Crosetto, Paolo; Filippin, Antonio

Working Paper

A theoretical and experimental appraisal of five risk elicitation methods

SOEPpapers on Multidisciplinary Panel Data Research, No. 547

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Crosetto, Paolo; Filippin, Antonio (2013) : A theoretical and experimental appraisal of five risk elicitation methods, SOEPpapers on Multidisciplinary Panel Data Research, No. 547, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/73671

SOEP Das Sozio-oekonomische Panel

547 2013

SOEPpapers
on Multidisciplinary Panel Data Research

A Theoretical and Experimental Appraisal of Five Risk Elicitation Methods

Paolo Crosetto and Antonio Filippin

DIW SOEP

SOEPpapers on Multidisciplinary Panel Data Research
at DIW Berlin

This series presents research findings based either directly on data from the German Socio-Economic Panel Study (SOEP) or using SOEP data as part of an internationally comparable data set (e.g. CNEF, ECHP, LIS, LWS, CHER/PACO). SOEP is a truly multidisciplinary household panel study covering a wide range of social and behavioral sciences: economics, sociology, psychology, survey methodology, econometrics and applied statistics, educational science, political science, public health, behavioral genetics, demography, geography, and sport science.

The decision to publish a submission in SOEPpapers is made by a board of editors chosen by the DIW Berlin to represent the wide range of disciplines covered by SOEP. There is no external referee process and papers are either accepted or rejected without revision. Papers appear in this series as works in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be requested from the author directly.

Any opinions expressed in this series are those of the author(s) and not those of DIW Berlin. Research disseminated by DIW Berlin may include views on public policy issues, but the institute itself takes no institutional policy positions.

The SOEPpapers are available at
http://www.diw.de/soeppapers

ISSN: 1864-6689 (online)

German Socio-Economic Panel Study (SOEP)
DIW Berlin
Mohrenstrasse 58
10117 Berlin, Germany

Contact: Uta Rahmann | soeppapers@diw.de

A Theoretical and Experimental Appraisal of Five Risk Elicitation Methods ☆

Paolo Crosetto[a], Antonio Filippin[b,c]

[a]*Max Planck Institute for Economics, Kahlaische Straße 10, 07745 Jena, Germany.*
[b]*University of Milan, Department of Economics, Via Conservatorio 7, 20122 Milano, Italy*
[c]*Institute for the Study of Labor (IZA), Schaumburg-Lippe-Str. 5-9, 53113 Bonn, Germany*

Abstract

We perform a comparative analysis of five incentivized tasks used to elicit risk preferences. Theoretically, we compare the elicitation methods in terms of completeness of the range of the estimates as well as their precision, the likelihood of triggering loss aversion, and problems arising when multiple choices are required. Using original data from a homogeneous population, we experimentally investigate the distribution of estimated risk preferences, whether they differ by gender, and the complexity of the tasks. We do so using both non-parametric tests and a structural model estimated with maximum likelihood. We find that the estimated risk aversion parameters vary greatly across tasks and that gender differences appear only when the task is more likely to trigger loss aversion.

JEL Classifications: C81; C91; D81

Keywords: Risk attitudes, Elicitation methods, Experiment

1. Introduction

Since uncertainty is a pervasive phenomenon in economic decisions, properly measuring attitudes toward risk is crucial in drawing conclusions from economic theory. Not surprisingly, experimental economists have proposed a strikingly long list of methods to experimentally measure risk preferences in a fast, simple, and reliable way. Unfortunately, the measured risk attitudes have been shown to be highly task-specific, calling for an analysis of the reasons for this lack of robustness. This paper aims at providing such an analysis by comparing five risk elicitation methods both theoretically and empirically.

Usually risk preferences are measured 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 descision (Charness and Gneezy, 2010; Gneezy and Potters, 1997). Alternatively, subjects might be asked to make *multiple* decisions between pairs or sets of

☆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 a seminar in Strasbourg for useful comments. All remaining errors are ours.

Contact: paolo.crosetto@gmail.com (Paolo Crosetto), antonio.filippin@unimi.it (Antonio Filippin)

risky lotteries presented in an established (Garcia-Gallego et al., 2010; Holt and Laury, 2002) or random way (Hey and Orme, 1994). Lotteries are sometimes presented by means of visual tasks without making explicit reference to probabilities (Crosetto and Filippin, 2013b; Lejuez et al., 2002; Slovic, 1966). The controlled design can also take the form of eliciting the certainty equivalent of some lotteries (Becker et al., 1964) or asking 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 incentivized choices within incentive compatible designs. A different, 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 could at least in part be explained by different research goals. For instance, different tasks should be used if the researcher wants to precisely analyze risk preferences *per se*, or if the aim is to control for risk preferences while analyzing choices in other contexts that nonetheless involve risk. It is hence important to understand the characteristics of each task in order to be able to choose the one more appropriate to the research goals at hand. 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 incentivized tasks that have been widely used in the literature, namely:

- the multiple price list, in its Holt and Laury (2002) incarnation (henceforth, HL);
- an ordered lottery choice task introduced by Eckel and Grossman (2002, 2008a) (EG);
- the Investment Game by Gneezy and Potters (1997) and Charness and Gneezy (2010) (CGP);
- the Balloon Analogue Risk Task by Lejuez et al. (2002) (Balloon), and
- the Bomb Risk Elicitation Task by Crosetto and Filippin (2013b) (BRET).

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

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

While many other risk elicitation mechanisms exist and have been extensively used,[1] we focus on the ones mentioned above as they are among the most commonly used; they do not

[1]For an extensive review, including other elicitation tasks with respect to those analyzed (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.

result in a high cognitive load on the subjects, and are fast and easy to implement. Therefore, such mechanisms are the most suitable 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 number of elicitation methods is considerable, even selecting only those more suitable to be used as a control, as we do in this paper. Not surprisingly, other scholars have already compared some of these tasks according to their implied coefficient of risk aversion, in some cases evaluating their degree of complexity (see Section 3 below). A low correlation between the behavior in different tasks is a recurrent finding. However, almost all these contributions entail a within-subject design. Although ensuring that individual heterogeneity is controlled for, this procedure suffers from severe drawbacks as long as the tasks involve a degree of uncertainty as is the case by construction when risk elicitation tasks are performed. In fact, apart from being subject to violation of the Reduction Axiom, a within-subject design is likely to induce some form of hedging across periods by non-risk-averse subjects that could determine a negative correlation across tasks. In other words, the low correlation between the behavior in different tasks could in part be an artifact of the design. Consequently, we have decided to adopt a pure between-subject design.

The originality of our paper stems from adopting a between-subject design while, at the same time, gathering data from a homogeneous pool of subjects instead of relying on a meta analysis as done, for instance, by Charness et al. (2013). Moreover, this paper is the first to include an in-depth comparison of the Bomb Risk Elicitation Task (Crosetto and Filippin, 2013b), and to the best of our knowledge, it contains the largest set of elicitation mechanisms. We compare the tasks in terms of their implied coefficients of risk aversion and, as sometimes done in the literature, of their degree of complexity and their precision. Moreover, we focus on the role played by a multiple decisions framework and the likelihood that the tasks trigger loss aversion. In fact, we find it rather striking that despite the growing evidence supporting the predictive power of Prospect Theory (Kahneman and Tversky, 1979), when it comes to measuring risk attitudes, the theoretical framework usually adopted to map the choices in the tasks is Expected Utility Theory (von Neumann and Morgenstern, 1944), which neglects the role of loss aversion. In contrast, we suggest there is a possibility that, even if entirely framed in the gain domain, the elicitation methods entail a reference point against which bad outcomes could be perceived as losses. We suggest that the existence of such reference points, besides generating biased estimates of the implied coefficient of risk aversion, correlates with the likelihood of observing gender differences in risk attitudes.

The outline of the paper is as follows. In Section 2, we describe the five risk elicitation tasks that we compare in this paper. After summarizing the results of other papers, in which comparison exercises have been performed, in Section 3, we compare and contrast the tasks' respective theoretical advantages and drawbacks in Section 4. The task parametrization and the experimental procedures are described in Section 5. The empirical comparison of the decisions made in the different tasks is presented in Section 6. Section 7 concludes.

2. Chosen risk elicitation tasks

2.1. Multiple price list: Holt and Laury (HL)

The multiple price list format is a simple procedure used to elicit values from a subject. Applied to risk, it consists of giving the subject an ordered list of binary choices between lot-

teries. The most widely known implementation has been provided by Holt and Laury (2002). Going by the number of citations, it is, to date, the most popular risk elicitation mechanism. In the HL task, subjects face a series of choices between pairs of lotteries, with one lottery being safer (i.e., with lower variance) than the other. The lottery pairs are ordered by increasing expected value. The set of possible outcomes is common to every choice, and the increase in expected value across lottery pairs is obtained by increasing 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 switching point captures the risk aversion of the subject. At the end of the experiment, one row is randomly chosen for payment, and the chosen lottery is played to determine the payoff.

	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 €

Table 1: The 10 lotteries chosen for the HL treatment

The parameters chosen in our experiment for the HL task can be seen in Table 1. The values are based on the baseline Holt and Laury (2002) values, doubled in order to offer salient payoffs and make them comparable with the other tasks. The expected value of the safe lottery increases from 3.28 to 4 and that of the risky lottery from 0.95 to 7.7 euro along the table. A risk-neutral subject should start with Option A and switch to B from the fifth choice on.

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 lottery preferred within a set of 5 lotteries characterized 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 make a single choice, i.e., to indicate their preferred 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.

	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 €

Table 2: The 5 lotteries chosen for the EG treatment

2.3. The Investment Game of Charness, Gneezy and Potters (CGP)

A different approach is the one introduced by Gneezy and Potters (1997) and refined by Charness and Gneezy (2010).[2] They propose a task in which the choice is framed as an investment decision. Instead of choosing among lotteries, subjects have to decide how to allocate a given endowment of 4 euro between a safe account and a risky lottery that, with 50% probability, yields 2.5 times the amount invested, zero otherwise. In other words, subjects must choose an amount k to invest in the lottery, thereby generating a set of lotteries:

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

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 higher than one for any k different from zero, a risk-neutral subject should invest all the endowment.

2.4. The Balloon Analog Risk Task (Balloon)

The Balloon Analogue Risk Task (Balloon, Lejuez et al., 2002) provides a pictorial rather than numerical representation of probabilities. A process of draw without replacement from an urn of unknown size is visualized as a balloon into which air can be pumped by the subject. The balloon is characterized by an explosion point, i.e., in the underlying urn with n balls, $n-1$ balls are safe, while one ball determines the explosion. Subject earn 0.1 euro each time they pump air into the balloon. At any moment, the subject can choose to stop inflating the balloon and collect the amount of money accumulated in the task. In case the balloon explodes, the earnings are reset to zero.[3]

[2] Charness and Gneezy (2012) summarize the results obtained in several experiments adopting this mechanism.

[3] The Balloon is similar to the Devil's Task, introduced by Slovic (1966) as a simple task to study risk aversion in children. More recently, a variation of the same task has been used by Harbaugh et al. (2002), again in a context with children. The task is simple as it does not rely on high numeracy skills.

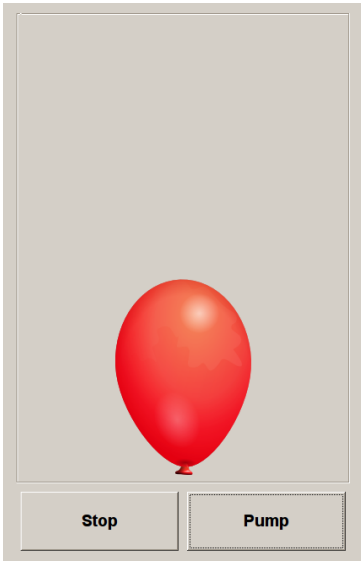

Figure 1: A screenshot of the Balloon task

The Balloon is ultimately an ambiguous task as the subjects are not informed of the probability of explosion at any pump, and it is impossible to visually inspect the state of the underlying urn (see Fig. 1). In order to increase the comparability between the Balloon and the BRET, which is described in the next section, we reduced the ambiguity by announcing the maximum number of pumps, equal to 100. Assuming that a subject is able to keep track of the number of times she has clicked and inflated the Balloon, the choice the subject faces after k pumps is between a safe amount $0.1 \cdot k$ and the lottery:

$$L^{k+1}_{Balloon} = \begin{cases} 0 & \frac{1}{100-k} \\ 0.1(k+1) & \frac{99-k}{100-k}. \end{cases}$$

If there were no ambiguity, a risk-neutral subject would then maximize her utility by choosing to stop at the 50^{th} pump.

2.5. The Bomb Risk Elicitation Task (BRET)

The BRET is a visual real-time risk elicitation task introduced by Crosetto and Filippin (2013b). 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 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 at 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) as shown to the subjects is reported in Figure 2. 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.

Unlike the Ballon task, the BRET transparently displays probabilities, since it is possible to visually appreciate how many boxes have been collected and how many are left. Moreover,

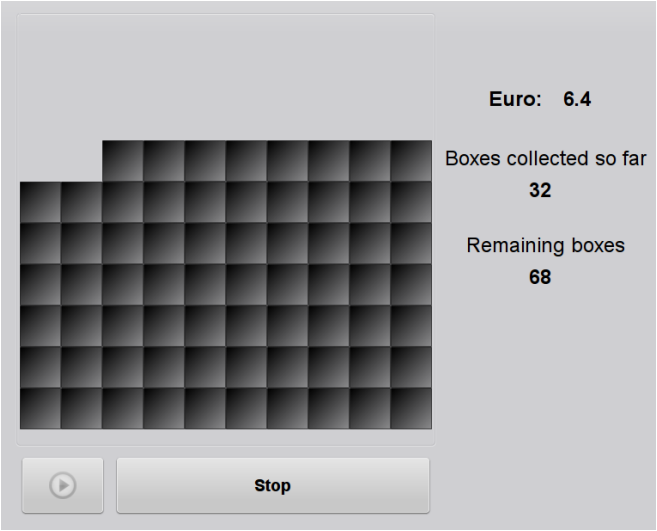

Figure 2: The BRET interface after 32 seconds

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 implement a choice in strategy method, avoiding the truncation of the data that would otherwise happen in case of a real-time notification such as in the Balloon task.

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

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

where k simultaneously drives the change of probabilities and amounts of money at stake, summarizing 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 $k = 50$.

2.6. Questionnaires

After having gone through one of the incentivized risk tasks, all the subjects in our experiment were exposed to two self-reported risk measures: the SOEP and the DOSPERT.

The SOEP measure consists of a direct question, extracted in 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 incentivized lottery-based tasks has been explored by Dohmen et al. (2011), who find that self-reported answers can represent a valid low-cost substitute for incentivized lottery schemes, although the fraction of variance explained is quite low (about 6%).

The Domain-Specific Risk-Taking Scale (Blais and Weber, 2006) has been developed by psychologists, reflecting the idea that elicited risk attitudes can vary across domains and in different contexts, and the fact that utility-based measures and estimates of coefficients seem to fail when extended to try and explain risk attitudes outside of the financial/gambling sphere. The DOSPERT is a 30-item risk questionnaire and a validated measure of the risk attitude as a personality trait. It takes into account 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.

3. Weak correlation between tasks in the literature

In the literature, many attempts to compare pairs or groups of risk elicitation mechanisms have already been made. A low correlation between the choices made in different tasks is a recurrent finding. After reviewing some of these contributions, we claim, in this section, that this evidence is likely to be driven by a within-subject experimental design, rather than signaling idiosyncratic and possibly inconsistent features of the different mechanisms.

Among the papers employing a within-subject design, Deck et al. (2010b) compare four common risk elicitation tasks: HL, EG, the Balloon, and a version of the 'Deal or Not Deal' TV show. They find a significant, though weak, correlation between the two 'static', table-based tasks (HL and EG, $\rho \cong 0.21$) and between the two dynamic, visual tasks (Balloon and DOND $\rho \cong 0.27$). No other significant correlation emerged between any other pair of tasks.

Bruner (2009) uses a multiple price-list task, in which subjects are asked to make a number of pairwise choices between a lottery increasing in expected value and a constant safe amount. The increase in the expected value is obtained in two different ways: increasing the reward while keeping the probability fixed, or increasing the probability of the best event while keeping the outcomes fixed. Bruner finds that subjects prefer the increase in probability to the increase in reward. Moreover, he finds that up to 58% of the subjects in his within-subject design reveal different risk preferences in the two different tasks.

Harbaugh et al. (2010) compare the price-based Becker-DeGroot-Marschack (BDM) mechanism with a choice-based procedure (whether six lotteries are preferred or not to their expected value with certainty), the goal being to test the fourfold pattern of risk attitudes put forward by Tversky and Kahneman (1992). Analyzing repeated individual choices, they report that preferences elicited with these two methods appear inconsistent, with almost half of the subjects flipping attitudes across tasks (from risk averse to risk seeking or vice versa).

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) with the aim of testing the stability of risk preferences across tasks. They find that the measured risk preferences are affected by the mechanism used but that the HL and EG tasks are indeed significantly correlated ($\rho \cong 0.4$). Moreover, about 20% of the subjects display stable preferences. However, the EG mechanism results in a much higher fraction of subjects being measured as

risk averse with respect to the HL: 75% *vs.* 54%. Reynaud and Couture (2012) also review the literature on the stability of risk preferences across tasks, reporting that, in general, correlations are rather low.

Deck et al. (2010a) try to rationalize such apparently inconsistent behavior and investigate whether different elicitation tasks (HL, EG, Balloon, and DOND) are mapping to different domains of uncertainty (health, financial, gambling, etc.), as measured by the DOSPERT questionnaire, but find little evidence supporting this explanation.

Dave et al. (2010) stress the trade-off between the complexity and precision of the task, comparing HL and EG risk elicitation mechanisms. They claim that the latter is much easier to understand and that this significantly reduces inconsistencies.

Harrison (1990) compares the BDM with the risk aversion coefficient implicit in first price auction bids using a between-subject approach and finding that the BDM displays stronger risk seeking. Isaac and James (2000) replicate this comparison, adopting a within-subject approach and obtaining similar results. Moreover, they find a negative correlation between choices, particularly for those who do not make risk-neutral evaluations, which suggests some sort of hedging across periods.

The comparisons above implicitly rely on the assumption that when facing multiple decisions under uncertainty, subjects maximize their utility in every period, thereby making the best choice every time, given stable underlying risk preferences. While well-grounded from a theoretical point of view, this argument could be contradicted by the evidence as long as subjects change their decision in different periods, even though, on average, behaving consistently with their risk preferences. In other words, while some subjects could make stable choices across periods, others could, for instance, decide to make a risk-averse decision in the first task and a risk-loving one in the second. If this is the case, the low correlation across task would be an artifact of the multiple decision framework rather than reflecting idiosyncratic features of different tasks.

To test this possibility, we can use the data of the Repeated Treatment of Crosetto and Filippin (2013b), in which the BRET was repeated five times and the issue of consistency across tasks could therefore not arise. 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. It is worth noting that the average choice is not significantly different than that of other subjects who played the same task in the one-shot mode.

Such a behavior is suboptimal from a theoretical point of view. However, the difference in the expected payoff when a subject plays according to her preference only on average and not in every period is not big, and it could be more than counterbalanced, for instance, by the benefit of avoiding the boredom that would arise from repeating the same choice. This interpretation suggests a further test, i.e., that the cost-benefit balance of the increased variance that arises when a subjects follows only on average her preferences could, in turn, be affected by risk attitudes. We find that this is indeed the case. Most of the subjects show a rather erratic behavior, and only about 30% of them display stable preferences.[4] Of these, 78% are risk averse, 22% risk neutral and only 4% risk loving.

Therefore, the low correlation between behavior in different tasks that emerges as a stylized

[4]We considered as stable choices that do not vary by more than 10 between the maximum and the minimum.

fact in the other contributions that compare elicitation methods in the literature could simply be a composition effect between a subgroup of subjects with a stable behavior and another subgroup which displays a possibly negative correlation across periods. Consequently, we decide to stick to a pure between-subject design.[5]

Concerning the tasks we are focusing on, there are some contributions in the literature that propose a comparison using a between-subject design. For instance, Charness and Viceisza (2011) elicit risk attitudes of 91 farmers in rural Senegal using HL and CGP besides asking the SOEP question. They find a large fraction of inconsistent choices in the HL task, with 40% of subjects making dominated choices and more than half switching more than once between safe and risky options. Moreover, women are more likely to be inconsistent. The CGP task seems to yield more consistent results, as the distribution of elicited risk preferences is in line with previous studies.

4. Theoretical comparison of the tasks

The tasks presented in Section 2 can be compared from a theoretical point of view along three dimensions. First, we analyze each task according to the completeness and precision of the range of risk preference it allows to identify. Second, we consider the potential role played by introducing multiple choices in the experimental session. Third, we investigate the possibility that endogenous reference points induce loss aversion, even though the task is entirely framed in the gain domain. The results of the theoretical comparison are summarized at the end of this section, in Table 3 on page 15.

4.1. Range and precision of the estimate

The different elicitation methods allow to classify participants in several categories, representing their different willingness to accept risk. 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 summarized by means of the coefficient of relative risk aversion r. Such a parametric assumption does not imply any loss of generality as long as it is used both to summarize the range of risk attitudes that can be identified by each task and to evaluate the precision of the estimate.

As far as the range of risk attitudes is concerned, two of the tasks above, the Investment Game and EG, can only estimate $r \leq 1$, that is, they cannot measure preference in the risk-seeking domain. In other words, they cannot distinguish risk neutrality from risk seeking (and, in case of EG, from a slight degree of risk aversion).[6] Charness and Gneezy (2012) claim that this is a minor problem because risk seeking preferences are seldom observed, although in our data about 20% of the subject pool is characterized by $r \geq 1$ (see Table 4 below).[7]

[5] Another reason suggesting a between-subject design is the fact that a withing -subject approach implements a compound lottery and is therefore prone to violations of the Reduction Axiom, see section 4.2 below.

[6] The version of the EG task implemented in Dave et al. (2010) features an additional lottery characterized 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 behavior 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.

[7] Moreover, there is evidence in other branches of the literature that the shape of the opportunity set may trigger demand effects insofar as subjects tend to avoid choices at the extremes (Bardsley, 2008; Filippin and Raimondi,

The Balloon Task allows, in principle, to disentangle any kind of risk attitude. However, the risk aversion coefficient for the Balloon task can be computed only assuming that the subjects know, or hold correct beliefs about, the details of the underlying urn. Moreover, this task can also provide a limited range of estimates due to the fact that the explosion of the balloon prevents the observation of the desired stopping point and, therefore, the measurement of the underlying risk preference. In other words, the Balloon task delivers truncated data.

In contrast, the HL and BRET allow to estimate a fairly complete range of preferences.

The precision of the estimate is also important because the lower the number of choices available, the larger the measurement error of the parameter that is introduced. With the HL mechanism one can classify subjects in 10 different categories, while greater simplicity of the EG task comes at the price of a coarser estimation allowing 5 categories only. In contrast, CGP, the BRET, and the Balloon allow to estimate risk attitudes almost continuously, with the Balloon delivering, however, truncated data.

Finally, the different tasks are also characterized by different ways of mapping choices to risk aversion parameters. To stress this fact, Figure 3 compares how each task maps choices to the parameter r of a CRRA when the domains are made somehow comparable.

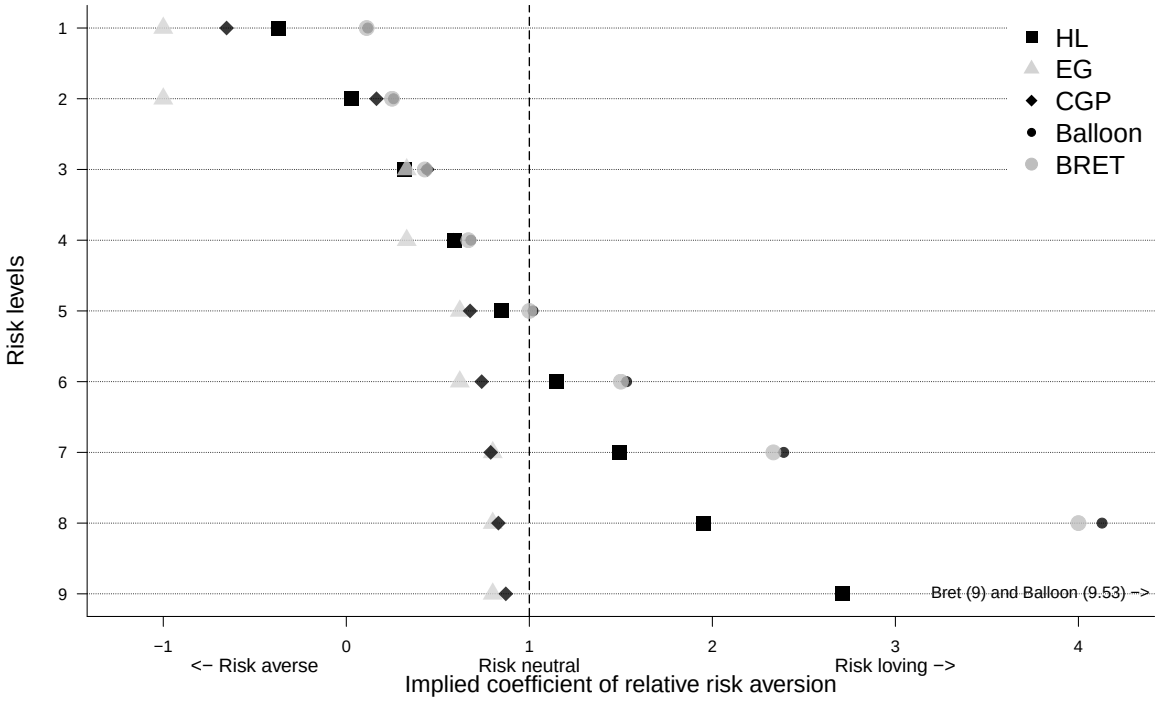

Figure 3: Comparison of r by task

In order to build the figure, we reparametrized each task as having ten choices, increasing

2012; List, 2007). If this is the case, the fraction of risk-neutral and risk-loving subjects would be underestimated by a task that does not allow to identify them.

in terms of risk seeking, which imply 9 cutoff points. This is straightforward insofar as the tasks feature almost continuous choices. For instance, CGP level 1 means the subject invested $\frac{1}{10}$ of her endowment in the risky asset, level 2 means $\frac{2}{10}$, and so on up to the cutoff point between levels 9 and 10, corresponding to an investment of $\frac{9}{10}$ of the endowment. Similarly, in the Balloon and BRET level 1 corresponds to 10 pumps (boxes), 2 to 20, etc. HL provides 10 categories according to the number of (meaningful) risky choices made by the subject.[8] In fact, although there are ten binary choices that, in principle, allow for a number of risky choices ranging from 0 to 10, zero risky choices is a dominated action, given that it implies that the subject turns down 7.7 euro for sure, preferring 4 euro for sure instead. As a consequence, there are only 10 meaningful choices that provide 9 cutoff values. The last elicitation method (EG) is more problematic and it requires some attention (and some assumptions) in order to represent it together with the other methods in Figure 3. In fact, EG allows for at most 5 different choices, thus implying only 4 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 neutrality, and risk loving at the same time.

As a result, we can display for each task how choices, summarized by each of the 9 cutoff points, map to the implied coefficient of relative risk aversion r. 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. Moreover, tasks characterized by a low number of categories will likely produce a large fraction of decisions in categories in which only one bound of r can be computed and for which it is therefore fairly impossible to assign a value. For these reasons we consider it more appropriate to use the median choice rather than the average to summarize and compare the risk attitudes in each task. Another advantage of this approach is that focusing on intermediate choices makes the different tasks more comparable in terms of r.

4.2. Single vs. multiple choice framework

A within-subject approach in which one task is paid at random at the end of the experiment *de facto* introduces a compound lottery.[9] Besides inducing erratic behavior, as already stressed in Section 3, this constitutes a serious issue because violations of the Reduction Axiom are commonly observed. For instance, Kaivanto and Kroll (2011) and Harrison and Swarthout (2012) show that subjects are not indifferent between real-money lotteries implemented with randomization devices that should instead be equivalent under the Reduction Axiom. To neglect this problem means to incur in what Harrison and Swarthout (2012) call "bipolar behavior," i.e., extreme pessimism about the (mixture) independence axiom when it comes to models evaluating pairs of lotteries, but extreme optimism when it comes to the application of that axiom to the veracity of the payment protocol.

Besides adding a reason to opt for a between approach in the comparison of several risk elicitation mechanisms, the same argument also prompts the administering of one-shot rather

[8] Usually, the HL task is summarized 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.

[9] The alternative of paying every task is usually not adopted because it entails wealth effects.

than repeated tasks. Note, however, that violations of the Reduction Axiom may also concern a one-shot task as long as it implements a compound lottery by entailing multiple choices of which only one at random is paid. Among the tasks we consider in this paper, this applies to the HL, while EG, the Investment Game, the Balloon, and the BRET imply a single choice and hence a simple lottery.

Given our main goal of comparing elicitation mechanisms meant to control for risk attitudes, it is unavoidable to have both the risk elicitation method and the main task in the same session. This obviously limits to the one-shot perspective and raises the problem of which payment protocol to implement. The pay-one-at-random protocol cannot be pursued, given that the main task *must* be incentivized. Hence, the only solution is either not to pay the elicitation task, with obvious concerns about the reliability of the results obtained, or to pay both tasks, i.e., allowing for the possibility of biases due to wealth effects. Apparently, there is no way out. However, the tasks in which uncertainty can be resolved in a different moment than when the choice has been made, i.e., all those analyzed in this paper except the Balloon, allow to reach a reasonable compromise. In fact, it is possible to implement the elicitation mechanism *before* the main task, but to resolve the uncertainty and thus determine the corresponding payoff only *afterwards*. In this way, the design would not implement a compound lottery as long as both tasks are rewarded, while, at the same time, wealth effects in the main task are minimized by the uncertainty of the payoff in the risk mechanism. From this point of view, the BRET is the mechanism that minimizes the predictability of the payoff since it has neither a safe option nor any minimum amount that can be earned for sure. Such a protocol should, in our opinion, be preferred to the other case in which the risk mechanism is played after the main task as long as the latter implies known or predictable payoffs because wealth effects during the experiment could bias the attitude toward uncertainty.

4.3. Loss aversion

All the elicitation mechanisms described in Section 2 empirically measure the likelihood of accepting or turning down lotteries that differ in terms of risk and reward. These choices are then often translated into coefficients of relative risk aversion using a CRRA utility function. However, while the precision of the tasks can be analyzed by assuming any functional form without loss of generality, this becomes more subtle when the quantitative assessment of subjects' risk attitudes is concerned.

By construction, any assumption about the form of the utility function affects the measured individual coefficient of risk aversion, which will obviously turn out to be biased whenever the specific functional form does not properly represent the underlying preferences. A less ambitious target could be to preserve the ranking of subjects in terms of their risk aversion.

It is therefore crucial that the elicitation mechanism does not trigger elements that could break the monotonicity between the choices made and the underlying degree of risk aversion. An immediate example of such a confounding factor is loss aversion: a higher loss aversion parameter could make a relatively risk-loving subject appear more conservative than a relatively risk-averse subject with low loss aversion.[10] Alternatively, we have to make sure that

[10] In Section 6.2 we discuss that what is often referred to as a gender gap in risk aversion should instead be regarded as a difference in loss aversion; in contrast, risk aversion does not significantly differ between males and females.

the parameters representing risk preferences are estimated using a theoretical framework that is consistent with the elicitation method.

For loss aversion to be relevant the prospect should, in principle, entail gains and losses. In most of the tasks, though, losses are almost never explicitly considered. However, it is possible that the opportunity set of a task endogenously determines a reference point against which uncertain outcomes, both higher (gains) and lower (losses), can be evaluated.

From this perspective, the CGP task uses a framework in which loss aversion can clearly be relevant. Subjects are endowed with an amount of money, and they have the opportunity of keeping the entire endowment by setting the amount invested in the risky lottery equal to zero. Such an endowment is therefore likely to act as a reference point. The amount invested, if positive, could then be evaluated as a loss in case of the negative outcome occurring.

The version of the EG task considered in this paper does not involve losses, strictly speaking, and in fact Eckel and Grossman (2008b) classify the task as 'gain only.' However, this task envisages a degenerate lottery with no uncertainty that can be considered as a reference point against which possible losses, implied by the other risky prospects in case the negative outcome occurs, can be compared.

The HL task is more difficult to assess. Subjects could consider the minimum amount of the safer lottery as a benchmark against which the worse outcome of the risky lottery could be evaluated as a loss. However, such an amount is less focal as it requires some elaboration on the part of subjects and is not shown directly to them as, e.g., in the CGP and EG tasks.

In principle, the Balloon task entails losses because subjects decide to put a sure amount at stake in exchange for a risky lottery every time they decide to inflate; moreover, since the gains at each inflating decision are negligible as compared to the losses, the game is played almost entirely in the loss domain. On the other hand, the probability of incurring a loss is relatively small, which is why gains are small but relevant; moreover, in the Balloon the focality of losses is likely to depend on how the task is framed in the mind of subjects. In fact, if subjects set a target of, say, 40 pumps, they should be aware that the amount earned up to any point before 40 has not been definitively earned but is continuously at stake; therefore, the explosion when pumping for, e.g., the twentieth time would probably not be framed as a loss. In other words, if a subject is undecided about what to do at some point, she would certainly frame the possible explosion as a loss; it does not necessarily have to be so when inflating the balloon is perceived as an intermediate step toward a predetermined goal.

Finally, the BRET does not provide any reference point because the resolution of uncertainty happens only after the decision is made. Compared to the Balloon, the main difference is that subjects are not notified about the explosion, and therefore the decision to proceed implies a comparison between uncertain amounts only. In the BRET there is thus no possibility of building a reference point against which worse outcomes could loom as losses. Observed choices can therefore be used to build a pure measure of risk aversion.

While representative of most real world gambles, tasks that trigger loss aversion can be problematic. When, as usually done in the literature, choices made when there is actually something at stake are mapped to a theoretical framework that disregards the role played by loss aversion, the estimated coefficients are biased by construction. On the one hand, imposing Expected Utility Theory delivers biased estimates by construction because the loss aversion parameter is omitted. On the other hand, assuming a theoretical framework that considers loss aversion (e.g., Prospect Theory) does not solve the problem either. The reason

	Precision (categories of r)	Parsimony (no. of choices)	Complete? (r range)	Ambiguity	Reduction	Loss aversion
HL	8	10	yes	no	suffers	mild anchor
EG	5	1	no	no	ok	strong anchor
CGP	Almost continuous	1	no	no	ok	strong anchor
Balloon	Almost continuous	$0 \leq n \leq 100$	yes (trunc)	yes	ok	strong anchor
BRET	Almost continuous	1	yes	no	ok	no anchor

Table 3: Summary of the theoretical comparison of the tasks

is that in order to identify the two parameters of risk and loss aversion, it would be necessary to find a context in which only one of the two matters. However, this is not feasible as long as a task intrinsically entails loss aversion even when framed in the gain domain.

5. Design of the experiment

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 444 subjects took part in 16 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 Jena. The experimental software for each of the five tasks and the questionnaires was programmed in Python (van Rossum, 1995).

For the reasons explained in Section 3 and 4.2 each subject took part in just one risk elicitation mechanism. In other words, the comparison of the risk aversion tasks was carried out in a pure between-subject design.

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.[11] These were then read aloud, and questions were answered on an individual basis. Once all questions had been answered, subjects were allowed to start with the risk elicitation mechanism, which constituted the main task of the experiment. After the main task, subjects were 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 perceived complexity of the task. All tasks except the Balloon involved some sort of randomization to compute the final payoffs. These randomizations were carried out manually using dices (for the HL, EG, CGP tasks) or draws from an urn (for the BRET), and all were, for the sake of transparency, publicly performed with the help of randomly selected subjects after everyone had completed the questionnaires. The subjects were then paid.

In order to improve the comparability across different mechanisms, we administered the tasks to similar groups of subjects and 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.[12] The sessions lasted on average less than thirty minutes.

[11] The English translation of the original German instructions is attached to this paper in Appendix A

[12] This is not the case for the Balloon, in which the expected payoff is 2.5 euro. The reason is that we collected the Balloon data at the very beginning of our project with to compare the results with the Explosion treatment of

	N	SOEP	Type of choice	Choice set	Median choice	Median r	Type classification (%)			
							Averse	Neutral	Lover	N.C.
HL	88	5.16	Risky choices	[0,10]	3/4	0.32	70.45	4.55	7.95	17.05
EG	88	5.18	Chosen lottery	[1,5]	2/3	0.33	81.81	*na*	*na*	18.18
CGP	86	5.27	Amount invested	[0,4]	2.5	0.75	80.23	*na*	*na*	19.77
Balloon	93	5.29	Stopping point	[0,100]	57	1.35	37.63	3.23	59.14	*0*
BRET	88	5.03	Stopping point	[0,100]	40	0.67	72.73	11.36	14.77	1.14

Table 4: Estimates of r and risk categories for all tasks

6. Experimental comparison

Introduction: Table 3 summarizes the main features of each task along the lines of Section 4. In this section, we further evaluate the tasks from an empirical point of view, focusing both on the measured risk attitudes, also adopting a gender perspective, and on their degree of complexity.

6.1. Estimated values

As already emphasized in Section 4.1, the tasks we are analyzing do not share a common codomain and are characterized by different ways of mapping choices to risk aversion parameters. The particular functional form assumed to analyze the data becomes hence crucial when comparing choices across elicitation methods. This problem is particularly severe when there are good reasons to believe that relevant dimensions are being omitted from the theoretical framework specified. We have argued above (Section 4.3) that this appears to be the case for the CGP and EG task with respect to loss aversion. If different tasks integrate or omit different dimensions, then comparisons become a difficult exercise. Elicitation methods can provide an ordinal ranking of subjects within the task according to their willingness to tolerate risk, but not a cardinal measure of their risk aversion.

As a consequence, choices made under different elicitation methods are not comparable, strictly speaking. However, following common practice in the literature and bearing in mind the underlying problems, we translate the choices into parameters r of a CRRA utility function x^r. Results are reported in Table 4, in which we report medians and distribution statistics instead of averages, for the reasons detailed in Section 4.1 above.

Subjects in all sessions are comparable in terms of their self-reported risk attitudes. In fact, the answers to the SOEP question are not significantly different in any of the treatments according to a Kruskal Wallis test (p-value 0.956).

Reporting the median choice in tasks with a low number of categories could result in the need to interpolate the data in order to get an unbiased picture. In both HL and EG, though, the median choice falls close to a cutoff point and therefore no interpolation is necessary. Among the tasks that should be comparable because they do not suffer from loss aversion,

the BRET in Crosetto and Filippin (2013b), before considering a systematic comparison of the BRET with other elicitation methods. We decided to retain the data nevertheless, because the only section where this is likely to affect the results is when the coefficient of risk aversion is measured, something that would be problematic even by doubling the amount at stake due to ambiguity, loss aversion, and truncation of the data.

we observe a higher measured risk aversion in the HL 0.32 than in the BRET 0.66. Keeping in mind the aforementioned caveat, results show that the estimated risk aversion is also strong in the EG task (0.33), while it is moderate in in CGP (0.75), and the median Balloon subject is a risk lover. We believe that the lower average amount at stake cannot account for such a striking difference, and that illusion of control should instead be considered as the main explanation (Langer, 1975).

Additional evidence can be gathered 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. Unfortunately, also in this case loss aversion acts as a confounding factor. When taking into account, however, that some of the subjects cannot be classified either because they make inconsistent choices (such as in the HL or in the BRET) or because the same choice does not allow to identify different risk attitudes (GCP and EG), we find that the Balloon reports the largest fraction of risk loving subjects.[13] As for the share of risk-averse subjects, this is rather similar in the other elicitation methods, although slightly lower in the BRET.

6.2. Gender effect

Many studies report gender differences in risk attitudes, showing that females are significantly more risk averse than men. Not surprisingly, this is the main message conveyed by the surveys available in the literature (Charness and Gneezy, 2012; Croson and Gneezy, 2009; Eckel and Grossman, 2008b). This large body of experimental literature is sometimes perceived as describing a stylized fact whose causes, rather than its the existence, should be investigated (Bertrand, 2011). However, as shown in detail by Crosetto and Filippin (2013a), this finding is not as ubiquitous as often reported, and the likelihood of gender differences being observed strongly correlates with the characteristics of the task used to elicit risk preferences.

For instance, Charness and Gneezy (2012) report that in the Investment Game gender differences are systematic and substantial. Males invest significantly more than females in almost all the experiments analyzed, and often such a difference is at least about $10-15\%$ of the initial endowment. Similar findings emerge with the EG task. However, the picture changes sharply for the other tasks. Both with the Balloon and the HL tasks, the occurrence of gender differences constitutes the exception rather than the rule. In the BRET, the absence of gender differences is a robust result, although the task is more recent and less evidence is available. The data used in Crosetto and Filippin (2013b) show that both in the baseline and control treatments the behavior of males and females does not differ.

The average choices by gender in our experiment are reported in Table 5. They display patterns that are in line with the findings in the literature basically for all tasks.

In the Investment Game, 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). Although our design is characterized by a lower expected payoff, we find

[13] We can safely attribute all subjects whose Balloon bursted to the category of risk lovers, since the explosion point had been randomly predetermined for all subjects by the computer at 62 pumps.

	Males		Females		Mann
	N	Mean choice	N	Mean choice	**Whitney**
HL	31	3.74	42	3.57	$\|p\|$=0.9090
EG	45	3.22	43	2.34	$\|p\|$=0.0050
CGP	37	2.93	49	2.25	$\|p\|$=0.0021
Balloon	32	54.03	61	48.78	$\|p\|$=0.0437
BRET	32	39.72	55	40.25	$\|p\|$=0.7913

Table 5: Choice by gender

similar results also for the EG method. In our experiment, the average choice is 3.22 for males vs. 2.34 for females, while in Eckel and Grossman (2008a) it is 3.63 and 2.95, respectively. We also observe gender differences reaching traditional significance levels in the Balloon task, which is not often observed in the literature.

In contrast, our replication of the HL task finds no gender differences, which is in line with the majority of the contributions analyzed in the literature. Similarly, in the BRET the behavior of men and women is indistinguishable.

Both our replications and the results in the literature reveal a notable correlation between the likelihood of observing gender differences in risk attitudes and the importance of loss aversion as presented in Section 4.3. In fact, gender differences appear systematically in the tasks in which subjects have the opportunity to avoid facing any risk by opting for a safe choice such as the CGP, EG and partially the Balloon. Note further that in Eckel and Grossman (2008a) there is an explicit test for the effect of loss aversion since in one of the treatments subjects could incur into minor losses by choosing one of the two riskier lotteries. Results are unchanged, and the absence of any treatment effect is a further indirect signal that loss aversion is at work even in the treatment framed in the gain domain.

Interestingly, gender differences tend to appear in the HL and in the BRET when loss aversion becomes salient. For instance, Schipper (2012) does not find gender differences when replicating the classic HL task, but he does find them when replicating it in the loss domain. Similarly, different versions of the HL task, involving a battery of choices between a safe option and increasingly favorable lotteries, are more likely to produce different risk attitudes by gender (see, among others, Sutter et al., 2013). Furthermore, the only treatment in which the BRET reveals gender differences is the condition in which the framing of the treatment induces loss aversion.

Results are thus in line with what has been suggested by Booij and de Kuilen (2009) and Crosetto and Filippin (2013b), i.e., that differences between males and females when facing decision under uncertainty are mainly driven by *loss* aversion rather than *risk* aversion.

6.3. Complexity

The comprehension of a task is crucial in order to obtain reliable data. This is true in general but particularly so when the task is performed by subjects with low numeracy skills. Complexity is thus one of the most relevant dimensions along which the tasks should be assessed.

A readily available proxy for the complexity of elicitation methods is the variance of the choices made, which can be assumed to decrease with a better understanding of the task, since a better understanding should reduce confusion and decision errors. However, the choices

	N	Mean	Mann-Whitney test results EG	CGP	Balloon	BRET
HL	88	3.30	n.s.	n.s.	**	*
EG	88	2.92	-	n.s.	***	n.s.
CGP	86	3.18	-	-	***	n.s.
Balloon	94	4.43	-	-	-	***
BRET	88	2.76	-	-	-	-

Mann-Whitney significance thresholds: *=0.1, **=0.05, ***=0.01

Table 6: Perceived complexity of the task, including inconsistent choices

collected under the different elicitation methods are not perfectly comparable in terms of simple variance – for several reasons, among them the number of alternatives available being the most evident. In our results, for instance, the BRET and the Balloon show the lowest standard deviation, although for the latter, due to truncated data, we can only observe a lower bound of the variance.

Another way of evaluating the complexity of the tasks is by counting the number of inconsistent choices. In HL we observe about 17% of the subjects switching more than once, which is in line with results already reported in the literature. In the BRET we also observe dominated choices in about 1% of the cases. The other tasks do not allow to detect inconsistent choices, so that, from this point of view, a comparison cannot be made. However, an evaluation of the random component in the decision based on observed choices can be performed for all the tasks estimating a structural model (see Section 6.4 below.)

We decided to include an alternative proxy for the understanding of the tasks by directly asking the subjects to report their perceived degree of complexity of the task on a scale from 1 (very simple) to 10 (very difficult). Although there is no consensus in the literature about the quantitative interpretation of qualitative scales, we believe that even a self-reported variable allows to derive some insights by means of non-parametric tests. Table 6.3 shows that subjects evaluate the Balloon as significantly more difficult than all other tasks, and that the BRET turns out to be considered significantly easier than HL. In contrast, all the other pairwise differences are not significant, showing that the BRET, CGP and EG are regarded as comparatively simple.

6.4. Maximum likelihood stimation

As subjects' choice might include stochastic elements, we estimated with maximum likelihood a structural model including a term to capture the noise, as done, among others, by Dave et al. (2010); Hey and Orme (1994); Holt and Laury (2002). We suppose that for each task the subject is an expected utility maximizer who can make an error in comparing the expected utility of the lotteries she faces. As in Hey and Orme (1994), we assume this error to be normally distributed. The resulting equation, determining the choice between a 'left' and a 'right' lottery, is hence given by

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

which implies that the probability of choosing the lottery on the right can be found by evaluating the cumulative distribution function of a standardized normal in

$$\frac{EU_R - EU_L}{\sigma}.$$

To proceed with the estimate, we need to express the tasks as a series of binary choices between pairs of lotteries. Apart from the HL task, where subjects directly evaluate 10 binary choices, this requires to transform the data. Such a goal can be achieved with the two dynamic elicitation methods (the BRET and Balloon), considering that, at any second, the decision maker faces the choice between stopping on the current lottery *vs.* waiting one more second and proceeding to the next, thereby revealing a chain of preference relations. For instance, a subject who decides to collect 40 boxes in the BRET reveals that he prefers collecting 2 boxes to 1, 3 to 2 and so on up to 40, that is preferred to 39. This amounts to implicitly assuming that preferences are single-peaked over the domain of feasible alternatives. As a consequence, we also input that 40 is preferred to 41, which is preferred to 42, and so forth. Hence, we recoded both tasks as potentially involving 100 binary choices.[14]

As far as the static methods are concerned (EG and CGP), we consider the revealed preferences of our subjects as the key to building the binary comparisons in a similar manner. If, for instance, the subject chose lottery x in EG, this would reveal that this lottery was preferred to all the other 4 lotteries. Similar to the transformation applied to the data by Dave et al. (2010), the assumption of single-peaked preferences implies in this case 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. As far as the Investment Game is concerned, we followed a similar logic. By choosing to invest in the risky option a share x of the endowment, the subject reveals that she prefers such a share to all possible alternatives. While the choice variable x is virtually continuous, no subject chose at a detail finer than one decimal point. Hence we transformed the CGP data into 40 choices, again extending the preference relation among non-chosen opportunities.

Whenever applicable, we also included inconsistent choices. This happens, for instance, when subjects display more than one switching point in HL or make dominated choices in the BRET.

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 maximizers characterized by CRRA preferences x^r, allowing both for a random component in the decision σ and for heterogeneity by gender of risk attitudes (r and r_{female}). We use a log likelihood function $L(r, r_{female}, \sigma | Y_i, F_i)$, in which we jointly estimate the risk aversion parameter by gender, and the variance of the error term given the individual choices Y_i, while F_i is a dummy assuming value 1 for female subjects. We performed the estimation with Stata, following Harrison (2008).[15]

Table 7 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 behavior of males and females is significantly different in CGP, EG, and in the

[14] In case of an explosion in the Balloon task we input as missing the preference relations after the explosion happens.

[15] Since choices covered just a small range of possibilities, as is usual with fast risk elicitation tasks used as controls, convergence was not achieved for HL and CGP, neither by letting the noise parameter σ depend on demographics nor adding age in the estimation of r. We hence opted for the minimal configuration detailed above.

	Log-likelihood	Coefficient	Estimate	St.Err.	p-value
HL	-391.25	r	.427	.064	.000
		r_{female}	-.061	.060	.310
		σ	.433	.090	.000
EG	-194.62	r	.694	.035	.000
		r_{female}	-.262	.057	.000
		σ	.206	.020	.000
CGP	-1546.79	r	.863	.014	.000
		r_{female}	-.093	.023	.000
		σ	.010	.001	.000
Balloon	-2243.81	r	1.13	.066	.000
		r_{female}	-.103	.042	.013
		σ	.345	.078	.000
BRET	-2584.71	r	.696	.089	.000
		r_{female}	.034	.049	.488
		σ	.104	.037	.006

Table 7: Maximum Liklihood structural model estimation

Balloon, while it is indistinguishable in HL and BRET. As far as complexity is concerned, it is the Investment game that displays the lowest noise parameter, followed by the BRET, while HL and the Balloon are confirmed as the most difficult tasks of those analyzed.

6.5. 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 task 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) but less, though still significantly, with the DOSPERT gamble ($\rho = 0.36$, p-value < 0.001) and investment sub-scales ($\rho = 0.31$, p-value < 0.001). 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 much the answers to the questionnaire answers are correlated with the incentivized choices.

First, we compute a battery of pairwise correlation coefficients between choices and questionnaires, also evaluating their significance level. Second, after running a linear regression of each choice on the observed demographics (age and gender) as a benchmark, we include each questionnaire separately in the regression, measuring the contribution of the last measure added to the adjusted R^2. In other words, this second indicator, that we name Δ adj. R^2, measures the increase in the variance of each choice that is explained by adding each questionnaire to the regression. Results are shown in Table 8, where the Δ 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 fairly low: the adjusted R^2 of the regressions (not reported) never increases by

		N	Soep	Dospert	Do-Investment	Do-Gamble
HL	Correlation	73	0.23*	0.25**	0.12	0.16
	Δ adj. R^2		≤ 0	≤ 0	≤ 0	
EG	Correlation	88	0.30***	0.30***	0.22**	0.33***
	Δ adj. R^2		2.7	1.6	1.9	
CGP	Correlation	86	0.13	0.17	0.36***	0.33***
	Δ adj. R^2		≤ 0	≤ 0	7.7	
Balloon	Correlation	93	0.37***	0.08	-0.03	0.12
	Δ adj. R^2		10	≤ 0	≤ 0	
BRET	Correlation	87	0.03	0.06	0.05	-0.01
	Δ adj. R^2		≤ 0	≤ 0	≤ 0	

Δ adj. R^2 expressed in extra percentage points. Significance thresholds: *=0.1, **=0.05, ***=0.01.

Table 8: Correlation with questionnaires and explained variance for each task

more than 10%. 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 Investment Game 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%) except for the general questions. In contrast, the Balloon seems to correlate mildly with the SOEP only, which increases the Δadj.R^2 by 10 points.

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.

Following previous findings in the literature on lottery choices (Grossman et al., 2006) 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 incentivized 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. The only exception is the Balloon in which the opposite happens, as women's answers to the SOEP question are more strongly correlated with choices than men's.

In general, the tasks that better correlate with the questionnaires seem to be those in which the importance of reference points, and hence of loss aversion, is higher, possibly reflecting the fact that when self-reporting their attitudes toward risk, subjects recall real-world experiences in which losses are a pervasive phenomenon. In line with this interpretation, it is worth noting that a version of the BRET, framed in order to induce a reference point, shows a $\rho = 0.2$ and significant (at 1%) correlation with the SOEP (see Crosetto and Filippin, 2013b).

7. Conclusions

In this paper, we present an appraisal of five risk elicitation methods both from a theoretical and an experimental point of view: a multiple price list a la Holt and Laury (2002) (HL), an ordered lottery choice a la Eckel and Grossman (2002) (EG), the Investment Game by Gneezy and Potters (1997) and Charness and Gneezy (2010), the Balloon Analogue Risk Task (Lejuez et al., 2002), and the Bomb Risk Elicitation Task (Crosetto and Filippin, 2013b) (BRET). The tasks have been chosen among those sufficiently simple and fast to be performed in order to control for risk attitudes in experimental sessions in the laboratory or in the field.

Although the tasks chosen were easy to understand, when analyzed they differed in their degree of perceived complexity. According to self-reported evaluations and choices made by the subjects, both HL and Balloon turn out to be the more difficult to comprehend. As for the other tasks, EG pays a high price for its simplicity in terms of precision of the estimate, as it allows to classify subjects in 5 categories only and without the possibility of disentangling slight risk aversion from risk neutrality or risk seeking. The BRET and the Investment Game can be rated as comparatively simple to understand while at the same time allowing to classify subjects precisely, given that the choices vary almost in the continuum. The Investment Game, however, spans only the risk aversion domain, while risk neutral and risk seeking behavior cannot be disentangled.

Multiple choices in a decision context involving uncertainty are problematic, given the evidence that subjects tend to hedge across periods instead of repeating the same choice, which would be optimal from a theoretical point of view. This finding induced us to perform a pure between-subject comparison of the tasks and avoid repeated choices. Given our goal of comparing elicitation mechanisms designed to control for risk attitudes, there is, by construction, a limit to the one-shot perspective since both the risk elicitation method and the main task must be performed and incentivized. Except for the Balloon, a possible solution is to play the elicitation mechanism before the main task but to resolve the uncertainty and thus determine the corresponding payoff only afterwards. Wealth effects in the main task are minimized by the uncertainty of the payoff in the risk mechanism. From this point of view, the BRET is the mechanism that minimizes the predictability of the payoff since it has neither a safe option nor any minimum amount that can be earned for sure.

The Investment Game and EG provide subjects with an opportunity set that contains a risk-free alternative, which is likely to act as a reference point against which unfavorable outcomes could be regarded as losses. Although framed in the gain domain, these elicitation methods are thus likely to trigger loss aversion, which becomes a confounding factor when evaluating risk aversion. If the goal is to control for an ordinal ranking of subjects according to their willingness to tolerate risk, these tasks can accomplish it. In contrast, if the goal is to derive a cardinal measure of risk aversion, these elicitation methods provide an estimate that is biased by loss aversion. From this point of view, in the BRET loss aversion should play no role, while in HL and the Balloon such a role, if any, appears negligible.

A direct comparison of the median coefficient of risk aversio,n as implied by subjects' choices, is possible only when confounding factors such as loss aversion do not play a role. Our data show that measured risk aversion is stronger in the HL than in the BRET, while in the Balloon the median subject is risk-lover. Looking instead at the distribution of subjects according to their risk attitudes, we find that apart from the Balloon in which the opposite happens, the share of risk averse subjects is high in all the tasks, although it is slightly lower

in the BRET. Our data also show another remarkable feature of the tasks characterized by loss aversion, which is that in these tasks gender differences in risk attitudes emerge systematically.

With few exceptions, choices in the tasks are orthogonal to the answers provided in the questionnaires: the correlation of the Investment Game with the investment domain of the DOPSERT, of the Balloon with the SOEP, and of EG with all the questionnaires are significant, even though the variance explained is low.

In this comparison, the BRET task turns out well placed along most dimensions, since it features a high number of risk categories within a complete range, entails a single choice, is easily understood, and does not provide reference points. All in all, however, the 'best' task is the one that is most in line with the experimenter's aims. what is important is to be aware of the different dimensions along which the tasks differ. For instance, EG and the Investment Game had better be used to control for risk attitudes in situations where losses are salient, while the BRET provides a pure measure of risk aversion. Budgetary reasons could lead the researcher to opt for a simple questionnaire, although the correlation with incentivized tasks is low. The simplest tasks should be preferred, particularly when working in low-numeracy contexts, and single-choice tasks should be preferred to multiple-choice ones for a many reasons. The choice of the right task always involves trade-offs, however, and it is up to the researcher to resolve them in the best way.

References

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

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

Bertrand, M. (2011). *New Perspectives on Gender*, volume 4 of *Handbook of Labor Economics*, chapter 17, pages 1543–1590. Elsevier.

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. and Weber, E. U. (2006). A Domain-Specific Risk-Taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, 1:33–47.

Booij, A. and de Kuilen, G. V. (2009). A parameter-free analysis of the utility of money for the general population under prospect theory. *Journal of Economic Psychology*, 30(4):651–666.

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

Charness, G. and Gneezy, U. (2010). Portfolio Choice And Risk Attitudes: An Experiment. *Economic Inquiry*, 48(1):133–146.

Charness, G. and Gneezy, U. (2012). Strong Evidence for Gender Differences in Risk Taking. *Journal of Economic Behavior & Organization*, 83(1):50–58.

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

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

Cox, J. C., Roberson, B., and Smith, V. L. (1982). *Theory and Behavior of Single Object Auctions*. Greenwich: JAI Press.

Crosetto, P. and Filippin, A. (2013a). And now for something completely different: Females are not more risk averse than males. *mimeo*.

Crosetto, P. and Filippin, A. (2013b). The 'bomb' risk elicitation task. *Journal of Risk and Uncertainty*, forthcoming.

Croson, R. and Gneezy, U. (2009). Gender Differences in Preferences. *Journal of Economic Literature*, 47(2):448–74.

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

Deck, C., Lee, J., and Reyes, J. (2010a). 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., and Rosen, C. (2010b). Measuring Risk Aversion on Multiple Tasks: Can Domain Specific Risk Attitudes Explain Apparently Inconsistent Behavior. Working Paper.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and 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. and 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. and 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. and Grossman, P. J. (2008b). *Men, Women and Risk Aversion: Experimental Evidence*, volume 1 of *Handbook of Experimental Economics Results*, chapter 113, pages 1061–1073. Elsevier.

Filippin, A. and Raimondi, M. (2012). The patron game: Individual determinants of the contribution to the public good.

Garcia-Gallego, A., Georgantzis, N., Jaramillo-Gutiérrez, A., and Parravano, M. (2010). The SGG risk elicitation task: Implementation and results. ThE Papers 10/07, Department of Economic Theory and Economic History of the University of Granada.

Gneezy, U. and Potters, J. (1997). An Experiment on Risk Taking and Evaluation Periods. *The Quarterly Journal of Economics*, 112(2):631–45.

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., Englestad, H., and Lugovskyy, O. (2006). Predicting Gamble Choice using a Domain-Specific Risk-Attitude Scale.

Harbaugh, W., Krause, K., and Vesterlund, L. (2002). Risk attitudes of children and adults: Choices over small and large probability gains and losses. *Experimental Economics*, 5(1):53–84.

Harbaugh, W., Krause, K., and Vesterlund, L. (2010). The Fourfold Pattern of Risk Attitudes in Choice and Pricing Tasks. *The Economic Journal*, 120(545):595–611.

Harrison, G. (2008). Maximum likelihood estimation of utility functions using Stata. *University of Central Florida, Working Paper*, pages 06–12.

Harrison, G. W. (1990). Risk Attitudes in First-Price Auction Experiments: A Bayesian Analysis. *The Review of Economics and Statistics*, 72(3):541–46.

Harrison, G. W. and Rutström, E. E. (2008). Risk Aversion in the Laboratory. In Cox, J. C. and Harrison, G. W., editors, *Risk Aversion in Experiments*, volume 12 of *Research in Experimental Economics*, pages 41–196. Emerald Group Publishing Limited.

Harrison, G. W. and Swarthout, J. T. (2012). The Independence Axiom and the Bipolar Behaviorist. Experimental Economics Center Working Paper Series 2012-01, Experimental Economics Center, Andrew Young School of Policy Studies, Georgia State University.

Hey, J. D. and Orme, C. (1994). Investigating Generalizations of Expected Utility Theory Using Experimental Data. *Econometrica*, 62(6):1291–1326.

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

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

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–91.

Kaivanto, K. and 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.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, 32(2):311–328.

Lejuez, C., Read, J., Kahler, C., Richards, J., Ramsey, S., Stuart, G., Strong, D., and Brown, R. (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.

Perugini, M., Tan, J. H. W., and 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. and Couture, S. (2012). Stability of risk preference measures: results from a field experiment on French farmers. *Theory and Decision*, 73(2):203–221.

Schipper, B. C. (2012). Sex Hormones and Choice under Risk. Working Papers 2012-07, University of California at Davis, Department of Economics.

Slovic, P. (1966). Risk-Taking in Children: Age and Sex Differences. *Child Development*, 37(1):169–176.

Sutter, M., Kocher, M. G., Rtzler, D., and Trautmann, S. T. (2013). Impatience and Uncertainty: Experimental Decisions Predict Adolescents' Field Behavior. *American Economic Review*, (forthcoming).

Tversky, A. and Kahneman, D. (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.

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

von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.

Wagner, G. G., Frick, J. R., and 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. and Deneffe, D. (1996). Eliciting von Neumann-Morgenstern Utilities When Probabilities Are Distorted or Unknown. *Management Science*, 42(8):1131–1150.

Appendix A. Experimental Instructions

The experimental instructions were originally drafted in English, then translated into German to enable us to run the experiments in the Max Planck Institute's lab in Jena, Germany. In what follows, we will report the original, English versions of the instructions used for all the tasks. The German versions are available upon request.
The instructions were made up of a first welcome screen, identical for every task, and by a second screen detailing the rules for every task.

First Screen

Welcome to the Experiment.
In the experiments all payoffs are expressed in euro.
For your punctuality you receive 2.5 euro.
The experiment consists of one short task, followed by a questionnaire.
Should you have any questions or need help, please raise your hand. An experimenter will then come to your place and answer your questions in private.

HL (modeled on instructions from Holt and Laury, 2002)

You will be asked to make 10 choices. Each decision is a paired choice between "Option A" and "Option B". For each decision row you will have to choose between Option A and Option B. You may choose A for some decision rows and B for other rows, and you may change your decisions and make them in any order.
Even though you will make ten decisions, only one of these will end up affecting your earnings. You will not know in advance which decision will be used. Each decision has an equal chance of being relevant for your payoffs.
Now, please look at Decision 1 at the top. Option A pays 4 euro if the throw of the ten sided die is 1, and it pays 3.2 euro if the throw is 2-10. Option B yields 7.7 euro if the throw of the die is 1, and it pays 0.2 euro if the throw is 2-10.
The other Decisions are similar, except that as you move down the table, the chances of the higher payoff for each option increase. In fact, for Decision 10 in the bottom row, the die will not be needed since each option pays the highest payoff for sure, so your choice here is between 4 or 7.7 euro.
To determine payoffs we will use a ten-sided die, whose faces are numbered from 1 to 10. After you have made all of your choices, we will throw this die twice, once to select one of the ten decisions to be used, and a second time to determine what your payoff is for the option you chose, A or B, for the particular decision selected.

EG (modeled on instructions from Eckel and Grossman, 2008a)

You will be asked to select from among five different gambles the one gamble you would like to play. The five different gambles will appear on your screen. You must select one and only one of these gambles. Each gamble has two possible outcomes (Event A or Event B), each happening with 50% probability.
Your earnings will be determined by: 1) which of the five gambles you select; and 2) which of the two possible events occur.
At the end of the experiment, we will roll a six-sided die to determine which event will occur. If a 1, 2, or 3 is rolled, then Event A will occur. If 4, 5, or 6 are rolled, then Event B will occur.

CGP (modeled on instructions from Gneezy and Potters, 1997)

You will be given 4 euros and will be asked to choose the portion of this amount (between 0 and 4 euros, in cents) that you wish to invest in a risky option. The money not invested is yours to keep.
There is a 50% chance that the investment in the risky asset will be successful. If it is successful, you receive 2.5 times the amount invested; if the investment is unsuccessful, you lose the amount invested.
The roll of a 6-sided die determines the value of the risky asset. You will be asked to choose 3 success numbers; if one of these numbers is rolled, the risky investment is successful; if not, it is not successful.
After the decisions are made the die will be rolled and then you will be paid the amount not invested plus 2.5 times the investment if it is successful and plus zero if it is not.

Balloon (following Lejuez et al., 2002)

In this experiment you will be shown a balloon. You can click on the button labeled "Pump" to increase the size of the balloon. You will accumulate 10 euro cents in a temporary bank each time you pump. At any point, you can stop pumping up the balloon and click on the button labeled "Stop". Clicking this button will transfer the accumulated money from your temporary bank to your account.
It is your choice to determine how much to pump up the balloon, but be aware that at some point the balloon will explode. The explosion point can range from the first pump to enough pumps to make the balloon fill the area dedicated to it on the screen.It takes 100 pumps to fill the entire area.
If the balloon explodes before you click "Stop", then all money in your temporary bank is lost and your earning is zero. If you click on "Stop" before the balloon explodes, you will earn the money accumulated in your temporary bank so far.

BRET (identical to Crosetto and Filippin, 2013b)

On the PC screen you will see a field composed of 100 numbered boxes.
You earn 20 euro cents for every box that is collected. Every second a box is collected, starting from the top left corner. Once collected, the box disappears from the screen, and your earnings are updated accordingly. At any moment you can see the amount earned up to that point.
Such earnings are only potential, however, because behind one of these boxes a time bomb is hidden that destroys everything that has been collected.
You do not know where the time bomb is. You only know that it can be in any place with equal probability. Moreover, even if you collect the bomb, you will not know it until the end of the experiment.
Your task is to choose when to stop the collecting process. You do so by hitting 'Stop' at any time.
At the end of the experiment, we will randomly determine the number of the box containing the time bomb by means of a bag containing 100 numbered tokens.
If you happen to have collected the box in which the time bomb is located, you will earn zero. If the time bomb is located in a box that you did not collect, you will earn the amount of money accumulated when hitting 'Stop'.
We will start with a practice round. After that, the paying experiment starts.