of the cognitive load is difficult to establish, it can be reasonably argued that task involving one or a few choices among clearly spelled out alternatives are less demanding than procedures implying dozens of choices, or mechanisms like the BDM requiring to expose the subjects to complicated instructions.

subject design, consistent with the focus on individual preference instability.[3] A *between subjects* design bars us from any conclusion about preference instability, but fits well with our aim of focusing on the bias introduced by the tasks themselves.[4]

We find that preferences estimations vary widely across tasks, even if the underlying population self-reported risk attitudes are comparable across treatments. The experimentally observed variation goes in the same direction as the one found through simulations, although sometimes its magnitude does not compare. The variance of the choices reflects instead very closely the pattern determined by the bias induced by the mechanics of the tasks.

Third, we investigate the possibility that the tasks elicit different types of preferences, rather than simply provide a different measure of the same preferences. We examine in detail the tasks to find out which of their characteristics might trigger different preferences. We find that some tasks feature a safe option, likely to induce certainty effects (Andreoni and Sprenger 2011, 2012) or to act as a focal reference point against which lower outcomes could be perceived as losses. The presence of a safe option appears to play a role from a gender perspective, too.

The outline of the paper is as follows. In Sect. 2 we describe the four risk elicitation tasks that we compare in this paper. The simulations are reported in Sect. 3. Section 4 reports results of the between subjects experimental replication of the tasks. Section 5 discusses the role played by the safe option and by other characteristics of the tasks and concludes.

2 Risk elicitation tasks

Our main goal is to compare the tasks in order to account for the differences observed in the literature rather than minimizing such differences, and therefore we choose the most common versions, which coincide with those originally published.[5]

[3] Using the data of the repeated treatment of Crosetto and Filippin (2013), in which the same task, the BRET, was repeated five times we find that the average choice is not significantly different than that of other subjects who played the same task in the one-shot mode. However, the correlation across periods turns out to be, on average, $\rho \cong 0.35$, ranging from $\rho \cong 0.01$ to $\rho \cong 0.6$, i.e., not much higher than what other contributions in the literature found using different elicitation methods. Hence, instability of results does not necessarily rely upon the use of different tasks.

[4] To the best of our knowledge the only papers in the literature using a between-subjects design are Charness and Viceisza (2011) and Harrison (1990). The former elicits risk attitudes of 91 farmers in rural Senegal using HL and GP besides asking the SOEP question, but the fraction of inconsistent choices above 70 % in the HL task makes the data not comparable. The latter finds that the BDM displays stronger risk seeking preferences as compared to risk aversion coefficients implicit in first price auction bids.

[5] The interested reader can find in Csermely and Rabas (2014) a comparison of different versions of a Multiple Price List that mimic several elicitation methods, including those analyzed here. The manipulations in Csermely and Rabas (2014) investigate the role played by some features of the payoffs and by fixed versus changing probabilities. In contrast, other characteristics such as the number of choices and how they map into coefficients of relative risk aversion are kept constant across MPL and therefore changed as compared to the original version of each task, with the exception of HL that is used as a benchmark. Their results differ considerably with respect to what usually found in the literature: for instance choices are on average significantly more risk averse in HL than in EG. Such a discrepancy indirectly shows that the framing of the elicitation methods play a role at least as important as that of the fundamentals defined by the underlying lotteries.

2.1 Multiple price list: Holt and Laury (HL)

The multiple price list format is a procedure used to elicit values from a subject. Applied to risk, it consists of giving the subject an ordered list of binary choices between lotteries. The most widely known implementation has been provided by Holt and Laury (2002), to date the most popular risk elicitation mechanism. In the HL task, subjects face a series of choices between pairs of lotteries, with Option A being safer than Option B (see Table 1).[6] The set of possible outcomes is common to every choice, and the lottery pairs are ordered by increasing expected value. The expected value increases from 3.28 to 4 euro for Option A and from 0.95 to 7.7 for Option B along the table thanks to the increase of the probability of the 'good' event. The subjects must make a choice for each pair of lotteries and, if consistent, should at some point switch to the risky option. The presence of multiple choices in HL generates the possibility of multiple switching. Such a behavior is usually deemed as inconsistent and is commonly observed, leading to data losses or to the need of a stochastic decision model. The switching point captures the risk aversion of the subject. A risk-neutral subject should start with Option A and switch to B from the fifth choice on. At the end of the experiment, one row is randomly chosen for payment, and the chosen lottery is played to determine the payoff.

2.2 Ordered lottery selection: Eckel and Grossman (EG)

In ordered lottery selection tasks, subjects are asked to pick one out of an ordered set of lotteries. This method was introduced in the literature by Binswanger (1981) to specifically measure risk preferences. A popular version is that proposed by Eckel and Grossman (2002, 2008a), in which subjects choose the preferred among a set of 5 lotteries characterised by a linearly increasing expected value as well as greater standard deviation. Differently from Holt and Laury (2002), the variation is obtained through manipulation of the outcomes of each lottery, keeping the probability of each outcome fixed at 50 %. Subjects are asked to choose one lottery. Then the lottery is played and the subject paid accordingly. The values used in the lab and the way they were presented to the subjects can be seen in Table 2. A risk-neutral subject should choose lottery 5, as it yields the higher expected value.

2.3 The investment game of Gneezy and Potters (GP)

A different approach is the one introduced by Gneezy and Potters (1997). They propose a task in which the choice is framed as an investment decision. Subjects have to decide how to allocate a given endowment of 4 euro between a safe account and a risky investment that yields 2.5 times the amount invested or zero with equal probability. In other words, the amount $k \in [0; 4]$ defines the following set of lotteries:

[6] The values are based on the baseline of Holt and Laury (2002), doubled to make them comparable with the other tasks.

Table 1 The ten lotteries chosen for the HL treatment

	Option A				Option B			
1	1/10	4 €	9/10	3.2 €	1/10	7.7 €	9/10	0.2 €
2	2/10	4 €	8/10	3.2 €	2/10	7.7 €	8/10	0.2 €
3	3/10	4 €	7/10	3.2 €	3/10	7.7 €	7/10	0.2 €
4	4/10	4 €	6/10	3.2 €	4/10	7.7 €	6/10	0.2 €
5	5/10	4 €	5/10	3.2 €	5/10	7.7 €	5/10	0.2 €
6	6/10	4 €	4/10	3.2 €	6/10	7.7 €	4/10	0.2 €
7	7/10	4 €	3/10	3.2 €	7/10	7.7 €	3/10	0.2 €
8	8/10	4 €	2/10	3.2 €	8/10	7.7 €	2/10	0.2 €
9	9/10	4 €	1/10	3.2 €	9/10	7.7 €	1/10	0.2 €
10	10/10	4 €	0/10	3.2 €	10/10	7.7 €	0/10	0.2 €

In our experimental replication

Table 2 The five lotteries chosen for the EG treatment

	Choice	Probability (%)	Outcome
1	A	50	4 €
	B	50	4 €
2	A	50	6 €
	B	50	3 €
3	A	50	8 €
	B	50	2 €
4	A	50	10 €
	B	50	1 €
5	A	50	12 €
	B	50	0 €

$$L_{GP} = \begin{cases} 4-k & \frac{1}{2} \\ 4+1.5k & \frac{1}{2}. \end{cases}$$

Similar to the EG task, in the Investment Game the choice of a larger fraction to be invested implies a change in the amount of money at stake, while the probabilities are not affected. Since the expected value of the task, equal to $4 + 0.25k$ is increasing in k a risk-neutral subject should invest all the endowment.

2.4 The bomb risk elicitation task (BRET)

The BRET is a visual real-time risk elicitation task introduced by Crosetto and Filippin (2013). Subjects face a 10×10 square in which each cell represents a box. They are told that 99 boxes are empty, while one contains a time bomb programmed

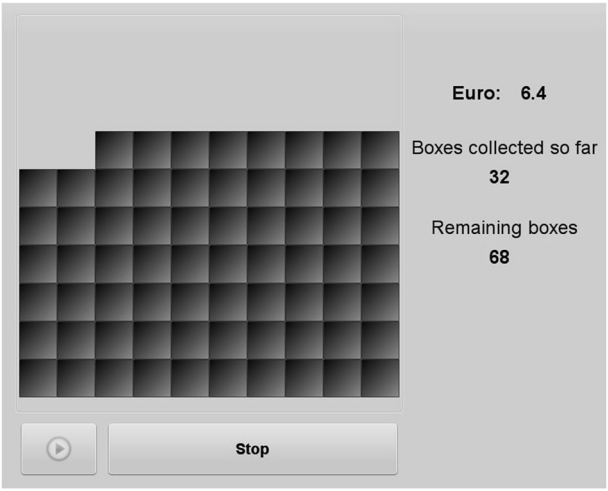


Fig. 1 The BRET interface after 32 seconds

to explode at the end of the task, i.e., *after* choices have been made. Below the square is a "Start" and a "Stop" button. From the moment the subject presses "Start" one box is automatically collected each second, starting from the upper left corner of the square. A screenshot of the task after 32 seconds (i.e., after 32 boxes have been collected) is reported in Fig. 1.

The BRET transparently displays probabilities, since it is possible to visually appreciate how many boxes have been collected and how many are left. Moreover, the subject is informed about the number of boxes collected at any point in time. Each time a box is collected, the subject's provisional account is credited with 20 additional euro cents. The subject can, at any moment, stop the drawing process by hitting the "Stop" button, thus determining the preferred number of boxes to be collected, $k \in [0, 100]$.

The position of the time bomb $b \in [1, 100]$ is determined after the choice is made by drawing a number from 1 to 100 from an urn. If $k_i^* \geq b$, it means that subject i collected the bomb which, by exploding, wipes out the subject's earnings. In contrast, if $k_i^* < b$, subject i leaves the minefield without the bomb and receives 20 euro cents for every box collected. The metaphor of the time bomb allows to avoid the truncation of the data that would otherwise happen in case of a real-time explosion of the bomb.

Subjects' decisions can be formalised as the choice of their favourite among the set of 101 lotteries, fully described both in terms of probabilities and outcomes by a single parameter $k \in [0, 100]$,

$$L_{BRET} = \begin{cases} 0 & \frac{k}{100} \\ 0.2k & \frac{100-k}{100}. \end{cases}$$

k summarizes the trade-off between the amount of money that can be earned and the likelihood of obtaining it. The degree of risk aversion negatively correlates with the choice of k and a risk-neutral subject should choose $k = 50$.

2.5 Questionnaires

After having gone through one of the four tasks the subjects were exposed to two self-reported risk measures: the SOEP and the DOSPERT. The SOEP measure consists of a direct question, extracted from the German Socio-Economic Panel Study (Wagner et al. 2007). It asks subjects to report, on a 0–10 scale, "How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?" The validity of this self-reported question in eliciting risk attitudes as compared to the results of incentivised lottery-based tasks has been explored by Dohmen et al. (2011), who claim that self-reported answers can represent a valid low-cost substitute for incentivised lottery schemes.

The Domain-Specific Risk-Taking Scale (Blais and Weber 2006) is a 30-item risk questionnaire spanning several different domains in which risk attitudes can play a role: ethical, financial (further decomposed into gambling and investment), health/safety, social, and recreational decisions. It has been developed by psychologists, reflecting the fact that utility-based measures and estimates of coefficients seem to fail when called to explain risk attitudes outside of the financial/gambling sphere.

3 The different measurement error induced by the mechanics of the tasks

The remarkably different features of the elicitation methods constitute a potentially important explanation for the wide variation observed in the measurement of risk preferences. Strikingly, the role played by the specific measurement error induced by each task has received little attention so far. We try to fill this gap in two ways. First, in Sect. 3.1 we analyse the range and precision of the elicitation methods from a theoretical point of view, empahsizing how they map choices into risk aversion parameter. Second, in Sect. 3.2 we evaluate the impact of such characteristics in a context of known preferences by means of a simulation exercise.

3.1 Theoretical determinants of task specific measurement error

The different elicitation methods allow us to classify participants in several categories, representing their different willingness to accept risk. The range of the estimate and the number of categories obviously correlate with the measurement error that each task implies on the underlying risk preferences. An elicitation mechanism that imposes a coarser classification of subjects or a truncated range of preferences automatically induces a higher measurement error. For the sake of simplicity, we assume that risk preferences are represented by a constant relative risk aversion (CRRA) utility function $u(x) = x^r$ so that risk attitudes can be summarised by means of the coefficient of relative risk aversion r.[7]

[7] For the correct formulation of the utility function when r is non-positive see Wakker (2008).

Range. As far as the type of risk attitudes is concerned, two of the tasks above, GP and EG, can only measure preferences in the risk-averse domain. They cannot distinguish risk neutrality from risk seeking (and from a slight degree of risk aversion).[8] Charness and Gneezy (2012) claim that this is a minor problem because risk seeking preferences are seldom observed, but in our experimental data about 20 % of the subject pool is characterised by $r \geq 1$. Moreover, a low fraction of risk-seeking subjects can be a function of the task itself as long as boundary effects matter (see below). In contrast, the HL and BRET allow to estimate a fairly complete range of preferences.

Precision. The lower the number of choices available, the larger the measurement error of the parameter that is introduced. HL can classify subjects in ten different categories, while the greater simplicity of the EG task comes at the price of a coarser estimation allowing five categories only. In contrast, GP and the BRET allow to estimate risk attitudes almost continuously.

Mapping from choices into $u(x) = x^r$. Besides in terms of range and precision, the tasks are also characterised by different ways in which they map choices into risk aversion parameters.

Every vertical bar in Fig. 2 separates the different choices in every task in the space of coefficient r of a CRRA power function. It is immediately evident that only HL is characterized by almost constant intervals. Figure 2 assumes that agents decide according to their 'true' risk preferences, while framing effects do not play any role. However, there is evidence in the literature that the shape of the opportunity set affects considerably the decisions, e.g., because subjects tend to avoid choices at the extremes of the opportunity set.[9]

Figure 3 internalizes such boundary effects, comparing how each task maps choices into the parameter r when the domains of the tasks are made comparable. Every task has been reparametrised as having ten choices, which imply nine cutoff points, increasing in terms of risk seeking preferences. Such a transformation is straightforward in the quasi-continuous tasks. For instance, GP cutoff 1 means the subject invested $\frac{1}{10}$ of her endowment in the risky asset, level 2 means $\frac{2}{10}$, and so on. Similarly, in the BRET cutoff 1 corresponds to 10 boxes, 2–20, and so on. EG is more problematic as it allows for five different choices only, thus implying only four cutoff points. In order to represent the EG task using the same 9-level scale, we replicate each cutoff point twice, except for the highest that is repeated three times, given that choosing the riskiest lottery is consistent with slight risk aversion, risk

[8] The version of the EG task implemented in Dave et al. (2010) features an additional lottery characterised by the same expected value as the fifth lottery, but by a higher variance. The additional choice reduces the problem because it allows to separate the behaviour of slightly risk-averse agents from that of risk seekers, but it does not solve it since a risk-neutral agent would still be indifferent between the two.

[9] Both Andersen et al. (2006) and Filippin and Crosetto (2015) show that the menu of lotteries available affects choices in the HL task. Crosetto and Filippin (2014) find that removing the first lottery in the EG results in the whole distribution of choices shifting towards more risk-loving decisions. There is similar evidence that subjects tend to avoid boundary choices in other branches of the literature, too (List 2007; Bardsley 2008).

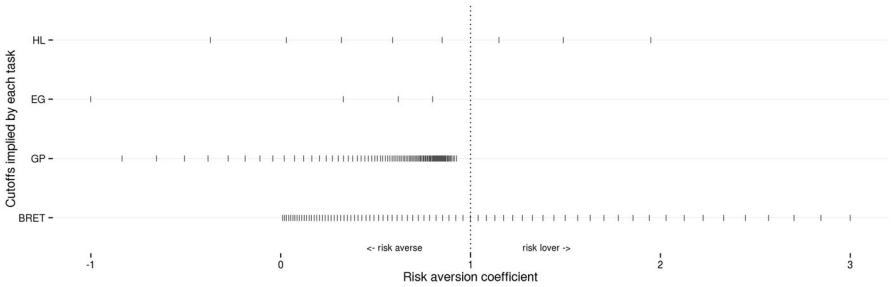


Fig. 2 *Vertical bars* separate the different choices in every task in the space of coefficient r of a CRRA power function

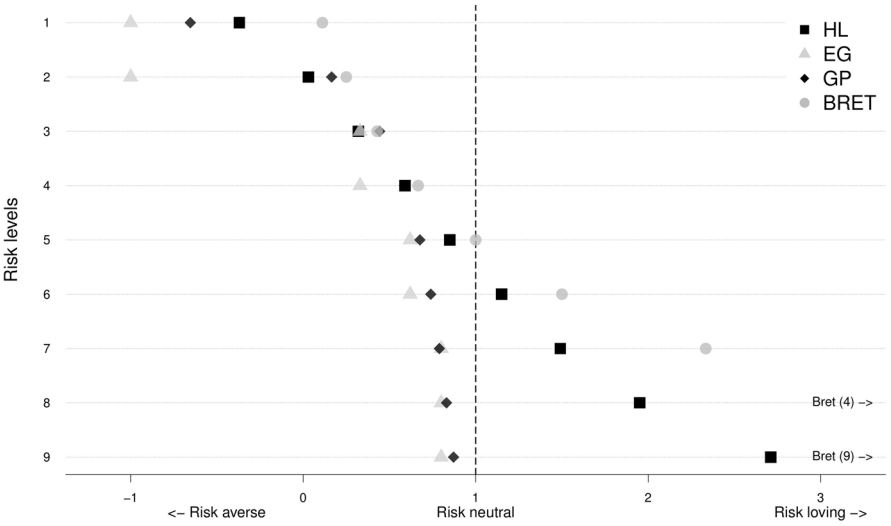


Fig. 3 Mapping of choices into the implied r by task. The figure assumes a CRRA power function $u(x) = x^r$. Risk levels are ordered from the safer to the riskier, i.e., for HL we count the number of risky choices. For clarity we limited the figure to the interval $(-1.5,3.5)$. Two values for the BRET (4 for risk level 8 and 9 for risk level 9) fall outside these bounds, and are reported in *brackets*

neutrality, and risk loving at the same time. HL already provides nine meaningful cutoff values.[10]

Figure 3 clearly shows that the shape of the function linking choices to r greatly differs across tasks. In some tasks, choices in the tails of the distribution imply extreme values in terms of r, thereby crucially affecting the measured average r.

[10] Making zero risky choices is a dominated action and is inconsistent with any degree of risk aversion, given that it implies that the subject prefers 4 euro for sure to 7.7 euro for sure. Note that usually the HL task is summarised by the number of *safe* choices. However, we prefer to use the number of *risky* choices for the sake of consistency with the other tasks so that in all the elicitation methods a higher choice represents lower risk aversion.

3.2 Simulation exercise

The impact of the aforementioned characteristics can be effectively summarized by means of a simulation exercise. We generated 100,000 virtual subjects, each characterised by a utility function $u(x) = x^r$, with $r \sim N(0.8, 0.3)$. The distribution was chosen to generate a realistic sample given the usual evidence in the lab, in which most of the mass is accounted for by moderate risk averse subjects, with a share of moderate risk lovers. In particular, this distribution generates the same share of risk-averse subjects ($\sim 73\,\%$) as in the large sample of HL replications collected by Filippin and Crosetto (2015).

Each of these virtual subjects is exposed to the four tasks parametrised exactly as described in Sect. 2. Using the data generated, we then retrieve the individual coefficient of risk aversion $\hat{r}$ assuming the correct utility function $u(x) = x^r$. We chose to assign the central value of r consistent with each interval.[11] Moreover, we assumed no inconsistencies in HL: no subject was allowed to switch multiple times or to submit dominated choices. This difference has to be taken into account in Sect. 4 when comparing the simulations with the experimental results, while the mapping between choices and $\hat{r}$ is obviously the same. This exercise allows us to numerically evaluate which bias in the measurement of risk preferences, if any, follows from distortions generated by the mere technical features of each task while estimating the correct preferences.

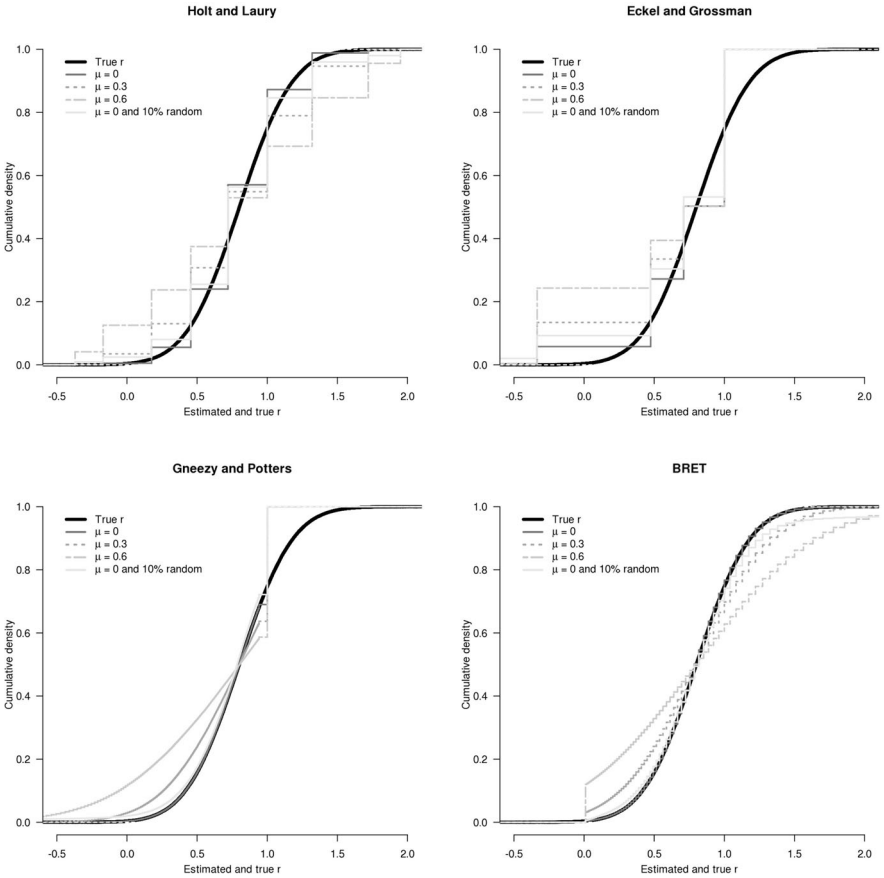
We run and report results of three sets of simulations. First, we assume totally deterministic preferences—that is, the virtual subjects act exactly as their r dictates them. This is done to measure the pure effect of the range and precision of the tasks on the parameter estimates.

Second, we assume stochastic preferences—that is, we add noise directly to the subject's r. The actual r_a followed by the subjects departs from their real r according to a normal noise with zero mean and variance μ: $r_a = r + \varepsilon$, $\varepsilon \sim N(0, \mu)$. We ran two noisy simulations, generating two distinct dataset: one with $\mu = 0.3$ and one with $\mu = 0.6$. Noisy preferences may induce subjects to make a choice different than that dictated by the true r, thereby making salient the shape of the function mapping choices into risk aversion parameters.

Finally, we repeat the same exercise imposing that a very large fraction ($10\,\%$) of the virtual subjects pick any choice out of the set offered by each task with equal probability, instead of following their r. This exercise simulates the presence of a share of confused subjects, in order to assess the robustness of the tasks to random behavior.

The results are plotted in Fig. 4. The figure shows, for each task, the empirical cumulative distribution of the true r (thick black line) and of the $\hat{r}$ retrieved from the choices made by the 100,000 virtual agents in the simulations. For the sake of visualization, the plots are shown only within the interval $r \in [-0.5, 2]$. The table

[11] Only one value of r can be meaningfully computed for the intervals at the extremes, the other being $r \to \pm\infty$. We assigned in these cases the only computable boundary of r for choices implying implausible degrees of risk aversion, and a value equal to one, i.e., risk neutrality, to the upper interval of EG and GP where different types of preferences converge.



	median	*mean*	*stdev*	*median*	*mean*	*stdev*	*median*	*mean*	*stdev*
True *r*	**0.800**	**0.801**	**0.300**	**0.800**	**0.801**	**0.300**	**0.800**	**0.801**	**0.300**
		$\mu = 0$			$\mu = 0.3$			$\mu = 0.6$	
HL	0.72	0.805	0.319	0.72	**0.805**	0.439	0.72	**0.804**	0.629
EG	0.71	0.743	0.339	0.71	0.666	0.444	0.71	0.562	0.548
GP	**0.799**	0.756	0.237	**0.801**	0.714	**0.312**	**0.803**	0.623	**0.457**
BRET	0.785	**0.801**	**0.299**	0.785	0.806	0.412	0.818	0.840	0.602
	μ = 0 and 10% random			*μ = 0.3 and 10% random*			*μ = 0.6 and 10% random*		
HL	0.72	**0.811**	**0.396**	0.72	**0.812**	**0.489**	0.72	**0.811**	0.649
EG	0.71	0.686	0.433	0.71	0.618	0.503	0.71	0.524	**0.580**
GP	**0.785**	0.533	8.62	0.779	0.494	8.62	0.771	0.411	8.63
BRET	0.818	1.12	3.72	**0.818**	1.13	3.73	**0.818**	1.16	3.75

Fig. 4 True and simulated coefficient of risk aversion by task. $N = 100{,}000$

below the figure shows the median, mean, and standard deviation of the same estimates, for each simulated condition.

Deterministic preferences. The results with deterministic preferences are shown in dark gray, solid line in the figure, and appear in the upper left panel in the table, in

which the task delivering the most precise estimate is shown in bold. The results are driven by the differences in the range and precision across tasks detailed above. EG and GP deliver downward biased estimates of the mean since they pool together all the risk lovers with the risk neutral (and slightly risk averse) subjects. EG and HL provide only a limited number of categories and both overestimate the variance, in particular the former. With respect to EG, HL is virtually linear in the mapping from choices to $\hat{r}$, and hence gives a more accurate prediction of the mean. The BRET and GP track the results by construction ($\forall r \in [0; 99]$ the former, $\forall r \in [-15.25; 0.93]$ the latter) thanks to the quasi-continuity of their choices. GP underestimates the variance because it collapses into $\hat{r} = 1$ the choices of all the subjects characterised by $r > 0.93$. GP also features a rather flat mapping for a vast region of choices: from investing 50 to 90 % the estimated risk aversion coefficient varies only from $r_{50} = 0.68$ to $r_{90} = 0.87$.

Stochastic preferences. The two sets of simulations featuring stochastic preferences ($\mu = 0.3$ and $\mu = 0.6$) are shown using two lighter shades of grey, dotted lines in Fig. 4, and detailed in the center and right upper panels of the table. When adding stochastic noise, the cumulative distribution functions tend to be flatter than for the real r.[12] The $\hat{r}$ estimated for each task tends to be biased according to how each task maps choices into r as shown in Figs. 2 and 3. EG and, to a more prominent degree, GP yield lower estimates for higher levels of noise, because they translate risk averse choices into exponentially decreasing $\hat{r}$. The BRET yields instead higher estimates for higher levels of noise due to the hyperbolic function between choices and r. Note that more dramatic effects would be observed by further increasing the variance in the data generating process. The HL task is instead virtually linear in the choice-r space, and this makes its estimate robust to stochastic decisions, at least if the analysis is restricted, as we do here, to consistent subjects.

Random subjects. The simulations including 10 % of random agents appear in the lower panel of the table below Fig. 4. The case for $\mu = 0$ is also drawn in light grey, solid line in the figure. Including a share of random subjects emphasizes the role played the shape of the map from choices to r (see Fig. 3 above). HL benefits from the linearity of the relation between choices and the parameter, proving to be robust to random behavior, once inconsistencies are assumed away. EG tends to underestimate r, because of a positive fraction of subjects that now choose the riskless lottery. On the other hand, GP and BRET are very sensitive to random behavior, because of the extreme values for r that are introduced by the outliers. The bounds imposed on Fig. 4 do not allow to visualise it, but the table makes it clear that, for both tasks, the variance increases dramatically and the point estimates drift in the expected direction.

Summing up, the addition of a stochastic component and of random behavior to the simulation exercise allows us to derive some interesting insights:

[12] Note however that adding a stochastic error is not equivalent to having a flatter normal that generates deterministic choices, because only in the first case we can test how each task reacts to preferences that are not perfectly defined.

1. The median choice is more appropriate than the mean to summarise and compare the risk attitudes in each task because intermediate choices are less dependent on task-specific distortions that characterise the tails. Moreover, tasks characterised by a low number of categories can produce a large fraction of decisions in categories in which only one bound of r is objectively defined.
2. Noisy decisions sharpen the differences in the estimated $\hat{r}$ across methods. HL is more robust to noise. However this advantage must be weighted against the loss of data due to inconsistent choices, which are likely to increase with the noisiness of the decision.
3. Adding random choices affects GP and BRET more than HL and EG, assuming that the incidence and intensity of random behavior, possibly coming from a low understanding of the task, is the same across methods.[13]

4 Experimental comparison

The experiment was carried out in the laboratory of the Max Planck Institute for Economics in Jena, Germany, from March to May 2012. A total of 350 subjects took part in 12 experimental sessions. Recruitment was carried out using Orsee (Greiner 2004) on the Jena subject pool, mainly composed of undergraduate students at the Friedrich Schiller University. The experimental software was programmed in Python (van Rossum 1995). Each subject took part in just one risk elicitation mechanism. 88 subjects participated in HL, EG and BRET, and 86 to GP sessions.

For all sessions and treatments a unique procedure was followed. Upon arrival at the lab, subjects were randomly assigned a seat and found on-screen instructions,[14] which were also read aloud. Once any questions had been answered on an individual basis, subjects were allowed to start with the risk elicitation mechanism, which constituted the main task of the experiment. Each subject went through the risk elicitation method only once. Subjects were then asked to complete the DOSPERT risk questionnaire and a further screen of questions, including the SOEP risk question, demographics, and a self-reported measure of understanding of the task. All tasks involved some sort of randomization to compute the final payoffs. These randomizations were carried out manually using dice (HL, EG, GP) or draws from an urn (BRET) after everyone had completed the questionnaires. For the sake of transparency the resolution of uncertainty was publicly performed with the help of randomly selected subjects. The subjects were then paid.

In order to improve the comparability across different mechanisms, we set the amounts at stake in such a way as to grant an expected earning in the order of 5 euro for a risk-neutral subject, plus the show-up fee of 2.5 euro. The sessions lasted on average <30 min.

[13] Our experimental data show that indeed no real subject in GP and BRET submitted such extreme choices. On the other hand, a much larger share of experimental subject choose the safe lottery in EG.

[14] The English translation of the original German instructions is attached in Online supplementary material.

4.1 Estimated values

Table 3 summarises the main results of our experimental comparison, reporting results only for subjects classified as consistent, i.e., never switching from risky to safe and not always choosing safe in the HL, and not submitting the dominated choices of 0 or 100 blocks in the BRET.

Subjects showing inconsistent behavior were 20 in HL (15 multiple switchers and five always choosing safe). This number is slightly higher (22 %) than on average (Filippin and Crosetto 2015, found 17.1 % of inconsistent choices in a vast sample of 6315 subjects over 54 published papers). The presence of a sizeable number of inconsistent subjects is a recurrent outcome of the HL method. This pattern may reflect a genuinely lower comprehensibility of the task, but it should also be noted that HL is the only method offering a solid possibility of filtering out the subjects with a bad comprehension of the task.[15]

Inconsistencies in the BRET consist of weakly dominated options (choosing either 0 or 100 boxes), something that we observe in 1 out of 88 observations. Such a frequency (1.2 %) is in line with our previous findings.

Note first that the subjects in all treatments are comparable in terms of their self-reported risk attitudes. The answers to the SOEP question are not significantly different in any of the tasks (Kruskal Wallis test, p value 0.905).

We translate the choices in each task into parameters r of a CRRA utility function x^r, reporting median, average and standard deviation by task.[16] Results show sizeable differences across aggregate choices in the different tasks.

We observe the higher measured risk aversion in the EG task (median 0.07, mean 0.09), followed by HL (median 0.45, mean 0.37), the BRET (median 0.66, mean 0.78), and GP (median 0.75, mean 0.73). Differences across tasks are highly significant: only two among all the pairwise comparisons do not display differences that reach traditional significance levels (BRET vs. GP and HL vs. EG).

Additional insights can be derived looking at the distribution of subjects according to their risk attitudes. This approach has the advantage of imposing weaker parametric assumptions since the fraction of subjects classified in the different categories does not depend on the specific functional form assumed for the utility function. Tasks are quite similar in this respect, with the BRET reporting a lower share of risk-averse subjects.

The last two rows of the table report the average choice and r found by other studies in the literature. We chose to refer, when possible, to meta-analyses or survey papers including more than one study. This turned out to be possible for HL

[15] Data from inconsistent subjects are sometimes used in the analyses adopting different techniques. The simplest approach is counting the number of safe choices irrespective of the inconsistency, and is not methodologically sound in our opinion. When building structural models that include a tremble, or when imposing only lower and upper bounds to the risk parameter, data of multiple switchers can instead be meaningfully exploited.

[16] Reporting the median choice in tasks with a low number of categories could result in the need to interpolate the data within the interval of risk aversion in which the median choice falls. In both HL and EG, though, the median choice falls by chance very close to a cutoff point and therefore no interpolation is necessary.

Table 3 Experimental results (consistent subjects only) and comparison with the literature

	HL	EG	GP	BRET
N	88	88	86	88
Inconsistent	20	–	–	1
Type of choice	# Risky choices	Chosen lottery	Amount invested	Stopping point
Choice set	[0, 10]	[1, 5]	[0, 4]	[0, 100]
Choice				
Median	3/4	2/3	2.5	40
Mean	3.91	2.79	2.54	40.05
St.Dev.	1.94	1.29	0.95	13.98
r				
Median	0.45	0.07	0.75	0.66
Mean	0.42	0.09	0.73	0.78
St.Dev.	0.49	0.66	0.18	0.51
% Risk averse	83.82	81.82	80.20	73.50
SOEP	5.29	5.27	5.16	5.03
DOSPERT	3.53	3.57	3.46	3.42
From the literature				
Mean choice	4.37	3.45	2.23	44.69
Mean r	0.60	0.47	0.70	0.81

Sources of data in 'from the literature' panel. HL: 48 studies, 5935 consistent subjects from (Filippin and Crosetto 2015); EG: 1 study, 256 subjects from (Eckel and Grossman 2008a); GP: 16 studies, 1955 subjects from (Charness and Gneezy 2012); BRET: 1 study, 2975 subjects (lower stakes and different treatments mixed together) from (Crosetto and Filippin 2013)

and GP. When a large survey was not available, we reported the results of the original paper introducing the task. Overall, our subjects proved to be on average considerably more risk averse in EG and HL, while in BRET and GP they were fairly in line with the average subject of other studies.

4.2 Comparison with the simulation results

The question that naturally follows is how do the results of the experiment compare with the results of the simulation exercise carried out in Sect. 3.2. In other words, are the idiosyncratic features of the task enough to account for the difference in the estimated risk preferences? This comparison is methodologically not trivial, since in the simulation we *know* the underlying data generating process, while in the experiment the original preferences are *unknown*. As long as the two underlying true distributions are not too different, however, the comparison can provide insights about task-specific differences that go over and beyond the bias introduced by the mere mechanics of the tasks.

Figure 5 compares the distribution of the $\hat{r}$ retrieved from the experimental data to the one generated by the simulations with $\mu = 0.6$, which better fits our data.[17]

[17] This set of parameters leads to a $\hat{r}$ distribution which is similar to what emerges by finding the best-fitting parameters from our experimental data $N(0.7123, 0.802)$.

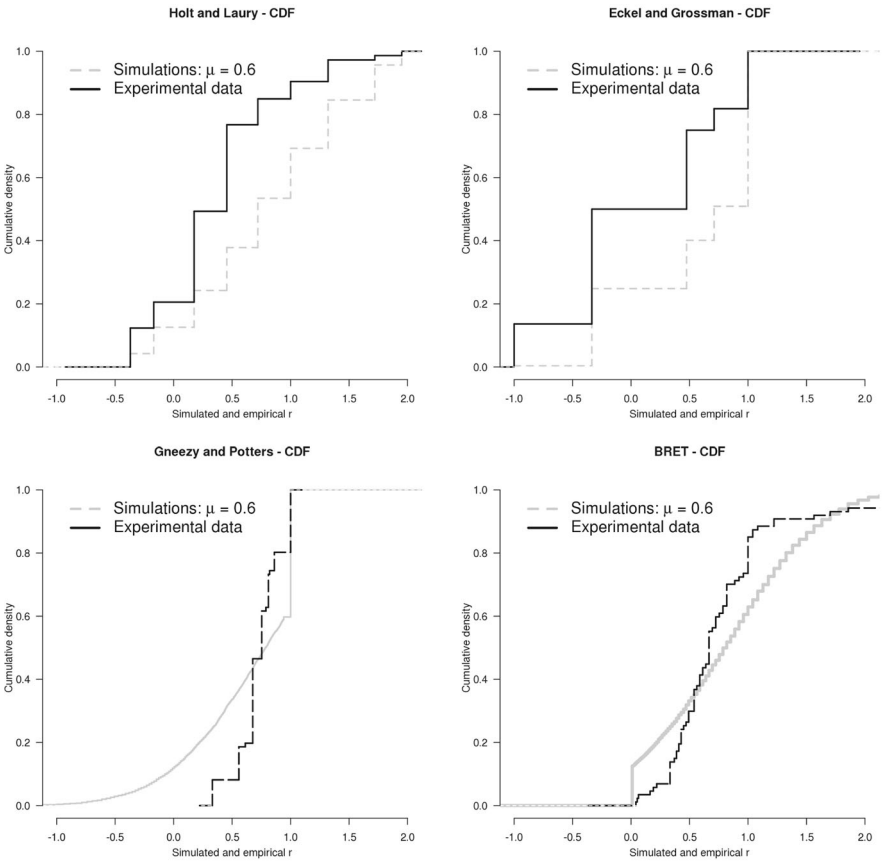


Fig. 5 Comparison between simulated and experimental results

Several insights can be derived from the visual comparison. First, our experimental subjects are more risk averse than the virtual ones in the case of HL and EG. In HL there is a mass of experimental subjects making only one risky choice, and in general a larger fraction of risk averse choices. In EG 13.7 % of subjects chose the degenerate safe lottery, while no virtual agent possessed such an extreme degree of risk aversion as that necessary to make the safe choice.

Second, in the GP and BRET we see a higher concentration of subjects in the central part of the $\hat{r}$ distribution, because experimental subjects did not submit choices as extreme as those submitted in the simulations. In BRET, choices ranged from 4 to 72 boxes in the experiment, and from 0 to 78 in the simulations. Consequently, the predicted mass at $\hat{r} = 0$ for the BRET, attracting all those characterized by $r < 0$ in the simulations, does not have a counterpart in the experimental data. Similarly, the lowest amount invested in GP was 1 euro, while this was zero in the simulations. As a result, the minimum estimated level of r is equal to 0.33. At the same time, the mass of risk loving subjects that should

concentrate at $\hat{r} = 1$ in GP is lower than what the simulations predict. While the first finding is consistent with the fact that extreme risk preferences are less frequent in the real world than hypothesized by the simulations, the second seems to support that subjects avoid extreme choices in the opportunity set.

In general, the comparison between the simulations and the experimental results shows that the mechanics of the task can in part explain the observed differences across tasks. The results are overall consistent in terms of ranking of the elicitation methods as far as the variance of the estimate is concerned, especially that EG tends to overestimate the variance, while GP underestimates it. In constrast, this is not so much the case as far as the average r is concerned. From this point of view, the experimental results (both ours and in the literature) seem to deliver systematically lower estimates of r in HL and in EG. This make us believe that different tasks likely elicit different types of preferences, something that we explore in the Discussion section.

4.3 Gender differences

Many studies report gender differences in risk attitudes, showing that females are significantly more risk averse than men. This is also the main message conveyed by the surveys available in the literature (Charness and Gneezy 2012; Eckel and Grossman 2008b; Croson and Gneezy 2009). These findings have been disputed, e.g., by Nelson (2014, 2015), while Filippin and Crosetto (2015) show in detail that the likelihood of observing gender differences strongly correlates with the task used. For instance, in GP gender differences are systematic and substantial. Males invest significantly more than females in almost all the experiments analysed, and often such a difference is about 10–15 % of the initial endowment. Similar findings emerge with the EG task. However, the picture changes sharply in other tasks. Using HL gender differences are the exception rather than the rule. In the BRET the behaviour of males and females does not differ.

Table 4 reports the average choices by gender in our experiment. In GP our sample of males invested significantly more than females. In percentage terms, they allocated to the risky asset 73.3 and 56.3 % of the endowment, respectively, i.e., fractions virtually identical to those reported for instance by Charness and Gneezy (2010). We find similar results also for the EG method. In our experiment, the average choice is 3.22 for males versus 2.34 for females, (3.63 vs. 2.95 in Eckel and Grossman (2008a)). In contrast, our replication of the HL task finds no gender

Table 4 Average choice by gender: higher numbers stands for lower risk aversion

	Males		Females		Mann Whitney
	N	Mean choice	N	Mean choice	
HL	31	3.74	42	3.57	$\|p\| = 0.9090$
EG	45	3.22	43	2.34	$\|p\| = 0.0050$
GP	37	2.93	49	2.25	$\|p\| = 0.0021$
BRET	32	39.72	55	40.25	$\|p\| = 0.7913$

differences, which is in line with the vast majority of the contributions in the literature, as emphasised by Filippin and Crosetto (2015). Similarly, in the BRET the behaviour of men and women is statistically indistinguishable.

4.4 Maximum likelihood estimation

Since subjects' choices might include stochastic elements, we estimate with maximum likelihood a structural model including a term to capture noise, as done, among others, by Dave et al. (2010), Hey and Orme (1994), Holt and Laury (2002), Harrison et al. (2007). We assume that for each task the subject is an expected utility maximiser who can make an error in comparing the expected utility of the lotteries she faces. We assume a Fechner error specification, as done in Hey and Orme (1994). This amounts to assuming that subjects make a normally distributed error when comparing the utilities of each option. The subjects choose between a 'left', safer, (L) and a 'right', riskier, (R) lottery based on the difference between their expected utilities, plus a normally distributed error term:

$$\Delta EU = EU_R - EU_L + \varepsilon, \text{ in which } \varepsilon \sim N(0, \sigma).$$

This implies that the probability of choosing lottery R is given by the cumulative distribution function of a standardised normal in $(EU_R - EU_L)/\sigma$.

To carry out the estimation we need to express the tasks as a series of binary choices between pairs of lotteries. In the HL task, where subjects directly evaluate ten binary choices, raw data can be directly used, including data from multiple switchers. Data from the other tasks need instead to be transformed. As done by Dave et al. (2010), we use the revealed preferences of our subjects as the key to building binary choices from tasks with single choices. This procedure implies assuming that preferences are single-peaked.

For EG, this implies that from the subject's choice of, say, lottery 4, we not only derive that 4 was preferred to 3 but also that 3 was preferred to 2, etc. The same logic applies to GP and the BRET. In GP, even if the choice variable is virtually continuous, no subject chose at a detail finer than one decimal point. Hence we transformed the GP data into 40 binary choices, applying the single-peaked property. In the BRET we assume that the decision maker at each second chooses between stopping on the current lottery versus waiting one more second and proceeding to the next. This allows us to recode the BRET as a task with 100 binary choices.

Given these data, we estimate separately for each task a structural model of choice using maximum likelihood and clustering standard errors by subject. We assume that subjects are expected utility maximisers characterised by CRRA preferences $u(x) = x^r$, allowing for heterogeneity by gender of both the random component in the decision σ and risk attitudes r. We performed the estimation with Stata, following Harrison and Rutström (2008).

Table 5 shows the results, which are in line with what is displayed in the previous sections. Estimated risk aversion is stronger in HL and EG once we consider gender differences in the latter. The behaviour of males and females is significantly

Table 5 Maximum Likelihood structural model estimation

	Log-likelihood	Coefficient	Estimate	SE	p value
HL	−391.15	r	0.441	0.080	0.000
		r_{female}	−0.092	0.118	0.435
		σ	0.456	0.138	0.001
		σ_{female}	−0.048	0.183	0.791
EG	−194.62	r	0.695	0.042	0.000
		r_{female}	−0.253	0.072	0.000
		σ	0.264	0.035	0.000
		σ_{female}	−0.136	0.066	0.039
GP	−1546.79	r	0.868	0.016	0.000
		r_{female}	−0.098	0.025	0.000
		σ	0.012	0.001	0.000
		σ_{female}	−0.004	0.002	0.014
BRET	−2584.71	r	0.673	0.106	0.000
		r_{female}	0.063	0.129	0.625
		σ	0.094	0.039	0.017
		σ_{female}	0.013	0.061	0.828

different in GP and EG, while it is indistinguishable in HL and BRET. The noise parameter σ is considerably higher in HL than in EG, GP, and BRET. This is mainly due to the inclusion of the multiple switchers in the analysis. An estimation carried out excluding the inconsistent subjects reveals for the HL a considerably lower noise ($\sigma = 0.30$). Nothing changes, instead, in the BRET when removing the inconsistent subject. It is worth stressing how the random component is lower for females in the few cases in which it is significant, something that seems to contradict the common wisdom about gender differences in numeracy.

4.5 Noise

In the simulations the introduction of a fixed amount of noise in the preferences of the virtual subjects generates specific biases in the risk preferences retrieved by each task. The simulations rely on the assumption that the tasks themselves do not have any influence on the noisiness of the choices. We now analyse whether this is indeed the case, or if instead the tasks, as a result of their different structure, are likely to generate additional amounts of noise.

One way of measuring the noisiness of the tasks is to look at the variance of the measured coefficient of risk aversion. The choices collected under the different elicitation methods are not directly comparable because of the differences in the number of alternatives available as well as in the likelihood of making choices implying extreme values of the coefficient of risk aversion. However, our experimental results display a pattern that does not systematically differ from the simulations (see Fig. 4; Table 3). Hence, this measure does not signal critical

situations, although it should be taken into account that the measure for HL includes the consistent subjects only.

The noisiness of a task might be reflected also in the level of observed inconsistencies. In HL we observe about 17 % of the subjects switching more than once and about 5 % always choosing the safe lottery. In the BRET we observe one dominated choice. The other tasks do not allow to detect inconsistent choices, so that, from this point of view, an exhaustive comparison cannot be made. The fact that HL generates a high number of inconsistent subjects looks as a double-edged sword. On the one hand it reveals a noisier nature of the task resulting in a sizeable loss of data. On the other hand, by signaling through inconsistent choices the subjects who most likely did not understand the task, it allows to focus on a cleaner subset of subjects in which choices do not look particularly noisier than in the other tasks.

The noisiness of a task may also be approximated by means of the noise parameter of the structural Maximum Likelihood estimations of Sect. 4.4. This approach has the advantage of providing a measure of the effect of inconsistent choices in HL. The GP task displays the lowest noise parameter, followed by EG and the BRET, while HL is characterized by the highest value. A high share of noise can be attributed in HL to inconsistent choices. Excluding inconsistent subjects reveals for HL a noise parameter more in line with the other tasks ($\sigma = 0.30$).

In principle, subjects' understanding of the task might influence the noisiness of the results. Existing studies (Dave et al. 2010; Charness and Viceisza 2011), for instance, report that HL is more difficult to understand than other tasks, particularly so when the task is performed by subjects with low numeracy. As a proxy for the understanding of the tasks we included a specific item in the questionnaire, directly asking the subjects to report how difficult they found the task on a Likert scale from very simple (1) to very difficult (10). Results show that the tasks are not perceived as significantly easier or difficult, with the exception of the BRET that is perceived as significantly easier than HL (at 10 %). When excluding inconsistent subjects in HL, however, the perceived difficulty of the two tasks becomes indistinguishable.

Summarizing, differently from the literature (see, for instance Dave et al. 2010) we do not find evidence of great differences in noise and understanding across methods. HL generates more inconsistent choices, but once these are eliminated it scores similarly to the other tasks. At least in our student sample there seems to be no gain in terms of understanding in exchange for the loss of precision implied by EG as compared to the GP and the BRET.

4.6 Correlation with questionnaires

After performing the risk elicitation method, all subjects answered both the SOEP risk attitude question as well as the DOSPERT questionnaire. This allowed us to test the correlation of choices made in the tasks with the questionnaires as well as that between the DOSPERT and the SOEP measures.

Across all tasks, the SOEP is highly and significantly correlated with the overall DOSPERT score ($\rho = 0.57$, p value < 0.001). SOEP correlates also significantly (in all cases, p value < 0.001) with all the DOSPERT subscales, as detailed in Table 6.

Table 6 Pearson's correlation coefficients of the SOEP and DOSPERT questionnaires

	Dospert-all	Do-recreational	Do-health	Do-gamble	Do-invest	Do-ethics	Do-social
SOEP	0.5702	0.5511	0.4084	0.3656	0.3125	0.2134	0.2017

All coefficients significant, p value <0.001

The DOSPERT subscales also correlate significantly among each other. Hence, subjects are overall consistent when self-reporting their risk attitudes, either directly (SOEP) or via questions on several domains of their lifestyle (DOSPERT).

For each task we test in two ways how the answers to the questionnaire are correlated with the incentivised choices. First, we compute a battery of pairwise Pearsons's correlation coefficients between choices and questionnaires. Second, after running a linear regression of the choices in each task on the observed demographics (age and gender) as a benchmark, we include each questionnaire separately in the specification,[18] measuring the increase of the adjusted R^2. Roughly speaking, this indicator, that we name $\Delta\, adj.R^2$, measures how much of the incentivized choices is explained by each questionnaire. Results are shown in Table 7, where the $\Delta\, adj.R^2$ is expressed in percentage points.

As expected, the two indexes are well aligned: all tasks show a low correlation, if any, with the questionnaires. The amount of variance explained is also fairly low: the adjusted R^2 of the regressions never increases by more than 10 %.[19] This is especially true for the BRET, which does not correlate with any measure. It is also true for the HL, which weakly correlates with the SOEP (at 10 %) and the general DOSPERT (at 5 %), though without any appreciable contribution to the explained variance of the choices. Only the EG task shows positive correlations across the board, although the magnitude of variance explained is low. The GP task significantly correlates with the investment and gambling sections of the DOSPERT, also displaying an appreciable increase in the amount of variance explained (about 7.7 %), but it does not correlate with the overall DOSPERT score nor with the SOEP. This result might be due to the fact that the GP task is the only one that has an investment frame.

Given that assigning a cardinal interpretation to the answers in the questionnaires may be seen as methodologically dubious we performed, as a robustness check, a similar exercise using Spearman's coefficients, which simply rely upon the rank of choices and answers, finding an even lower significance of correlations.

Such low correlations are commonly found in the literature.[20] Deck et al. (2013) report of a failed attempt of explaining differences in behavior across several tasks, including HL, Balloon and a custom version of EG, with the relevant domain of the

[18] The Do-Investment and the Do-Gamble are added together because belonging to the same domain.

[19] No task correlates significantly with the Dospert subscales for *health-related*, *ethical* and *recreational* risk, and hence we do not include them in Table 7.

[20] An exception is the aforementioned contribution of Dohmen et al. (2011), who compare the SOEP question with an incentivised lottery scheme. Also in their case, however, the fraction of variance explained is fairly low (about 6 %).

Table 7 Correlation with questionnaires and explained variance for each task

	N	Indicator	Soep	Dospert	Do-investment	Do-gamble	Do-social
HL	73	Correlation	0.23*	0.25**	0.12	0.16	0.15
		Δ adj. R^2	≤ 0	≤ 0	≤ 0		≤ 0
EG	88	Correlation	0.30***	0.30***	0.22**	0.33***	0.30***
		Δ adj. R^2	2.7	1.6	1.9		5.2
GP	86	Correlation	0.13	0.17	0.36***	0.33***	−0.13
		Δ adj. R^2	≤ 0	≤ 0	7.7		2.2
BRET	87	Correlation	0.03	0.06	0.05	−0.01	−0.06
		Δ adj. R^2	≤ 0	≤ 0	≤ 0		≤ 0

Δ adj. R^2 expressed in extra percentage points

Significance thresholds * 0.1, ** 0.05, *** 0.01

DOSPERT (in particular, Do-Gamble and Do-Invest). Anderson and Mellor (2009) find that correlation is low between HL and a set of hypothetical questions, taken from Barsky et al. (1997), about accepting a new job with an uncertain wage. The BRET results are qualitatively similar to those detailed in Crosetto and Filippin (2013). The EG task has been found to show somewhat higher correlation with various other measures (Ball et al. 2010), but there is no correlation with survey measures of risk perceptions and risk taking over several domains (e.g., safety, injury, insurance). The GP task has not frequently been run alongside questionnaires, but Charness and Viceisza (2011) reported dramatically different behavior in the SOEP as compared to GP in a field experiment in rural Senegal.

Following previous findings in the literature on lottery choices (Grossman and Lugovskyy 2011) and public good games (Perugini et al. 2010), we also tested whether there are gender differences in the way questionnaire answers correlate with choices in incentivised tasks. In line with existing evidence, we find that when such differences are present, the correlation is usually stronger for males, while it is either not significant or much weaker for females.

5 Discussion and conclusion

Despite decades of research and the development of many different methods for eliciting risk preferences, it is still unclear which task yields their most correct estimate, and, to some extent, the nature of what the tasks are measuring. The rather bleak picture that a reader gets from the literature is one with low consistency, low if any correlation either with questionnaires or with behavior in other experiments, and a general feeling of indecision as to what is actually being measured.

In this paper we performed an in-depth comparison of four risk elicitation methods, with the aim of investigating if *the tasks themselves* might be heterogeneous enough to generate (at least some of) the observed instability of behavior. We focus on simple and fast tasks that are usually run in order to control

for risk attitudes in experimental sessions in the laboratory or in the field. We include a multiple price list a la Holt and Laury (2002), an ordered lottery choice a la Eckel and Grossman (2002), the Investment Game by Gneezy and Potters (1997), and the Bomb Risk Elicitation Task (Crosetto and Filippin 2013).

The results of a simulation exercise show that the tasks as they are commonly implemented generate systematically biased estimates *by construction*, explaining at least some of the variability. This bias is due to three factors: the range of risk preferences that the task can identify, the precision of the estimate, and the shape of the function mapping the choices in the task into risk aversion parameters. GP and, to a less prominent degree, EG yield downward biased estimates of the coefficient of risk aversion r. This happens on the one hand because they they cannot identify risk loving preferences and on the other hand because for higher level of noise they translate risk averse choices into extremely low values of r. The BRET yields instead higher estimates for higher levels of noise due to the hyperbolic function between choices and r. HL is instead virtually linear in the choice-r space, and this makes its estimate robust to stochastic decisions. Factoring this finding into the analysis reduces the observed heterogeneity of estimates, but does not eliminate it. In fact, an experimental replication, carried out between subjects and with a strict one-shot procedure, shows that the simulations are able to predict quite accurately the standard deviation of results in the various tasks, especially that EG tends to overestimate the variance, while GP underestimates it.

The relative ranking of the estimated risk aversion across tasks is also detected – with EG biasing results to a lower mean, and the BRET to a higher one. However, the estimates yielded by the tasks vary over and above what is predicted by the simulations, even if our subjects' self-reported risk attitudes were comparable across treatments. In particular, the experimental results (both ours and in the literature) seem to deliver systematically lower estimates of r in HL and in EG. A possible interpretation is that the tasks, because of their different characteristics, elicit different preferences, besides providing a different measure of the same preferences.

Table 8 summarizes the main dimensions along which the tasks differ. The effects of precision, number and completeness of choices have been analyzed by means of the simulations. The tasks, though, vary along two more dimensions that might have behavioral consequences not captured by the simulations: the availability of a safe option and a single- versus multiple-choice environment.

Table 8 Summary of the main differences among tasks

	Precision (categories of r)	Parsimony (no. of choices)	Completeness (r range)	Compound lottery	Safe option
HL	10	10	Yes	Yes	Mildly focal
EG	5	1	No	No	Focal
GP	Almost continuous	1	No	No	Focal
BRET	Almost continuous	1	Yes	No	No

We now briefly analyze these two characteristics with the aim of rationalizing at least partly the differences between the simulations and the experiment.

Availability of a riskless alternative. Some of the tasks include a focal safe alternative in their choice set, while others do not. The presence of a safe option might affect the behavior for several reasons. First, it could induce certainty effects. Second, it could act as a reference point against which uncertain outcomes, both higher (gains) and lower (losses), can be evaluated. Alternatively, a safe amount could increase the salience of regret (Loomes and Sugden 1982). Whatever the ultimate cause, the presence of a safe option is likely to induce failures of Expected Utility Theory (see, for instance, Andreoni and Sprenger 2011, 2012; Camerer 1992; Harless and Camerer 1994; Starmer 2000). This happens because subjects seems to disproportionally prefer a certain outcome as compared to what the independence axiom of EUT would predict.

Among the task discussed in this paper, EG and GP feature a clear safe option and the BRET does not. HL does not strictly speaking include a safe option, but the lower outcome of the safer lottery can play a similar role, since it represents the minimum amount that can be earned with probability one (provided that Option A is chosen).

The differences along this dimension could partially explain the discrepancies between the simulations and the experiments, in particular for the EG task. The degenerate lottery that yields 4 euro with certainty is never observed in the deterministic and stochastic simulations. In the simulations including random decisions it is observed, as expected, about 2 % of the times. In contrast, 13.7 % of our experimental subjects in the EG treatment choose the safe option. Within a CRRA framework, such a choice implies an incredibly low coefficient $(r<-1)$.

Another intriguing result is that both our replications and the results in the literature reveal a clear correlation between the likelihood of observing gender differences in risk attitudes and the availability and focality of a safe option in the choice set. In fact, gender differences appear systematically only in the tasks in which subjects have a clear opportunity to avoid any risk by opting for a safe choice such as the GP and EG. This observation led us to investigate experimentally whether the availability of a riskless alternative is what causes the behavior of males and females to differ, and in a companion paper we display evidence weakly supporting this conjecture (Crosetto and Filippin 2014). Significant gender differences emerge when adding a safe option to the BRET and to HL. In contrast, such differences do not disappear when removing the riskless alternative from the menu of lotteries in EG, despite females being significantly more likely than males to choose the degenerate lottery when it is available.

Compound versus simple lotteries. Experimental subjects have been shown not to be indifferent between lotteries implemented with different randomization devices that should instead be equivalent under the Reduction Axiom (Kaivanto and Kroll 2011). Moreover Kaivanto and Kroll (2012) show that subjects are more risk averse when faced with the gamble framed as a compound rather than simple lottery. This results might explain the lower estimate that characterizes HL in the experimental data as compared to the simulation exercise. In fact, HL is built on 10 choices and

implements a compound lottery: a random draw selects the row to be paid, and a second random draw selects the outcome of the chosen lottery.

The results of Kaivanto and Kroll (2012) also suggest the administering of one-shot rather than repeated tasks. Unfortunately, it is often impractical to implement pure one-shot experiments, despite theoretical and empirical implications of the payment protocol (Harrison and Swarthout 2014). Moreover, when using risk elicitation tasks as controls in other experiments, it is unavoidable to have both the risk elicitation method and the main task in the same session. In this case the pay-one-at-random protocol cannot be pursued, given that the main task *must* be incentivised. Hence, the only solution is either not to pay the risk task, with obvious concerns about the reliability of the results obtained, or to pay both tasks introducing biases due to wealth effects.

Limited to those risk elicitation methods that require one choice only, these problems can be solved implementing the elicitation mechanism *before* the main task, but resolving the uncertainty and thus determining the corresponding payoff only *afterwards*. This method does not implement a compound lottery as long as both tasks are rewarded, and minimises wealth effects in the main task since the payoff of the risk task is not yet determined. From this point of view, the BRET is the mechanism that minimises the predictability of the payoff since it has neither a safe option nor any minimum amount that can be earned for sure.

Summarizing, despite all the analysis, we cannot single out one best task. A single all-purpose task that delivers an estimate of risk preferences that is valid in all applications might well not exist. We might nonetheless use the results of our analysis to provide a small list of points to guide researchers in choosing the best risk elicitation task according to their needs.

A risk elicitation task must be simple to understand, to avoid adding noise to the data, especially in contexts in which the numeracy of the subjects is an issue. In this case, HL might be troublesome. Once inconsistent are excluded HL shows similar noise levels of other tasks, and in our subject pool it is not perceived as more difficult than the competition. Nonetheless, it does result in data loss of the order of 15 to 20 %. As far as the other tasks are concerned, there seems to be no gain in terms of understanding for the loss of precision implied by EG as compared to GP and the BRET, at least within a pool of subjects predominantly composed by students and therefore with relatively high numeracy.

Another relevant feature of a risk elicitation task is its correlation with self-reported risky behaviors.[21] In fact, budgetary reasons could lead the researcher to opt for a simple questionnaire. Unfortunately, the performance of the tasks is rather poor across the board, given that the correlation with self-reported answers is low in the few circumstances in which they are not orthogonal. The task performing best along this dimension is EG. For applications to the financial domain, GP seems a good compromise, as it shows some correlation with self-reported financial risk. Moreover, EG and GP share many important features (range of the estimates,

[21] Ideally, the tasks should capture relevant features of real-life risky behaviour, something that we do not investigate in this paper, but that is definitely an interesting line of future reseach.

availability of a safe option, comprehension) but the latter has a clear relative advantage in terms of precision.

Whether in a particular decision environment the availability of a safe option is a relevant feature or not, together with the importance of disentangling preferences of risk loving subjects, should guide the choice between GP and EG on the one hand and HL and the BRET on the other. In particular, if the goal is that of estimating risk aversion in an Expected Utility framework the tasks entailing a safe option should be turned down in favor of either the BRET or HL. The former is better located in the understandability-precision trade-off and it is particularly indicated in auction experiments, as it is isomorphic to a first price auction against an opponent who bids uniformly. The latter delivers more robust estimates in case of uncertain preferences once we are ready to accept its inherent data loss.

Acknowledgments We are grateful to the Max Planck Institute of Economics (Jena) for financial and logistic support and to Denise Hornberger, Nadine Marmai, Florian Sturm, and Claudia Zellmann for their assistance in the lab. We would like to thank the members of the ESA mailing list for useful references and participants to seminars in Strasbourg, Middlesex, Paris 1 Sorbonne, MPI Jena, DIW Berlin, INRA Rennes and Göttingen as well as the audience of the IMEBE conference in Madrid and the BEELAB conference in Florence and two anonymous referees for useful comments. All remaining errors are ours.

References

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

Andersen, S., Harrison, G., Lau, M., & Rutström, E. (2006). Elicitation using multiple price list formats. *Experimental Economics*, *9*, 383–405. doi:10.1007/s10683-006-7055-6.

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

Andreoni, J., & Sprenger, C. (2011). Uncertainty equivalents: Testing the limits of the independence axiom. NBER Working Papers 17342, National Bureau of Economic Research, Inc.

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

Ball, S., Eckel, C., & Heracleous, M. (2010). Risk aversion and physical prowess: Prediction, choice and bias. *Journal of Risk and Uncertainty*, *41*(3), 167–193.

Bardsley, N. (2008). Dictator game giving: Altruism or artefact? *Experimental Economics*, *11*(2), 122–133.

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

Becker, G., DeGroot, M., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, *9*, 226–236.

Binswanger, H. P. (1981). Attitudes toward risk: Theoretical implications of an experiment in rural India. *The Economic Journal*, *91*(364), 867–890.

Blais, A. R., & Weber, E. U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, *1*, 33–47.

Bruner, D. (2009). Changing the probability versus changing the reward. *Experimental Economics*, *12*(4), 367–385.

Camerer, C. F. (1992). Recent tests of generalizations of expected utility theory. In W. Edwards (Ed.), *Utility theories: Measurements and applications studies in risk and uncertainty* (pp. 207–251). Boston, MA: Kluwer Academic Publishers.

Charness, G., & Gneezy, U. (2010). Portfolio choice and risk attitudes: An experiment. *Economic Inquiry*, *48*(1), 133–146.

Charness, G., & Gneezy, U. (2012). Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization*, *83*(1), 50–58.

Charness, G., & Viceisza, A. (2011). Comprehension and risk elicitation in the field: Evidence from rural Senegal. IFPRI discussion papers 1135, International Food Policy Research Institute (IFPRI)

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

Cox, J. C., Roberson, B., & Smith, V. L. (1982). *Theory and behavior of single object auctions*. Greenwich: JAI Press.

Crosetto, P., & Filippin, A. (2013). The 'bomb' risk elicitation task. *Journal of Risk and Uncertainty*, *47*(1), 31–65.

Crosetto, P., & Filippin, A. (2014). Experimental evidence on the cause of gender differences in risk attitudes. mimeo

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

Csermely, T., & Rabas, A. (2014). How to reveal people's preferences: Comparing time consistency and predictive power of multiple price list risk elicitation methods. Department of economics working papers, Vienna University of Economics and Business, Department of Economics.

Dave, C., Eckel, C., Johnson, C., & Rojas, C. (2010). Eliciting risk preferences: When is simple better? *Journal of Risk and Uncertainty*, *41*(3), 219–243.

Deck, C., Lee, J., & Reyes, J. (2010) Personality and the consistency of risk taking behavior: Experimental evidence. Working Papers 10–17, Chapman University, Economic Science Institute.

Deck, C., Lee, J., Reyes, J. A., & Rosen, C. C. (2013). A failed attempt to explain within subject variation in risk taking behavior using domain specific risk attitudes. *Journal of Economic Behavior & Organization*, *87*, 1–24.

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*(3), 522–550.

Eckel, C. C., & Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, *23*(4), 281–295.

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

Eckel, C. C., & Grossman, P. J. (2008b). *Men, women and risk aversion: Experimental evidence* (Vol. 1, pp. 1061–1073, chap 113). Handbook of experimental economics results. Amsterdam: Elsevier.

Filippin, A., & Crosetto, P. (2015). A reconsideration of gender differences in risk attitudes. *Management Science*, forthcoming.

Garcia-Gallego, A., Georgantzis, N., Jaramillo-Gutiérrez, A., & Parravano, M. (2012). The lottery-panel task for bi-dimensional parameter-free elicitation of risk attitudes. Technical report, Department of Economic Theory and Economic History of the University of Granada.

Gneezy, U., & Potters, J. (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics*, *112*(2), 631–645.

Greiner, B. (2004). The online recruitment system ORSEE 2.0—A guide for the organization of experiments in economics. Working Paper Series in Economics 10, University of Cologne, Department of Economics.

Grossman, P. J., & Lugovskyy, O. (2011). An experimental test of the persistence of gender-based stereotypes. *Economic Inquiry*, *49*(2), 598–611.

Harbaugh, W., Krause, K., & Vesterlund, L. (2010). The fourfold pattern of risk attitudes in choice and pricing tasks. *The Economic Journal*, *120*(545), 595–611.

Harless, D. W., & Camerer, C. F. (1994). The predictive utility of generalized expected utility theories. *Econometrica*, *62*(6), 1251–1289.

Harrison, G., & Swarthout, J. (2014). Experimental payment protocols and the bipolar behaviorist. *Theory and Decision*, *77*(3), 423–438.

Harrison, G. W. (1990). Risk attitudes in first-price auction experiments: A Bayesian analysis. *The Review of Economics and Statistics*, *72*(3), 541–546.

Harrison, G. W., & Rutström, E. E. (2008). Risk aversion in the laboratory. In J. C. Cox & G. W. Harrison (Eds.), *Risk aversion in experiments, research in experimental economics* (Vol. 12, pp. 41–196). Bradford: Emerald Group Publishing Limited.

Harrison, G. W., Lau, M. I., & Rutström, E. E. (2007). Estimating risk attitudes in Denmark: A field experiment. *Scandinavian Journal of Economics*, *109*(2), 341–368.

Hey, J. D., & Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica*, *62*(6), 1291–1326.

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

Isaac, R., & James, D. (2000). Just who are you calling risk averse? *Journal of Risk and Uncertainty*, *20*(2), 177–187.

Kaivanto, K., & Kroll, E. B. (2011). Negative recency, randomization device choice, and reduction of compound lotteries. Working Paper Series in Economics 22, Karlsruhe Institute of Technology (KIT), Department of Economics and Business Engineering.

Kaivanto, K., & Kroll, E. B. (2012). Alternation bias and reduction in st. petersburg gambles: An experimental investigation. Technical report, Lancaster University.

Lejuez, C., Read, J., Kahler, C., Richards, J., Ramsey, S., Stuart, G., et al. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, *115*, 482–493.

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *Economic Journal*, *92*(368), 805–824.

Nelson, J. A. (2014). Are women really more risk-averse than men? A re-analysis of the literature using expanded methods. *Journal of Economic Surveys*, *29*(3), 566–585.

Nelson, J. A. (2015). Not-so-strong evidence for gender differences in risk taking. *Feminist Economics*, forthcoming.

Perugini, M., Tan, J. H. W., & Zizzo, D. J. (2010). Which is the more predictable gender? Public good contribution and personality. *Economic Issues Journal Articles*, *15*(1), 83–110.

Reynaud, A., & Couture, S. (2012). Stability of risk preference measures: Results from a field experiment on French farmers. *Theory and Decision*, *73*(2), 203–221.

Slovic, P. (1966). Risk-taking in children: Age and sex differences. *Child Development*, *37*(1), 169–176.

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

van Rossum, G. (1995). Python reference manual. CWI Report CS-R9525.

Wagner, G. G., Frick, J.R., & Schupp, J. (2007) The german socio-economic panel study (soep): Scope, evolution and enhancements. SOEPpapers on Multidisciplinary Panel Data Research 1, DIW Berlin, The German Socio-Economic Panel (SOEP).

Wakker, P. P., (2008) Explaining the characteristics of the power (CRRA) utility family. *Health Economics*, *17*(12), 1329–1344.

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