A Reconsideration of Gender Differences in Risk Attitudes ☆

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

[a] *University of Milan, Department of Economics, Via Conservatorio 7, 20122 Milano, Italy*
[b] *Institute for the Study of Labor (IZA), Schaumburg-Lippe-Str. 5-9, 53113 Bonn, Germany*
[c] *INRA, UMR 1215 GAEL, University of Grenoble, 38000 Grenoble, France*

Abstract

This paper reconsiders the wide agreement that females are more risk averse than males. We survey the existing experimental literature, finding that significance and magnitude of gender differences are task-specific. We gather data from 54 Holt and Laury (2002) replications, involving more than 7000 subjects. Gender differences appear in less than 10% of the studies, and are significant but negligible in magnitude once all the data are pooled. We exclude that this result is driven by noisier HL data. Gender differences appear to correlate with the presence of a safe option and fixed probabilities in the elicitation method.

JEL Classifications: C81; C91; D81

Keywords: *Gender, Risk, Survey*

☆We are grateful to the Max Planck Institute of Economics (Jena) for financial and logistic support and to Janna Heider for excellent research assistance. We would like to thank all the authors that kindly contributed their data, the members of the ESA mailing list for useful references, Ainhoa Aparicio Fenoll, Tore Ellingsen, Andrea Ichino for useful suggestions, together with the participants, in 2013, to: IMEBE, Madrid; BEELAB workshop, Florence; ESA World, Zurich; SIE Conference, Bologna; MPI Autumn Workshop, Jena; in 2014, to: ASFEE, Besançon; FUR, Rotterdam; EALE, Ljublijana; Workshop on Gender, Alicante; and to seminars in EUI Florence, Milan, Oxford, Göttingen, Prague and Padua. All remaining errors are ours.

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

January 30, 2015

1. Introduction

Gender differences in risk preferences are often regarded as a stylised fact in the economics and psychology literature. Many studies as well as the available meta-analyses find that women display a more risk averse behaviour than men when confronted with decisions under risk. In economics, for instance, surveys made by Eckel and Grossman (2008c) and Croson and Gneezy (2009) find mostly supporting evidence and investigate the robustness of this result along several dimensions, such as the characteristics of the subject pool, the strength of incentives, the gain *vs.* loss domain, the abstract *vs.* contextual framework. These surveys, though, are based on a relatively small sample of studies (16 and 10, respectively, 3 of which in common) given the variety of designs covered. As noted by Charness and Gneezy (2012) and Holt and Laury (2014), the differences in the methods used to measure the preferences can act as an additional source of heterogeneity. Consequently, Charness and Gneezy (2012) focus on a single elicitation method, the Investment Game, and find strong evidence that females are less willing to take risk. In psychology, Byrnes et al. (1999) provide a meta-analysis including 150 studies, using a broad definition of risk, from smoking to driving to gambling, and analyzing self- reported, incentivised, as well as observed choices. The study finds that males take more risks than females in most of the risk categories, even though the magnitude of the effect is usually small, seldom significant, and some studies find contrary evidence.

Despite the apparently wide agreement that females are more risk averse than males, we believe that the evidence supporting this view cannot be considered conclusive for two reasons. First, there are important branches of the literature still largely unexplored. For instance, the Holt and Laury (2002) (henceforth, HL) task has never been the subject of a comprehensive analysis by gender, despite being by far the most popular elicitation method in economics according to the number of citations. Moreover, in the Bomb Risk Elicitation Task (Crosetto and Filippin, 2013a) no gender difference was found. Second, no attempt has been made yet to investigate whether and how the elicitation methods play a role in shaping the observed results by gender. Risk attitudes are a latent construct that can only be indirectly and imperfectly measured: their measurement is by construction a combination of the latent preferences *and* the measurement error induced by the tool used to elicit them. Crosetto and Filippin (2013b) analyse to what extent, and in which direction the measured risk preferences are shaped by the characteristics of the elicitation task adopted. In this paper we aim to extend this exercise along a gender dimension.

We first provide a thorough survey of the literature, finding mixed results. We then focus on the unexplored HL task. Unfortunately, only a small fraction of contributions explicitly report about gender differences, because HL is usually a companion task in unrelated experimental studies. Only twenty papers, out of the more than five hundred citing Holt and Laury (2002), provide data on the gender breakdown of risk preferences. Contrary to the widespread consensus, out of these twenty papers, only three report significant gender differences in risk preferences.

This striking result, combined with the presence of large amounts of uncharted HL gender data, spurred us to directly contact the authors of the 94 published HL replications. We collected the data of 54 published studies, corresponding to almost eight thousand subjects, and reduced them to a common comparable format.

The resulting dataset increases dramatically the information as compared to that available in published results and allows us to provide conclusive evidence about gender differences in HL. The results consistently show that gender differences are the exception rather than the rule in HL replications. Men and women display a similar behaviour, and when a difference can be detected it is usually small.

The large amount of comparable data also allows us to greatly increase the statistical power of the analysis. Moreover, access to all microdata allows us to exploit the data of subjects making inconsistent choices using a structural model estimated with maximum likelihood. The results on the pooled data show a comeback of significant gender differences, but the magnitude of the effect turns out to be economically unimportant. Differences amount to one sixth of a standard deviation, less than a third of the effect found by other elicitation methods analysed in this paper (e.g., by Charness and Gneezy, 2012; Eckel and Grossman, 2008b).

Our results indicate that the frequency and the importance of gender differences reflect specific characteristics of the elicitation methods over and above true differences in the underlying (and latent) risk attitudes. Importantly, such a heterogeneity of the gender pattern is not due to the fact that HL induces more noise than other tasks, something that, if true, would make it more difficult to detect the same differences in the underlying preferences. Observing a gender gap not only depends on the task being contextual or not (Eckel and Grossman, 2008a), on it having to do with risk or with uncertainty (Wieland and Sarin, 2012), or on the choices being incentivised, self-reported or observed (Byrnes et al., 1999). Even restricting the analysis to the narrow domain of incentivised lottery choice tasks currently used in experimental economics, gender differences depend on the details of the task. We single out two characteristics that jointly correlate with the likelihood of observing gender differences: a) the presence of a safe option within the choice set, and b) the use of lotteries with $50\% - 50\%$ fixed probabilities in tasks that generate the menu of lotteries changing the amounts at stake.

Published results as well as our dataset do not allow us to further investigate and to disentangle the effect of each of these two characteristics. Nevertheless, we believe that this paper provides a leap forward in the understanding of gender differences in risk preferences from two points of view. First, it makes clear that, instead of being treated as a fact, gender differences should be analysed jointly with the characteristics of the task used to elicit risk preferences. Second, it greatly restricts the set of possible determinants. In a companion paper we analyse this issue by means of controlled experiments, but without finding conclusive evidence (Crosetto and Filippin, 2014). The availability of a riskless alternative allows to rationalise gender differences in some but not in all the cases, while the $50\% - 50\%$ fixed probabilities do not play a significant role.

The outline of the paper is as follows. Section 2 summarises the state of the art in the literature about gender differences in risk preferences and presents the survey of the few HL published results by gender. Section 3 describes the characteristics of the dataset of HL replications we built and use. Section 4 analyses our dataset, first paper by paper and then pooling the data, using both descriptive statistics and structural modeling allowing for errors in the choices. Section 5 discusses which characteristics of the task could trigger the stark difference in behaviour observed, identifying some candidates, and Section 6 concludes.

2. Literature Review

There are more risk elicitation methods than can be mentioned here. Our ambition is not that of providing an exhaustive survey of the results by gender across the different tasks used to measure risk preferences. In contrast, the goal of this section is to summarise the state of the art in the risk and gender literature. Consequently, we limit our analysis to three representative and widely used methods: the Investment Game, introduced by Gneezy and Potters (1997), an Ordered Lottery Selection task proposed by Eckel and Grossman (2002, 2008b), and the Holt and Laury (2002) task, the most cited and replicated risk elicitation method.

In the Investment Game (henceforth IG) subjects decide how to allocate a given endowment E between a safe account and a risky lottery that yields with 50% probability 2.5 times the amount invested, zero otherwise. The task is framed as an investment decision, and a risk neutral subject should invest all of her endowment, since the marginal return of the risky option is greater than one.

In the Eckel and Grossman task (henceforth EG) subjects make a single choice, picking one out of an ordered set of lotteries. This method has been first introduced in the literature to specifically measure risk preferences by Binswanger (1981). In the EG implementation subjects are faced with 5 lotteries characterised by a linearly increasing expected value as well as greater standard deviation (see Table 1). The task is not framed, and a risk neutral subject should choose lottery 5, since it has the highest expected value.

	Choice	**Probability**	**Outcome**
1	A	50%	16 $
	B	50%	16 $
2	A	50%	24 $
	B	50%	12 $
3	A	50%	32 $
	B	50%	8 $
4	A	50%	40 $
	B	50%	4 $
5	A	50%	48 $
	B	50%	0 $

Table 1: The 5 lotteries of the orginal Eckel and Grossman (2002) paper

The Holt and Laury (2002) (henceforth HL) risk elicitation method constitutes the most widely known implementation of a multiple price list format applied to risk. The subjects face a series of choices between pairs of lotteries, with one lottery safer (i.e., with lower variance) than the other. At the end of the experiment, one row is randomly chosen for payment, and the chosen lottery is played to determine the payoff. 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 rows is obtained by increasing the probability of the 'good' event (see Table 2).

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

	Option A				Option B			
1	1/10	2 $	9/10	1.6 $	1/10	3.85 $	9/10	0.1 $
2	2/10	2 $	8/10	1.6 $	2/10	3.85 $	8/10	0.1 $
3	3/10	2 $	7/10	1.6 $	3/10	3.85 $	7/10	0.1 $
4	4/10	2 $	6/10	1.6 $	4/10	3.85 $	6/10	0.1 $
5	5/10	2 $	5/10	1.6 $	5/10	3.85 $	5/10	0.1 $
6	6/10	2 $	4/10	1.6 $	6/10	3.85 $	4/10	0.1 $
7	7/10	2 $	3/10	1.6 $	7/10	3.85 $	3/10	0.1 $
8	8/10	2 $	2/10	1.6 $	8/10	3.85 $	2/10	0.1 $
9	9/10	2 $	1/10	1.6 $	9/10	3.85 $	1/10	0.1 $
10	10/10	2 $	0/10	1.6 $	10/10	3.85 $	0/10	0.1 $

Table 2: The 10 lotteries of the original Holt and Laury (2002) paper

point captures their degree of risk aversion. A risk-neutral subject should start with Option A, and switch to B from the fifth choice on. The higher the number of safe choices, the stronger the degree of risk aversion. Never choosing the risky option or switching from B to A are not infrequent and are regarded as inconsistent choices when modeling the choices without including a stochastic component.

That women are more risk averse than men is often considered a stylised fact in the economic literature. This finding is confirmed by some surveys (Croson and Gneezy, 2009; Eckel and Grossman, 2008a).[1] Among the risk elicitation tasks analysed in this paper, this state of the art is well captured by both the IG and EG tasks. Both tasks have already been object of a survey from a gender perspective, and females have been shown to consistently display a significantly more risk averse average behaviour.

Charness and Gneezy (2012) report that in the IG the gender gap is rather systematic and quite sizable. Males invest more than females in most of the experiments analysed, and such a difference is usually about $10 - 15\%$ of the initial endowment (Charness and Genicot, 2009; Charness and Gneezy, 2004, 2010; Dreber and Hoffman, 2007; Dreber et al., 2010; Ertac and Gurdal, 2012; Fellner and Sutter, 2009; Gong and Yang, 2012; Langer and Weber, 2004). Significant differences, but lower than 10% in size, appear in Haigh and List (2005), Bellemare et al. (2005), and Crosetto and Filippin (2013b), while Gneezy et al. (2009) is the only contribution in which a gender gap does not appear. Such a result is robust to the context (lab vs. field) in which data have been gathered as well as to other features (amounts at stake, geographical location, type of subjects).

Similar findings emerge in the EG task, with sizable gender differences appearing both in the original experiment and in later replications (Arya et al., 2012; Ball et al., 2010; Crosetto and Filippin, 2013b; Dave et al., 2010; Eckel et al., 2009, 2011; Grossman and Eckel, 2009; Wik et al., 2004). Cleave et al. (2010) find a gender gap in a wide sample but not in a subsample that participated to later experiments, but it is, to the best of our knowledge, the only

[1] Surveys also stress how some characteristics of the experiments make gender differences more likely to appear. For instance, they are usually less likely to be found in contextual experiments (Eckel and Grossman, 2008a; Schubert et al., 1999).

exception.

The results obtained using these two elicitation methods are clear-cut: women display, on average, a significantly more risk averse behaviour. The question is, however, whether these two tasks simply capture a regularity that holds in general, or whether instead the observed results are a function of some characteristics of these two elicitation methods. If this is the case, results should not be replicated at all or would be replicated to a clearly different extent using a sufficiently different elicitation method.

The perfect example is provided by HL, which is the most popular risk elicitation method in the literature, but whose replications have never been systematically analysed along a gender dimension.

A survey of the literature reveals that gender differences are only rarely found in this case. Despite the fact that more than five hundred published papers cite Holt and Laury (2002),[2] only 20 of them report the breakdown of results by gender. Out of these 20, only 3 report significant gender differences, 2 provide mixed evidence as in the original contribution, while 15 find no significant difference.

The three papers reporting a significant gender difference are Agnew et al. (2008) using an unmodified low stake HL task, Dave et al. (2010), using the 20X high stake HL treatment, and Brañas-Garza and Rustichini (2011), implementing a non-incentivised version with 9 choices.

The contributions reporting mixed results find a significant effect only for a subsample, or only through one and not all statistical methods. Already in the original HL article a gender gap appears only in the low but not in the high stake treatment. In Chen et al. (2013) significant gender differences do not emerge in the unconditional distribution of choices, but choices become significantly different (at 10%) when controlling for other observable characteristics (age, race, academic major and number of siblings). Menon and Perali (2009) on the other hand find, within one study, females to be significantly more risk averse in one sample and significantly less in another.

The list of the 15 studies in which the behaviour of males and females does not differ includes the first replication of the original task, (Harrison et al., 2005), Anderson and Freeborn (2010); Carlsson et al. (2012) in the field and Andersen et al. (2006); Baker et al. (2008); Chakravarty et al. (2011); Drichoutis and Koundouri (2012); Eckel and Wilson (2004); Ehmke et al. (2010); Harrison et al. (2013); Houser et al. (2010); Mueller and Schwieren (2012); Ponti and Carbone (2009); Viscusi et al. (2011) and Masclet et al. (2009) in the lab.

Summarizing, the frequency of significant gender differences sharply changes according to the elicitation method used. Significant gender differences appear using the EG and IG tasks, while they do not using HL.

This instability of results supports the view that a latent construct like risk attitudes can only be indirectly measured and what is observed heavily depends on the characteristics of the risk elicitation procedure used. Applied to differences of risk preferences along a gender perspective, this argument implies that the stylised fact describing females as more

[2] According to the database Scopus, queried on January 2013, 528 articles cited Holt and Laury (2002). See below section 3 for details about these papers.

risk averse than males could be less solid than what it appears at first glance and definitely requires further investigation.

The evidence in this section is based on the twenty studies that provide in their published version information about gender differences. Such evidence cannot be regarded as conclusive, however, due to both the small size of the available sample, as compared to the overall number of published HL replications, and to problems of data comparability across papers. This spurred us to collect the original data of the HL replications, with the aim of covering the largest possible number of studies. The details and results of this exercise are described in the next section.

3. The Dataset

In this section we describe and analyse our dataset, composed of a large sample of HL replications. The direct collection of the original data proved necessary for several reasons.

First, few studies replicating HL report gender results. Collecting the original data allows us to increase the size and representativeness of the sample analysed. The final dataset includes data from 36 articles that did not report gender results.[3]

Second, papers are heterogeneous in the way they report their results. Comments about gender differences are not always accompanied by quantitative results. When results are published, they take different and not comparable forms, such as parametric or non-parametric tests of equality in mean or median, or coefficients in multivariate regressions. Moreover, inconsistent choices are treated in different ways and constitute an additional source of heterogeneity. Collecting the data we can reduce to a common metric a large body of potentially heterogeneous literature.

The final dataset covers 54 published and 9 unpublished papers, twice as many as all the previous survey papers in the experimental economics literature combined.

3.1. Getting the data

For published papers we queried the Scopus bibliographic database, tracking all papers that cited Holt and Laury (2002). We ran our query on January 31st, 2013, finding 528 citing papers. We included some unpublished studies, either issue of a request on the Economics Science Association discussion group or papers we came across at conferences. This resulted in the identification of 26 additional contributions, that we treat separately.

We examined closely all the 555 papers in the resulting pool to check whether the authors had replicated the HL experiment, in its original version or with some small variations of the design. Among the experimental replications, we restricted the range of possible departures from the original HL to be included in the dataset. We regard as comparable the multiple choice lists in which the amount at stake is held constant while the increase in the expected value of the lotteries is obtained through a higher probability of the good outcome.[4] Within

[3] In principle, published results by gender could be the output of a process of selective reporting. In contrast, no evidence of outcome reporting bias is found in the HL replications (for details see Crosetto et al., 2014). The explanation for the low reporting rate seems to be the fact that the task is performed as a control, and therefore gender differences in risk preferences are of little interest to the authors.

[4] In order to keep our search within tractable limits we do not include the so-called Outcome Scale version of a multiple price lists in which an increasing safe amount is compared with a fixed 50/50 lottery, or, in general, any

these boundaries a multiple price list can take many different forms. For instance, we include tasks in which the number of choices is different than 10, or in which the amounts at stake differ as compared to the original HL.

The results of this exercise are detailed in Table 3. We could not access, either in electronic or in paper form, 48 studies. Out of the remaining contributions, we found 118 published and 17 unpublished studies replicating the HL mechanism as described above, while 21 further papers, 16 published and 5 unpublished, used a modified version of HL, involving a safe amount instead of the safe lottery. These papers are analysed separately in Section 5.

Articles citing Holt and Laury (2002) as of Jan 31st, 2013	**Published** **529**	**Not published** **26**
Not accessible	48	-
Not replicating Holt and Laury (2002)	347	4
Using an HL version with a safe option	16	5
Replicating Holt and Laury (2002)	**118**	**17**
of which:		
Duplicate dataset	8	0
Not keeping track of gender or single gender	16	0
Universe of reference	**94**	**17**
of which:		
No response or data not shared	40	8
Final dataset	**54**	**9**
of which:		
Microdata (shared or available online)	48	6
Summary statistics (shared or published)	6	3

Table 3: Building the dataset of HL replications

We directly contacted the authors of all the replications, asking them for a set of summary statistics and significance tests, or, if possible, for the original data. We sent a first email (in two batches, on March 15th and March 28th, 2013) to the corresponding authors, and two reminders (on July 7th and on September 17th, 2013, the latter to all authors of the papers) to those not having answered previous messages.

Whenever the same dataset was used in two or more studies we counted it only once, including the other references in the 'Duplicate dataset' category. 16 studies could not be used, either because they involved a single-gender sample, or because the gender of the subjects was not recorded. Subtracting these particular cases leads to a universe of 111 HL replications, 94 published and 17 unpublished, suitable for gender analysis.

task in which outcomes change and probability are fixed (see, among others, Abdellaoui et al., 2011; Andersson et al., 2013; Dohmen and Falk, 2011; Dohmen et al., 2010, 2011; Eriksen et al., 2011; Falk et al., 2006; Masatlioglu et al., 2012; Sapienza et al., 2009; Sutter et al., 2013).

Altogether, for more than half of the relevant papers we could get either the microdata or exhaustive summary statistics. Our final dataset includes data from 54 published and 9 unpublished papers, for a total of 8713 subjects.[5]

3.2. Building a homogeneous dataset

The datasets of the replications differ along several dimensions, from the purpose and the design of the experiment to the exact format of the multiple price list. Moreover, datasets differ in terms of which control variables are recorded, and in the way in which 'inconsistent' choices (multiple switchers, dominated choices) are treated.

Although we try to follow the common sense rule of keeping all the information available, making datasets comparable requires to take decisions that inherently encompass a degree of arbitrariness. The decisions and assumptions we made in building the dataset are detailed here.

3.2.1. Design of the replications

In case of a within-subject design in which the subjects completed more than one HL price list under different conditions (e.g. alone *vs.* in groups, with different frames, with different amounts at stake) we just kept the data from the first HL table the subjects were exposed to, provided that the task was performed by the subject alone. This reduced the number of observations but also the problems induced by other possible confounds such as order effects or serial correlation.

For studies employing a between-subject design, we used all observations. When the study included different experimental conditions accompanied by the HL task – usually used as a control for risk attitudes – we used all data as well. Changes in the HL task administered in the different treatments are infrequent and of marginal importance; nonetheless, we take them into account through the variable `treatment`.

In general the rules described above allowed us to easily process the replications and include them in the dataset. In some cases, though, the inclusion proved harder and ad-hoc rules necessary.[6]

3.2.2. Level of detail of the data

Datasets come in four formats. We deal with this heterogeneity including the variable `detail`.

[5] The number of contributions replicating HL among those currently classified as 'no response' is likely to be lower than the 48 (40 published and 8 not published) reported in Table 3. In fact in about the 30% of the cases we had to exclude the paper from the sample because of a sufficiently different design or missing gender information. Assuming a similar distribution in the residual category, we can reasonably expect the real number of missing dataset to be in the order of thirty. This would also imply that the current coverage rate is downward biased, and that it is likely to be already in the order of two thirds.

[6] In the case of Andersen et al. (2010), we faced a dataset with three different price lists, between subjects. One of the lists was a standard 'symmetric' one, while the other two where asymmetric ('SkewLow' and 'SkewHigh'). The asymmetric price lists featured 6 choices each, and the choices did not cover the whole probability range. They instead, in the case of 'SkewLow', covered probabilities 0.1, 0.2, 0.3, 0.5, 0.7, and 1. In order to include this paper, we rescaled the choices to 10, assuming continuity of preferences – i.e., a subject in the missing choice 0.6 has been assumed to make the same choice, safe or risky, made in choice 0.5. In case of a gap of two choices and of a different choice in the two observed extremes, we assigned one choice as safe, and one as risky.

The most complete datasets provide us with data for each and every binary choice the subjects made (`detail` = ′full′). Other datasets record the number of safe choices of every subject and a dummy variable indicating whether they switched only one or multiple times – this behaviour is usually labeled as ′inconsistent′ (`detail` = ′partial′). For these datasets we can reconstruct the binary choices of the consistent (single-switchers) only, while for multiple-switchers we cannot tell which choices were made in which lotteries, and we have to treat their binary choices as missing. Third, five datasets report only the number of individual safe choices, but no information about inconsistent behaviour. In order not to lose these observations, by default we assume that the authors sent us data for single-switchers only.[7] Finally, in some cases we only obtained summary statistics of the results, including the average number of safe choices by gender, and the results of statistical tests (`detail`= ′summary′). In this case we cannot retrieve any information about inconsistent behaviour, nor reconstruct the subjects′ binary choices.

	Detail	*Consistent subjects*			*Inconsistent subjects*		
		Males	Females	Total	Males	Females	Total
Microdata	*full*	2119	2205	4324	411	502	913
# of safe choices + consistency	*partial*	504	408	912	64	98	162
# of safe choices only		375	324	699	3	1	4
Summary statistics (shared/published)	*summary*	413	359	772	-	-	
Total		3411	3296	**6707**	478	601	**1079**

Table 4: Subjects in the sample by consistency and type of data. Published papers only. The four subjects that are classified as inconsistent when only the number of safe choices is known are those who choose the safe lottery when the good outcome is certain.

The breakdown of the number of consistent and inconsistent subjects in our dataset by gender and by detail of the data is provided in Table 4. This Table is the key to identifying the different samples used in different parts of the paper. For instance, while the description of results by paper (Section 4.1) relies upon all the information available, the analysis of microdata (Section 4.2) cannot include ′summary′ data, and the structural model estimation allowing for error (Section 4.3) can instead rely upon the ′full′ datasets only.

3.2.3. Variables included in the analysis

We shrank the number of variables of interest to get a minimum common ground for all papers and avoid having dozens of paper-specific demographics or controls. This meant including in the final dataset only the information concerning:

- The *subjects*. The dataset includes a unique identification number for every participant (`subject`), his choice in every binary lottery (`safechoice`) conditional on the completeness of the data received as explained above. This is the information we exploit to build the dependent variable used to proxy the risk attitude of the agents, i.e., the total number of safe choices.[8] Data also contain a variable summarizing whether the partic-

[7] We know from correspondence with the authors that for two of these papers (Rosaz, 2012; Rosaz and Villeval, 2012) the data cover single-switchers only. In the other cases we cannot tell, but results do not change if we exclude these three datasets from the analysis.

[8] Several features of the multiple price list need to be taken into account in order to obtain a comparable

ipant made inconsistent choices, and some individual controls such as female and age, though the latter is not always available.

- The *format of the multiple price list*. The papers included in our analysis greatly differ in the specific features of the multiple price list adopted. Examples of such differences are: a) the number of binary choices (numchoices)and consequently the change in the probability of the good outcome from one row to the next, b) the support of the probability spanned ([0.1 : 1] is the most common version, but [0 : 0.7] is also rather frequent, and we include other domains as well) and c) the variance of the outcomes. All these features are summarised by the variables Av1 Av2 Bv1 Bv2 , storing the values of lotteries A (safe) and B (risky), expressed in experimental units, and Ap1 Ap2 Bp1 Bp2, storing the probabilities of the four outcomes, for every decision.
- The *procedure of the task*. There are two variables keeping track of whether the subjects' consistency was forced or subjects were instead free to switch more than once from option A to option B, and whether the decisions were proposed following the increasing likelihood of the good outcome or instead in a randomorder. Regarding the structure of the incentives, we keep track of whether choices were incentivised or hypothetical and of the exchange rate from experimental currency to dollars. By multiplying the amounts seen at screen by the exchange rate we can also compute the realmoney at stake in the experiment as the expected value of the 50/50 lottery A.
- The *characteristics of the experiment*. Some studies focus explicitly on measuring risk preferences directly for different subpopulations and in different contexts, or study the task itself or different versions of it, or else contribute mainly from a theoretical point of view to the understanding of decisions under risk. Other studies focus on other topics, like auctions, strategic games, tournaments, and use the HL task just as a control for risk preferences. We built the variable control to take this difference into account. Moreover, especially for the papers in which HL was used as a control, we record in the variable treatment the fact that the HL data might have been associated to different treatments in the core part of the experiment.

The summary statistics of the variables included in the dataset, for the cases in which they are informative, are detailed in Table 5.

4. Results

In this section we analyse our dataset of HL replications from a gender perspective. We first analyse each paper separately, finding that an overwhelming majority of papers do not

measure of risk aversion across studies. For instance, making 6 safe choices in a classic HL task as that described in Table 2 implies that the subject switches to the risky option when the probability of the good outcome is 0.7. In contrast, making 6 safe choices in the version of the task like that implemented by Harrison et al. (2007) corresponds to switching when the probability of the good outcome is equal to 0.35, due to the fact that in this case there are 20 choices and the change in probability between each row is 5% instead of 10%. Therefore, we parametrise the number of safe choices to the probability of switching in order to impose a common metric. In the example above in Harrison et al. (2007) we assign a number of safe choices equal to 3 to a subject who switches when the probability of the good outcome is equal to 0.35.

Variable	Type	Description			
		Source of data			
ID	integer	Unique ID for the *paper*			
detail	categorical	See section 3.2.2			
		Subjects characteristics and choices			
			Min	*Mean*	*Max*
subject	integer	Unique ID for each subject in the dataset			
safechoice	dummy	1 if safe lottery A chosen, 0 if risky lottery B	0	0.573	1
inconsistent	dummy	1 if multiple switches *or* dominated choices	0	0.149	1
female	dummy	1 if female, 0 if male	0	0.499	1
age	integer	Age in years	0	26.12	84
		Format of the multiple price list			
			Min	*Mode*	*Max*
decision	integer	Decision row number			
numchoices	integer	Number of rows in the HL table	7	10	20
Av1	float	High outcome of (safer) lottery A	1	2	125000
Av2	float	Low outcome of (safer) lottery A	0.8	1.6	100000
Bv1	float	High outcome of (riskier) lottery B	1.90	3.85	240625
Bv2	float	Low outcome of (riskier) lottery B	0.05	0.1	6250
Ap1	float	Probability of high outcome of lottery A			
Ap2	float	Probability of low outcome of lottery A			
Bp1	float	Probability of high outcome of lottery B			
Bp2	float	Probability of low outcome of lottery B			
		Procedure of the task			
			Min	*Mean*	*Max*
forced	dummy	1 if consistency was forced, 0 otherwise	0	0.007	1
random	dummy	1 if decisions in random order, 0 otherwise	0	0.063	1
incentivised	dummy	1 if task paid with money, 0 otherwise	0	0.905	1
exchange	float	Exchange rate ECU/$	1	37.69	2500
realmoney	float	Expected value ($) of Option A (50% − 50%)	0	25.5	274.8
		Characteristics of the experiment			
			Min	*Mean*	*Max*
control	dummy	1 if task used as control, 0 otherwise	0	0.547	1
treatment	integer	Treatment in the original paper (*not* in the HL)	1	1.566	13

Table 5: Description of the dataset, published and unpublished papers

find significant gender differences. We then pool the data to increase the statistical power and to explore how the characteristics of the task and of the subjects affect the measured risk preferences.

4.1. Paper by paper

The first step of the analysis is to consider each paper separately, as done in meta-analyses. In this section we focus our attention to consistent choices (i.e., to subjects switching once and not choosing dominated options), including both published and unpublished papers to give the vastest possible overview of the literature. For each paper we compute the average number of safe choices by gender, the p-value of a non-parametric Mann-Whitney test, and the Cohen's *d* (Cohen, 1988) as a measure of the magnitude of the effect. Cohen's *d* is a measure of the size of an effect that is independent of the sample size. It is computed as

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

where X_m and X_f are the average male and female number of safe choices and s is the pooled standard deviation. The *d* is positive if females are more risk averse than males and negative if the opposite is true. Cohen (1988) indicated thresholds for interpreting his *d*: as long as the discussion is related to aggregate differences, 0.2 is a small effect, 0.5 is a medium effect, and from 0.8 on there can be said to be a large effect.[9]

Results are detailed in Table 6, and graphically displayed in Figure 1, which includes only the papers for which we have full detail. Figure 1 shows the mean choice by gender and its confidence intervals, as well as the p-value of the Mann-Whitney test. In both the Table and the Figure, unpublished results are reported separately. In Table 6 papers are listed alphabetically, and significant results are highlighted. In Figure 1 papers are sorted according to the strength of their results supporting the stylised fact that women are more risk averse. The upper part of each panel contains the papers in which females are more risk averse than males, sorted by decreasing significance. In the lower part of the figure are instead listed the papers (12 published, 2 unpublished) in which the average female is *less* risk averse than the average male, sorted by increasing significance.

In 40 published and 6 unpublished papers females show a more risk averse average behaviour than males, as far as point estimates are concerned. However the difference is in the majority of cases not significant. Males are more risk averse than females in 13 published and 3 unpublished papers, and this difference is never significant. When looking together at the whole dataset of published and working papers, only around 12.6% (8 out of 63) of the HL replications display significant gender differences, a result that is even weaker than the already weak evidence of a gender difference that emerged in the survey made in Section 2. This fraction decreases to about 9.25% (5 out of 54) restricting the analysis to the published studies only.

While test statistics tell us if an effect can be said to apply out of sample and to the whole population, effect size statistics tell us how substantial this effect is, irrespective of sample

[9] To be able to interpret the effect at the individual level – i.e., predicting with high accuracy a subject's gender observing his or her risk aversion only – a Cohen's *d* of 2 or more is needed, with a value of 4 meaning almost absolute discriminability (Nelson, 2014).

size. Applying the aforementioned thresholds to our data, including both published and unpublished papers, we find that 23 papers find a small effect, and 3 a medium effect. At the same time 5 papers find a small and 1 a medium effect in the opposite direction (i.e., males more risk averse than females). 22 papers find a null effect (Cohen's $d < 0.2$) in either direction.

These descriptive statistics immediately show that gender differences in risk attitudes are not an ubiquitous phenomenon. In contrast, using the HL task they appear as the exception rather than the rule. This finding is clearly at odds with the common wisdom in the literature that females are more risk averse than males. However, before drawing any conclusion we have to make sure that we are not observing a false negative: failing to detect an effect cannot be directly interpreted as the proof of its absence. In what follows we will come back to this point, starting from the next section in which we merge the microdata.

4.2. Merging the datasets

The goal of this section is to derive additional insights by merging all the available microdata rather than analyzing them separately. This approach has many advantages. First, it allows us to boost the statistical power of our test, thereby almost eliminating the likelihood of observing a false negative. Second, it makes possible to provide a precise quantitative estimate of the magnitude of gender differences using the HL task. Third, it gives the opportunity to identify the determinants of the number of safe choices over and above the role played by gender. Fourth, the panel structure of the dataset grants the opportunity of controlling for any paper-specific characteristic, both observable and unobservable. A byproduct of this exercise is also to deliver a precise quantitative estimate of the main findings in the HL in general. However, before pursuing these goals we deal with an important feature of the HL task, i.e. that of generating inconsistent choices.

4.2.1. Inconsistent observations

One of the features of the HL task is that it generates a significant fraction of choices that cannot easily be interpreted. In particular, an expected utility maximiser should switch once (and only once) from Option A to Option B. It is commonly found instead that a fraction of subjects do not conform to this behaviour, switching from Option B to Option A. This can be the consequence of going back and forth from Option A to B, or starting from B and then moving to A. In both cases, such a pattern is not consistent with the behaviour of an expected utility maximiser and for this reason such choices are usually defined as inconsistent. This is not the only way in which the behaviour seems to contradict the predictions implied by the axioms of Expected Utility Theory. For instance, choosing Option A when the good outcome is sure violates monotonicity, and the same happens when choosing Option B in the versions of HL containing a row in which the bad outcome is certain.

However, observing similar patterns does not necessarily imply a violation of the axioms underlying expected utility, as the subjects could simply be consistent with this model but at the same times making mistakes. We test what happens when accepting this view by estimating a structural model with a stochastic component in Section 4.3 below. The goal of this section is instead to describe the pattern of inconsistent choices, trying also to shed some light on their determinants and consequences. We do so exploiting all the information we have concerning inconsistent choices, including also the 'partial' datasets. In contrast,

Article	N_m	N_f	safe$_m$	safe$_f$	Mann-Whitney	Cohen's d	detail
Abdellaoui et al. (2011)	21	15	4.90	5.20	0.66	0.15	full
Andersen et al. (2008)	117	122	5.89	5.83	0.62	-0.03	full
Andersen et al. (2010)	65	24	6.25	6.71	0.55	0.26	partial
Baker et al. (2008)	25	11	5.28	5.63	0.56	-	summary
Barrera and Simpson (2012)	32	66	5.31	5.44	0.80	0.08	full
Bauernschuster et al. (2010)	67	107	6.18	6.55	0.22	0.25	full
Bellemare and Shearer (2010)	60	24	4.18	4.92	0.06	0.34	full
Brañas-Garza and Rustichini (2011)	53	92	4.49	4.67	0.65	0.07	full
Carlsson et al. (2012)	105	108	5.82	5.39	0.26	-0.17	full
Casari (2009)	40	38	5.35	5.82	0.30	0.34	full
Chakravarty et al. (2011)	32	5	6.31	6.60	0.72	0.17	full
Chen et al. (2013)	26	46	6.15	6.28	0.35	0.10	full
Cobo-Reyes and Jimenez (2012)	32	44	4.50	5.23	0.29	0.34	full
Dave et al. (2010)	353	449	6.13	6.60	**0.00**	0.25	full
Deck et al. (2012)	27	20	6.30	5.75	0.31	-0.31	full
Dickinson (2009)	72	54	4.82	4.46	0.18	-0.23	partial
Drichoutis and Koundouri (2012)	20	37	4.45	5.32	0.28	0.31	full
Duersch et al. (2012)	104	96	4.38	5.28	**0.00**	0.58	partial
Eckel and Wilson (2004)	133	99	5.30	5.50	0.30	-	summary
Eckel and Wilson (2006)	118	80	5.25	5.49	0.28	0.14	partial
Ehmke et al. (2010)	170	175	5.26	5.58	no	-	summary
Fiedler and Glöckner (2012)	11	18	6.55	7.78	0.12	0.72	full
Fiore et al. (2009)	21	19	6.24	6.00	0.23	-0.16	full
Glöckner and Hilbig (2012)	93	66	5.45	5.70	0.45	0.14	partial
Glöckner and Pachur (2012)	15	23	6.87	6.74	0.59	-0.08	full
Grijalva et al. (2011)	43	34	4.42	5.09	0.24	0.35	partial
Harrison et al. (2005)	72	80	5.43	5.89	0.07	0.32	full
Harrison et al. (2007)	14	7	3.50	1.79	0.22	-0.61	full
Harrison et al. (2013)	68	22	6.13	6.09	0.95	-0.02	full
Holt and Laury (2002)	114	85	5.95	6.33	0.13	0.23	full
Houser et al. (2010)	123	71	6	6.21	no	-	summary
Jacquemet et al. (2008)	47	40	5.79	6.25	0.29	0.28	partial
Jamison et al. (2008)	55	75	5.55	6.20	**0.01**	0.44	full
Lange et al. (2007a)	68	53	5.34	5.83	0.09	0.30	partial
Lange et al. (2007b)	97	75	5.27	5.55	0.19	0.19	partial
Levy-Garboua et al. (2012)	29	25	6.07	5.68	0.36	-0.24	full
Lusk and Coble (2005)	38	9	5.58	4.78	0.43	-0.44	full
Masclet et al. (2009)	39	40	5.10	5.38	0.75	0.14	full
Mueller and Schwieren (2012)	55	61	5.29	5.43	0.93	0.09	full
Nieken and Schmitz (2012)	131	156	5.27	5.46	0.53	0.11	full
Pogrebna et al. (2011)	27	30	5.22	5.57	0.68	0.21	partial
Ponti and Carbone (2009)	21	12	5.33	5.75	0.82	0.16	full
Rosaz (2012)	47	65	5.70	5.71	0.87	0.00	partial
Rosaz and Villeval (2012)	138	141	5.16	5.40	0.32	0.14	partial
Ryvkin (2011)	21	21	5.86	5.76	1.00	-0.05	full
Schram and Sonnemans (2011)	90	47	5.83	5.51	0.30	-0.22	full
Schunk (2009)	14	25	7.00	6.00	0.22	-	summary
Shafran (2010)	31	33	4.55	5.15	0.16	0.40	full
Slonim and Guillen (2010)	74	42	5.09	5.74	0.07	0.38	full
Sloof and van Praag (2010)	39	47	5.08	5.45	0.19	0.28	full
Szrek et al. (2012)	80	118	5.15	5.86	**0.03**	0.34	full
Viscusi et al. (2011)	71	49	5.79	5.82	no	-	summary
Wakolbinger and Haigner (2009)	71	60	5.27	5.73	**0.05**	0.27	full
Yechiam and Hochman (2013)	5	6	5.00	5.00	0.93	0.00	full
Working papers							
Crosetto and Filippin (2013b)	30	38	6.13	6.05	0.70	-0.05	full
Deck et al. (2010)	18	21	6.75	6.88	0.74	-	summary
Delnoij (2013)	52	65	5.67	6.60	**0.00**	0.62	full
He et al. (2011)	100	100	4.48	5.25	**0.05**	-	summary
Kocher et al. (2011)	97	49	5.40	5.84	**0.03**	0.28	full
Kocher et al. (2013)	157	126	5.62	5.95	0.07	0.21	partial
Laury (2005)	17	9	5.88	5.77	0.87	-	summary
Niemeyer et al. (2013)	13	5	5.31	6.00	0.84	0.33	full
Schipper (2012)	110	78	4.68	4.67	0.78	-0.01	full

Table 6: Results by gender of the HL replications – consistent subjects only

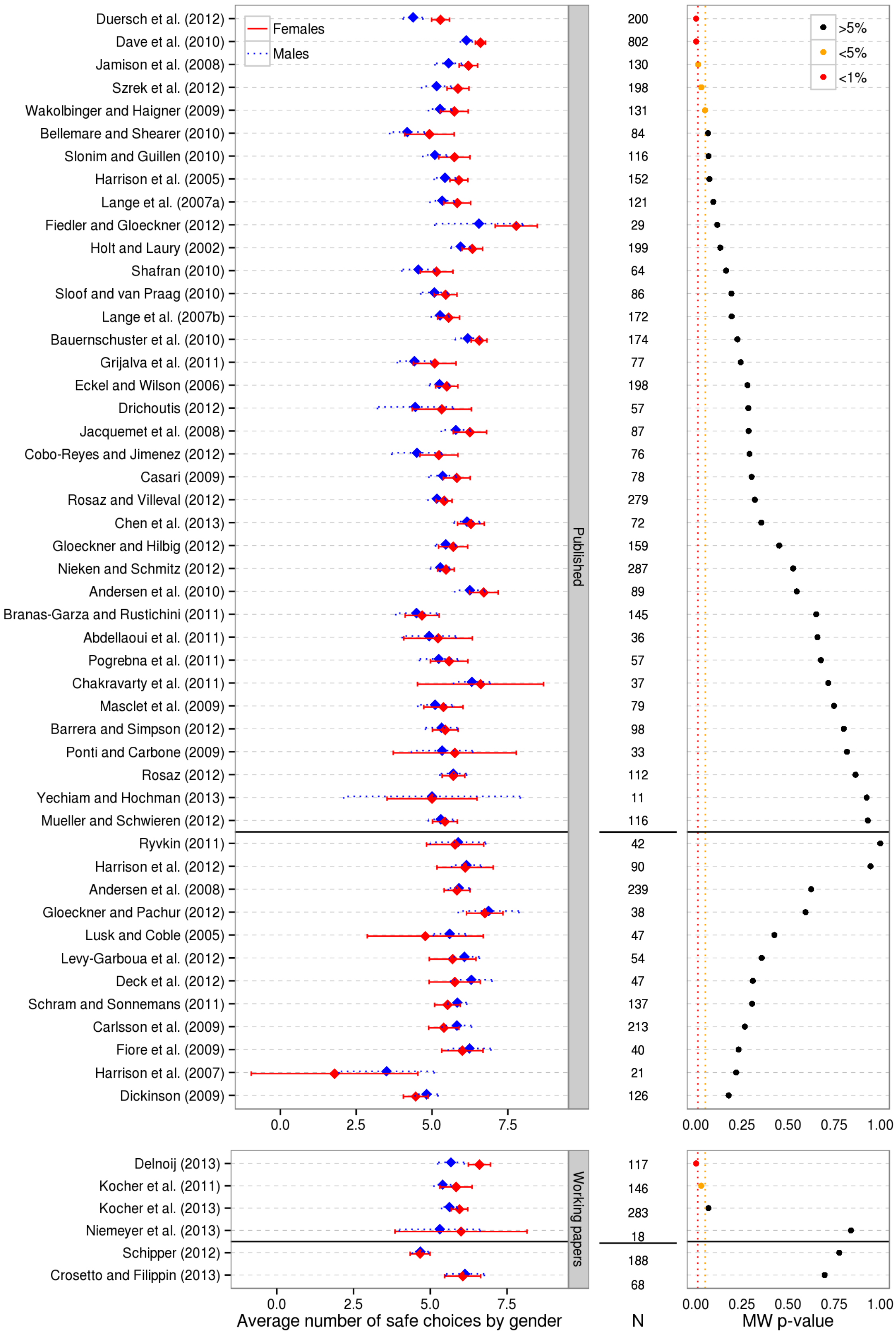

Figure 1: Gender gap in risk taking across HL replications

we cannot rely upon the papers about which we only have descriptive statistics. We limit our analysis to published papers.

The absolute frequency of inconsistent choices has already been summarised in Table 4. In Table 7 we provide a more detailed picture showing a breakdown by gender and type of inconsistency. Table 7 displays the number of inconsistent choices, overall and by gender, out of the total number that can be potentially observed for each type of inconsistency. For instance, multiple switching cannot be observed in papers in which a single switching decision is imposed by design. Always choosing the safer (riskier) lottery is a dominated action only if there is a choice in which the good outcome has probability one (zero).[10]

	Inconsistent choices		*% of inconsistent subjects*		
	Number	out of	Males	Females	Total
Switching from B to A	973	6962	12.1	15.8	14.0
Always Option A	100	6334	1.8	1.3	1.6
Always Option B	6	383	1.4	1.7	1.5
Total	**1079**	-			

Note. For each type if inconsistency the maximum number of observations (out of) has been computed separately, including only the studies in which each event can possibly happen.

Table 7: Summary statistics of inconsistent subjects by type and gender

Multiple switching is the most common type of inconsistent behaviour, observed about 14% of the times. Females are significantly more likely to be inconsistent (Fisher Exact test $p < 0.001$) and this at first glance seems to aim at numeracy as a possible explanation. These differences survive also in a multivariate framework in which other possible determinants are included. In particular, presenting the lotteries in random order dramatically increases the fraction of inconsistencies. The number of choices in the price list also significantly increases inconsistencies, although to a much lower extent, while the presence of monetary incentives significantly reduces them.

Inconsistent subjects make on average 5.15 safe choices, without significant gender differences (Mann Whitney test, $p = 0.67$). This number is lower than that of consistent subjects (5.63), and significantly so (Mann Whitney test, $p < 0.001$). At first glance this seems to suggest that inconsistent subjects tend to systematically bias downward the number of safe choices. However, a more careful interpretation suggests that inconsistent subjects simply tend to make choices that are closer to a random decision, which in the framework of the HL task coincides with choosing each option half of the times. This interpretation is in line with Andersson et al. (2013), who claim that the positive correlation between risk aversion and IQ that has been emphasised, among others, by Dohmen et al. (2010), is an artefact of the format of the price list.

Dominated choices are much less frequent. Gender in this case does not help explaining the results, and neither do the other determinants, with the exception of monetary incentives, that affect behaviour on the expected direction.

[10] We consider this to be the case when the probability of the good outcome is zero, but also for one paper in which the lowest probability of the good outcome is 1%. Strictly speaking this is not a direct violation of consistency but an expected utility maximiser should be characterised by an unbelievably high risk aversion coefficient to choose the safe lottery in this case.

4.2.2. Estimate of gender differences and determinants of the number of safe choices

In this Section we analyse the risk attitudes of consistent subjects only. Besides greatly simplifying the estimated decision making process, this approach has the advantage of allowing us to analyse the whole sample of microdata, as we must give up only the papers with 'summary' data.[11] This is important because the higher variance in HL implementation details granted by the whole sample of published papers helps to better identify the determinants of the choices.

	Mean	St.Dev	N
Data (`detail` $! =$ 'summary')	**5.63**	**1.91**	**5935**
Males	5.47	1.89	2998
Females	5.78	1.91	2937
Data (`detail` = 'full')	**5.73**	**1.96**	**4324**
Males	5.59	1.94	2119
Females	5.87	1.97	2205

Table 8: Summary statistics of safe choices, published papers, consistent subjects only

Table 8 shows that on average males make a lower number of safe choices, while variance is similar. Thanks to the high number of observations gender differences turn out to be statistically significant (Mann Whitney test, $p < 0.001$) in both samples. The Cohen's d on the pooled sample is $d = 0.163$, a tiny 16% of a standard deviation, even below the threshold of 0.2 used to identify a small effect. To give an example of how small this is, consider that if we compared two random persons, and assuming normal distribution of risk preferences, we would have a 54.3% chance of being correct when saying that the more risk averse of the two is a woman, against a 50% rate if we just randomised our answer.

For the sake of comparison, we run a similar exercise using data for the Investment Game and for the Eckel and Grossman task. For the IG we use the Cohen's ds computed by Nelson (2013) for all the studies included in the survey paper by Charness and Gneezy (2012). For the EG task we use the data provided by the papers replicating the task, when available. In both cases we add the Cohen's d computed from our own data presented in Crosetto and Filippin (2013b). The average effect size coincides for the two elicitation methods and it is equal to $d = 0.55$, three and a half times the effect found in HL.[12] This effect is still not huge, but classifiable as a medium effect at the aggregate level.

A significant gender gap is found in the HL task only when considering a vast sample, but it is negligible in size. In both IG and EG it is found even in small samples and it is three and a half times as large.

The next step is to try to identify the determinants of the number of safe choices by

[11] Estimates including inconsistent subjects can be performed only for the subset of papers for which we have 'full' data, as done via structural model estimation in Section 4.3.

[12] In order to make the two measures comparable, we compute the Cohen's d for each paper in our dataset, and we then compare the mean and distribution of this measure with the mean and distribution of the papers for which we have enough data - 16 papers for the IG and 6 papers for the EG. The Cohen's d for HL, computed from our data, turns out to be $d_{HL} = 0.13$, significantly different from $d_{IG} = 0.55$ (Mann-Whitney, p-value < 0.001) and $d_{EG} = 0.55$ (Mann-Whitney, p-value = 0.003).

means of a regression analysis, whose results are reported in Table 9.

	Dep. var.: number of safe choices			
	(1)	(2)	(3)	(4)
female	.311***	.315***	.280***	.288***
realmoney		.013***	.020***	
realmoney^2/100		-.0.04***	-.0.07***	
exchange/100		.010	-.002	
randomorder		.361***	.311***	
fixed effects	no	no	no	yes
R^2	.007	.019	.024	.095
N	5935	5935	4324	5935

Note. Fixed effects at the replication level

Table 9: Determinants of the number of safe choices

The unconditional gender difference is of 0.31 safe choices, and is significant (Column 1). In Column 2 we present our preferred specification.[13] Gender differences barely change when relevant factors are controlled for. On the other hand, we find that incentives matter. Subjects tend to be more risk averse when the incentives increase, although less so at the margin. We also find that the money illusion induced by inflating the experimental payoffs (given the same amount of money at stake) has no effect. In contrast, administering the lotteries in random order significantly increases the average number of safe choices on top of increasing the likelihood of observing an inconsistent behaviour, as observed in the previous section.

In Column 3 we estimate the same specification but restricting the sample to the papers for which we have ’full’ detail. We perform this exercise for the sake of comparability with what will be shown in Section 4.3, where also inconsistent subjects are included in the analysis, requiring the availability of all their binary choices. Results barely change, and in particular the gender gap decreases only slightly at 0.28.

The panel dimension of our dataset allows to control for any observable and unobservable characteristic common to each replication. Column 4 reports the results of a fixed effect specification. Females make on average 0.288 safe choices more than males, confirming by and large what found above.

The results of this section show that in HL task the choices of males and females are not identical. However, this difference can be detected in a significant way only when the statistical power of the test is high, and it is economically unimportant in terms of magnitude. This evidence is clearly different from what emerges for instance in the Investment Game or in the EG task. Hence, evidence based on those two tasks only cannot be regarded as suf-

[13] There are different formats of the HL implemented, but variation is low. Many papers are exact replications of HL. This generates problems of collinearity when including many controls at the same time. For instance, we do not have enough variance to meaningfully estimate the effect of the support of probability spanned by the HL list together with administering the lotteries in random order. Similarly, we cannot interact the features of the HL task with gender. On the other hand, there is no gender difference in the reaction to the amount of money at stake and in the random order of the lottery. Hence we do not include these interactions even if technically possible.

ficient to attribute the different observed behaviour to actual differences in the underlying risk attitudes. The characteristics of the risk elicitation mechanism affect systematically the measured risk preferences, and do not simply add some noise. Along the gender dimension the influence of the features of the task is so important to affect the behaviour at the aggregate level. The problem becomes then to disentangle the task *vs.* underlying preferences conundrum. We start in the next section trying to exploit all the information we have concerning the decision process, i.e., also including possible mistakes in a structural model that includes a stochastic component.

4.3. Structural estimation with Maximum Likelihood

To assess the effect of gender on choices while controlling for both a level of noise in decision making and the variations in the characteristics of the task we build a stochastic choice model and estimate it through maximum likelihood. For this exercise we restrict our focus to published papers for which we have the 'full' microdata, and we can make use of both consistent and inconsistent subjects. This leaves us with 5237 subjects.

We build our estimation using the error specification of Holt and Laury (2002), and using the script provided by Harrison (2008). We assume that subjects are expected utility maximisers characterised by CRRA preferences $U(x) = x^r$, and that they can make an evaluation error μ when comparing the utility of the two lotteries. The probability of choosing the safe lottery (Option A) is

$$Prob(A) = \frac{EU_A^{\frac{1}{\mu}}}{EU_B^{\frac{1}{\mu}} + EU_A^{\frac{1}{\mu}}}, \quad \text{and } EU_i = \sum_j p_j(x_j)^r,$$

in which A is the safe lottery, B the risky lottery, and μ is the noise parameter. It is easily shown that $Prob(A)$ converges to $\frac{1}{2}$ as $\mu \to \infty$, whilst, as $\mu \to 0$, it goes to 1 when $EU_A > EU_B$ and to 0 when $EU_A < EU_B$.

Given the above assumptions, we can write the log-Likelihood function as

$$LogLik = \begin{cases} \ln Prob(A) & \text{if choice is } \textit{safe} \\ \ln 1 - Prob(A) & \text{if choice is } \textit{risky} \end{cases},$$

and then estimate separately for each paper and jointly over all the dataset a structural model of choice using maximum likelihood and clustering standard errors by subject.[14] We allow for heterogeneity by gender of both r and μ, and we also include some controls when estimating both parameters: a dummy for the random order of the choices (`randomorder`), the money illusion induced via the experimental exchange rate (`exchange`), and the (possibly quadratic) effect of the actual amounts of money at stake (`realmoney`). Results are shown in Table 10 and are in line with what found in the previous section using regression analysis.[15]

The estimated risk parameter is $r = 0.64$, in line with the estimation in the original HL paper for the 1× treatment, with females showing a significantly higher risk aversion.

[14] The estimate paper by paper gives similar results to the ones detailed in Table 6 and is not reported.

[15] Coefficients have opposite sign in Table 9 and Table 10, because the dependent variable in the former is the number of safe choices, while in the latter the dependent variable is the risk aversion parameter. Given the utility function employed, a higher r implies a lower risk aversion.

CRRA specification $u(x) = x^r$				
		Coeff.		*St.Err.*
r	constant	0.640	***	(0.0179)
	female	-0.0633	**	(0.0203)
	realmoney/100	-0.457	***	(0.1028)
	realmoney^2/100	0.00158	***	(0.0003)
	randomorder	-0.0950	*	(0.0392)
	exchange/1000	0.00348		(0.0313)
μ	constant	0.229	***	(0.0073)
	female	-0.0135		(0.0085)
	realmoney/100	-0.19	***	(0.0247)
	realmoney^2/100	0.000658	***	(0.0000)
	randomorder	0.012		(0.0160)
	exchange/1000	0.00861		(0.1160)
N decisions				52735
n subjects				5237
Log-likelihood				-23494.025
Wald χ^2 p-value				0.000

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. t statistics in parentheses.

Table 10: Maximum Likelihood CRRA estimation, $u(x) = x^r$

Presenting the lotteries in a random order has a significant effect, increasing risk aversion. Higher amounts at stake display a similar pattern, although decreasing at the margin. In contrast, inflating numbers on screen by increasing the exchange rate between the experimental currency unit and dollars (or euros) does not significantly affect the choices.

The average level of noise is $\mu \sim 0.22$, higher than what found in Dave et al. (2010).[16] Females display a similar μ as males, providing some evidence against numeracy as a possible explanations of gender differences. In fact, if numeracy played a role in explaining the results along a gender dimension, the lower understanding of the task should have been reflected by a significantly stronger role played by confusion and captured by μ_{female}. Interestingly, and according to intuition, increasing the stakes slightly reduces noise in a concave way. In contrast, the experimental currency inflation has no effect. Strikingly, displaying the lottery choices in random order does not significantly increase the confusion coefficient.

5. Gender differences and the characteristics of the task

The analysis carried out above shows that the likelihood of observing gender differences differs systematically across elicitation methods. The question then becomes why this is the case and which characteristics of the tasks drive such a result.

[16] This is to be expected given the higher heterogeneity in terms of designs, list length, domains, and stakes in our dataset.

Higher noisiness of HL. It has been argued (Charness and Viceisza, 2011; Dave et al., 2010, among others) that HL is a relatively demanding task from a cognitive point of view. Being more difficult to understand than other methods, HL might elicit noisier signals. The noise could then blur the evidence and lead to the observed lack of significant gender differences in small to medium samples.

While relevant from a logical point of view (a noisier signal would both make differences less likely to be significant and reduce the Cohen's *d*) this argument fails empirically. HL indeed generates a high number of inconsistent choices, but this is a double-edged sword. On the one hand, the presence of large shares of inconsistent subjects is a sign of the cognitive complexity of the task; on the other hand, inconsistencies allow the researcher to single out and exclude the subjects who did not understand the task, yielding a cleaner dataset. In fact, there are several pieces of evidence consistently showing that HL is *not* noisier than the other methods analysed once the inconsistent subjects are excluded.

First, inconsistent choices are indeed the channel through which complexity affects the results in HL. Replicating the maximum likelihood estimation of Table 10 with consistent subjects only, we find a lower estimate of noise of $\mu = 0.159$, with the other results qualitatively unchanged. Moreover, pooling the sample by gender but interacting the μ with a dummy for inconsistent choices we estimate the constant at $\mu = 0.146$ and the coefficient on the dummy at 0.521, both significant at $< 0.01\%$.

Second, we compare the Signal to Noise Ratio (SNR) of the tasks, defined as the mean of a signal divided by its standard deviation. If HL were noisier, it should display a lower SRN than the other tasks. This is not the case. The SNR in our dataset of HL replications is equal to 3.34, higher than the average of the replications of the SNR of the Investment Game (2.06) and the EG task (2.41).[17] These results are confirmed by a replication in a homogeneous subject pool Crosetto and Filippin (2013b), with SNR of 3.27 for HL, 2.67 for IG and 2.16 for EG.

Third, in Crosetto and Filippin (2013b) we simulate with virtual agents (not affected by complexity and with a known distribution of risk preferences) the effect of the mechanics of the different tasks on the measured risk attitudes. If HL induces noisier choices, we should observe a sizeable discrepancy of the variance of choices in the human relative to the virtual subjects. This is not the case. The standard deviation of measured preferences in the simulations is similar to that obtained by consistent human subjects.

Theoretical comparison. Having excluded that the pattern of gender differences stems from a different precision in measuring risk attitudes, we move to a quick comparison of the methods described in Section 2 from a theoretical point of view. The goal is to identify the features that correlate systematically with the observation of gender differences.

Apart from the number of choices, the tasks differ along three main lines: *a*) the lotteries being generated by changes in probabilities rather than outcomes; *b*) the truncation of the domain of risk preferences covered by the task; *c*) the availability of a safe (risk-free) option among the set of alternatives.

[17] We use data from Nelson (2013) for the IG, and our computations for EG. Note that since we do not have the microdata of the replications of IG and EG, we cannot compute the SNR of the pooled samples. However, the distribution of the SNR of the individual replications of both IG and EG is significantly different than that of the HL replications (Mann-Whitney, $p < 0.001$).

The Investment Game and the EG task are similar as far as these theoretical characteristics are concerned. They both generate lotteries varying the amounts at stake, while probabilities are kept fixed at 50%. Moreover, both tasks can identify only different degrees of risk aversion and cannot disentangle risk-loving from risk-neutral behaviour. In the IG risk-neutral as well as risk-loving subjects should invest their entire endowment. In the EG task lottery 5 yields the highest expected value and should be the preferred alternative of risk-neutral and risk-loving subjects alike. Finally, both elicitation methods include risk-free alternatives. EG includes a degenerate lottery with no uncertainty that is equivalent to a safe choice, while in IG subjects have the opportunity of securing any amount between zero and the whole endowment.

The HL task differs from IG and EG along all three dimensions. First, lotteries are generated changing probabilities over fixed outcomes. Second, HL measures preferences both in the risk averse and in the risk loving domain. Third, the choice set does not include a riskless alternative. The subject must incur some risks as the degenerate lottery in row number ten of the original HL is played with 10% probability only. It can be argued that the role of the risk-free alternative might be played by the low amount of the safe lottery, that can be secured by always choosing Option A (except in row 10). Whether such an amount can be considered as a risk-free alternative is disputable, but it is definitely less focal than in the other two elicitation methods. In fact, it is not directly shown to the subjects, it requires some elaboration to be identified, and its salience is likely diluted by the existence of multiple choices, which induce row-by-row comparisons.[18]

The joint presence of these three factors (safe option, truncation of the domain, change in probabilities *vs* change in amounts at stake with fixed 50% probability) seem to correlate with the likelihood of observing gender differences in risk preferences. Evidence from the Bomb Risk Elicitation Task (Crosetto and Filippin, 2013a), a task sharing the three characteristics with HL, goes in the same direction, since no gender differences are found.

The next step is to try to disentangle the role of each of these factors.

Our data allows us to exclude that the observed pattern of gender differences depend on the truncation of the opportunity set. This could be true only if females were more risk seeking in the risk loving domain. A task that covers only the risk aversion domain could hide gender differences in the risk-loving domain, thereby delivering a upward biased estimate of females' risk aversion. Our dataset allows us to directly test and exclude this possibility. In fact, in HL females appear slightly more risk averse uniformly, i.e., also in the risk loving domain. Further evidence supporting this claim is provided by a different task, namely the Outcome Scale method, consisting of a multiple price list with an increasing safe option against the same 50/50 lottery. The Outcome Scale method has hence two features in common with EG and IG, while, similarly to HL, it covers the entire domain of preferences. A gender gap is a recurrent finding also with the Outcome Scale method. For instance, gender differences are found by Dohmen et al. (2011); Sapienza et al. (2009); Sutter et al. (2013), with Cohen's d in the range of ~ 0.35, while no differences are reported by Dohmen et al. (2010); Masatlioglu et al. (2012).

[18] We tried to estimate an endogenous reference point à la Koszegi and Rabin (2007). This turned out not to be possible due to identification problems, since several combinations of the reference point and the loss and risk aversion parameter could generate the same data.

The availability of a safe option within the set of alternatives has been shown to increase the likelihood of observing violations of expected utility theory (Andreoni and Sprenger, 2012; Camerer, 1992; Harless and Camerer, 1994; Starmer, 2000), and therefore a possibility is that the impact of certainty effects differs by gender.

The literature offers the possibility of testing this explanation only indirectly. Some HL replications use a slightly modified version of the HL task in which subjects choose repeatedly between a safe amount and risky lotteries characterised by fixed amounts and differing probabilities. We collected data from 15 studies using versions of HL that broadly fit into this category. Within these studies, gender differences emerge more frequently when a safe option is available (in 20% rather than 9.5% of the papers); when pooling the data and interacting gender and the availability of a safe option in a joint regression, though, results do not support a significant role of the safe option.

Unfortunately, such an exercise cannot be considered a meaningful and direct tests of the safe option conjecture. The majority of the safe-option papers differ from the standard HL in several dimensions. The number of options is higher (15) and limited to the range of probability $[0.3 - 1]$ of the good outcome to occur. Moreover, the fixed amount is usually lower than the expected value of the $50\% - 50\%$ risky lottery, it is kept constant across all rows and is therefore different from the expected value of the corresponding Option A in the classic HL. Finally, since the 15 papers we collected come from the same group of authors employing the same design (it is the case of, among others, Cason et al., 2010; Price and Sheremeta, 2011; Sheremeta, 2010), the variance in the sample comes mainly from differences *across* rather than *within* the two subsamples, hence being confounded with the availability of a safe option itself.

The role of fixed probability is even harder to ascertain given the existing literature. A study by Bruner (2009) tests two different HL tables, one with changing stakes and one with changing probabilities, but unfortunately no information on gender is available. Another recent paper (Andersson et al., 2013) employs an Outcome Scale method without a safe option, effectively replicating a HL method with fixed probabilities, but finds significant gender differences in one of the two experiments of the study, and not in the other. On the other hand, other tasks combine varying probability and varying outcomes with the absence of a safe option, as is the case of the Bomb Risk Elicitation Task (BRET, Crosetto and Filippin, 2013a), in which no gender difference appears.

To directly test the role of a safe option and of 50-50 probabilities in triggering gender differences, in a companion paper (Crosetto and Filippin, 2014) we ran clean HL, EG and BRET replications with and without a safe option, and EG with and without 50-50 probabilities. We find suggestive but not conclusive evidence. The availability of a riskless alternative allows to rationalise gender differences in the HL and the BRET, but not in EG, where women are more subject to a certainty effect but a gender gap persists in the no-safe version; the $50\% - 50\%$ fixed probabilities do not appear to play a significant role.

6. Discussion and conclusions

In the economics literature there is a wide agreement that females are more risk averse than males. In this paper we reconsider this issue, complementing the existing literatures with several findings.

First, we show that the emergence of gender differences appears to be task-specific. While gender differences are a constant finding of both the Investment Game (Gneezy and Potters, 1997) and of the Eckel and Grossman (2002) task, they do not appear in the Bomb Risk Elicitation Task (Crosetto and Filippin, 2013a). A thorough survey of the literature shows that gender differences are the exception rather than the rule also in the most widely used risk elicitation task, Holt and Laury (2002).

Second, we provide the largest analysis of HL replications to date. Since HL is usually employed as a companion task in experiments focusing on other topics, the number of papers directly reporting gender results is small relative to the number of replications. By gathering the original data from the authors we built a dataset of 54 published papers involving more than seven thousand subjects and covering more than half of all the HL published replications. Analysing the dataset we find that gender differences appear only in less than ten percent of the published papers. This striking difference is not due to a different average sample size, and we can also exclude that it is an artefact of a greater complexity of the HL task.

The creation of a comparable dataset of HL replications allows us to merge the data and reach several further goals.

First, we can provide a reliable estimate of the typical results obtained with the HL task. The average unconditional choice corresponds to an Arrow-Pratt coefficient of risk aversion equal to $\rho = 0.36$. Incentives increase risk aversion in a significant and concave way. Inconsistent choices are commonly found, and characterise on average 14% of the subjects. We find that females are more likely to display an inconsistent behaviour than males, but the choices of inconsistent subjects do not differ by gender. Second, we shed light on the pattern of inconsistent choices estimating a stochastic choice model with maximum likelihood. This procedure provides some evidence against numeracy as a possible explanation of the gender pattern. Third and foremost, merging the replications allows us to boost the statistical power when testing the existence of gender differences, virtually eliminating the possibility of facing a false negative. Doing so, significant differences are indeed detected. Their magnitude is however economically unimportant, about a sixth of a standard deviation, three times lower than what found for instance in the Gneezy and Potters (1997) Investment Game or in the Eckel and Grossman (2002) lottery choice task.

Heterogeneity in risk preferences across tasks and domains has already been observed in the literature. Our striking finding is that systematic differences exist at the aggregate level even across incentivised tasks that consist essentially of one ingredient only: choices among lotteries. The main difference between our results and the stated view in the literature is that we link the likelihood of observing gender differences with the features of the task used to elicit risk preferences. This is an interesting result *per se* because it proves that there is a structure behind the finding of gender differences in risk attitudes. At the same time, if the measured risk preferences depend on the elicitation tasks, it is natural to ask why this is the case and which task gets closer to the *true* value of risk preferences.

We do not provide a final answer to this question, but we draw a map of the features of the different tasks that might trigger different behaviour by gender. We can rule out that the observed gender pattern is due to the different domain of preferences (risk-averse, risk-loving) investigated by the risk elicitation methods. The characteristics that correlate with the emergence of gender differences are restricted to *a*) the availability of a safe option among the set of alternatives, and *b*) the use of $50 - 50$ lotteries that vary only in the amounts at stake. The first determinant is likely to trigger certainty effects, and it is known in the literature that safe options increase the likelihood of observing violation of the predictions of Expected Utility Theory. The second factor prevents misperceptions of probabilities from playing a role. Unfortunately, the studies available in the literature and an *ad hoc* experimental investigation carried out in a companion paper, though hinting to a prominent role of risk-free options and certainty effects, do not yield conclusive results. Further research is needed to properly explain when and in which sense males are more risk tolerant than females and what is the theoretical framework more suitable to represent this fact.

References

Abdellaoui, Mohammed, Ahmed Driouchi, and Olivier L'Haridon, "Risk aversion elicitation: reconciling tractability and bias minimization," *Theory and Decision*, 2011, *71*, 63–80.

Agnew, Julie R., Lisa R. Anderson, Jeffrey R. Gerlach, and Lisa R. Szykman, "Who Chooses Annuities? An Experimental Investigation of the Role of Gender, Framing, and Defaults," *The American Economic Review*, 2008, *98* (2), pp. 418–422.

Andersen, Steffen, Glenn Harrison, Morten Lau, and E. Rutström, "Elicitation using multiple price list formats," *Experimental Economics*, 2006, *9*, 383–405.

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

—, —, Morten Igel Lau, and E. Elisabet Rutström, "Preference heterogeneity in experiments: Comparing the field and laboratory," *Journal of Economic Behavior & Organization*, 2010, *73* (2), 209 – 224.

Anderson, LisaR. and BethA. Freeborn, "Varying the intensity of competition in a multiple prize rent seeking experiment," *Public Choice*, 2010, *143*, 237–254.

Andersson, Ola, Jean-Robert Tyran, Erik Wengström, and HÃěkan J. Holm, "Risk Aversion Relates to Cognitive Ability: Fact or Fiction?," Working Papers 2013:9, Lund University, Department of Economics April 2013.

Andreoni, James and Charles Sprenger, "Risk Preferences Are Not Time Preferences," *American Economic Review*, 2012, *102* (7), 3357–76.

Arya, Shweta, Catherine Eckel, and Colin Wichman, "Anatomy of the credit score," *Journal of Economic Behavior & Organization*, 2012, *forthcoming.*

Baker, Ronald J., Susan K. Laury, and Arlington Walton Williams, "Comparing Small-Group and Individual Behavior in Lottery-Choice Experiments," *Southern Economic Journal*, 2008, *75* (2), 367–382.

Ball, Sheryl, Catherine Eckel, and Maria Heracleous, "Risk aversion and physical prowess: Prediction, choice and bias," *Journal of Risk and Uncertainty*, 2010, *41* (3), 167–193.

Barrera, Davide and Brent Simpson, "Much Ado About Deception: Consequences of Deceiving Research Participants in the Social Sciences," *Sociological Methods & Research*, 2012, *41* (3), 383–413.

Bauernschuster, Stefan, Peter Duersch, JÃűrg Oechssler, and Radovan Vadovic, "Mandatory sick pay provision: A labor market experiment," *Journal of Public Economics*, 2010, *94* (11-12), 870 – 877.

Bellemare, Charles and Bruce Shearer, "Sorting, incentives and risk preferences: Evidence from a field experiment," *Economics Letters*, 2010, *108* (3), 345 – 348.

—, Michaela Krause, Sabine Kroger, and Chendi Zhang, "Myopic loss aversion: Information feedback vs. investment flexibility," *Economics Letters*, June 2005, *87* (3), 319–324.

Binswanger, Hans P., "Attitudes Toward Risk: Theoretical Implications of an Experiment in Rural India," *The Economic Journal*, 1981, *91* (364), pp. 867–890.

Brañas-Garza, Pablo and Aldo Rustichini, "Organizing Effects of Testosterone and Economic Behavior: Not Just Risk Taking," *PLoS ONE*, 2011, *6* (12), e29842.

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

Byrnes, James P, David C Miller, and William D Schafer, "Gender differences in risk taking: A meta-analysis.," *Psychological bulletin*, 1999, *125* (3), 367.

Camerer, Colin F., "Recent Tests of Generalizations of Expected Utility Theory," in Ward Edwards, ed., *Utility Theories: Measurements and Applications*, Studies in Risk and Uncertainty, Boston, MA: Kluwer Academic Publishers, 1992, pp. 207–251.

Carlsson, Fredrik, Haoran He, Peter Martinsson, Ping Qin, and Matthias Sutter, "Household decision making in rural China: Using experiments to estimate the influences of spouses," *Journal of Economic Behavior & Organization*, 2012, *84* (2), 525–536.

Casari, Marco, "Pre-commitment and flexibility in a time decision experiment," *Journal of Risk and Uncertainty*, 2009, *38*, 117–141.

Cason, Timothy N., William A. Masters, and Roman M. Sheremeta, "Entry into winner-take-all and proportional-prize contests: An experimental study," *Journal of Public Economics*, October 2010, *94* (9-10), 604–611.

Chakravarty, Sujoy, Glenn W. Harrison, Ernan E. Haruvy, and Elisabet E. Rutström, "Are You Risk Averse Over Other People's Money?," *Southern Economic Journal*, 2011, *77* (4), 901 – 913.

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

— and Garance Genicot, "Informal Risk Sharing in an Infinite-Horizon Experiment," *Economic Journal*, 2009, *119* (537), 796–825.

— and Uri Gneezy, "Gender, Framing, and Investment," Technical Report 2004.

— and — , "Portfolio Choice And Risk Attitudes: An Experiment," *Economic Inquiry*, 2010, *48* (1), 133–146.

— and — , "Strong Evidence for Gender Differences in Risk Taking," *Journal of Economic Behavior & Organization*, 2012, *83* (1), 50–58.

Chen, Yan, Peter Katuščák, and Emre Ozdenoren, "Why Can't a Woman Bid More Like a Man?," *Games and Economic Behaviour*, 2013, *77* (1), 181–213.

Cleave, Blair L., Nikos Nikiforakis, and Robert Slonim, "Is There Selection Bias in Laboratory Experiments?," Department of Economics - Working Papers Series 1106, The University of Melbourne 2010.

Cobo-Reyes, RamÃşn and Natalia Jimenez, "The dark side of friendship: 'envy'," *Experimental Economics*, 2012, *15*, 547–570.

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

Crosetto, Paolo and Antonio Filippin, "The 'Bomb' Risk Elicitation Task," *Journal of Risk and Uncertainty*, August 2013, *47* (1), 31–65.

— and — , "A Theoretical and Experimental Appraisal of Five Risk Elicitation Methods," Jena Economic Research Papers 2013-009, Friedrich-Schiller-University Jena, Max-Planck-Institute of Economics February 2013.

— and — , "Experimental Evidence on the Cause of Gender Differences in Risk Attitudes," *mimeo*, 2014.

— , — , and Janna Heider, "A Study of Outcome Reporting Bias Using Gender Differences in Risk Attitudes," *CESifo Economic Studies*, 2014, *forthcoming.*

Croson, Rachel and Uri Gneezy, "Gender Differences in Preferences," *Journal of Economic Literature*, June 2009, *47* (2), 448–74.

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

Deck, Cary, Jungmin Lee, and Javier Reyes, "Personality and the Consistency of Risk Taking Behavior: Experimental Evidence," Working Papers 10-17, Chapman University, Economic Science Institute 2010.

— , — , — , and Chris Rosen, "Risk-Taking Behavior: An Experimental Analysis of Individuals and Dyads," *Southern Economic Journal*, 2012, *79* (2), 277–299.

Delnoij, Joyce, "To bid or to buy? Heterogeneous bidders' preferences over auction mechanisms," 2013. Unpublished, presented at IMEBE conference.

Dickinson, DavidL., "The Effects of Beliefs Versus Risk Attitude on Bargaining Outcomes," *Theory and Decision*, 2009, *66*, 69–101.

Dohmen, Thomas and Armin Falk, "Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender," *American Economic Review*, September 2011, *101* (2), 556–90.

— , — , David Huffman, and Uwe Sunde, "Are Risk Aversion and Impatience Related to Cognitive Ability?," *American Economic Review*, June 2010, *100* (3), 1238–60.

— , — , — , — , JÃijrgen Schupp, and Gert G. Wagner, "Individual Risk Attitudes: Measurement, Determinants, And Behavioral Consequences," *Journal of the European Economic Association*, 2011, *9* (3), 522–550.

Dreber, Anna and M. Hoffman, "2D:4D and Risk Aversion: Evidence that the Gender Gap in Preferences is Partly Biological," mimeo 2007.

— , David G. Rand, Nils Wernerfelt, Justin R. Garcia, J. Koji Lum, and Richard Zeckhauser, "Dopamine and Risk Choices in Different Domains: Findings among Serious Tournament Bridge Players," Working Paper Series rwp10-034, Harvard University, John F. Kennedy School of Government 2010.

Drichoutis, Andreas C. and Phoebe Koundouri, "Estimating risk attitudes in conventional and artefactual lab experiments: The importance of the underlying assumptions," *Economics - The Open-Access, Open-Assessment E-Journal*, 2012, *6* (38), 1–15.

Duersch, Peter, JÃČÂűrg Oechssler, and Radovan Vadovic, "Sick pay provision in experimental labor markets," *European Economic Review*, 2012, *56* (1), 1 – 19.

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

— and — , "Chapter 113 Men, Women and Risk Aversion: Experimental Evidence," 2008, *1*, 1061 – 1073.

— and — , "Forecasting risk attitudes: An experimental study using actual and forecast gamble choices," *Journal of Economic Behavior & Organization*, 2008, *68* (1), 1–17.

__ **and** __ , *Men, Women and Risk Aversion: Experimental Evidence*, Vol. 1 of *Handbook of Experimental Economics Results*, Elsevier,

__ **and Rick K. Wilson**, "Is trust a risky decision?," *Journal of Economic Behavior & Organization*, 2004, *55* (4), 447 – 465.

__ **, Mahmoud A. El-Gamal, and Rick K. Wilson**, "Risk loving after the storm: A Bayesian-Network study of Hurricane Katrina evacuees," *Journal of Economic Behavior & Organization*, 2009, *69* (2), 110–124.

__ **, Philip J. Grossman, Cathleen A. Johnson, Angela De Oliveira, Christian Rojas, and Rick K. Wilson**, "On the Development of Risk Preferences: Experimental Evidence," Working Paper Series 2008-5, CBEES 2011.

Eckel, CatherineC. and RickK. Wilson, "Internet cautions: Experimental games with internet partners," *Experimental Economics*, 2006, *9*, 53–66.

Ehmke, Mariah, Jayson Lusk, and Wallace Tyner, "Multidimensional tests for economic behavior differences across cultures," *The Journal of Socio-Economics*, 2010, *39* (1), 37 – 45.

Eriksen, Kristoffer W., Ola Kvaløy, and Trond E. Olsen, "Tournaments with Prize-setting Agents*," *The Scandinavian Journal of Economics*, 2011, *113* (3), 729–753.

Ertac, Seda and Mehmet Y. Gurdal, "Deciding to Decide: Gender, Leadership and Risk-Taking in Groups," *Journal of Economic Behavior & Organization*, 2012, *83* (1), 24–30.

Falk, Armin, David Huffman, and Uwe Sunde, "Self-Confidence and Search," Discussion paper, IZA - Forschungsinstitut zur Zukunft der Arbeit – Institute for the Study of Labor 2006.

Fellner, Gerlinde and Matthias Sutter, "Causes, Consequences, and Cures of Myopic Loss Aversion - An Experimental Investigation," *Economic Journal*, 2009, *119* (537), 900–916.

Fiedler, Susann and Andreas Glöckner, "The dynamics of decision making in risky choice: An Eye-tracking Analysis," *Frontiers in Psychology*, 2012, *3* (335).

Fiore, Stephen M., Glenn W. Harrison, Charles E. Hughes, and E. Elisabet Rutström, "Virtual experiments and environmental policy," *Journal of Environmental Economics and Management*, 2009, *57* (1), 65 – 86.

Glöckner, A. and T. Pachur, "Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory," *Cognition*, 2012.

Glöckner, Andreas and BenjaminE. Hilbig, "Risk is relative: Risk aversion yields cooperation rather than defection in cooperation-friendly environments," *Psychonomic Bulletin & Review*, 2012, *19* (3), 546–553.

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

__ **, Kenneth L. Leonard, and John A. List**, "Gender Differences in Competition: Evidence from a Matrilineal and a Patriarchal Society," *Econometrica*, 2009, *77* (5), 1637–64.

Gong, Binglin and Chun-Lei Yang, "Gender differences in risk attitudes: Field experiments on the matrilineal Mosuo and the patriarchal Yi," *Journal of Economic Behavior & Organization*, 2012, *83* (1), 59–65.

Grijalva, Therese, Robert P. Berrens, and W. Douglass Shaw, "Species preservation versus development: An experimental investigation under uncertainty," *Ecological Economics*, 2011, *70* (5), 995 – 1005.

Grossman, Philip J. and Catherine C. Eckel, "Loving the Longshot: Risk Taking with Skewed Gambles," Economics Seminar Series 10, St. Cloud State University 2009.

Haigh, Michael S. and John A. List, "Do Professional Traders Exhibit Myopic Loss Aversion? An Experimental Analysis," *Journal of Finance*, 2005, *60* (1), 523–534.

Harless, David W and Colin F Camerer, "The Predictive Utility of Generalized Expected Utility Theories," *Econometrica*, 1994, *62* (6), 1251–89.

Harrison, Glenn W., "Maximum likelihood estimation of utility functions using Stata," *University of Central Florida, Working Paper*, 2008, pp. 06–12.

__ **, Eric Johnson, Melayne M. McInnes, and E. Elisabet Rutström**, "Risk Aversion and Incentive Effects: Comment," *American Economic Review*, June 2005, *95* (3), 897–901.

Harrison, Glenn W, John A List, and Charles Towe, "Naturally Occurring Preferences and Exogenous Laboratory Experiments: A Case Study of Risk Aversion," *Econometrica*, 2007, *75* (2), 433–458.

Harrison, Glenn W., Morten I. Lau, E. Elisabet RutstrÃűm, and Marcela Tarazona-GÃşmez, "Preferences over social risk," *Oxford Economic Papers*, 2013, *65* (1), 25–46.

He, Haoran, Peter Martinsson, and Matthias Sutter, "Group Decision Making Under Risk: An Experiment with Student Couples," Working Papers 2011-27, Faculty of Economics and Statistics, University of Innsbruck 2011.

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

Holt, Charles A. and Susan K. Laury, "Chapter 4 - Assessment and Estimation of Risk Preferences," in Mark Machina and Kip Viscusi, eds., *Handbook of the Economics of Risk and Uncertainty*, Vol. 1 of *Handbook of the Economics of Risk and Uncertainty*, North-Holland, 2014, pp. 135 – 201.

Houser, Daniel, Daniel Schunk, and Joachim Winter, "Distinguishing trust from risk: An anatomy of the investment game," *Journal of Economic Behavior & Organization*, 2010, *74* (1-2), 72 – 81.

Jacquemet, Nicolas, Jean-Louis Rullière, and Isabelle Vialle, "Monitoring optimistic agents," *Journal of Economic Psychology*, 2008, *29* (5), 698 – 714.

Jamison, Julian, Dean Karlan, and Laura Schechter, "To deceive or not to deceive: The effect of deception on behavior in future laboratory experiments," *Journal of Economic Behavior & Organization*, 2008, *68* (3-4), 477 – 488.

Kocher, Martin G., Ganna Pogrebna, and Matthias Sutter, "Other-regarding preferences and management styles," *Journal of Economic Behavior & Organization*, 2013, *88* (0), 109 – 132.

__ , Julius Pahlke, and Stefan T. Trautmann, "Tempus Fugit: Time Pressure in Risky Decisions," Discussion Papers in Economics 12221, University of Munich, Department of Economics 2011.

Koszegi, Botond and Matthew Rabin, "Reference-Dependent Risk Attitudes," *American Economic Review*, 2007, *97* (4), 1047–1073.

Lange, Andreas, John A. List, and Michael K. Price, "A fundraising mechanism inspired by historical tontines: Theory and experimental evidence," *Journal of Public Economics*, 2007, *91* (9), 1750 – 1782.

__ , __ , and __ , "USING LOTTERIES TO FINANCE PUBLIC GOODS: THEORY AND EXPERIMENTAL EVIDENCE*," *International Economic Review*, 2007, *48* (3), 901–927.

Langer, T. and M. Weber, "Does Binding or Feedback Influence Myopic Loss Aversion? An Experimental Analysis," mimeo 2004.

Laury, Susan K., "Pay One or Pay All: Random Selection of One Choice for Payment," Research Paper Series, Andrew Young School of Policy Studies 2005.

Levy-Garboua, Louis, Hela Maafi, David Masclet, and Antoine Terracol, "Risk aversion and framing effects," *Experimental Economics*, 2012, *15*, 128–144.

Lusk, Jayson L. and Keith H. Coble, "Risk Perceptions, Risk Preference, and Acceptance of Risky Food," *American Journal of Agricultural Economics*, 2005, *87* (2), 393–405.

Masatlioglu, Yusufcan, Sarah Taylor, and Neslihan Uler, "Behavioral mechanism design: evidence from the modified first-price auctions," *Review of Economic Design*, 2012, *16*, 159–173.

Masclet, David, Nathalie Colombier, Laurent Denant-Boemont, and Youenn Lohéac, "Group and individual risk preferences: A lottery-choice experiment with self-employed and salaried workers," *Journal of Economic Behavior & Organization*, 2009, *70* (3), 470 – 484.

Menon, Martina and Federico Perali, "Eliciting Risk and Time Preferences in Field Experiments: Are They Related to Cognitive and Non-Cognitive Outcomes? Are Circumstances Important?," *Rivista Internazionale di Scienze Sociali*, 2009, *117* (3), 593–630.

Mueller, Julia and Christiane Schwieren, "Can personality explain what is underlying women's unwillingness to compete?," *Journal of Economic Psychology*, 2012, *33* (3), 448 – 460.

Nelson, Julie A., "Not-So-Strong Evidence for Gender Differences in Risk Taking," Working Papers 19, University of Massachusetts Boston, Economics Department November 2013.

__ , "Are Women Really More Risk-Averse than Men? A Re-Analysis of the Literature Using Expanded Methods," *Journal of Economic Surveys*, 2014.

Nieken, Petra and Patrick W. Schmitz, "Repeated moral hazard and contracts with memory: A laboratory experiment," *Games and Economic Behavior*, 2012, *75* (2), 1000 – 1008.

Niemeyer, Claudia, J. Philipp Reiss, and Abdolkarim Sadrieh, "Reducing risk in experimental games and individual choice," Technical Report, Karlsruhe Institute of Technology 2013.

Pogrebna, Ganna, DavidH. Krantz, Christian Schade, and Claudia Keser, "Words versus actions as a means to influence cooperation in social dilemma situations," *Theory and Decision*, 2011, *71*, 473–502.

Ponti, Giovanni and Enrica Carbone, "Positional learning with noise," *Research in Economics*, 2009, *63* (4), 225 – 241.

Price, Curtis R. and Roman M. Sheremeta, "Endowment effects in contests," *Economics Letters*, 2011, *111* (3), 217 – 219.

Rosaz, JULIE, "BIASED INFORMATION AND EFFORT," *Economic Inquiry*, 2012, *50* (2), 484–501.

Rosaz, Julie and Marie Claire Villeval, "Lies and biased evaluation: A real-effort experiment," *Journal of Economic Behavior & Organization*, 2012, *84* (2), 537 – 549.

Ryvkin, Dmitry, "Fatigue in Dynamic Tournaments," *Journal of Economics & Management Strategy*, 2011, *20* (4), 1011–1041.

Sapienza, Paola, Luigi Zingales, and Dario Maestripieri, "Gender differences in financial risk aversion and career choices are affected by testosterone," *Proceedings of the National Academy of Sciences*, 2009.

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

Schram, Arthur and Joep Sonnemans, "How individuals choose health insurance: An experimental analysis," *European Economic Review*, 2011, *55* (6), 799 – 819.

Schubert, R., M. Brown, M. Gysler, and H.W. Brachinger, "Financial decision-making: are women really more risk-averse?," *The American Economic Review*, 1999, *89* (2), 381–385.

Schunk, Daniel, "Behavioral heterogeneity in dynamic search situations: Theory and experimental evidence," *Journal of Economic Dynamics and Control*, 2009, *33* (9), 1719 – 1738.

Shafran, Aric P., "Interdependent security experiments," *Economics Bulletin*, 2010, *Vol. 30 no.3*, 1950–1962.

Sheremeta, Roman M., "Experimental comparison of multi-stage and one-stage contests," *Games and Economic Behavior*, 2010, *68* (2), 731 – 747.

Slonim, Robert and Pablo Guillen, "Gender selection discrimination: Evidence from a Trust game," *Journal of Economic Behavior & Organization*, 2010, *76* (2), 385 – 405.

Sloof, Randolph and C. Mirjam van Praag, "The effect of noise in a performance measure on work motivation: A real effort laboratory experiment," *Labour Economics*, 2010, *17* (5), 751 – 765. <ce:title>European Association of Labour Economists 21st annual conference, Tallinn, Estonia, 10-12 September 2009</ce:title>.

Starmer, Chris, "Developments in Non-expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk," *Journal of Economic Literature*, 2000, *38* (2), 332–382.

Sutter, Matthias, Martin G. Kocher, Daniela Raetzler, and Stefan T. Trautmann, "Impatience and Uncertainty: Experimental Decisions Predict Adolescents' Field Behavior," *American Economic Review*, 2013, *103* (forthcoming), 510–31.

Szrek, Helena, Li-Wei Chao, Shandir Ramlagan, and Karl Peltzer, "Predicting (un)healthy behavior: A comparison of risk-taking propensity measures.," *Judgment & Decision Making*, 2012, *7* (6), 716 – 727.

Viscusi, W., Owen Phillips, and Stephan Kroll, "Risky investment decisions: How are individuals influenced by their groups?," *Journal of Risk and Uncertainty*, 2011, *43* (2), 81–106.

Wakolbinger, Florian and Stefan Daniel Haigner, "Peer advice in a tax-evasion experiment," *Economics Bulletin*, 2009, *29* (3), 1653–1669.

Wieland, Alice and Rakesh Sarin, "Gender Differences in Risk Aversion: A Theory of When and Why," mimeo 2012.

Wik, Mette, Tewodros Aragie Kebede, Olvar Bergland, and Stein Holden, "On the measurement of risk aversion from experimental data," *Applied Economics*, 2004, *36* (21), 2443–2451.

Yechiam, Eldad and Guy Hochman, "Loss-aversion or loss-attention: The impact of losses on cognitive performance," *Cognitive Psychology*, 2013, *66* (2), 212 – 231.