Journal of Risk and Uncertainty
https://doi.org/10.1007/s11166-022-09400-0

Safe options and gender differences in risk attitudes

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

Abstract
Gender differences in risk attitudes have recently been shown to be context-dependent rather than ubiquitous. We manipulate three widely used risk elicitation tasks to test whether the presence of a safe option among the set of alternatives can explain the heterogeneity of the findings. We find that the availability of a safe option induces significant effects in two out of three tasks. Despite the well-known instability of elicited risk preferences, we show with a structural model that the effect on risk attitudes is rather stable across tasks, but not sufficiently strong to reach traditional significance levels.

Keywords Gender differences · Risk attitudes · Experiment · Safe option

JEL Classifications C81 · C91 · D81

A large body of literature in experimental economics and psychology reports gender differences in risk attitudes. Consistent with the evidence reported by some surveys (Eckel & Grossman, 2008b; Croson & Gneezy, 2009; Byrnes et al., 1999) a consensus has developed that women are more risk averse than men in non-strategic decision making over risk. Recent contributions have nonetheless found these findings to be less robust than previously thought (Nelson, 2014; Nelson, 2016) and context-dependent. For instance, women tend to behave in a less risk averse way when exposed to same-gender environments, be it in the context of female high schools (Booth & Nolen, 2012) or tutorials at the college level (Booth et al., 2014). Even within the more specific realm of choices over lotteries, the likelihood of observing gender differences in risk attitudes heavily depends on the elicitation method

✉ Paolo Crosetto
paolo.crosetto@inrae.fr

Antonio Filippin
antonio.filippin@unimi.it

[1] GAEL, INRAE, CNRS, Grenoble INP, Univ. Grenoble-Alpes, Grenoble 38000, France

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

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

Published online: 14 January 2023

adopted, as recently shown by Filippin and Crosetto (2016). Gender differences are nearly always found using some tasks, whereas only seldom or never in others.

Gender differences in behavioral traits like risk attitudes may have evolutionary roots. Trivers (1972) develops a theory of sexual selection based on parental investment and Dekel and Scotchmer (1999) extend the theory to risk aversion. The number of offspring a female is expected to have is nearly linear in the amount of resources available. Men, on the other hand, are characterized by a lower parental investment. Therefore, by competing over mates they can exploit a convex mapping from resources to reproductive success. Such differences might have driven the evolution of risk-seeking behavior in men but not in women.[1]Dreber and Hoffman (2010) acknowledge that the mechanisms triggering gender differences in humans has not yet been clearly identified. Not surprisingly, the evolutionary origin of differences in risk attitudes can be subject to alternative interpretations. For instance, Schmidt et al. (2021) argue that the difference in parental investment made more important for men to be relatively better off than their peers, while women were more concerned with the absolute outcome. Consequently, Schmidt et al. (2021) argue that the influence of social comparison on a risky decision should be higher for men than for women.

The literature in economics started only recently to look for an explanation for why women appear to be more risk averse than men. Besides the social comparison channel (Schmidt et al., 2021), an explanation has been proposed by Charness et al. (2018) who argue that the complexity of the task blurs the observability of gender differences. Charness et al. (2018) disassemble the Holt and Laury (2002) task administering the list of choices one at the time in a between-subject design. They report that gender differences are observed only in (some of) the single-choice treatments, concluding that this finding should be ascribed to the reduced complexity of the decision. While all our experimental conditions are constant in terms of social comparison, the potential implications of the role played by complexity are discussed in Sect. 3.

Filippin and Crosetto (2016) argue that two characteristics correlate with the likelihood of observing gender differences: a) the availability of a safe option among the set of alternatives, and b) the presence of fixed $50\% - 50\%$ probabilities. Unfortunately, these features do not change independently across tasks. Moreover, although all risk elicitation tasks are ultimately built on choices among lotteries, they may differ in several other aspects such as the number of choices, their mathematical vs. visual representation, or the interval of preferences that can be estimated. Hence, it is not possible to identify the role played by the different features of the task using the results already available in the literature, and an experimental investigation becomes necessary.

[1] Studies involving apes are used to address the origins of human behavior. Several contributions show that risk attitudes are indeed shaped by evolution, although without providing evidence along a gender perspective. These studies find for instance that bonobos are more risk averse than chimpanzees (Rosati & Hare, 2013; Heilbronner et al., 2008). The two species derive from a common ancestor but evolved differently, with chimpanzees developing a riskier foraging strategy. Differences in risk taking between the two species are nowadays observed within captive populations, i.e. among subjects that are fed and do not need to undertake any foraging strategy.

In this paper, we test whether the availability of a riskless alternative, defined as a degenerate lottery in which a positive amount of money can be obtained with certainty, induces gender-specific behavior and can account for the observed heterogeneity of results. To isolate the role of riskless alternatives we carry out a set of controlled experiments involving 1085 subjects in which we add a safe option to (or remove from) three well-known and widely used tasks, keeping all else equal: a multiple price list (Holt & Laury, 2002), an ordered lottery selection (Eckel & Grossman, 2002), and the Bomb task (Crosetto & Filippin, 2013). Replicating the manipulation proves necessary since risk elicitation tasks vary in several dimensions and map choices into risk aversion parameters in different ways. Moreover, risk attitudes elicited with different tasks display a notoriously low consistency (Isaac & James, 2000; Reynaud & Couture, 2012; Deck et al., 2013; Crosetto & Filippin, 2016; Menkhoff & Sakha, 2017; Pedroni et al., 2017). We hence build new ad-hoc treatments in which the safe option is carefully manipulated over several tasks.

To the best of our knowledge this is one of the first attempts to systematically analyze and provide causal evidence about a determinant of gender differences in risk attitudes, and it is the first to do so across different risk elicitation tasks.[2] We find evidence that the availability of a riskless alternative indeed makes gender differences more prominent. The effect is consistent in sign over the three different risk elicitation methods once results are compared using a structural estimation. The magnitude of the effect is weak, however, so that differences in the choices with and without the safe option do not reach traditional significance levels in a difference-in-difference structural estimation. We conclude that the presence of a safe option is not the only factor affecting the emergence of a gender gap in risk elicitation. Nonetheless, safe options seem to play a role that needs to be further tested.

The structure of the paper is as follows. We introduce the general structure of our experimental design in Sect. 1. Sections 2.1, 2.2, and 2.3 present in detail the baseline version of each task and the corresponding manipulation, and then test the presence of gender differences using both non-parametric, and, when supported by the data, parametric tests. Section 3 compares the results across risk elicitation methods using a structural estimation approach. Section 4 concludes.

1 Methods

Risk attitudes are a latent construct, and as such can only be indirectly and imperfectly measured. Given the heterogeneous features of the different elicitation methods, it is not surprising that measures of risk attitudes tend to show low correlations across domains and tasks. Such an instability of results has been emphasized also

[2] The careful reader may have noticed that the Allais paradox is per se a test of the role of a riskless alternative. The experimental literature using the Allais paradox, however, is not informative towards our research goal because usually results are not displayed by gender. The few exceptions are Petit et al. (2011) who find that women are more prone to the paradox and choose more often the safe alternative (N = 938, of which 611 women), and Da Silva et al. (2013) who do find on the contrary that men are more prone to the paradox (N = 120).

along a gender dimension. This evidence imposes several key requirements in order to obtain a clean test of the effect of safe option on risk taking by gender. In particular, we need to a) exogenously manipulate the presence of a safe option in a task *ceteris paribus*, i.e. keeping its structure unchanged; and b) replicate the exercise in more than one task, because the heterogeneity of results renders the generalization of results from a single elicitation method a questionable exercise.[3]

The first question to answer is therefore which elicitation methods should be used. We believe that in order to be robust, our findings must rely upon tasks delivering clear and different results along a gender perspective. Significant gender differences are a systematic finding in an ordered lottery choice task *à la* (Eckel & Grossman, 2002, henceforth:*EG*) and in the Investment Game by Gneezy and Potters (1997). In contrast, gender differences are rarely found and, when found, small in magnitude in the most widely used risk elicitation task, (Holt & Laury, 2002, henceforth:*HL*), as documented by the meta-analysis of Filippin and Crosetto (2016). Finally, the behavior of men and women is indistinguishable when preferences are elicited with the Bomb Risk Elicitation Task (Crosetto & Filippin 2013, henceforth:*BRET*).

The likelihood of observing gender differences strongly correlates with the presence and focality of a safe alternative. *EG* and the Investment Game by Gneezy and Potters (1997) present a focal safe option in the form of a degenerate lottery yielding the same payoff irrespective of the random event. *HL* does not provide an explicit safe option but allows the subject to get a minimum payoff with probability 1. Finally, the *BRET* does not allow the subjects to earn any positive amount of money with probability one.

The Investment Game and *EG* share many characteristics. They both feature fixed $50\% - 50\%$ probabilities, they both cannot identify risk neutral and risk loving preferences, and in both tasks, gender differences are a nearly ubiquitous finding. Since in the Investment Game a safe option is present but virtually never chosen, and given the similarity of the two tasks, we decided to focus on *EG* (together with *HL*, and *BRET*).

We create new versions of each task either introducing (*HL* and *BRET*) or removing (*EG*) a safe option. Our aim is to reduce changes to a minimum, in order to preserve all the idiosyncratic characteristics of each task but still be able to causally identify the role played by the riskless alternative. Towards this goal we assume that agents are characterized by classic CRRA preferences:

$$U(x) = \frac{x^{1-\rho}}{1-\rho},$$

where $\rho \neq 1$ represents the coefficient of relative risk aversion and $U(x) = \log x$ when $\rho = 1$. This assumption a) allows us to build a treatment version of each task that is isomorphic to the baseline condition under the null assumption that the safe option is irrelevant; b) helps to make results comparable across tasks.

[3] Building an ad-hoc task would possibly allow us to test the effect of a safe option in a cleaner manner, but the heterogeneity of results in this literature would prevent any generalization. Therefore, we believe that the first test needs to be done with the same tasks that have been used to build the current consensus.

Table 1 Distribution of the 1085 subjects by task and gender

Task	N_{task}	Version	$N_{condition}$	Men	Women
Holt and Laury (2002)	344	Baseline: *HL*	179	84	95
		Treatment: *HLsafe*	165	79	86
Bomb Risk Elicitation Task	462	Baseline: *BRET*	271	106	165
		Treatment *BRETsafe*	191	73	118
Eckel and Grossman (2002)	279	Baseline: *EGsafe*	145	67	78
		Treatment: *EGnosafe*	134	57	77

Another source of heterogeneity in the results might stem from the repetition of the choice. It has been shown that (part of) the subjects make choices that are even negatively correlated over time (Isaac & James, 2000).[4] We hence opt for a pure between-subject experiment in which each subject participates in only one experimental condition. A grand total of 1085 subjects took part to our six conditions (3 tasks times 2 treatments design). The distribution of subjects by condition and the breakdown by gender are detailed in Table 1.

The sample sizes have been obtained by weighing two different principles: matching existing samples for the baseline conditions, and getting enough power to detect an effect in the treatment conditions under our hypothesis. Related literature shows us that the gender effect in the BRET is nearly zero ($d = 0.01$, Crosetto & Filippin, 2013), it is $d = 0.17$ in HL (Filippin & Crosetto, 2016) and $d = 0.46$ in EG (Nelson, 2014). Under our conjecture that the presence and focality of a safe option drives gender differences in risk attitudes, we expected hence the effect in the *safe* version of the BRET and HL to reach EG-like levels of $d \sim 0.45$. Under this assumption, we need 75 subjects per condition to detect an effect at 80% power.

1.1 Experimental procedures and details

The experimental sessions were run in 2014 (*BRETsafe*, *HLsafe*, *HL* extra sessions) at the Laboratory of the Max Planck Institute of Economics, and in 2016 (*EGnosafe* and *EGsafe* extra sessions) at the Laboratory of the Friedrich Schiller University, both in Jena, Germany.[5]

[4] Crosetto and Filippin (2013) show that a roller-coaster behavior is observed even when repeating the same task several times. Menkhoff and Sakha (2017) report that if the subjects fail to properly reduce the compound lottery generated by within-subjects designs, instability and inconsistencies are to be expected. Pedroni et al. (2017) show large within-subjects inconsistencies on a large sample across seven risk elicitation tasks.

[5] The data of the baseline *BRET* are the same as in Crosetto and Filippin (2013). Part of the baseline *HL* and *EGsafe* data are the same as in Crosetto and Filippin (2016). However, given the focus on gender of this study, we needed to increase the overall sample size to support a gender comparisons. We hence planned three additional *HL* and *EGsafe* sessions. The treatment conditions are entirely original data.

The experimental procedures were identical for all tasks.[6] Subjects entered the laboratory, and instructions were both read aloud and available on screen. The English translation of the original instructions in German is available in Appendix 1. Control questions about the experimental procedure and tasks were asked, and subjects were allowed to continue only after having replied correctly to all questions. Then the subjects faced the task, one shot.[7]

After all subjects had completed the task, they were exposed to a short questionnaire including demographics and a self-reported measure of the perceived complexity of the task. The randomization of the assignment to the six conditions should guarantee a balanced distribution of risk attitudes. However, in order to allow us to control for possible unbalances in mid-size samples like those we gathered, we exposed the subjects to the SOEP self-reported measure of attitude toward risk (introduced in the risk elicitation literature by Dohmen et al., 2011).

2 Results

2.1 Experiment 1: Multiple price list

2.1.1 Methods

Baseline condition: The classic Holt and Laury task (HL) The multiple price list format is a general procedure used to elicit values from a subject. Applied to risk, it consists of giving the subject an ordered list of binary choices between lotteries. The most widely known implementation has been provided by Holt and Laury (2002), which is, to date, the most popular risk elicitation mechanism according to the number of citations.

In the *HL* task, subjects face a series of choices between pairs of lotteries, ordered by increasing expected value. The set of possible outcomes is common to every choice, with one lottery safer (i.e., with lower variance) than the other. The increase in expected value across lottery pairs is obtained by increasing the probability of the "good" event (see Table 2). At the end of the experiment, one row is randomly chosen for payment, and the chosen lottery is played to determine the payoff.

The subjects make a choice for each pair of lotteries, switching at some point from the safe to the risky option as the probability of the good outcome increases. The switching point captures their degree of risk aversion. For instance, a risk-neutral subject should start with Option A, and switch to B from the fifth choice on. Never choosing the risky option or switching from B to A are not infrequent and are regarded as inconsistent choices. They can be rationalized only adding a stochastic component in the decision process.

[6] The custom experimental software for each task, written in Python, is available upon request. The full dataset and the scripts used to generate all results of this paper are available online at https://github.com/paolocrosetto/Safe_options_risk_attitudes_gender_data_and_analysis.

[7] A trial run of the task was provided for the *BRET* and *BRETsafe* tasks. The presence of such a trial does not affect the results, see Crosetto and Filippin (2013).

Table 2 The standard Holt and Laury task

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

In the *HL* condition, subjects choose between lotteries characterized by uncertain outcomes with the exception of row 10 in which two sure amounts are compared. The lottery played at the end of the experiment is selected randomly. Therefore, there is no way subjects can avoid uncertainty so that this condition can be used to build a pure measure of risk aversion in which certainty effects play no role.

Treatment condition: HLsafe We introduce a safe option in the *HL* task by replacing Option A with a sure amount (see Table 3). Note that the amount changes across rows, in order to eliminate any difference in the fundamentals between the two treatments, except the availability of a safe choice. Every lottery proposed as Option A in Table 2 has been replaced with its certainty equivalent for an agent characterized by a CRRA utility function and a risk aversion parameter such that she would switch to Option B in that row. For instance, a subject should switch in the 6^{th} row if his risk aversion coefficient is ($\rho \in [0.15, 0.41]$). Hence, the safe amount in the sixth row (3.7) has been derived as the certainty equivalent of the lottery (4 € with $p = 0.6$; 3.2 € with $p = 0.4$) assuming ρ equal to the midpoint of the interval [0.15, 0.41]. Under

Table 3 The Holt and Laury task with a safe option

	Option A		Option B		
1	3.3 €	1/10	7.7 €	9/10	0.2 €
2	3.4 €	2/10	7.7 €	8/10	0.2 €
3	3.5 €	3/10	7.7 €	7/10	0.2 €
4	3.5 €	4/10	7.7 €	6/10	0.2 €
5	3.6 €	5/10	7.7 €	5/10	0.2 €
6	3.7 €	6/10	7.7 €	4/10	0.2 €
7	3.7 €	7/10	7.7 €	3/10	0.2 €
8	3.8 €	8/10	7.7 €	2/10	0.2 €
9	3.9 €	9/10	7.7 €	1/10	0.2 €
10	4 €	10/10	7.7 €	0/10	0.2 €

the null assumption that the safe option does not matter, the two conditions *HL* and *HLsafe* are isomorphic.[8]

The goal of this treatment is to exogenously manipulate the presence of a safe option within a set of otherwise equivalent alternatives. Such a manipulation can be considered rather weak, however. In fact, the multiple price format is also likely to induce a comparison of risky alternatives across rows. Moreover, the necessity of maintaining a direct comparability across conditions imposes to substitute each Option A with a different safe amount. Both factors are likely to dilute the impact of the introduction of a safe choice in every row.

2.1.2 Results

In *HL,* a higher degree of risk aversion induces the decision maker to choose Option A for a higher probability of the good outcome. The switching point from Option A to Option B is, therefore, used to measure risk aversion. Unfortunately, it is not infrequent to observe that participants never choose Option B (i.e. they prefer 4 € to 7.7 €) or switch from Option B to Option A. Such patterns of decision are regarded as inconsistent because they cannot be rationalized by a deterministic expected utility maximizer. In our data, such patterns of decision are displayed by 28 subjects (18 women, 10 men) in the Baseline *HL* and by 6 subjects (equally divided by gender) in *HLsafe*.[9] In this section, we present the descriptive statistics and the basic results removing the inconsistent subjects. The behavior of multiple switchers can be analyzed in a structural model including a stochastic component (see Sect. 3 below).

Figure 1 displays the distribution of the switching point by gender and treatment. No clear pattern can be identified in the Baseline *HL* (left panel). In contrast, in the *HLsafe* treatment it clearly emerges that men are more represented than women in the risk seeking and risk neutral domains (switching points 3 to 5), while the opposite occurs in the risk aversion domain (switching points 6 to 7).

The main advantage of the Holt and Laury task is that it has been built on the expected utility model. Therefore, the different switching points map almost linearly into coefficients of relative risk aversion (see Crosetto & Filippin 2016, Fig. 3). As a result, using the average choice to summarize the results does not introduce distortions in the analysis.

Table 4 shows that significant gender differences appear only when introducing a safe option. In the Baseline *HL,* the average switching points of men and women do not significantly differ according to a Mann Whitney test ($p = .18$). This evidence is perfectly in line with the meta-analysis of Filippin and Crosetto (2016), showing

[8] Other methods used to compute an amount row by row as similar as possible to the corresponding lottery in Treatment 1 would deliver virtually identical results. For instance, using the expected value of the lottery would deliver slightly different amounts in only two out of ten rows. In contrast, using the same amount in all the ten lotteries would change the underlying incentives across conditions.

[9] The fraction of inconsistent subjects in our data (15.6%) is in line with the literature, as reported by Filippin and Crosetto (2016).

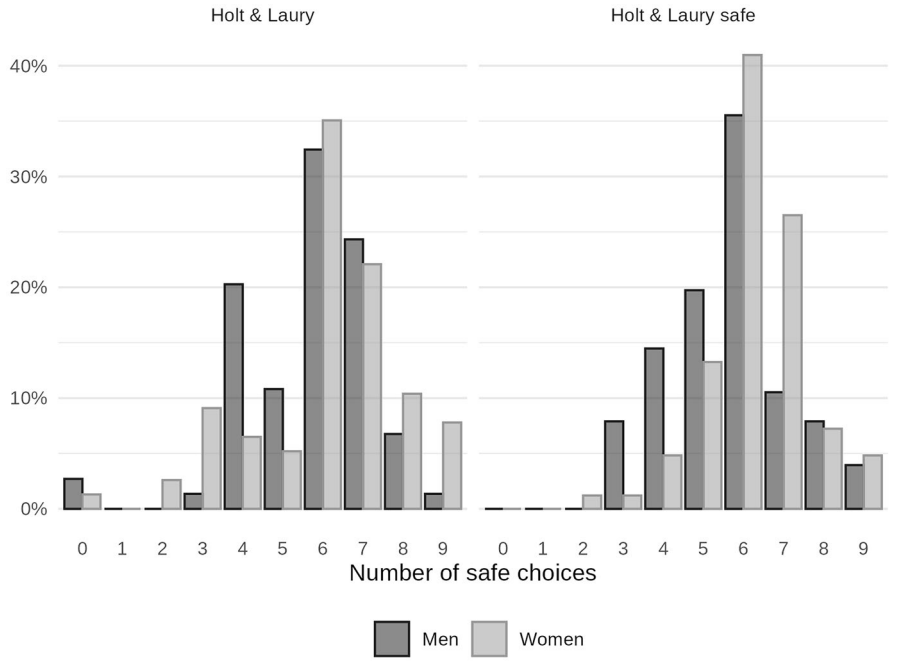

Fig. 1 Distribution of the switching point by gender and treatment in the Holt and Laury task

that the point estimate typically differs by gender but does not reach traditional significance levels given the usual sample size of lab experiments. Strictly speaking, we cannot conclude that the two groups are characterized by the same degree of risk aversion because what we observe could be a false negative.

In the *HLsafe* treatment, the difference in the average switching point is instead highly significant ($p = .004$). We cannot reject normality for the *HLsafe* treatment, and we can thus run a t-test, that is also highly significant ($p = 0.008$). Furthermore, the magnitude of the difference is also larger, as captured by the Cohen's *d* (Cohen,

Table 4 Average switching point by gender and treatment in the Holt and Laury task

		N	Average # Safe choices	Std. Dev.	Cohen's d	Mann Whitney	Norm. test[o]	t-test
Holt & Laury	Men	74	5.73	1.52	.185	.183	0.006	
	Women	77	6.04	1.80				
	Diff. (M-W)		*-0.31*					
Holt & Laury safe	Men	76	5.66	1.47	.427	.004	0.754	0.008
	Women	83	6.24	1.26				
	Diff. (M-W)		*-0.58*					

[o]Jarque-Bera skewness-kurtosis normality test

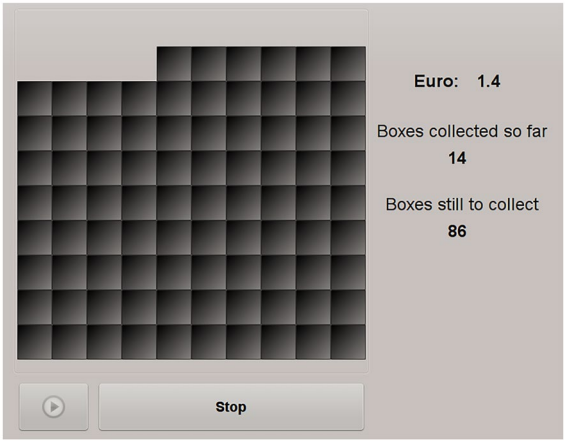

Fig. 2 The Bomb Risk Elicitation Task interface after 14 seconds

1988).[10] Hence, the comparison across treatments shows that the introduction of a riskless alternative qualitatively affects gender differences in the Holt and Laury task. We show in Sect. 3 below that this effect is not strong enough to translate into significant differences across treatments in a structural model, though.

2.2 Experiment 2: The Bomb Risk Elicitation Task (BRET)

2.2.1 Methods

Baseline condition: Standard BRET Our baseline condition uses the dynamic version of the *BRET*, a risk elicitation task introduced by Crosetto and Filippin (2013). Subjects face a 10×10 square in which each cell represents a box. Ninety-nine boxes are empty, while one contains a time bomb. Every second one box is automatically collected. (see Fig. 2).

The subjects have to decide how many boxes to collect, i.e. $k^* \in [0, 100]$, by clicking the stop button. 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

[10] Cohen's d is a measure of the size of an effect that is independent of the sample size. It is computed as:

$$d = \frac{\bar{X}_f - \bar{X}_m}{\sigma},$$

where $\bar{X}_m$ and $\bar{X}_f$ are the average group choices and σ is the pooled standard deviation. Cohen (1988) indicates thresholds for interpreting his d: referring to aggregate differences, 0.2 should be considered a small effect, 0.5 a medium effect, and from 0.8 on a large effect.

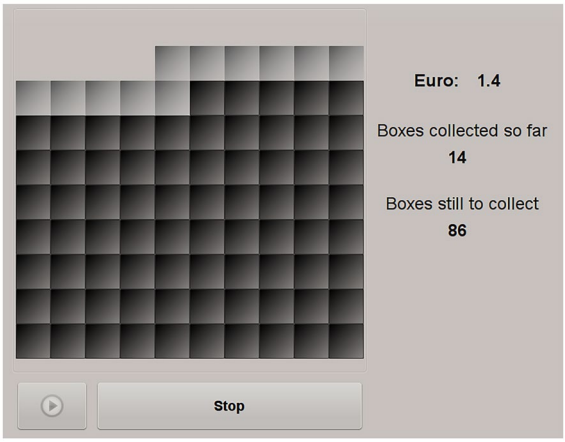

Fig. 3 The Safe-Bomb Risk Elicitation Task interface after 14 seconds

out his earnings. In contrast, if $k_i^* < b$, subject i leaves the minefield without the bomb and receives 10 euro cents for every box collected.

The *BRET* interface provides a visual representation of probabilities that allows subjects to keep track of how many boxes have been collected and how many are left. Subjects' decision can be formalized as the choice of their favorite among the set of 101 lotteries fully described by the parameter $k \in [0, 100]$, which summarizes the trade-off between the amount of money that can be earned and the likelihood of obtaining it:

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

The degree of risk aversion negatively correlates with the choice of k and a risk-neutral subject should choose $k^* = 50$. The *BRET* does not provide safe options as the only amount that can be secured with certainty is zero, by chosing $k = 0$ or $k = 100$. Hence, the choice of k implies a comparison between uncertain amounts only and it can therefore be used to build a pure measure of risk aversion in which certainty effects play no role.

Treatment condition: BRETsafe In the *BRETsafe* condition, a riskless alternative is made available by preventing the time bomb to be in the first 25 boxes. In other words, by choosing $k \leq 25$ subjects can secure a positive amount without incurring any risk. Figure 3 displays the graphical interface of the *BRETsafe* after 14 seconds.

For instance, by choosing $k = 20$, the subject earns for sure the value of 20 boxes (2 euro) because the time bomb can only be in $b \in [26, 100]$. In contrast, if the choice is $k = 40$, the underlying lottery implies earning either 4 euro with

probability $(100-40)/75$ or nothing with probability one fifth $((40-25)/75)$. More generally, each lottery is then characterized by:

$$L^k = \begin{cases} \left[k \text{ with prob. } 1 \right. & \text{if } k \leq 25 \\ \left[\begin{array}{l} 0 \text{ with prob. } \frac{k-25}{75} \\ k \text{ with prob. } \frac{100-k}{75} \end{array} \right. & \text{if } 25 \leq k \leq 100 \end{cases}.$$

Note that for $k \geq 25$ the expected utility in the *BRETsafe* condition is a linear transformation of the Baseline under the reasonable assumption that $u(0) = 0$. Therefore, an expected utility maximizer should make the same choice in the two conditions as long as his optimal choice is $k \geq 25$. The only effect of the safe manipulation is that of inducing the more risk averse subjects to choose the highest safe option $k = 25$.[11] Any choice $k < 25$ violates the monotonicity assumption and would be irrational. As a result, we can expect to observe a slightly higher average choice in the *BRETsafe*. In any case, we are not interested in a point prediction of the average behavior across treatments, but only in the different effect that a safe alternative can induce along a gender dimension.

2.2.2 Results

As explained above, the number of boxes k captures the degree of risk aversion in the *BRET*. In this section, we present the descriptive statistics and the basic results eliminating two subjects making dominated options - one stopping after one box and one collecting all 100 boxes, both in the Baseline *BRET* treatment.[12]

Figure 4 shows a kernel density of the choices by gender and treatment. In the Baseline *BRET* (left panel), the two distributions nicely overlap with the exception that women tend to make more disperse choices. Looking at the *BRETsafe* (rigth panel) two things are immediately evident. First, women choose the safe option $k = 25$ more often than men, though not significantly so (Fisher exact test, $p = .542$). Second, the distribution of men is now shifted to the right as compared to that of women. In particular, men are now overrepresented in the risk loving domain.

Under a consequentialist approach choices should be affected only through the likelihood of opting for the riskless alternative. However, a second and indirect effect can occur through choices not directly affected by the manipulation of the safe option. This second effect could not be detected using binary choices such as in the Allais paradox. However, in a more complex environment, like that implemented in this paper, things are more subtle. The introduction of a safe option, even though not directly chosen, can affect subjects' risk tolerance over the whole set of alternatives,

[11] Assuming a CRRA utility function only the subjects characterized by $\rho \geq .658$ should opt for the safe option.

[12] Results are robust to the inclusion of these two subjects, that we wish nonetheless to exclude since they submitted clearly dominated choices.

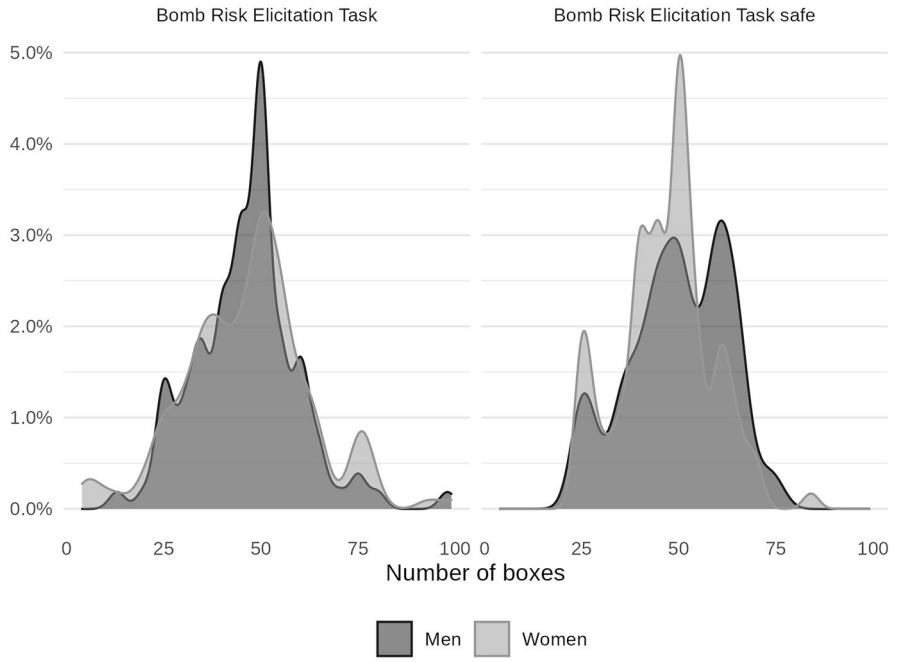

Fig. 4 Kernel density (bandwidth adjustment 0.7) of the choices by gender and treatment in the Bomb Risk Elicitation Task

inducing them to choose a different lottery than what would have been chosen otherwise.[13] Analyzing this kind of effects goes however beyond the goal of this paper, as the focus of our research is to test whether women are more risk averse when a safe option is available. The comparison of the choices in the two experimental conditions indicates that this is the case also in the *BRETsafe*, at the same time suggesting that the indirect effect of the safe option (i.e., through choices other than the safe one) may even be prevalent.

The *BRET* entails 101 possible choices and the kernel density above provides a partial picture because it necessarily smooths the actual distribution. Hence, in Table 5 we report the average choices, which confirm that introducing a safe option in the *BRET* generates significant gender differences. In the baseline version of the task, the behavior of men and women is indistinguishable, and differently from the *HL* case here the point estimate is virtually identical. In the *BRETsafe,* women turn out to be relatively more risk averse than men according both to a Mann-Whitney test ($p = .079$) and also to

[13] The intuition is pretty similar to the Give vs. Take manipulation in Dictator Games. Bardsley (2008) and List (2007) show that the possibility of taking affects not only those whose choice was truncated by the lower bound of zero in the Give framework. In contrast, the whole distribution, including the counterparts of those who give a positive amount, shifts towards more selfish decisions once taking is a practicable alternative.

Table 5 Average number of boxes chosen by gender and treatment in the Bomb Risk Elicitation Task

		N	Average Choice	Std. Dev.	Cohen's d	Mann Whitney	Norm. test[o]	t-test
Bomb Risk Elicitation Task	Men	105	46.38	13.3	-.018	.658	.025	
	Women	164	46.65	16.3				
	Diff. (M-W)		-.27					
Bomb Risk Elicitation Task safe	Men	73	49.79	12.7	.254	.079	.526	.097
	Women	118	46.72	11.8				
	Diff. (M-W)			3.07				

[o]Jarque-Bera skewness-kurtosis normality test

a t-test ($p = .045$) performed as the normality of the distributions is not rejected. The Cohen's *d* shows that the effect of the safe option is small in size (about 0.25). In Sect. 3 we show that, like in the case of *HL*, the magnitude of this effect is not sufficiently strong to detect significant differences across the two conditions.

2.3 Experiment 3: Ordered lottery selection

2.3.1 Methods

Baseline condition: The classic Eckel and Grossman task (EGsafe) In ordered lottery selection tasks, subjects make a single choice picking one out of an ordered set of lotteries. This method has been first introduced in the literature to measure risk preferences by Binswanger (1981). A popular version is the one proposed by Eckel and Grossman (2002, 2008a), which has often been referred to in the literature about gender differences. In the original *EG* task, subjects make their choice from a set of five lotteries characterized by a linearly increasing expected value as well as a larger and larger variance (see Table 6, panel (a)).

The risk-reward trade-off is induced by manipulating the outcomes of each lottery while keeping the probability of each outcome fixed at 50%. A risk-neutral subject should choose lottery 5, as it yields the higher expected value. Increasing degrees of risk aversion induce choices with lower expected returns. Crucially towards our goal, the menu of choices includes a degenerate lottery that if chosen allows the subjects to secure a positive amount – in this case 4 Euro – without incurring any risk.[14]

[14] The EG task has later been proposed in a version including a 6^{th} option, with the same expected value as the 5^{th} but with a larger variance, designed to identify risk seeking subjects. The two versions do not significantly differ for the purpose of our study.

Table 6 Variations of the Eckel & Grossman task used in the paper

	Event	Probability	Outcome
(a) Eckel & Grossman			
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 €
(b) Eckel & Grossman nosafe			
1	A	50%	4.5 €
	B	50%	3.6 €
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 €

Treatment condition: EGnosafe In case of the *EG* task, the experimental manipulation simply amounts to proposing a version of the task in which the safe choice is replaced with an equivalent risky choice (see Table 6, panel (b)). Like in the case of the Holt and Laury task, the amounts have been chosen in such a way to be isomorphic to the original task for agents characterized by a CRRA utility function. In both conditions, the cutoff level of risk aversion that makes the agent switch from Lottery 1 to Lottery 2 is $\rho = 2$. If the availability of a safe alternative triggers gender differences, we should observe that the behavior of men and women is more similar in this treatment than in the baseline condition.

2.3.2 Results

Subjects' decisions in each treatment are summarized in Fig. 5. The Baseline *EGsafe* (right panel) displays a large difference in the average choice of men and women, as usually found in the literature. Part of the gap derives from women choosing more often the safe option: 20.5% against only about 6%, a difference that is statistically

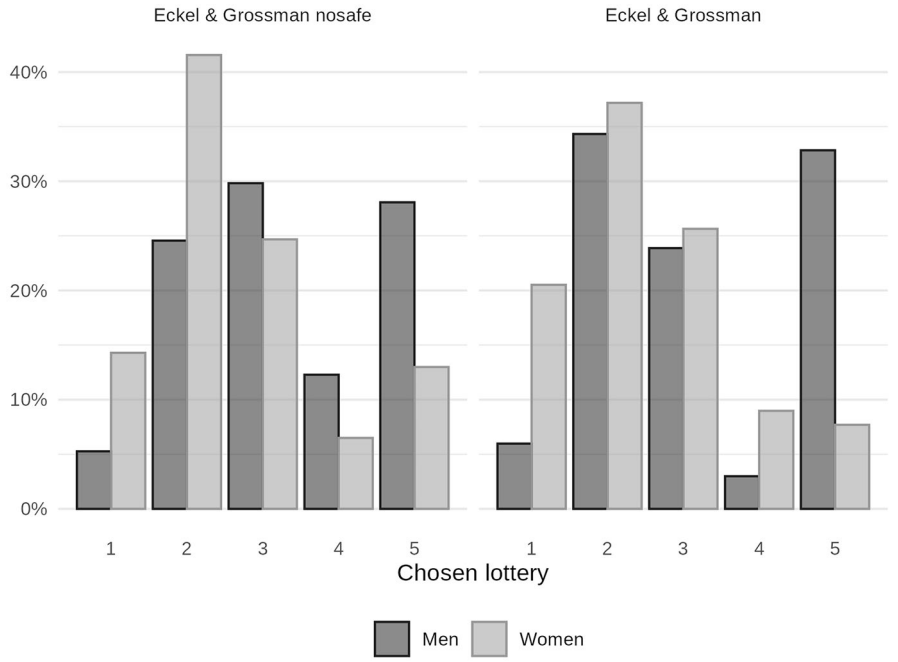

Fig. 5 Distribution of lottery choices by gender and safe option in the Eckel & Grossman task

significant according to a two-sided Fisher exact test ($p = .020$).[15] In contrast, men are disproportionately more likely to choose the riskiest alternative (32.8% vs. 7.7%). Not surprisingly, the distribution of choices by gender turns out to be significantly different according to a Mann-Whitney test ($p = 0.001$).

Removing the riskless alternative in the *EGnosafe* treatment affects the distribution of the choices, without however changing the overall picture. The equivalent (according to EU theory) safer lottery in this case involves a small risk and is chosen by a lower fraction of women (14.3%). This fraction is still higher than that of men (which does not change), though in this case not significantly so according to a two-sided Fisher exact test ($p = .156$). As already noted for the *BRETsafe*, the availability of a safe option may significantly affect the decisions even without immediately translating into a certainty effect. In the *EGnosafe,* we can see that women are relatively more likely to choose lottery 5 than in the Baseline EG. However, in the *EGnosafe* gender differences are still evident, as confirmed by a Mann-Whitney test ($p = 0.001$).

[15] The reason why gender differences in the certainty effect are significant here but not in the *BRETsafe* may be due to the different salience of the safe option. In the *EGsafe* task the safe option is one out of only five alternatives, and moving from lottery 1 to lottery 2 makes a clear difference in terms of risk incurred. In contrast, in the *BRET* the salience of the riskless alternative is likely diluted by the fact that the risk of the bad outcome increases at a very low rate (1.33% per additional box.)

In case of the *EG* task computing the average choice is not a meaningful exercise for two reasons. The first is that the menu of lotteries is ordinal but has no cardinal meaning. The second is that unfortunately the mapping from choices to coefficients of relative risk aversion is highly non-linear. Therefore, a quantitative assessment of the effect of the experimental manipulation is left to the structural model in the next section. Nevertheless, a bird's-eye view of the results is enough to note that removing the safe option does not eliminate gender differences. While the introduction of a riskless alternative induces gender differences in the elicitation methods where usually they are not observed (*HL* and *BRET*), the reverse pattern does not hold in the *EG* task.

3 Maximum likelihood estimation

The evidence gathered separately with our experiments is informative but not conclusive. On the one hand, results in the *HLsafe* and *BRETsafe* conditions show that the availability of a safe option plays a role in inducing observable gender differences in risk attitudes. On the other hand, the findings of the *EGnosafe* treatment indicate that there are other determinants at work because women behave in a more risk averse manner also without the safe option.

The choices across the three tasks are not directly comparable, but the fact that all tasks rely upon monetary lotteries allows us to use the coefficient of risk aversion as a common metric. However, specific features of the elicitation methods as well as additional sources of heterogeneity need to be taken into account in order to properly assess the role played by the availability of a riskless alternative.

The first concern is to exclude that differences across experimental conditions may be due to a different degree of complexity. Subjects' choice across treatments and/or conditions may be influenced by their comprehension of the task in a direction that correlates with the presence of the riskless alternative. In our opinion complexity may affect the results in two ways. The first channel is direct. The complicated nature of a decision may induce more risk averse choices, while a simpler environment could be more hospitable for risk taking. The second transmission

Table 7 Perceived difficulty of the task across elicitation method, experimental condition, and gender

Task	Gender	Mean (st.dev)		Diff	WRST*
		No safe	Safe		*p*-value
Bomb Risk Elicitation Task	M	1.71 (2.27)	1.49 (2.03)	-0.22	0.83
	W	2.91 (2.87)	2.07 (2.3)	-0.84	0.05
Holt & Laury	M	2.06 (2.1)	1.32 (1.7)	-0.74	0.01
	W	2.53 (2.29)	2.19 (2.24)	-0.34	0.23
Eckel & Grossman	M	1.46 (1.59)	2.03 (2.15)	0.57	0.21
	W	2.1 (2.1)	2.72 (2.86)	0.62	0.60

*Wilcoxon rank-sum test

Table 8 Standard deviation of the choices by elicitation method, experimental condition, and gender

Task	Gender	No safe	Safe	Diff
Bomb Risk Elicitation Task	M	13.34	12.70	-0.64
	W	16.29	11.81	-4.48
Holt & Laury	M	1.61	1.47	-0.14
	W	1.84	1.26	-0.59
Eckel & Grossman	M	1.27	1.38	0.11
	W	1.20	1.15	-0.06

mechanism is instead indirect and operates by increasing the likelihood of making mistakes. By increasing the variance of the choices, complexity would decrease the likelihood of detecting significant differences across groups. The introduction of a safe alternative may indeed reduce the complexity of the task, as shown by Taylor (2016) who finds that the availability of a certain option reduces errors for the lowest-ability subjects.

Given the research question of this paper, complexity is relevant as long as it has a different impact on men and women. Charness et al. (2018) find that the likelihood of observing gender differences increases in a simpler environment, namely deconstructing the HL task in a between-subject design. The HL task is indeed a rather complex environment, and therefore the results in Charness et al. (2018) do not seem to explain why men and women usually behave in a similar manner in a simple elicitation method like the BRET. However, they definitely call for a proper investigation of the potential effect of complexity in our setting.

In the experiment we asked subjects in each experimental condition to report the perceived degree of complexity. Table 7 reports the average results by task and gender testing whether the safe-choice manipulation induces a significant change in the perceived difficulty of the task. The scale of the variable ranges from 1 to 10: all the experimental conditions are perceived as very easy on average. Even if some significant differences across groups emerge, it is not possible to recognize any systematic pattern. The safe option seems to make the HL and the BRET easier, while the experimental manipulation operates in the opposite direction in the EG task.

Table 8 focuses instead on the variance of the choices, under the assumption that a simpler environment should reduce the number of mistakes. Also, in this case some differences in the conjectured direction can be observed in the HL and BRET, but not in EG.[16] However, the changes do not match closely those in Table 7, so that a correlation between complexity and the presence of a safe option is not apparent. The absence of a systematic pattern suggests the use of a unifying framework such as a structural estimation in which a non-profit-maximizing behavior triggered by complexity can be explicitly modeled.

[16] Note that here statistical significance cannot be computed as we have just one observation per condition.

Table 9 Descriptive statistics of the Socio-Economic panel risk question, by task, condition and gender

Task	Condition		SOEP mean	SOEP st.dev.	MW* p-value
Holt & Laury	Baseline	M	5.56	1.98	< .001
		W	4.55	2.02	
	safe	M	5.08	2.26	.003
		W	4.1	1.87	
Bomb Risk Elicitation Task	Baseline	M	5.36	1.95	.036
		W	4.84	2.06	
	safe	M	5.15	2.31	.013
		W	4.34	1.89	
Eckel & Grossman	nosafe	M	5.39	2.29	.029
		W	4.57	2.02	
	Baseline	M	5.39	2.06	.179
		W	4.95	2.19	

*Mann-Whitney test

An additional potential confounding factor besides complexity is probability weighting. When subjects maximize their objective function holding a distorted perception of probabilities, they may change their choices across treatments in a direction that correlates with the presence of the riskless alternative. The tasks used in our experiment are not ideal to estimate the shape of a probability weighting function. However, borrowing the functional form and the value of the parameters by gender estimated in Fehr-Duda et al. (2006), we find that probability weights cannot rationalize our results. Women should be more likely to choose the safe option in the *EGsafe* than the corresponding lottery in the *EGnosafe*, as we indeed observe. However, a similar prediction should apply to men, something contradicted by our evidence. In the *HLsafe,* the effect of distorting objective probabilities should also be similar for men and women, while in the *BRETsafe* the prediction is counterintuitive because characterized by a discontinuity in the choices not mirrored by the data.[17]

The third concern to consider is the possible heterogeneity of subjects across conditions, particularly in terms of average risk preferences. We can control whether groups are balanced exploiting the self-reported degree of risk tolerance collected through the SOEP question. Table 9 shows that in general women report to be more risk averse than men, as usually found in the literature. This difference emerges across all our experimental conditions with the exception of the *EGsafe,* in which, they are not statistically different from men. While this result seems to violate the assumption that groups are ex ante equal thanks to the randomization procedure, it could also explain why the experimental manipulation appears to be less effective in the EG condition.

[17] Using cumulative weights would instead be an uninformative exercise. The distortion of probabilities would imply a sort of optimism or pessimism (i.e. the better outcome perceived not as likely as the worse one), which cannot be disentangled from risk aversion when dealing with binary lotteries (l'Haridon and Vieider, 2016).

The specific inner working of each elicitation method, complexity, and heterogeneity may, in principle, affect the measurement of a latent construct such as risk attitudes. A structural estimation provides an encompassing framework in which all these relevant factors can be taken into consideration, while at the same time capturing the regularities across methods. In order to measure the impact of the safe option in a richer and unifying environment we, therefore, estimate a random-parameter structural model with maximum likelihood (ML) following Apesteguia and Ballester (2018).[18]

This procedure requires, first, to make assumptions about the form of the utility function, the stochastic component, and the coding of the experimental choices. We assume the standard CRRA utility function:

$$U(x) = \frac{x^{1-\rho}}{1-\rho},$$

in which ρ represents the coefficient of relative risk aversion. In random parameter models, subjects make mistakes while comparing the utility of two lotteries because their preferences are fuzzily defined around a true value. A random error distorts subjects' risk aversion parameter $\hat{\rho} = \rho + \varepsilon$. We assume, along with Apesteguia and Ballester (2018), that ε follows a logistic distribution.

By denoting the two lotteries under consideration as "left" (L) and "right" (R), the probability of subjects choosing R has the closed form:

$$Pr(R) = \frac{e^{\lambda\rho(L,R)}}{e^{\lambda\rho(L,R)} + e^{\lambda\rho}},$$

in which $\rho(L,R)$ is the risk aversion parameter ρ that makes subjects indifferent between L and R and λ is the precision parameter of the logistic distribution of the error. To account for trembles, we further assume that subjects make the choice dictated by the above model with probability $(1-\kappa)$, while they make the opposite, non-preferred choices with a tremble probability κ. Both λ and κ capture subject confusion: λ by estimating the fuzziness of the risk aversion parameter, and κ by estimating the probability that subjects make mistakes unaccountable by the estimated parameter and its imprecision. The two parameters are, therefore, well-suited to capture different effects of complexity.

Using a Random Parameter Model with a tremble allows us to include in the sample the subjects switching back from the riskier to the safer option in *HL*, a choice that cannot be rationalized in a deterministic version of Expected Utility. We instead retain the monotonicity axiom, and therefore keep out of the sample the subjects making dominated choices, such as the safer option in row 10 of *HL*, or choosing 100 boxes in the *BRET*.

[18] Apesteguia and Ballester (2018) criticize the widely used random utility model (see for instance Harrison, 2008) on the ground that it shows non-monotonicity problems of the predicted choice in the degree of risk aversion of a subject. The debate on the point is ongoing, see for instance Conte and Hey (2018). We choose a random parameter model for our exercise in light of the considerations of Apesteguia and Ballester (2018), but a random utility Fecher-error model yields the same qualitative results.

To apply this estimation procedure, we need the data to be shaped as binary choices among lotteries. The *HL* task does not require any change, since it is already built on comparisons of pairs of lotteries. Data in the *EG* and in the *BRET* need instead a transformation. Following Dave et al. (2010) and Crosetto and Filippin (2016), we reshape the *EG* and the *BRET* as implicitly containing a series of choices among pairs of lotteries. In the *EG* task, we thus interpret a choice of, say, lottery 3 as implying not only that 3 was preferred to all the other lotteries, but also that 3 was preferred to 2, and 2 to 1. In other words, we assume preferences to be smooth and single-peaked. Similarly, in the *BRET*, we interpret a subject having chosen for instance 35 boxes as preferring 1 box to 0, then 2 to 1, then 3 to 2, and so on, up to the point where he preferred 35 boxes to 34, then, 35 to 36, 36 to 37, etc. The transformed data consists of four comparisons of pairs of lotteries for each *EG* subject, and 100 such comparisons for each *BRET* subject.

We run separate regressions for each task because the different number of choices would bias the estimation towards the results of the BRET, that has 10 times more choices than HL and 25 times more than EG. For each condition, we hence estimate a CRRA random-parameter model, and we let the parameter ρ, as well as the precision parameter λ and the tremble parameter κ depend on gender.

We now comment step by step on the motivation of this specification. The test of our research question is the effect of the safe option in triggering gender differences. Therefore, we allow the risk aversion parameter ρ to vary by gender. If the safe option triggers gender differences, we should observe the ρ_{woman} dummy to be significant only in the safe-option conditions. We control for the SOEP self-reported risk attitude in the EG task to correct the potential randomization failure signaled by the unusual mean and distribution of the SOEP variable for that task.[19] A higher imprecision of the estimated parameter (as captured by λ) or a higher probability of trembling (as captured by κ) can be taken as different indicators of the impact of noise in risk taking. To test the insight by Charness et al. (2018) that gender differences could be blurred by the complexity of the task, we interact these parameters with the gender dummy. To take into account sample heterogeneity, we cluster standard errors at the individual level. Results are reported in Table 10.

Focusing on our research question, the results in Table 10 go in a similar direction as the summary statistics of Sects. 2.1-2.3 above. In two cases – *HL* and *BRET* – the introduction of a riskless alternative induces observable and significant gender differences, while this is not the case in *EG*. However, the ML estimations allows us to derive additional insights. First, in *EGnosafe* the size of the ρ_{woman} coefficient is lower than in *EGsafe*. Furthermore, a comparison across tasks shows that the availability of a riskless alternative causes a similar increase of the gender difference, with the exception of HL where this change is small. Whilst it is true that a change in the

[19] One could argue that including the SOEP question in the model is not a good idea because it conceptually coincides with what we aim to estimate, i.e. risk aversion. While this variable has been shown to correlate with risk preferences elicited in an incentivized manner, the variance explained is indeed very low (Dohmen et al., 2011). Hence, there is no risk of over-controlling. The effect of including the SEOP control is that of measuring the effect of the safe-option manipulation net of the subjects' self-representation of their risk tolerance.

curvature starting from different levels cannot be given a straightforward interpretation, it is reassuring that the observed effect is not of a different order of magnitude across tasks. This rather stable effect is remarkable, given the fact that risk elicitation measures are very volatile and task-dependent.

The ρ_{SOEP} coefficient, used to control for the heterogeneity of the underlying risk attitudes in medium samples in EG, is shows some differences across the safe option manipulation. In both cases the point estimate is negative, significantly so only in the *safe* version, indicating that self-reported risk tolerance contributes in the expected direction, but it is small, thus showing that risk attitudes vary over and above what captured by the subjects' representation of their risk attitudes.

We see no clear pattern in the noise parameters capturing a potential effect of complexity on risk attitudes. The only remarkable fact is a high probability of a tremble in HL – something to be expected since about 1 out of 7 subjects makes inconsistent choices. However, neither precision λ nor tremble κ show any consistent pattern by gender. The absence of any pattern of noise by gender suggests that complexity does not play a significant role in rationalizing gender differences in risk attitudes, and that the reduced complexity induced by the presence of the safe option cannot explain our results.

The evidence presented in Table 10 is again suggestive, but not conclusive. Detecting a difference in significance across two conditions does not imply that the difference itself is significant (Gelman & Stern, 2006). The appropriate test of our main hypothesis is hence given by a difference-in-difference model, in which we jointly estimate the effect of the safe option on women with respect to men for each task.

Within the same specification outlined above for Table 10, we generate dummy variables for women and for the safe condition, and then include the complete set of interactions for each parameter. The woman×safe term summarizes the differential effect on women in the safe condition with respect to the baseline case of men in the non-safe condition. Under our hypothesis, that safe options trigger or magnify gender differences in risk attitudes, the $\rho_{woman \times safe}$ coefficient ought to be positive and significant across tasks. Results are shown in Table 11.

The estimations show that the $\rho_{woman \times safe}$ parameter is indeed positive for all tasks, but approaching (without reaching) traditional significe levels only in the *BRET*. The effect of a safe option is consistent but not sufficiently strong to induce significant differences. The results of the interactions for the precision and tremble parameters (not reported) confirm the findings of the previous estimation by condition, i.e. that there is no detectable effect of complexity.

The lack of significance of the $\rho_{woman \times safe}$ interaction could be due to the fact that the tasks we used are not an optimal test bed for our conjecture. Indeed, these tasks were not designed to investigate the role played by a riskless alternative and the underlying determinants. We chose to use them, even knowing their limits, because of the role they played in creating the current consensus about gender differences in risk attitude. Hence, the first necessary step was to build on these tasks despite their shortcomings. The current results, while not conclusive, show that the presence of a riskless alternative has some bite in explaining gender differences in risk attitudes. The logical next step will be to use a comprehensive set of lotteries designed on purpose to test our conjecture with more power.

Table 10 Maximum likelihood structural estimation, random parameter model, by task and condition

Parameter	(1) HL	(2) HL safe	(3) BRET	(4) BRETsafe	(5) EGnosafe	(6) EG
ρ	0.59***	0.47***	0.10**	-0.097	0.75***	1.10***
	(0.05)	(0.05)	(0.04)	(0.07)	(0.2)	(0.3)
ρ_{female}	0.16*	0.18***	-0.032	0.17**	0.20**	0.34*
	(0.09)	(0.06)	(0.06)	(0.08)	(0.09)	(0.2)
ρ_{SOEP}					-0.052*	-0.11**
					(0.03)	(0.05)
λ	2.15***	2.31***	1.58***	1.26***	1.39***	1.47***
	(0.2)	(0.2)	(0.09)	(0.1)	(0.3)	(0.3)
λ_{female}	-0.24	0.33	-0.24**	0.28**	0.46*	0.51
	(0.3)	(0.4)	(0.1)	(0.1)	(0.3)	(0.8)
κ	0.81***	0.82***	0.014*	-0.0016	0.053*	0.062
	(0.01)	(0.008)	(0.008)	(0.002)	(0.03)	(0.04)
κ_{female}	-0.0056	0.019*	0.0085	0.0053	0.053	0.11**
	(0.02)	(0.01)	(0.01)	(0.004)	(0.04)	(0.05)
Observations	1790	1650	27000	19100	536	580

HL Holt & Laury, *BRET* Bomb Risk Elicitation Task, EG: Eckel & Grossman

ρ: risk aversion, λ: precision, κ: trembling

Standard errors in parentheses – * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Table 11 Maximum likelihood structural estimation, random parameter model, by task

Parameter	(1) HL	(2) BRET	(3) EG
ρ	0.59***	0.10**	0.85***
	(0.05)	(0.04)	(0.2)
ρ_{female}	0.16*	-0.032	0.19*
	(0.09)	(0.06)	(0.1)
ρ_{safe}	-0.13*	-0.20**	0.040
	(0.07)	(0.08)	(0.10)
$\rho_{female \times safe}$	0.067	0.20*	0.082
	(0.1)	(0.1)	(0.2)
ρ_{SOEP}			-0.071**
			(0.03)
λ	2.15***	1.58***	1.26***
	(0.2)	(0.09)	(0.3)
κ	0.81***	0.014*	0.053*
	(0.01)	(0.008)	(0.03)
Observations	3440	46100	1116

Standard errors in parentheses

HL Holt & Laury, *BRET* Bomb Risk Elicitation Task, *EG* Eckel & Grossman

ρ: risk aversion, λ: precision, κ: trembling* $p<0.10$, ** $p<0.05$, *** $p<0.01$

* $p<0.10$, ** $p<0.05$, *** $p<0.01$

4 Conclusion

We investigate experimentally the role played by the availability in the choice set of a riskless alternative, defined as a degenerate lottery, in which a positive amount of money can be obtained with certainty. We carefully manipulate three well-known and widely used risk elicitation methods, adding to or removing from the menu of choices a riskless alternative *ceteris paribus*. The results of our between-subject design provide some evidence that the availability of a safe option plays a role.

When subjects face the original *HL* multiple price list featuring choices between two lotteries, one safer and one riskier, the observed decisions of men and women do not differ significantly. In contrast, significant gender differences emerge when we manipulate exogenously the *HL* task by substituting each safer lottery with an equivalent (under EU) sure amount.

Some changes of behavior appear in the *BRET* when the original task is manipulated introducing the possibility of securing a positive amount. However, gender differences (both in the average choice and in the frequency of choosing the safe option) do not reach traditional significance levels.

Our design also includes an *EG* task, that in the commonly used version includes a safe option, which we substitute in our treatment condition with an equivalent (under EU) lottery with small variance. Significant gender differences in the frequency of choice of the less risky alternative are observed only when it is a safe option. However, removing the riskless alternative does not cause the gender differences to disappear when considering the whole distribution of choices.

We then make the choices across tasks comparable mapping them onto CRRA risk-aversion coefficients and estimating a structural random parameter model with maximum likelihood. The results confirm that gender differences in risk attitudes are qualitatively affected by the availability of a riskless alternative. The comparison across tasks shows that the safe option affects gender differences in the same direction and by a similar order of magnitude. An increase of women's risk aversion as compared to men emerges in this case in the *EG* task, too. The stability of the effect is particularly remarkable given that the measures of risk preferences are usually extremely volatile and task-dependent. The estimation of a structural model also allows us to address a possible criticism, namely that the presence of a sure amount instead of a lottery may affect gender differences through a different channel, i.e. reducing the complexity of the task (Charness et al., 2018). Our data show no evidence that the availability of a riskless alternative affects the complexity of the task.

The effect of a safe option, however, is not sufficiently strong to be significant in a *diff-in-diff* framework. In other words, the change in the significance of the results by gender does not translate into significant differences across conditions. Summarizing, our evidence does not exhaustively explain gender differences in risk attitudes. Nonetheless, it shows that the presence of a safe option has the potential to mark a leap forward in the understanding of this empirically debated aspect of decision under risk. A proper test of this conjecture requires to purposely build a menu of lotteries, for instance as done for risk and ambiguity by (Hey & Pace, 2014; Hey & Orme, 1994). This approach would also allow us to test the ultimate cause of

the effect of a riskless alternative, by generating sufficiently different predictions for different underlying determinants, such as certainty effects (Andreoni & Sprenger, 2012), reference point and probability weighting (Kahneman & Tversky, 1979), Salience (Bordalo et al., 2012) and Regret Theory (Loomes & Sugden, 1982). However, the first test of the role played by a safe option required the use of classic elicitation methods for two reasons. First, it was necessary to compare the results against a sufficiently established consensus. Second, the notorious inconsistency of risk measurements mandated the replication of the results across several elicitation tasks. By showing that the availability of a safe option consistently triggers or magnifies gender differences in several elicitation methods, we therefore believe that this paper identifies a promising line for future research.

Appendix. Experimental instructions

HL and HLsafe

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 {*HL*: 4 euro if the throw of the ten sided die is 1, and it pays 3.2 euro if the throw is 2-10; *HLsafe*: 3.3 euro in any case}. 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.

BRET and BRETsafe

On a sheet of paper on your desk you see a square composed of 100 numbered boxes. Behind one of these boxes hides a mine; all the other 99 boxes are free from mines. You do not know where this mine lies. You only know that the mine can be in any place between {*BRET*: 1; *BRETsafe*: 26} and 100 with equal probability.

You earn 10 eurocents for every box that is collected. After you 'Start' in the corresponding square on your screen, 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 hides the time bomb that destroys your earnings in case it is collected. You do not know where this time bomb lies. You only know that the time bomb {*BRET*: can be in any place between 1 and 100 with equal probability; *BRETsafe*: is not in the boxes from number 1 to 25, while it can be in any place between 26 and 100 with equal probability}. Moreover, even if you collect the time 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 {*BRET*: 100 tokens numbered from 1 to 100 ; *BRETsafe*: 75 tokens numbered from 26 to 100 }.

If you happen to have harvested the box where the mine is located - i.e. if your chosen number is greater than or equal to the drawn number - you will earn zero. If the mine is located in a box that you did not harvest - i.e. if your chosen number is smaller than the drawn number - you will earn in euro an amount equivalent to the number you have chosen divided by ten.

We will start with a practice round. After that, the paying experiment starts.

EG and EGnosafe

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.

Acknowledgements We gratefully acknowledge the financial support of the Max Planck Institute of Economics, of the Grenoble Applied Economics laboratory, and of the Small Grants program of the Einaudi Institute for Economics and Finance (EIEF). We thank Eicke Hauck, Ivan Soraperra, Iolanda Viceconte, and Claudia Zellmann for their help in running the experiments. We also thank Maria De Paola, Frank Heinemann, Doron Sonsino, and participants to seminars at the University of Göttingen, Université Paris 1 Sorbonne and University of Economics in Prague, as well as to participants to the 2013 CESifo workshop in Venice, the 8^{th} Nordic Conference in Behavioral and Experimental Economics in Stockholm, the 2013 EUI Alumni Conference in Florence, and the 2014 IMEBESS conference in Oxford for useful suggestions. All remaining errors are ours.

References

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

Apesteguia, J., & Ballester, M. A. (2018). Monotone stochastic choice models: The case of risk and time preferences. *Journal of Political Economy, 126*, 74–106.

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

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

Booth, A., Cardona-Sosa, L., & Nolen, P. (2014). Gender differences in risk aversion: Do single-sex environments affect their development? *Journal of Economic Behavior & Organization, 99*, 126–154.

Booth, A. L., & Nolen, P. (2012). Gender differences in risk behaviour: Does nurture matter? *The Economic Journal, 122*, F56–F78.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2012). Salience theory of choice under risk. *The Quarterly Journal of Economics, 127*, 1243–1285.

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, *125*, 367.

Charness, G., Eckel, C., Gneezy, U., & Kajackaite, A. (2018). Complexity in risk elicitation may affect the conclusions: A demonstration using gender differences. *Journal of Risk and Uncertainty, 56*, 1–17.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates.

Conte, A., & Hey, J. D. (2018). *Rehabilitating the Random Utility Model. A comment on Apesteguia and Ballester (2018)*. Discussion Papers 18/12 Department of Economics, University of York.

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

Crosetto, P., & Filippin, A. (2016). A theoretical and experimental appraisal of four risk elicitation methods. *Experimental Economics, 19*, 613–641.

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

Da Silva, S., Baldo, D., & Matsushita, R. (2013). Biological correlates of the allais paradox. *Applied Economics, 45*, 555–568.

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

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

Dekel, E., & Scotchmer, S. (1999). On the evolution of attitudes towards risk in winner-take-all games. *Journal of Economic Theory, 87*, 125–143.

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

Dreber, A., & Hoffman, M. (2010). *Biological basis of sex differences in risk aversion and competitiveness*. Technical Report Citeseer.

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

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

Eckel, C. C., & Grossman, P. J. (2008b). Men, women and risk aversion: Experimental evidence. Chapter 113. *Handbook of Experimental Economics Results, 1*, 1061–1073.

Fehr-Duda, H., Gennaro, M. D., & Schubert, R. (2006). Gender, financial risk, and probability weights. *Theory and Decision, 60*, 283–313.

Filippin, A., & Crosetto, P. (2016). A reconsideration of gender differences in risk attitudes. *Management Science, 62*, 3138–3160.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60*, 328–331.

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

Harrison, G. W. (2008). Maximum likelihood estimation of utility functions using Stata. *University of Central Florida, Working Paper* (pp. 06–12).
Heilbronner, S. R., Rosati, A. G., Stevens, J. R., Hare, B., & Hauser, M. D. (2008). A fruit in the hand or two in the bush? divergent risk preferences in chimpanzees and bonobos. *Biology Letters, 4*, 246–249.
Hey, J. D., & Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica, 62*, 1291–1326.
Hey, J. D., & Pace, N. (2014). The explanatory and predictive power of non two-stage-probability theories of decision making under ambiguity. *Journal of Risk and Uncertainty, 49*, 1–29.
Holt, C., & Laury, S. (2002). Risk aversion and incentive effects. *American Economic Review, 92*, 1644–1655.
Isaac, R., & James, D. (2000). Just who are you calling risk averse? *Journal of Risk and Uncertainty, 20*, 177–187.
Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–291.
l'Haridon, O., & Vieider, F. (2016). *All Over the Map: Heterogeneity of Risk Preferences across Individuals, Prospects, and Countries*. Economics & Management Discussion Papers em-dp2016-04 Henley Business School, Reading University.
List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy, 115*, 482–493.
Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *Economic Journal, 92*, 805–824.
Menkhoff, L., & Sakha, S. (2017). Estimating risky behavior with multiple-item risk measures. *Journal of Economic Psychology, 59*, 59–86.
Nelson, J. A. (2014). Are women really more risk-averse than men? A re-analysis of the literature using expanded methods. *Journal of Economic Surveys, 29*, 566–585.
Nelson, J. A. (2016). Not-so-strong evidence for gender differences in risk taking. *Feminist Economics, 22*, 114–142.
Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., & Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behaviour*, *1*, 803.
Petit, E., Tcherkassof, A., & Gassmann, X. (2011). *Anticipated regret and self-esteem in the Allais paradox*. Cahiers du GREThA 2011-25 Groupe de Recherche en Economie Theorique et Appliquee.
Reynaud, A., & Couture, S. (2012). Stability of risk preference measures: Results from a field experiment on french farmers. *Theory and Decision, 73*, 203–221.
Rosati, A. G., & Hare, B. (2013). Chimpanzees and bonobos exhibit emotional responses to decision outcomes. *PLOS ONE, 8*, 1–14.
Schmidt, U., Friedl, A., Eichenseer, M., & Lima de Miranda, K. (2021). Social comparison and gender differences in financial risk taking. *Journal of Economic Behavior & Organization, 192*, 58–72.
Taylor, M. P. (2016). Are high-ability individuals really more tolerant of risk? A test of the relationship between risk aversion and cognitive ability. *Journal of Behavioral and Experimental Economics, 63*, 136–147.
Trivers, R. (1972). Parental investment and sexual selection. *Sexual Selection & the Descent of Man, Aldine de Gruyter, New York* (pp. 136–179).