DISCUSSION PAPER SERIES

IZA DP No. 8184

A Reconsideration of Gender Differences in Risk Attitudes

Antonio Filippin
Paolo Crosetto

May 2014

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

A Reconsideration of Gender Differences in Risk Attitudes

Antonio Filippin
University of Milan
and IZA

Paolo Crosetto
INRA, UMR 1215 GAEL, University of Grenoble

Discussion Paper No. 8184
May 2014

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit organization supported by Deutsche Post Foundation. The center is associated with the University of Bonn and offers a stimulating research environment through its international network, workshops and conferences, data service, project support, research visits and doctoral program. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

IZA Discussion Paper No. 8184
May 2014

ABSTRACT

A Reconsideration of Gender Differences in Risk Attitudes*

This paper reconsiders the wide agreement that females are more risk averse than males providing a leap forward in its understanding. Thoroughly surveying the experimental literature we first find that gender differences are less ubiquitous than usually depicted. Gathering the microdata of an even larger sample of Holt and Laury replications we boost the statistical power of the test and show that the magnitude of gender differences, although significant, is economically unimportant. We conclude that gender differences systematically correlate with the features of the elicitation method used and in particular the availability of a safe option and fixed probabilities.

JEL Classification: C81, C91, D81

Keywords: gender, risk, survey

Corresponding author:

Antonio Filippin
University of Milan
Department of Economics
Via Conservatorio 7
20122 Milano
Italy
E-mail: antonio.filippin@unimi.it

* 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, and the participants to the IMEBE 2013 Conference in Madrid, the 2013 BEELAB workshop in Florence, the 2013 ESA World Conference in Zurich, the 2013 SIE Conference in Bologna and the MPI 2013 Autumn Workshp in Jena for useful suggestions. All remaining errors are ours.

1. Introduction

Gender differences in risk preferences are often seen as a stylized fact in the economics and psychology literature. Most of the studies and meta-analyses find that women display a more prudent behavior 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), 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, incentivized, 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.

Although there is a 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. Second, no attempt has been made yet to investigate whether and how the elicitation methods play a role in shaping the observed results, over and above possibly increasing the variance in the results. 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. Whether, to what extent, and in which direction the observed results are driven by the characteristics of the elicitation task adopted is an interesting and yet unexplored question.

The goal of this paper is to try to fill these gaps. We first provide a thorough survey of the literature concerning several elicitation methods finding mixed results. In particular we focus on the unexplored HL task, finding that only twenty papers provide enough information to be included in the survey out of more than five hundreds studies that cite Holt and Laury (2002). This is due to the fact that the HL procedure is widely used as a companion task in experimental sessions in which the core treatments deal with other topics involving uncertainty. As a consequence, only a small fraction of contributions explicitly report about gender differences. We hence move beyond a simple meta-analysis to carry out a wide-range investigation based on the largest possible set of microdata, generated by directly asking for data the authors of all the 94 published HL replications.

We collected the data of 54 published studies, more than half of the universe of replications, corresponding to almost eight thousand subjects. This sample increases dramatically the information available through published results, both by increasing the number of studies analyzed and by making them directly comparable. Besides shedding some light on

behavior in the HL task in general, this considerably large set of microdata allows us to provide conclusive evidence about gender differences using this elicitation method. The striking and consistent result is that a gender gap is the exception rather than the rule in HL replications: men and women display similar behavior, and when a difference can be detected it is usually small to negligible.

The large amount of comparable data also allows us to merge them in order 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 analyzed 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. 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 incentivized, self-reported or observed (Byrnes et al., 1999). Even restricting the analysis to the narrow domain of incentivized 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.

Our data and to the best of our knowledge also the rest of the literature available do not allow us to further disentangle the effect of each of these two characteristics. More research by means of controlled experiments is needed to identify which of these two features ultimately triggers gender differences as well as the theoretical framework most suitable to capture this phenomenon. 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 analyzed jointly with the characteristics of the task used to elicit risk preferences. Second, it greatly restricts the set of possible determinants.

The outline of the paper is as follows. Section 2 summarizes the state of the art in the literature about gender differences in risk aversion and presents the survey of the few HL published results by gender. Section 3 describes the building and characteristics of the dataset of HL replications we built and use. Section 4 analyzes 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 behavior 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 summarize the

state of the art in the risk and gender literature, while at the same time illustrating possible advancements. Consequently, we limit our analysis to three representative and widely used methods: the Investment Game, introduced by Gneezy and Potters (1997) and refined by Charness and Gneezy (2010), 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 task.

In the Investment Game 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 (henceforth EG) task 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 characterized 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 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

	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

to A are not infrequent and are regarded as inconsistent choices when modeling the choices without including a stochastic component.

That female are more risk averse than males is often deemed a stylized fact in the economic literature, as emphasized by many papers. This finding is confirmed by some surveys (Croson and Gneezy, 2009; Eckel and Grossman, 2008a), which also stress how some characteristics of the experiments make gender differences more likely to appear. Gender differences may depend on the context, e.g., they are usually not found in contextual experiments (Eckel and Grossman, 2008a; Schubert et al., 1999). Among the risk elicitation tasks analyzed in this paper, this state of the art is well captured by both the Gneezy and Potters (1997) and the Eckel and Grossman (2002) tasks. Both tasks have been already analyzed in the literature from a gender perspective, across different studies and with different subject pools. Females have been shown to consistently display a significantly more prudent average behavior.

Charness and Gneezy (2012) report that in the Investment Game the gender gap is rather systematic and quite sizable. Their findings have been critically reviewed by Nelson (2013), and found to be less significant and robust than reported. Even after Nelson's reconsideration, males invest more than females in most of the experiments analyzed, and such a difference is most of the time 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 many other features like the amount of money at stake, the geographical location, the type of subjects (students vs. professionals), etc.

Similar findings emerge with the Eckel and Grossman (2002) task, with sizable gender differences appearing both in the experiment presenting the task as well as 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) find a gender gap among

peasant households in Zambia. 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 exception.

The results obtained using these two elicitation methods are clear-cut: Females undoubtedly display, on average, a significantly more prudent behavior. 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. Absent any strong difference in the underlying latent construct, what looks like a robust result could instead be driven by the elicitation method. 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 Holt and Laury (2002), which is the most popular risk elicitation method in the literature, but whose replications have never been systematically analyzed along a gender dimension. A survey of the literature reveals that gender differences are only rarely found. Already in the original Holt and Laury (2002) article a gender gap appears only in the low stake but not in the high stake treatment. Despite the fact that more than five hundred published papers cite Holt and Laury (2002),[1] 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 difference.

The three papers finding 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-incentivized version with 9 choices.

The two contributions reporting mixed results find a significant effect only for a subsample, or only through one and not all statistical methods. 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 risk averse in another.

The list of contributions in which the behavior of males and females does not differ significantly is longer, starting with the first replication of the original task (Harrison et al., 2005). The list includes 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. The results obtained with EG task and the Investment Game

[1] 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.

display systematic gaps, while the HL task results tend to support the view that males and females are characterized by similar risk attitudes.[2]

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 stylized fact describing females as more 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 contributions that replicate the HL task and that either provide direct information about gender differences or contain the statistical information necessary to derive them. Such evidence cannot be regarded as conclusive, however, due to both the size of the available sample and to problems of comparability, as explained below. For this reason, we prefer not to base our claims on a meta-analysis only. In the next section we provide a comprehensive analysis of the HL task based on a large sample of microdata of all the articles that replicated the HL task.

3. The Dataset

In this section we build and analyze data from a large sample of HL replications, also including the articles that directly replicate a version of the HL task without reporting gender results.

The reason is manifold. First, this allows us to increase the size and the representativeness of the sample analyzed. As seen above, few papers replicating HL report results about gender differences. While in principle this fact could be a result of selective reporting, no evidence of outcome reporting bias is found in the HL replications, and the likely explanation for the low reporting rate is the fact that the task is performed only as a control, and therefore gender differences in risk preferences are of little interest to the authors (for details see Crosetto et al., 2013).

Second, this allows us to reduce to a common metric a large body of potentially heterogeneous literature. 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.

Therefore, we decided that instead of putting forward a meta-analysis of the published results it was worth trying to collect the microdata of all the replications of the HL task in order to gather an unbiased sample containing the largest possible number of comparable data points.

The result is a large dataset, covering 63 papers, three times as many papers as the ones covered by the simple survey of the literature of Section 2, and twice as many as all the previous survey papers in the experimental economics literature combined.

[2]The same pattern of findings is replicated using a homogeneous pool of subjects by Crosetto and Filippin (2013b).

3.1. Getting the data

For published papers we queried the Scopus bibliographic database, tracking all papers that cited the original Holt and Laury (2002) study. We ran our query on January 31st, 2013. The query resulted in 529 papers: the original Holt and Laury (2002) and 528 papers citing it. As far as unpublished works are concerned, we came across some studies at the conferences we attended while others were signaled to us by the Economics Science Association discussion group. 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, or had simply cited the paper. Among the experimental replications, we restrict the range of possible departures from the original HL that we include 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.[3] Within these boundaries a multiple price list can take many different forms. For instance, we consider tasks in which the number of choices is different than 10, or in which the amounts at stake differ as compared to the original Holt and Laury (2002).

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 analyzed separately in Section 5.

We directly contacted the authors of all the replicating papers, asking them for a set of summary statistics and significance tests, or, if possible, to share with us their microdata. 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.

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.[4]

[3] 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 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)

[4] The number of contributions replicating HL among those currently classified as 'no response' is very likely to be lower than the 48 (40 published and 8 not published) reported in Table 3. In fact, processing the data we collected, it turned out that in about the 30% of the cases we had to exclude the paper from the sample, because of a sufficient different design from the original HL or because of missing gender information. Assuming a similar distribution in the residual category, this implies that we can reasonably expect the real number of

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

3.2. Building a homogeneous dataset

The datasets received 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 in this subsection.

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 ensured a dataset free from order effects or other confounds.[5]

For studies employing a between-subject design, we used all observations. Moreover, when the study included several different treatments concerning the main focus of the paper

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 very likely to be already in the order of two thirds.

[5]Repeated observations for each subject are likely characterized by a great instability as shown for instance by Crosetto and Filippin (2013b); Harrison and Swarthout (2012).

but just one HL task – usually used as a control for risk attitudes – we used all data and kept track of the different underlying conditions through the variable `treatment`. Changes in the HL task administered in the different treatments are infrequent and of marginal importance, but we take them into account in the control variables.

In general the rules described above allowed us to easily reformat the data 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 basically four formats. We deal with this heterogeneity including the variable `detail`.

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 behavior 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 behavior. 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 behavior, nor reconstruct the subjects' binary choices.

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 mistakes (Section 4.3) can instead rely upon the 'full' datasets only.

[6]For instance, 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.

[7]The four subjects that are classified as inconsistent for this category in Table 4 are those who choose the safe lottery when the good outcome is certain. As far as the missing information for this category of studies is concerned 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 really tell from the data we have if multiple-switchers were or not included, but results do not change if we exclude these three dataset from the analysis.

	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.

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 participant made `inconsistent` choices, and some individual controls such as `female` and `age`, though the latter was not always present;

- 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 summarized 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*: Whether the subjects' consistency was `forced` or subjects were 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 `incentivized` 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

[8] See below for an explanation of how a comparable measure has been built across different versions of the multiple price list.

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 keep track of this difference. 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.

Many of the features of the multiple price list have to be taken into account in order to obtain a comparable measure of risk aversion across papers. Usually the proxy for risk aversion adopted in the HL task is the number of safe choices. 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 parametrize the number of safe choices to the probability of switching in order to impose a common metric. For instance, in the example above in Harrison et al. (2007) to a subject who switches when the probability of the good outcome is equal to 0.35 we assign a number of safe choices equal to 3.

4. Results

In this section we analyze our dataset of HL replications from a gender perspective. We first analyze each paper separately, finding that an overwhelming majority of papers do not find any gender difference. We then pool the data, to give details on the risk preferences measured by the HL in general, and how they are affected by characteristics of the task or the subjects.

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 restrict attention to consistent choices (i.e., to subjects switching once and not choosing dominated actions), including both published and unpublished papers to give the vastest overview of the literature possible. For each paper we can compute the mean number of safe choices by gender, as well as the results of a non-parametric Mann-Whitney test, and of Cohen's d (Cohen, 1988) as a measure of effect size. 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 stylized 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

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
`incentivized`	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

by increasing 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 decreasing significance.

The information provided by the p-values of the test and by Cohen's d is complementary. Cohen's d is a measure of effect size, independent of the sample size, and it is computed as

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

where X_m is the average male number of safe choices, X_f is the average female number of safe choices, and s is the pooled standard deviation. The d is positive if females are more risk averse than males, but is free to be negative, in which case it means that the average female is less risk averse than the average male.

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 size. 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. If the interpretation wants to be carried over to differences at the individual level – i.e., predicting with high accuracy if a subject is male or female observing his or her risk aversion only – Cohen's ds of 2 or more are needed, with a value of 4 meaning almost absolute discriminability (Nelson, 2012).

Applying these 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.

In 40 published and 6 unpublished papers females report a more prudent average behavior than males, as far as point estimates are concerned. However, this evidence in favor of a more prominent risk aversion in women is weak, as the differences are 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.

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 because not detecting an effect cannot be directly interpreted as the proof of its absence. The fact of gathering the microdata makes possible such an exercise.

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. Doing so allows us to boost the statistical

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

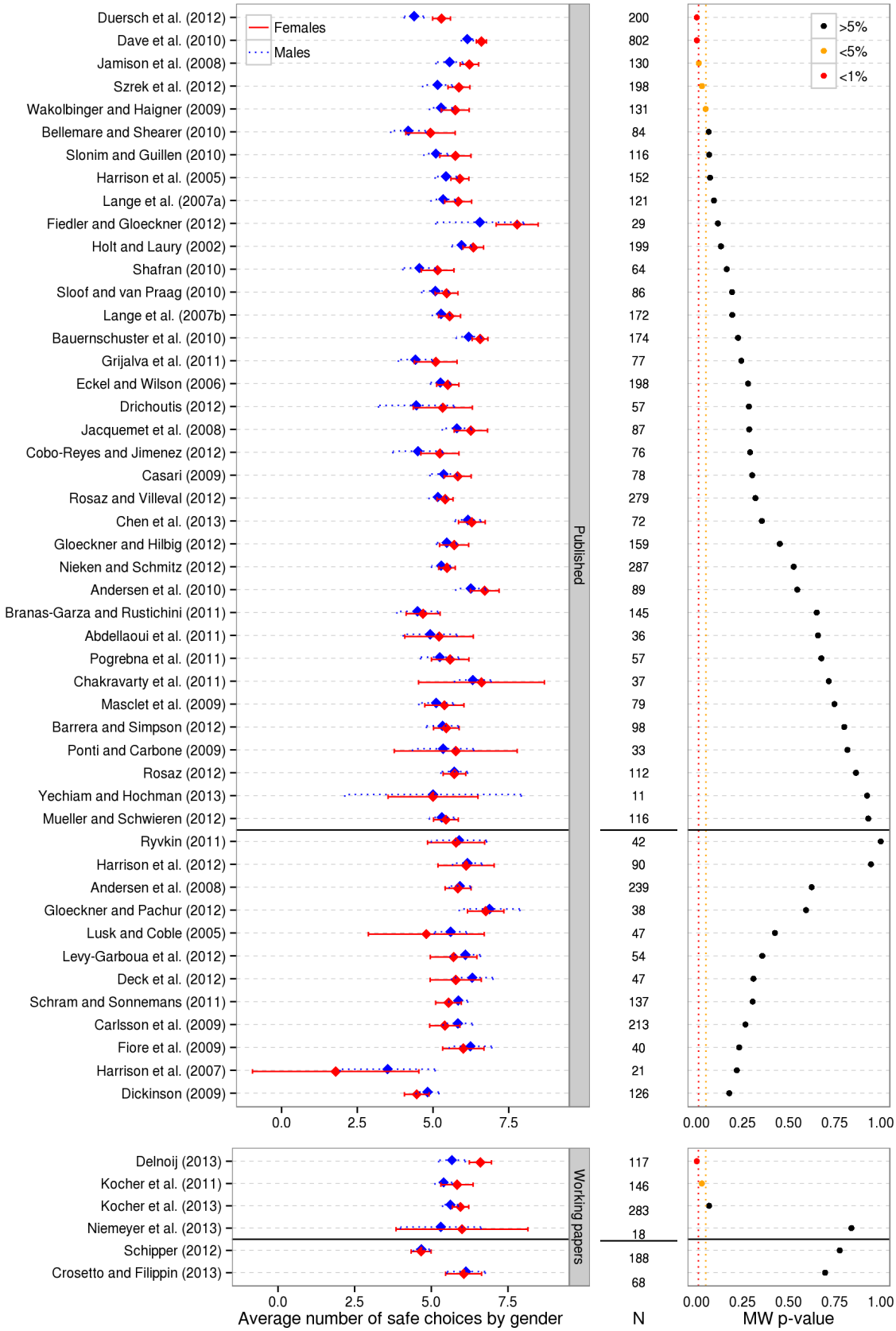

Figure 1: Gender gap in risk taking across HL replications

power of our test, thereby almost eliminating the likelihood of observing a false negative. Even better, it makes possible to provide a precise quantitative estimate of the importance of gender differences using the HL task. Moreover, it gives the opportunity to identify the determinants of the number of safe choices over and above the role played by gender. Finally, 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 shed some light on a feature of the HL task, i.e. inconsistent choices. In this section we cannot rely upon the papers about which we only have descriptive statistics.

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 maximizer 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 behavior, 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 behavior of an expected utility maximizer and for this reason it is usually defined as inconsistent.

This is not the only way in which the behavior can 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.

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 (see Section 4.3 below).

In this Section our goal is to describe the pattern of inconsistent choices, trying also to shed some light on their determinants and consequences. Therefore, we exploit all the information we have concerning inconsistent choices, including also the 'partial' datasets.

The overall frequency of inconsistent choices has already been summarized in Table 4. In Table 7 we provide a more detailed picture showing a breakdown by gender and type of inconsistency. The Table displays the number of inconsistent choices, overall and by gender, out of the total number that can be potentially observed. 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).[9]

Multiple switching is the most common type of inconsistent behavior, 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

[9]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 maximizer should be characterized by an unbelievably high risk aversion coefficient to choose the safe lottery in this case.

	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

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 suggests that the positive correlation between risk aversion and IQ emphasized, among others, by Dohmen et al. (2010), could be an artifact of the format of the price list, as already argued by Andersson et al. (2013).

Dominated choices (choosing always safe or always risky when certain outcomes exist) are much less frequent. Gender in this case does not help explaining the results, and neither do the other determinants, with the exception of incentives that also reduce the likelihood of making a safe choice when the good outcome is certain.

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

In this Section we analyze the risk attitudes of consistent subjects only. Estimates including inconsistent subjects can be performed only for the papers for which we have 'full' data, as done via structural model estimation in Section 4.3. Besides greatly simplifying the estimated decision making process, this approach has the advantage of allowing us to analyze a larger sample of data. This is important, even on the face of very similar average choices (Table 8), since the higher variance in HL implementation details due to an extended sample helps to better identify the determinants of the choices.

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 53% chance of being correct when saying that the more risk averse of the two is a woman, against a 50% rate if we just randomized our answer.

	Mean	St.Dev	N
Whole sample	**5.63**	**1.91**	**5935**
Males	5.47	1.89	2998
Females	5.78	1.91	2937
Microdata (`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, consistent subjects only

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 *d*s 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.[10] This effect is still not huge, but classifiable as a medium effect at the aggregate level.

The significant gender gap, hence, is found in the HL task only when considering a vast sample but it is still negligible in size, while in both the Investment Game and the EG task it is found even in small samples, and is three and a half times as large, on average.

The next step is to try to identify the determinants of the number of safe choices as reported in Table 9. At the same time we can verify whether the unconditional gender differences (0.31 safe choices, Column 1) are robust.

In Column 2 we present our preferred specification trying to emphasize the determinants of the number of safe choices among the controls available.[11] Gender differences barely change (0.3 safe choices) even when relevant factors are controlled for. For instance, we find that not surprisingly incentives matter. Subjects tend to behave in a more risk averse way 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 additional 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 behavior, although the evidence is based on four papers only.

[10] 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)

[11] There are different formats of the HL implemented, but the variation is not so high. Most of the paper are true replications of the HL task. For this reasons we have problems of collinearity when trying to include 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. 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 doable.

	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

Table 9: Determinants of the number of safe choices

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 shown in Section 4.3, where also inconsistent subjects are included in the analysis thanks to the availability of all their binary choices. Results barely change, and in particular the gender gap in the average number of safe choices is stable around 0.3.

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 using the HL task the choices of males and females are not identical. However, they can be detected in a significant way only when the statistical power of the test is high enough because such differences are 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, showing that the features of the risk elicitation mechanism affect the measured behavior over and above the effect of adding some noise. Along the gender dimension the influence of the features of the task is systematic, to the point that it affects the behavior at the aggregate level. The likelihood of observing gender differences is strikingly lower in the HL than in the EG and IG task. Hence, evidence based on those two tasks only cannot be regarded as sufficient to attribute the different observed behavior to actual differences in the underlying risk attitudes. 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 random utility structural 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 maximizers characterized 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 is

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

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

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

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

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.[12] We allow for heterogeneity by gender of both r and μ, and we also control the effects on both parameters of the actual amounts at stake (`realmoney`), also allowing for a quadratic effect, of inflating the numbers on screen via an experimental exchange rate (`exchange`), and of a random order of the choices (`randomorder`). Results can be seen in Table 10.

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, $u(x) = x^r$

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

The estimated risk parameter is $r = 0.64$, in line with the estimation in the original HL paper for the $1\times$ treatment. For what concerns r, females show a significantly higher risk aversion and increasing the stakes significantly increases risk aversion, even though to a slightly decreasing rate. Presenting the lotteries in a random order has a significant effect, increasing risk aversion. Inflating numbers on screen by increasing the exchange rate between the experimental currency unit and dollars (or euros) does not affect the estimates. These results are in line with what found in the previous section using regression analysis.[13]

The average level of noise is $\mu \sim 0.22$, higher than what found in Dave et al. (2010).[14] Females display a similar μ as males and this evidence rules out numeracy from the possible explanations of gender differences. In fact, HL is considered a relatively demanding task from the cognitive point of view. 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. Interestingly, and according to intuition, increasing the stakes slightly reduces noise. On the other hand, displaying the lottery choices in random order has no significant effect, and similarly no effect has the experimental currency inflation.

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.

To start with, it has been argued (Charness and Viceisza, 2011; Dave et al., 2010, among others) that HL is more difficult to understand than other methods. Hence, HL could elicit noisier signals making differences more difficult to be detected. The different gender pattern could then simply be driven by a lower precision in the estimates that characterizes the HL method. We can rule out this possibility comparing the signal to noise ratio (SNR) of the tasks. 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).[15] Similar evidence emerges comparing the SNRs obtained by Crosetto and Filippin (2013b) in a replication of the three tasks in a homogeneous subject pool (HL: 3.27; IG: 2.67; EG: 2.16).

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 observed pattern.[16]

[13] Coefficients have opposite sign in Table 9 and Table 10, because in one case the dependent variable is the number of safe choices, and in the other the risk aversion parameter. Given the utility function employed, a lower r implies more risk aversion.

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

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

[16] See Crosetto and Filippin (2013b) for a thorough comparison of some elicitation methods including those

Apart from the number of choices, the tasks differ along three main lines: a) the menu of 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 option among the set of alternatives.

The Investment Game and the EG task are very similar as far as these theoretical characteristics are concerned. For instance, they both rely upon a change in the amount of money at stake in order to generate the opportunity set from which subjects choose their preferred lottery. Probabilities of the (good and bad) outcomes are instead kept fixed at 50%. Moreover, they are both characterized by a set of possible choices that allows to identify only different degrees of risk aversion. In fact, in the Investment Game risk-neutral as well as risk-loving subjects should invest their entire endowment because the expected return of the risky option is larger than one. In the EG task lottery 5 yields the highest expected value and is the preferred alternative of risk-neutral and risk-loving subjects alike. Finally, both elicitation methods include a risk free alternative. In the Investment Game subjects have the opportunity of securing the entire endowment by investing nothing, while the EG task includes a degenerate lottery with no uncertainty that is equivalent to a safe choice.

The HL task differs with respect to the Investment Game and the EG along all three dimensions. First, the variance across the set of lotteries is obtained modifying the probability distribution. In particular, the good outcome becomes more and more likely while the amounts at stake are kept constant. 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 out with 10% probability only. One could argue that the role of the riskless alternative might be played by the minimum amount of the safer lottery. In other words, by always choosing Option A (except maybe in row 10) the subject can be sure to earn 3.2 euro avoiding less favorable outcomes. Whether such an amount can be considered as a riskless alternative is disputable, but in any case it is definitely less focal than in the other two elicitation methods. In fact, it requires some elaboration by the subjects to be identified as it is not shown directly to them as a specific choice.[17] Moreover, its salience is likely to be diluted by the multiple choice dimension, which induces a comparison across risky alternatives across rows.

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) clearly correlates with the likelihood of observing gender differences in risk preferences. The next step is to try to disentangle the role of each of these factors.

Our data allows us to exclude that gender differences depend on the domain of analysis. The observed pattern of the gender gap could be an artefact of the truncation of the opportunity set as long as female are more risk seeking in the risk loving domain, something that would be hidden by a task that scans only the risk aversion domain, thereby delivering biased estimates. Our dataset allows to directly test and exclude this possibility. In fact,

used in this paper.

[17] We tried to estimate an endogenous reference point a 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.

in the HL females appear slightly more risk averse uniformly, i.e., also in the risk loving domain. Further evidence supporting this claim is provided by the Outcome Scale method, which implements a multiple price list of an increasing safe option against the same 50/50 lottery. The Outcome Scale method has hence two features in common with the EG task and the Investment Game, 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, this is 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). Gender differences are therefore triggered by the availability of a safe option within the set of alternatives and/or by the change in outcomes *vs* change in probabilities.

The availability of a safe option within the set of alternatives has already 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 importance of certainty effects differs by gender. The literature offers the possibility of indirectly testing this explanation. In fact, some contributions use a slightly modified version of the HL task in which instead of facing two risky lotteries, subjects choose repeatedly between a safe amount and a risky lottery. We collected also the contributions that implemented this version of the HL task in order to investigate whether the presence of a safe option increases the magnitude of gender differences.

Table 11 replicates the analysis of Section 4.1 when a safe choice is available. Gender differences emerge more frequently when a safe option is available (in 20% rather than 9.5% of the papers). Of course, we do not dare interpreting such a slightly higher frequency as more than merely suggestive evidence in favor of this interpretation.

Article	N_m	N_f	safe$_m$	safe$_f$	Mann-Whitney	Cohen's d	detail
Burfurd et al. (2012)	127	91	5.54	5.74	0.38	0.16	full
Cason et al. (2010)	192	63	4.11	3.86	0.39	-0.22	partial
Cason et al. (2012a)	181	113	4.21	4.43	0.08	0.21	partial
Cason et al. (2012b)	182	102	3.98	4.28	0.11	0.27	partial
Evans et al. (2009)	76	42	5.88	5.83	0.91	-0.03	full
Gangadharan and Nemes (2009)	29	19	5.17	5.58	0.28	0.25	full
McIntosh et al. (2007)	32	22	4.47	4.41	0.98	-0.04	full
Price and Sheremeta (2011)	60	27	3.98	4.17	0.36	0.17	partial
Samak (2013)	32	13	4.36	5.00	0.07	0.61	full
Schildberg-Hörisch and Strassmair (2012)	196	120	4.92	5.44	**0.00**	-	summary
Sheremeta (2010b)	101	39	4.05	4.05	0.63	-0.00	partial
Sheremeta and Zhang (2010)	71	34	4.19	4.19	1.00	0.00	partial
Sheremeta (2010a)	129	73	3.98	4.38	**0.00**	0.37	partial
Sheremeta (2011)	174	64	4.28	4.48	0.11	0.19	partial
Zhang and Casari (2012)	73	27	4.49	5.04	**0.03**	0.43	full

Table 11: Results by gender of the HL replications

Another exercise that can be performed to test this conjecture is to include these studies in a regression together with those analyzed in the previous sections, and to see whether the interaction between gender and the availability of a safe option is positively and significantly correlated with the number of safe choices. Unfortunately, such an exercise cannot be considered a meaningful and direct tests of this conjecture for several reasons. First, 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. In fact, the

design of the task in most of the safe-choice experiments comes from the same group of authors and it is not directly comparable with the classic HL. The number of options is higher (15) and limited to the range of probability $[0.3 - 1]$ of the good outcome to occur. Second and foremost, in the safe-choice experiments the fixed amount is usually lower than the expected value of the $50\% - 50\%$ risky lottery, and it is kept constant across all the choices and is therefore different from the expected value of the corresponding Option A in the classic HL. Hence, this evidence cannot be regarded as conclusive. A proper test of this conjecture by means of a controlled experiment is left for future research. In any case, results of a similar regression do not detect significant differences associated to the availability of a safe alternative.

The literature does not offer many papers that allow us to tell apart the effects on gender differences of the safe option and of the change in outcomes rather than in probabilities. 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 changing 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 (Crosetto and Filippin, 2013a), in which no gender difference appears.

Summing up, the results in the literature seem to indicate that when a safe option is available in the choice set, and when the tasks employ 50/50 lotteries and change the amounts at stake to generate variation in the expected values, then gender differences in risk preferences are usually found. The data we collected and the evidence present in the literature do not allow us, though, to disentangle which of the two characteristics of the task is crucial for the emergence of gender differences. The absence of both (HL, Bomb task) seem to lead to a similar behavior of males and females. More research and the development of *ad hoc* tests in a controlled environment are needed to shed more light on the issue.

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.

A thorough survey of the literature concerning several elicitation methods offers indeed mixed results. In particular, we focus on the most widely used risk elicitation task (Holt and Laury, 2002), the results of which have never been thoroughly analyzed along a gender perspective. We find that using this task results sharply differ from the consensus, with significant gender differences being the exception rather than the rule.

The HL task is usually employed as a companion task in experiments focusing on other topics. Hence, the number of papers directly reporting gender results is small relative to the number of replications. We decided to move beyond a simple meta-analysis based on a survey of the published results, to carry out a systematic investigation by gathering the largest possible set of microdata. Our dataset contains 54 published studies, an amount that

is way larger than the number of papers included in previous surveys of risk and gender and covering more than half of all the HL replications. Using the HL task gender differences are the exception rather than the rule, as they appear only in less than ten percent of published papers, while they are a regular finding with other elicitation methods. This striking difference is not due to a different average numerosity of the sample, and we can also exclude that it is an artefact of a greater complexity of the HL task.

The possibility of merging the microdata allows us to reach several 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 $\rho = 0.24$. Concerning the determinants of the behavior we find that incentives increase risk aversion in a significant and concave way. Inconsistent choices are a quite recurrent phenomenon, characterizing 14% of the subjects. We also find that females are more likely to display an inconsistent behavior 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 random utility structural model with maximum likelihood. This procedure provides 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, but their magnitude is economically unimportant, at about a sixth of a standard deviation. Note that this magnitude is 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. While heterogeneity in measured risk preferences has already been observed in the literature along many dimensions, the striking fact that we show is that strong differences exist even across incentivized tasks that consist essentially of one ingredient only: choices among lotteries.

The striking difference between our results and the stated view in the literature shows that the likelihood of observing gender differences crucially depends on the features of the task used to elicit risk preferences. This is an interesting result *per se* because it proves that gender differences in risk attitudes cannot be treated as a fact and need to be reconsidered. 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 is the one getting closer to the *true* value of risk preferences.

The paper starts to provide an answer to this question by drawing a first map of the features of the different tasks that might trigger such a different behavior. We can rule out that the observed gender pattern is due to the different support of preferences investigated by the risk elicitation methods. The characteristics that correlate with the emergence of gender differences are instead a) the availability of a safe option among the set of alternatives and b) manipulating the expected value of the lotteries changing the outcomes at stake while keeping their probability fixed at 50%. 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, available data do not allow us to disentangle the two effects in order to identify the ultimate cause of gender differences, something that requires the design of controlled experiments.

We believe that this paper provides a leap forward in the understanding of decision

under risk along a gender perspective, by bringing new evidence to the debate and by providing a map of which characteristics of the task might trigger a different gender behavior. However, 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, M., Driouchi, A., L'Haridon, O., 2011. Risk aversion elicitation: reconciling tractability and bias minimization. Theory and Decision 71, 63–80.

Agnew, J. R., Anderson, L. R., Gerlach, J. R., Szykman, L. R., 2008. Who chooses annuities? an experimental investigation of the role of gender, framing, and defaults. The American Economic Review 98 (2), pp. 418–422.

Andersen, S., Harrison, G., Lau, M., Rutström, E., 2006. Elicitation using multiple price list formats. Experimental Economics 9, 383–405.

Andersen, S., Harrison, G. W., Lau, M. I., Rutström, E. E., 2008. Eliciting Risk and Time Preferences. Econometrica (3), 583–618.

Andersen, S., Harrison, G. W., Lau, M. I., Rutström, E. E., 2010. Preference heterogeneity in experiments: Comparing the field and laboratory. Journal of Economic Behavior & Organization 73 (2), 209 – 224.

Anderson, L., Freeborn, B., 2010. Varying the intensity of competition in a multiple prize rent seeking experiment. Public Choice 143, 237–254.

Andersson, O., Tyran, J.-R., Wengström, E., Holm, H. J., Apr. 2013. Risk aversion relates to cognitive ability: Fact or fiction? Working Papers 2013:9, Lund University, Department of Economics.

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

Arya, S., Eckel, C., Wichman, C., 2012. Anatomy of the credit score. Journal of Economic Behavior & Organization forthcoming.

Baker, R. J., Laury, S. K., Williams, A. W., 2008. Comparing small-group and individual behavior in lottery-choice experiments. Southern Economic Journal 75 (2), 367–382.

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

Barrera, D., Simpson, B., 2012. Much ado about deception: Consequences of deceiving research participants in the social sciences. Sociological Methods & Research 41 (3), 383–413.

Bauernschuster, S., Duersch, P., Oechssler, J., Vadovic, R., 2010. Mandatory sick pay provision: A labor market experiment. Journal of Public Economics 94 (11-12), 870 – 877.

Bellemare, C., Krause, M., Kroger, S., Zhang, C., June 2005. Myopic loss aversion: Information feedback vs. investment flexibility. Economics Letters 87 (3), 319–324.

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

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

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

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

Burfurd, I., Gangadharan, L., Nemes, V., 2012. Stars and standards: Energy efficiency in rental markets. Journal of Environmental Economics and Management 64 (2), 153 – 168.

Byrnes, J. P., Miller, D. C., Schafer, W. D., 1999. Gender differences in risk taking: A meta-analysis. Psychological bulletin 125 (3), 367.

Camerer, C. F., 1992. Recent tests of generalizations of expected utility theory. In: Edwards, W. (Ed.), Utility Theories: Measurements and Applications. Studies in Risk and Uncertainty. Kluwer Academic Publishers, Boston, MA, pp. 207–251.

Carlsson, F., He, H., Martinsson, P., Qin, P., Sutter, M., 2012. Household decision making in rural china: Using experiments to estimate the influences of spouses. Journal of Economic Behavior & Organization 84 (2), 525–536.

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

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

Cason, T. N., Savikhin, A. C., Sheremeta, R. M., 2012a. Behavioral spillovers in coordination games. European Economic Review 56 (2), 233 – 245.

Cason, T. N., Sheremeta, R. M., Zhang, J., 2012b. Communication and efficiency in competitive coordination games. Games and Economic Behavior 76 (1), 26 – 43.

Chakravarty, S., Harrison, G. W., Haruvy, E. E., Rutström, E. E., 2011. Are you risk averse over other people's money? Southern Economic Journal 77 (4), 901 – 913.

Charness, G., Genicot, G., 2009. Informal Risk Sharing in an Infinite-Horizon Experiment. Economic Journal 119 (537), 796–825.

Charness, G., Gneezy, U., 2004. Gender, Framing, and Investment. Tech. rep.

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

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

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

Chen, Y., Katuščák, P., Ozdenoren, E., 2013. Why Can't a Woman Bid More Like a Man? Games and Economic Behaviour 77 (1), 181–213.

Cleave, B. L., Nikiforakis, N., Slonim, R., 2010. Is There Selection Bias in Laboratory Experiments? Department of Economics - Working Papers Series 1106, The University of Melbourne.

Cobo-Reyes, R., Jimenez, N., 2012. The dark side of friendship: 'envy'. Experimental Economics 15, 547–570.

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

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

Crosetto, P., Filippin, A., Feb. 2013b. 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.

Crosetto, P., Filippin, A., Heider, J., 2013. A Study of Outcome Reporting Bias Using Gender Differences in Risk Attitudes. CESifo Working Paper Series 4466.

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

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

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

Deck, C., Lee, J., Reyes, J., Rosen, C., 2012. Risk-taking behavior: An experimental analysis of individuals and dyads. Southern Economic Journal 79 (2), 277–299.

Delnoij, J., 2013. To bid or to buy? heterogeneous bidders' preferences over auction mechanisms, unpublished, presented at IMEBE conference.

Dickinson, D., 2009. The effects of beliefs versus risk attitude on bargaining outcomes. Theory and Decision 66, 69–101.

Dohmen, T., Falk, A., September 2011. Performance pay and multidimensional sorting: Productivity, preferences, and gender. American Economic Review 101 (2), 556–90.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., June 2010. Are risk aversion and impatience related to cognitive ability? American Economic Review 100 (3), 1238–60.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., Wagner, G. G., 2011. Individual Risk Attitudes: Measurement, Determinants, And Behavioral Consequences. Journal of the European Economic Association 9 (3), 522–550.

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

Dreber, A., Rand, D. G., Wernerfelt, N., Garcia, J. R., Lum, J. K., Zeckhauser, R., 2010. 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.

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

Duersch, P., Oechssler, J., Vadovic, R., 2012. Sick pay provision in experimental labor markets. European Economic Review 56 (1), 1 – 19.

Eckel, C., Wilson, R., 2006. Internet cautions: Experimental games with internet partners. Experimental Economics 9, 53–66.

Eckel, C. C., El-Gamal, M. A., Wilson, R. K., 2009. Risk loving after the storm: A Bayesian-Network study of Hurricane Katrina evacuees. Journal of Economic Behavior & Organization 69 (2), 110–124.

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

Eckel, C. C., Grossman, P. J., 2008a. Chapter 113 men, women and risk aversion: Experimental evidence 1, 1061 – 1073.

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

Eckel, C. C., Grossman, P. J., 2008c. Men, Women and Risk Aversion: Experimental Evidence. Vol. 1 of Handbook of Experimental Economics Results. Elsevier, Ch. 113, pp. 1061–1073.

Eckel, C. C., Grossman, P. J., Johnson, C. A., De Oliveira, A., Rojas, C., Wilson, R. K., 2011. On the Development of Risk Preferences: Experimental Evidence. Working Paper Series 2008-5, CBEES.

Eckel, C. C., Wilson, R. K., 2004. Is trust a risky decision? Journal of Economic Behavior & Organization 55 (4), 447 – 465.

Ehmke, M., Lusk, J., Tyner, W., 2010. Multidimensional tests for economic behavior differences across cultures. The Journal of Socio-Economics 39 (1), 37 – 45.

Eriksen, K. W., Kvaløy, O., Olsen, T. E., 2011. Tournaments with prize-setting agents*. The Scandinavian Journal of Economics 113 (3), 729–753.

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

Evans, M. F., Vossler, C. A., Flores, N. E., 2009. Hybrid allocation mechanisms for publicly provided goods. Journal of Public Economics 93 (1-2), 311 – 325.

Falk, A., Huffman, D., Sunde, U., 2006. Self-confidence and search. Discussion paper, IZA - Forschungsinstitut zur Zukunft der Arbeit – Institute for the Study of Labor.

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

Fiedler, S., Glöckner, A., 2012. The dynamics of decision making in risky choice: An eye-tracking analysis. Frontiers in Psychology 3 (335).

Fiore, S. M., Harrison, G. W., Hughes, C. E., Rutström, E. E., 2009. Virtual experiments and environmental policy. Journal of Environmental Economics and Management 57 (1), 65 – 86.

Gangadharan, L., Nemes, V., 2009. Experimental analysis of risk and uncertainty in provisioning private and public goods. Economic Inquiry 47 (1), 146–164.

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

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

Gneezy, U., Leonard, K. L., List, J. A., 2009. Gender Differences in Competition: Evidence from a Matrilineal and a Patriarchal Society. Econometrica 77 (5), 1637–64.

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

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

Grijalva, T., Berrens, R. P., Shaw, W. D., 2011. Species preservation versus development: An experimental investigation under uncertainty. Ecological Economics 70 (5), 995 – 1005.

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

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

Harless, D. W., Camerer, C. F., 1994. The predictive utility of generalized expected utility theories. Econometrica 62 (6), 1251–89.

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

Harrison, G. W., Johnson, E., McInnes, M. M., Rutström, E. E., June 2005. Risk Aversion and Incentive Effects: Comment. American Economic Review 95 (3), 897–901.

Harrison, G. W., Lau, M. I., RutstrÃűm, E. E., Tarazona-GÃşmez, M., 2013. Preferences over social risk. Oxford Economic Papers 65 (1), 25–46.

Harrison, G. W., List, J. A., Towe, C., 2007. Naturally Occurring Preferences and Exogenous Laboratory Experiments: A Case Study of Risk Aversion. Econometrica 75 (2), 433–458.

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

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

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

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

Jacquemet, N., Rullière, J.-L., Vialle, I., 2008. Monitoring optimistic agents. Journal of Economic Psychology 29 (5), 698 – 714.

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

Kocher, M. G., Pahlke, J., Trautmann, S. T., 2011. Tempus Fugit: Time Pressure in Risky Decisions. Discussion Papers in Economics 12221, University of Munich, Department of Economics.

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

Koszegi, B., Rabin, M., 2007. Reference-dependent risk attitudes. American Economic Review 97 (4), 1047–1073.

Lange, A., List, J. A., Price, M. K., 2007a. A fundraising mechanism inspired by historical tontines: Theory and experimental evidence. Journal of Public Economics 91 (9), 1750 – 1782.

Lange, A., List, J. A., Price, M. K., 2007b. Using lotteries to finance public goods: Theory and experimental evidence*. International Economic Review 48 (3), 901–927.

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

Laury, S. K., 2005. Pay one or pay all: Random selection of one choice for payment. Research paper series, Andrew Young School of Policy Studies.

Levy-Garboua, L., Maafi, H., Masclet, D., Terracol, A., 2012. Risk aversion and framing effects. Experimental Economics 15, 128–144.

Lusk, J. L., Coble, K. H., 2005. Risk perceptions, risk preference, and acceptance of risky food. American Journal of Agricultural Economics 87 (2), 393–405.

Masatlioglu, Y., Taylor, S., Uler, N., 2012. Behavioral mechanism design: evidence from the modified first-price auctions. Review of Economic Design 16, 159–173.

Masclet, D., Colombier, N., Denant-Boemont, L., Lohéac, Y., 2009. Group and individual risk preferences: A lottery-choice experiment with self-employed and salaried workers. Journal of Economic Behavior & Organization 70 (3), 470 – 484.

McIntosh, C. R., Shogren, J. F., Dohlman, E., 2007. Supply response to countercyclical payments and base acre updating under uncertainty: An experimental study. American Journal of Agricultural Economics 89 (4), 1046–1057.

Menon, M., Perali, F., 2009. 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 117 (3), 593–630.

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

Nelson, J., Sep. 2012. Are women really more risk-averse than men? INET Research Notes 12, Institute for New Economic Thinking (INET).

Nelson, J. A., Nov. 2013. Not-so-strong evidence for gender differences in risk taking. Working Papers 19, University of Massachusetts Boston, Economics Department.

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

Niemeyer, C., Reiss, J. P., Sadrieh, A., 2013. Reducing risk in experimental games and individual choice. Tech. rep., Karlsruhe Institute of Technology.

Pogrebna, G., Krantz, D., Schade, C., Keser, C., 2011. Words versus actions as a means to influence cooperation in social dilemma situations. Theory and Decision 71, 473–502.

Ponti, G., Carbone, E., 2009. Positional learning with noise. Research in Economics 63 (4), 225 – 241.

Price, C. R., Sheremeta, R. M., 2011. Endowment effects in contests. Economics Letters 111 (3), 217 – 219.

Rosaz, J., 2012. Biased information and effort. Economic Inquiry 50 (2), 484–501.

Rosaz, J., Villeval, M. C., 2012. Lies and biased evaluation: A real-effort experiment. Journal of Economic Behavior & Organization 84 (2), 537 – 549.

Ryvkin, D., 2011. Fatigue in dynamic tournaments. Journal of Economics & Management Strategy 20 (4), 1011–1041.

Samak, A. C. S., 2013. An experimental study of reputation with heterogeneous goods. Decision Support Systems 54 (2), 1134 – 1149.

Sapienza, P., Zingales, L., Maestripieri, D., 2009. Gender differences in financial risk aversion and career choices are affected by testosterone. Proceedings of the National Academy of Sciences.

Schildberg-Hörisch, H., Strassmair, C., 2012. An experimental test of the deterrence hypothesis. Journal of Law, Economics, and Organization 28 (3), 447–459.

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

Schram, A., Sonnemans, J., 2011. How individuals choose health insurance: An experimental analysis. European Economic Review 55 (6), 799 – 819.

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

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

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

Sheremeta, R. M., 2010a. Expenditures and information disclosure in two-stage political contests. Journal of Conflict Resolution 54 (5), 771–798.

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

Sheremeta, R. M., 2011. Contest design: An experimental investigation. Economic Inquiry 49 (2), 573–590.

Sheremeta, R. M., Zhang, J., 2010. Can groups solve the problem of over-bidding in contests? Social Choice and Welfare 35,

175–197.

Slonim, R., Guillen, P., 2010. Gender selection discrimination: Evidence from a trust game. Journal of Economic Behavior & Organization 76 (2), 385 – 405.

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

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

Sutter, M., Kocher, M. G., Raetzler, D., Trautmann, S. T., 2013. Impatience and Uncertainty: Experimental Decisions Predict Adolescents' Field Behavior. American Economic Review 103 (forthcoming), 510–31.

Szrek, H., Chao, L.-W., Ramlagan, S., Peltzer, K., 2012. Predicting (un)healthy behavior: A comparison of risk-taking propensity measures. Judgment & Decision Making 7 (6), 716 – 727.

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

Wakolbinger, F., Haigner, S. D., 2009. Peer advice in a tax-evasion experiment. Economics Bulletin 29 (3), 1653–1669.

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

Wik, M., Kebede, T. A., Bergland, O., Holden, S., 2004. On the measurement of risk aversion from experimental data. Applied Economics 36 (21), 2443–2451.

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

Zhang, J., Casari, M., 2012. How groups reach agreement in risky choices: an experiment. Economic Inquiry 50 (2), 502–515.