Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Reconsideration of Gender Differences in Risk Attitudes

Antonio Filippin, Paolo Crosetto

To cite this article:
Antonio Filippin, Paolo Crosetto (2016) A Reconsideration of Gender Differences in Risk Attitudes. Management Science

Published online in Articles in Advance 16 Feb 2016

. http://dx.doi.org/10.1287/mnsc.2015.2294

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–23
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2015.2294

A Reconsideration of Gender Differences in Risk Attitudes

Antonio Filippin
Department of Economics, University of Milan, 20122 Milano, Italy; and Institute for the Study of Labor (IZA), 53113 Bonn, Germany, antonio.filippin@unimi.it

Paolo Crosetto
Institut National de la Recherche Agronomique and Grenoble Applied Economics Laboratory, Université Grenoble Alpes, F-38000 Grenoble, France, paolo.crosetto@gmail.com

This paper reconsiders the wide agreement that females are more risk averse than males. We survey the existing experimental literature, finding that significance and magnitude of gender differences are task specific. We gather data from 54 replications of the Holt and Laury risk elicitation method, involving about 7,000 subjects. Gender differences appear in less than 10% of the studies and are significant but negligible in magnitude once all the data are pooled. Results are confirmed by structural estimations, which also support a constant relative risk aversion representation of preferences. Gender differences correlate with the presence of a safe option and fixed probabilities in the elicitation method.

Keywords: gender; risk; survey; meta-analysis
History: Received November 25, 2014; accepted July 10, 2015, by John List, behavioral economics. Published online in *Articles in Advance*.

1. Introduction

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

Despite the apparently wide agreement that females are more risk averse than males, we believe that the evidence supporting this view cannot be considered conclusive for two reasons. First, there are important branches of the literature still largely unexplored. For instance, the Holt and Laury (2002) (henceforth HL) task has never been the subject of a comprehensive analysis by gender, despite being by far the most popular elicitation method in economics according to the number of citations. Moreover, in the Bomb Risk Elicitation Task (Crosetto and Filippin 2013) no gender difference emerges. Second, no attempt has been made yet to investigate whether and how the elicitation methods play a role in shaping the observed results by gender. Risk attitudes are a latent construct that can only be indirectly and imperfectly measured: their measurement is by construction a combination of the latent preferences *and* the measurement error induced by the tool used to elicit them. Crosetto and Filippin (2016) analyse to what extent, and in which direction, the measured risk preferences are shaped by the characteristics of the elicitation task adopted. In this paper we also try to extend their exercise along a gender dimension.

We first provide a thorough survey of the literature, finding mixed results. We then focus on the

unexplored HL task. Unfortunately, only a small fraction of contributions explicitly report about gender differences, because HL is usually a companion task in unrelated experimental studies. Only 20 papers, out of the more than 500 citing Holt and Laury (2002), provide data on the gender breakdown of risk preferences. Contrary to the widespread consensus, out of these 20 papers, only 3 report significant gender differences in risk preferences.

This striking result, combined with the presence of large amounts of uncharted HL gender data, spurred us to directly contact the authors of the 94 published HL replications. We collected the data of 54 published studies, corresponding to about 7,000 subjects, and reduced them to a common comparable format.

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

The large amount of comparable data also allows us to greatly increase the statistical power of the analysis. Moreover, access to all microdata allows us to exploit the data of subjects making inconsistent choices using a structural model estimated with maximum likelihood. The results on the pooled data show a comeback of significant gender differences, both using linear and panel regressions and in the maximum-likelihood estimation, that indicates constant relative risk aversion (CRRA) as the supported representation of preferences. The magnitude of the effect turns out however to be economically unimportant. Differences amount to one-sixth of a standard deviation, less than a third of the effect found by other elicitation methods analysed in this paper (e.g., by Charness and Gneezy 2012, Eckel and Grossman 2008b).

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

Published results as well as our data set do not allow us to further investigate and disentangle the effect of each of these two characteristics. Nevertheless, we believe that this paper provides a leap forward in the understanding of gender differences in risk preferences from two points of view. First, it makes clear that, instead of being treated as a fact, gender differences should be analysed jointly with the characteristics of the task used to elicit risk preferences. Second, it greatly restricts the set of possible determinants and proposes two candidate explanations.

The outline of this paper is as follows. Section 2 summarises the state of the art in the literature about gender differences in risk preferences and presents the survey of the few HL published results by gender. Section 3 describes the characteristics of the data set of HL replications we built and use. Section 4 analyses our data set, first paper by paper and then pooling the data, using both descriptive statistics and structural modeling allowing for errors in the choices. Section 5 discusses which characteristics of the task could trigger the stark difference in behaviour observed, identifying some candidates. 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 all the tasks used to measure risk preferences. In contrast, our goal is to summarise the state of the art in the risk and gender literature. Consequently, we limit our analysis on the Holt and Laury (2002) task, the most cited and replicated risk elicitation method, and on two other methods that besides being widely used are also those on which the evidence on gender differences has been mainly based: the Investment Game, introduced by Gneezy and Potters (1997), and an Ordered Lottery Selection task proposed by Eckel and Grossman (2002, 2008b).

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

In the Eckel and Grossman task (henceforth EG) subjects make a single choice, picking one out of an

Table 1 The 5 Lotteries of the Orginal Eckel and Grossman (2002) Paper

	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 2 The 10 Lotteries of the Original Holt and Laury (2002) Paper

	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$

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

In the Holt and Laury (2002) risk elicitation method 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).[1]

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 option B from the fifth choice on. The higher the number of safe choices, the stronger the degree of risk aversion. Never choosing the risky option or switching from B to A are not infrequent and are regarded as inconsistent choices when modeling the choices without including a stochastic component.

That women are more risk averse than men is often considered a stylised fact in the economic literature. This finding is confirmed by some surveys (Eckel and Grossman 2008a, Croson and Gneezy 2009).[2] Both the IG and EG tasks have already been surveyed from a gender perspective, and females have been shown to consistently display a significantly more risk-averse average behaviour.

Charness and Gneezy (2012) report that in the IG the gender gap is rather systematic and quite sizable. Males invest more than females in most of the experiments analysed, and such a difference is usually about 10%–15% of the initial endowment (Dreber et al. 2011, Charness and Gneezy 2010, Charness and Genicot 2009, 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 (2016), whereas 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 versus field) in which data have been gathered as well as to other features (amounts at stake, geographical location, type of subjects).

Similar findings emerge in the EG task, with sizable gender differences appearing both in the original experiment and in later replications (Eckel et al. 2009, Dave et al. 2010, Ball et al. 2010, Grossman and Eckel 2015, Arya et al. 2013, Crosetto and Filippin 2016, Wik et al. 2004). Cleave et al. (2013) 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.

[1] The application by Holt and Laury (2002) of a multiple price list is not the first to elicit risk preferences in decisions under risk. Earlier contributions could be found, for instance, in Cohen et al. (1987) and Tversky and Kahneman (1992). However, at the present time, the HL task constitutes the most widely known implementation of the multiple price list approach applied to risk, thereby making it the ideal candidate for our exercise. To keep our search within tractable limits, we do not include other well-established versions of a multiple price lists, such as the so-called Outcome Scale, in which an increasing safe amount is compared with a fixed 50/50 lottery, or, in general, any task in which outcomes change and probabilities are fixed (see, among others, Abdellaoui et al. 2011; Sapienza et al. 2009; Falk et al. 2006; Eriksen et al. 2011; Dohmen and Falk 2011; Dohmen et al. 2010, 2011; Sutter et al. 2013; Masatlioglu et al. 2012; and Andersson et al. 2016).

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

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

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

A survey of the literature reveals that gender differences are only rarely found in this case. Despite the fact that more than 500 published papers cite Holt and Laury (2002),[3] 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, and 15 find no significant difference.

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

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

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

Summarizing, the frequency of significant gender differences sharply changes according to the elicitation method used. Significant gender differences appear systematically using the EG and IG tasks, whereas they do not appear using HL or the Bomb Risk Elicitation Task.

This instability of results supports the view that a latent construct like risk attitudes can only be indirectly measured and what is observed heavily depends on the characteristics of the risk elicitation procedure used. Applied to differences of risk preferences along a gender perspective, this argument implies that the stylised fact describing females as more risk averse than males could be less solid than what it appears at first glance and definitely requires further investigation.[4]

The evidence in this section is based on the 20 studies that provide in their published version information about gender differences. Such evidence cannot be regarded as conclusive, however, because of both the small size of the available sample, as compared to the overall number of published HL replications, and to problems of data comparability across papers. This spurred us to collect the original data of the HL replications, with the aim of covering the largest possible number of studies. The details and results of this exercise are described in the next section.

3. The Data Set

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

First, few studies replicating HL report gender results. Collecting the original data allows us to increase the size and representativeness of the sample analysed. The final data set includes data from 33 articles that did not report gender results.[5] The final data set covers, therefore, 54 published (plus 9 unpublished) papers, twice as many as all the previous survey papers in the experimental economics literature combined.

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

[4] A similar recommendation can be found in Nelson (2015b), who focuses instead on the magnitude of the gender differences analyzing the results in the IG task.

[5] In principle, published results by gender could be the output of a process of selective reporting, as suggested for instance by Nelson (2015a). In contrast, no evidence of outcome reporting bias is found in the HL replications (for details, see Crosetto et al. 2015). An explanation for the low reporting rate is that the task is often performed as a control, and therefore gender differences in risk preferences are of little interest to the authors.

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

3.1. Getting the Data

For published papers we queried the Scopus bibliographic database, tracking all papers that cited Holt and Laury (2002). We ran our query on January 31, 2013, finding 528 citing papers. We also included some unpublished studies, either signalled to us by the Economics Science Association discussion group or that we came across at conferences, resulting in 26 additional contributions.[6]

We closely examined all the 555 papers in the resulting pool to check whether the authors had replicated the HL experiment, in its original version or with some small variations of the design. Among the experimental replications, we restricted the range of possible departures from the original HL to be included in the data set. 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. Within these boundaries, a multiple price list can take many different forms. For instance, we include tasks in which the number of choices is different than 10, or in which the amounts at stake differ as compared to the original HL.

The results of this exercise are detailed in Table 3. We could not access, either in electronic or in paper form, 48 studies. Out of the remaining contributions, we found 118 published and 17 unpublished studies replicating the HL mechanism as described above, whereas 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 surveyed separately in §5.

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

Table 3 Building the Data Set of HL Replications

Articles citing Holt and Laury (2002) as of January 31, 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 data set	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 data set	54	9
Of which:		
Microdata (shared or available online)	48	6
Summary statistics (shared or published)	6	3

Whenever the same data set was used in two or more studies, we counted it only once, including the other references in the "duplicate data set" category. Sixteen 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 data set includes data from 54 published and 9 unpublished papers, for a total of 7,654 and 927 subjects, respectively.[7]

3.2. Building a Homogeneous Data Set

The data sets of the replications differ along several dimensions, from the purpose and the design of the experiment to the exact format of the multiple price list. Moreover, data sets 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 data

[6] Given that the subsample of unpublished papers is not representative and might suffer from severe self-selection issues, we treat it separately whenever possible, and we exclude it in §4.2 and in §4.3 where we derive aggregate results.

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

sets comparable requires decisions that inherently encompass a degree of arbitrariness. The decisions and assumptions we made in building the data set are detailed here.

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

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

In general, the rules described above allowed us to easily process the replications and include them in the data set. In some cases, though, the inclusion proved harder and ad hoc rules were necessary.[8]

3.2.2. Level of Detail of the Data. The data sets come in four formats. We deal with this heterogeneity including the variable `detail`.The most complete data sets provide us with data for each and every binary choice the subjects made (`detail` = "full"). Other data sets record the number of safe choices of every subject and a dummy variable indicating whether they switched only one or multiple times—this behaviour is usually labeled as "inconsistent" (`detail` = "partial"). For these data sets we can reconstruct the binary choices of the consistent (single switchers) only, whereas 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 data sets report only the number of individual safe choices, but no information about inconsistent behaviour. To not lose these observations, by default we assume that the authors sent us data for single switchers only.[9] Finally, in some cases we only obtained summary statistics of the results, including the average number of safe choices by gender, and the results of statistical tests (`detail` = "summary"). In this case we cannot retrieve any information about inconsistent behaviour, nor reconstruct the subjects' binary choices.

The breakdown of the number of consistent and inconsistent subjects in our data set 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, whereas the description of results by paper (§4.1) relies upon all the information available, the analysis of microdata (§4.2) cannot include "summary" data, and the structural model estimation allowing for error (§4.3) can instead rely upon the "full" data sets 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 data set only the following information:

- The *subjects*. The data set 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.[10] Data also contain a variable summarizing whether the participant made `inconsistent` choices, and some individual controls such as `female` and `age`, though the latter is not always available.
- The *format of the multiple price list*. The papers included in our analysis greatly differ in the specific features of the multiple price list adopted. Examples of such differences are (a) the number of binary

[8] In the case of Andersen et al. (2008), we faced a data set with three different price lists, between subjects. One of the lists was a standard "symmetric" one, whereas the other two were asymmetric ("SkewLow" and "SkewHigh"). The asymmetric price lists featured six choices each, and the choices did not cover the whole probability range. For instance, the "SkewLow," covered probabilities 0.1, 0.2, 0.3, 0.5, 0.7, and 1. We keep only the symmetric treatment of Andersen et al. (2008).

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

[10] Several features of the multiple price list need to be taken into account to obtain a comparable measure of risk aversion across studies. For instance, making six 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 six 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, because in this case there are 20 choices and the change in probability between each row is 5% instead of 10%. Therefore, we parametrise the number of safe choices to the probability of switching in order to impose a common metric. In the example above in Harrison et al. (2007) we assign a number of safe choices equal to three to a subject who switches when the probability of the good outcome is equal to 0.35.

Table 4 Subjects in the Sample by Consistency and Type of Data (Published Papers Only)

		Consistent subjects			Inconsistent subjects		
	Detail	Males	Females	Total	Males	Females	Total
Microdata	Full	2,057	2,139	4,196	409	500	909
No. of safe choices + Consistency	Partial	504	408	912	64	98	162
No. of safe choices only	Partial	375	324	699	3	1	4
Summary statistics (shared/published)	Summary	413	359	772	—	—	
Total		3,349	3,230	6,579	475	596	1,075

Note. The four subjects that are classified as inconsistent when only the number of safe choices is known are those who choose the safe lottery when the good outcome is certain.

choices (`numchoices`) and consequently the change in the probability of the good outcome from one row to the next; (b) the support of the probability spanned ([0.1:1] is the most common version, but [0:0.7] is also rather frequent, and we include other domains as well); and (c) the variance of the outcomes. All these features are summarised by the variables `Av1` `Av2` `Bv1` `Bv2`, storing the values of lotteries A (safe) and B (risky), expressed in experimental units, and `Ap1` `Ap2` `Bp1` `Bp2`, storing the probabilities of the four outcomes, for every `decision`.

- The *procedure of the task*. There are two variables keeping track of whether the subjects' consistency was `forced` or subjects were instead free to switch more than once from option A to option B, and whether the decisions were proposed following the increasing likelihood of the good outcome or instead in a `randomorder`. Regarding the structure of the incentives, we keep track of whether choices were `incentivised` or hypothetical and of the `exchange` rate from experimental currency to dollars. By multiplying the amounts seen at the screen by the exchange rate we can also compute the `realmoney` at stake in the experiment as the expected value of the 50/50 lottery A.
- The *characteristics of the experiment*. Some studies focus explicitly on measuring risk preferences directly for different subpopulations and in different contexts, or study the task itself or different versions of it, or else contribute mainly from a theoretical point of view to the understanding of decisions under risk. Other studies focus on other topics, like auctions, strategic games, tournaments, and use the HL task just as a control for risk preferences. We built the variable `control` to take this difference into account. Moreover, especially for the papers in which HL was used as a control, we record in the variable `treatment` the fact that the HL data might have been associated to different treatments in the core part of the experiment.

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

4. Results

In this section, we analyse our data set of HL replications from a gender perspective. We first analyse each paper separately, finding that an overwhelming majority of papers do not find significant gender differences. We then pool the data to increase the statistical power and to explore how the characteristics of the task and of the subjects affect the measured risk preferences.

4.1. Paper by Paper

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

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

where $\bar{X}_m$ and $\bar{X}_f$ are the average male and female number of safe choices and s is the pooled standard deviation. The d is positive if females are more risk averse than males and negative if the opposite is true. Cohen (1988) indicated thresholds for interpreting his d: as long as the discussion is related to aggregate differences, 0.2 is a small effect, 0.5 is a medium effect, and from 0.8 on there can be said to be a large effect.[11]

Results are detailed in Table 6, and graphically displayed in Figure 1, which includes only the papers for which we have full or partial detail. Figure 1 shows

[11] To be able to interpret the effect at the individual level—i.e., predicting with high accuracy a subject's gender observing his or her risk aversion only—a Cohen's d of 2 or more is needed, with a value of 4 meaning almost absolute discriminability (Nelson 2015a).

Table 5 Description of the Data Set (Published Papers Only)

Variable	Type	Description			
		Source of data			
ID	Integer	Unique ID for the *paper*			
detail	Categorical	See §3.2.2			
		Subjects' characteristics and choices			
			Min	Mean	Max
subject	Integer	Unique ID for each subject in the data set			
safechoice	Dummy	1 if safe lottery A chosen, 0 if risky lottery B	0	0.569	1
inconsistent	Dummy	1 if multiple switches *or* dominated choices	0	0.161	1
female	Dummy	1 if female, 0 if male	0	0.500	1
age	Integer	Age in years	0	27.68	84
		Format of the multiple price list			
			Min	Mode	Max
decision	Integer	Decision row number			
numchoices	Integer	Number of rows in the HL table	9	10	20
Av1	Float	High outcome of (safer) lottery A	1	2	125,000
Av2	Float	Low outcome of (safer) lottery A	0.8	1.6	100,000
Bv1	Float	High outcome of (riskier) lottery B	1.90	3.85	240,625
Bv2	Float	Low outcome of (riskier) lottery B	0.05	0.1	6,250
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	0.008	1
random	Dummy	1 if decisions in random order, 0 otherwise	0	0.071	1
incentivised	Dummy	1 if task paid with money, 0 otherwise	0	0.896	1
exchange	Float	Exchange rate experimental currency unit/$	1	42.37	2,500
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	0.537	1
treatment	Integer	Treatment in the original paper (*not* in the HL)	1	1.566	13

the mean choice by gender and its confidence intervals, as well as the p-value of the Mann-Whitney test on the equality of the two distributions. In both the table and the figure, unpublished results are reported separately. In Table 6, papers are listed alphabetically, and significant results are shown in bold. In Figure 1, papers are sorted according to the strength of their results supporting the stylised fact that women are more risk averse. The upper part of each panel contains the papers in which females are more risk averse than males, sorted by decreasing significance. In the lower part of the figure, the papers (12 published, 2 unpublished) in which the average female is *less* risk averse than the average male are listed and sorted by increasing significance.

In 41 published and 6 unpublished papers, females show a more risk-averse average behaviour than males, as far as point estimates are concerned. However, the difference is in the majority of cases not significant. Males are more risk averse than females in 12 published and 2 unpublished papers, and this difference is never significant. When looking together at the whole data set 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 §2. This fraction decreases to about 9.25% (5 out of 54) restricting the analysis to the published studies only.[12]

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

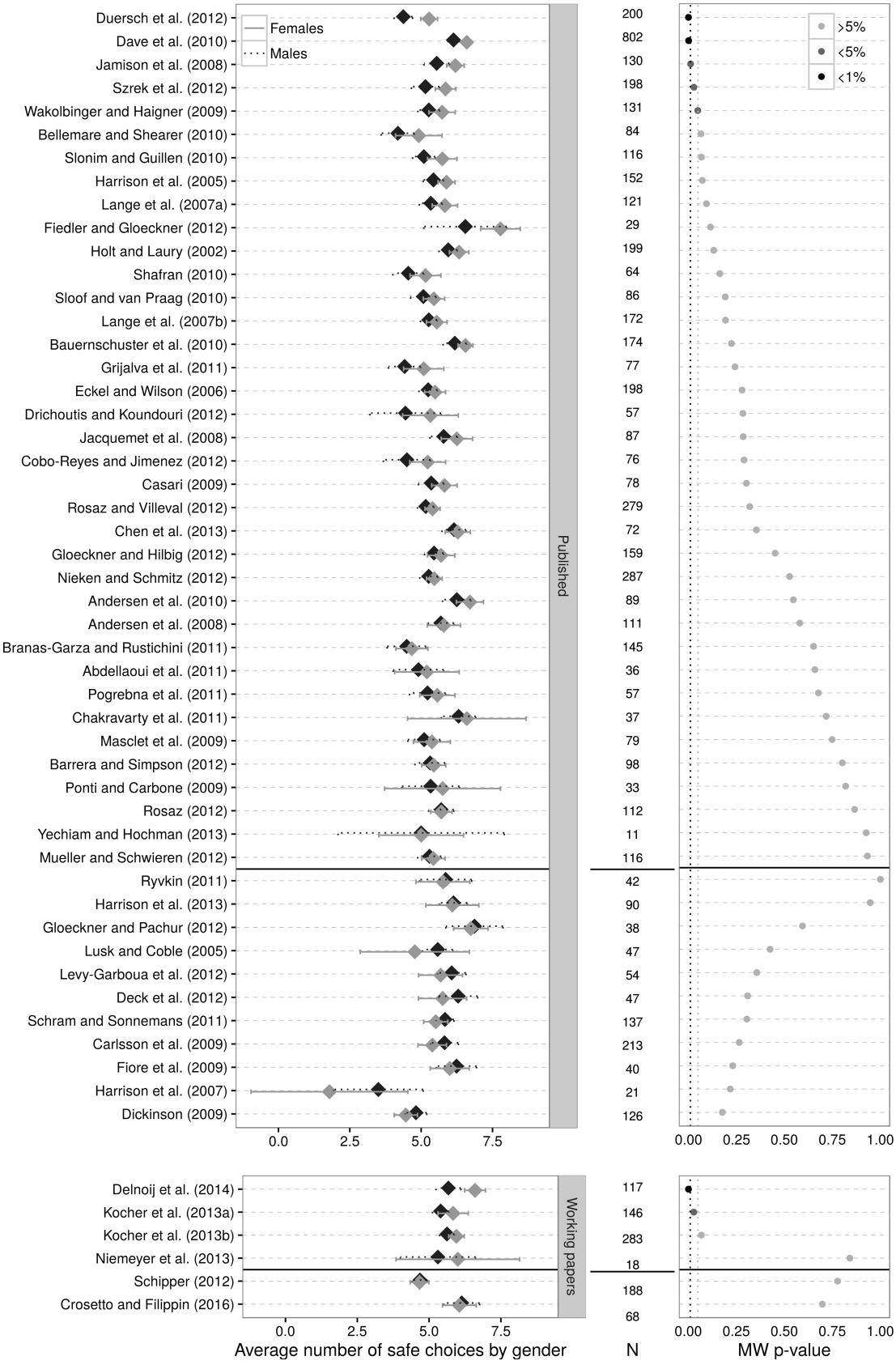
[12] We also carried out maximum-likelihood structural estimations paper by paper for the cases in which full information is available. The results are by and large consistent with the Mann-Whitney results, with just 3 out of 37 papers finding significant gender differences. See the results in Table A.2 in the appendix; the methodology applied to derive the table is described in detail in §4.3.

Table 6 Results by Gender of the HL Replications (Consistent Subjects Only)

Article	N_m	N_f	Safe$_m$	Safe$_f$	Mann-Whitney	Cohen's d	Detail
Published							
Abdellaoui et al. (2011)	21	15	4.90	5.20	0.66	0.15	Full
Andersen et al. (2008)	55	56	5.69	5.79	0.58	0.05	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	Full
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	6.00	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: Not published as of January 31, 2013							
Crosetto and Filippin (2016)	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 et al. (2014)	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. (2013a)	97	49	5.40	5.84	**0.03**	0.28	Full
Kocher et al. (2013b)	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

Note. Significance at 5% of gender differences emphasized in bold.

Figure 1 **Gender Differences Across HL Replications (Papers with Full and Partial Detail, Consistent Subjects)**



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

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

4.2. Merging the Data Sets

The goal of this section is to derive additional insights by merging all the available microdata rather than analyzing them separately.[13] This approach has many advantages. First, it allows us to boost the statistical power of our test, thereby almost eliminating the likelihood of observing a false negative. Second, it makes possible to provide a precise quantitative estimate of the magnitude of gender differences using the HL task. Third, it gives the opportunity to identify the determinants of the number of safe choices over and above the role played by gender. Fourth, the panel structure of the data set grants the opportunity of controlling for any paper-specific characteristic, both observable and unobservable. A byproduct of this exercise is also to deliver a precise quantitative estimate of the main findings in the HL in general. However, before pursuing these goals, we deal with an important feature of the HL task, i.e., that of generating inconsistent choices.

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

However, observing similar patterns does not necessarily imply a violation of the axioms underlying expected utility, as the subjects could simply be consistent with this model but at the same time making mistakes. We test what happens when accepting this view by estimating a structural model with a stochastic component in §4.3.

The goal of this section is instead to describe the pattern of inconsistent choices, also trying to shed some light on their determinants and consequences. We do so by exploiting all the information we have concerning inconsistent choices, including also the 'partial' data sets. In contrast, we cannot rely upon the papers about which we only have descriptive statistics.

The absolute frequency of inconsistent choices has already been summarised in Table 4. In Table 7 we provide a more detailed picture showing a breakdown by gender and type of inconsistency. Table 7 displays the number of inconsistent choices, overall and by gender, out of the total number that can be potentially observed for each type of inconsistency. For instance, multiple switching cannot be observed in papers in which a single switching decision is imposed by design. Always choosing the safer (riskier) lottery is a dominated action only if there is a choice in which the good outcome has probability one (zero).[14]

Multiple switching is the most common type of inconsistent behaviour, observed about 10% of the time. Females are significantly more likely to be inconsistent (Fisher exact test $p < 0.001$) and, at first glance, this might be consistent with the literature finding gender differences in numeracy (see, for instance, Niederle and Vesterlund 2010). These differences survive also in a multivariate framework in which other possible determinants are included. In particular, presenting the lotteries in random order dramatically increases the fraction of inconsistencies. The number

[13] Since the unpublished papers are a small and nonrepresentative sample, their presence could bias the results of the merged data set. Hence, in this and in all further sections we restrict our attention to published papers only.

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

Table 7 Summary Statistics of Inconsistent Subjects by Type and Gender

	Inconsistent subjects		% of inconsistent subjects		
	Number	Out of	Males	Females	Total
Multiple or inverse switching	703	6,825	8.8	11.8	10.3
Dominated choices	102	6,882	1.8	1.2	1.5
Switch *and* dominated	270	6,825	3.6	4.3	4.0
Total	1,075	—	14.1	17.3	15.8

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

of choices in the price list also significantly increases inconsistencies, although to a much lower extent, whereas the presence of monetary incentives significantly reduces them.

Inconsistent subjects make on average 5.16 safe choices, without significant gender differences (Mann-Whitney test, $p = 0.67$). This number is lower than that of consistent subjects (5.62), 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 time. This interpretation is in line with Andersson et al. (2016), who claim that the positive correlation between risk aversion and IQ that has been emphasised, among others, by Dohmen et al. (2010), is an artifact of the format of the price list.

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

4.2.2. Ordinary Least Squares Estimate of Gender Differences. In this section we analyse the risk attitudes of consistent subjects only. Besides greatly simplifying the estimated decision-making process, this approach has the advantage of allowing us to analyse the largest possible sample of microdata, because we must give up only the papers with "summary" data.[15] The higher variance in HL implementation details granted by the whole sample of published papers helps to better identify the determinants of the choices.

Table 8 shows that on average males make a lower number of safe choices, whereas variance is similar. Thanks to the high number of observations, gender differences turn out to be statistically significant

Table 8 Summary Statistics of Safe Choices (Published Papers, Consistent Subjects Only)

	Mean	Std. dev.	*N*
Data (`detail` = "partial" + "full")	5.62	1.89	5,807
Males	5.46	1.89	2,936
Females	5.78	1.89	2,871
Data (`detail` = "full")	5.73	1.95	4,196
Males	5.58	1.93	2,057
Females	5.88	1.95	2,139

(Mann-Whitney test, $p < 0.001$) in both samples. The Cohen's d on the pooled sample is $d = 0.17$, a tiny 17% 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 when comparing two random persons, and assuming normal distribution of risk preferences, there would be a 54.78% chance of being correct when saying that the more risk averse of the two is a woman, against a 50% chance for a random answer.

For the sake of comparison, we run a similar exercise using data for the IG and for the Eckel and Grossman task. For the IG, we use the Cohen's d's computed by Nelson (2015b) 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 data presented in Crosetto and Filippin (2016). The average effect size coincides for the two elicitation methods and it is equal to $d = 0.55$, more than three times the effect found in HL.[16] This effect is still not huge, but classifiable as a medium effect at the aggregate level.

Summarizing, a significant gender gap is found in the HL task only when considering a vast sample, but it is negligible in size. In both IG and EG it is found

[15] Estimates including inconsistent subjects can be performed only for the subset of papers for which we have "full" data, as done via structural model estimation in §4.3. Table 8 shows that restricting the sample to "full" data would not significantly change the picture.

[16] To make the two measures comparable, we compute the Cohen's d for each paper in our data set, 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 $= 0.003$).

Table 9 Determinants of the Number of Safe Choices

	Dependent variable: *Number of safe choices*							
	(1)		(2)		(3)		(4)	
	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.
constant	5.460***	0.035	5.300***	0.039	5.350***	0.049	5.470***	0.034
female	0.321***	0.049	0.326***	0.049	0.294***	0.059	0.298***	0.048
realmoney			0.013***	0.002	0.020***	0.002		
realmoney/^{2}/100			−0.004***	0.000	−0.007***	0.000		
exchange/100			0.011	0.220	−0.002	0.009		
randomorder			0.359***	0.005	0.309**	0.129		
Fixed effects	No		No		No		Yes	
R^2	0.007		0.019		0.025		0.098	
N	5,807		5,807		4,196		5,807	
Detail	Full + Partial		Full + Partial		Full		Full + Partial	

Note. Fixed effects at the replication level.
** and *** denote statistical significance at the 5% and 1% levels, respectively.

even in small samples and it is more than three times as large.

The first step is to try to identify the determinants of the number of safe choices, and their difference along a gender perspective, by means of a regression analysis, whose results are reported in Table 9. The unconditional gender difference is of 0.32 safe choices, and is significant (column (1)). In column (2) we present our preferred specification. Gender differences barely change when relevant factors are controlled for.[17] On the other hand, we find that incentives matter. Subjects tend to be more risk averse when the incentives increase, although less so at the margin. We also find that the money illusion induced by inflating the experimental payoffs (given the same amount of money at stake) has no effect. In contrast, administering the lotteries in random order significantly increases the average number of safe choices on top of increasing the likelihood of observing an inconsistent behaviour, as observed in the previous section.

In column (3) of Table 9 we estimate the same specification but restrict the sample to the papers for which we have "full" detail. We perform this exercise for the sake of comparability with what will be shown in §4.3, where inconsistent subjects are also included in the analysis, requiring the availability of all their binary choices. Results barely change, and in particular the gender gap decreases only slightly at 0.29.

The panel dimension of our data set allows us to control for any observable and unobservable characteristic common to each replication. Column (4) of Table 9 reports the results of a fixed-effect specification. Females make on average 0.298 safe choices more than males, confirming by and large what is found in §4.1.

The results of this section show that in HL the choices of males and females are not identical. However, this difference can be detected in a significant way only when the statistical power of the test is high, and it is economically unimportant in terms of magnitude. This evidence is clearly different from what emerges, for instance, in the IG or in the EG task. Hence, evidence based on those two tasks only cannot be regarded as sufficient to attribute the different observed behaviour to actual differences in the underlying risk attitudes. The characteristics of the risk elicitation mechanism affect systematically the measured risk preferences, and do not simply add noise. Along the gender dimension the influence of the features of the task is so important to affect the behaviour at the aggregate level. The problem becomes then to disentangle the task versus underlying preferences conundrum. We do so in the next section exploiting all the information we have concerning the decision process, i.e., also including rationalizable mistakes in a structural model that includes a stochastic component.[18]

4.3. Structural Estimation with Maximum Likelihood

To identify possible gender differences in the shape of risk preference, we make use of a structural model,

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

[18] We exclude subjects making transparent errors, i.e., when a lower amount, for certain, is preferred to a higher one. Results are qualitatively similar including them, or excluding only the choices in which transparent errors are made.

which does not impose a deterministic choice between lotteries. We assume that the subject is an expected utility maximiser who can make an error (ε) in comparing the expected utility of the available lotteries. We employ a Fechner representation of stochastic decisions. In the case of binary choices, this means that the subject will choose the lottery displayed on the right whenever $EU_R > EU_L + \varepsilon$. The probability of choosing the lottery right (R) is therefore given by

$$\Pr(R) = \Pr(\varepsilon < EU_R - EU_L). \tag{1}$$

Different parametric assumptions on the shape of the error distribution identify different choice models. Previous research has shown that the error structure assumed has relevant effects on the estimated risk preferences (Blavatskyy and Pogrebna 2010, Stott 2006). Therefore, we will investigate gender differences in the estimated preferences using three different error specifications: probit, logit, and Luce.

The *probit* specification assumes that the error is normally distributed, as done, for instance, in Hey and Orme (1994). From the assumption that $\varepsilon \sim N(0, \mu^2)$, it follows that ε/μ follows a unit normal distribution. Therefore, the probability of choosing R can be evaluated using the cumulative distribution function of a standardised normal $\Phi()$:

$$\Pr(R) = \Phi\left(\frac{EU_R - EU_L}{\mu}\right), \tag{2}$$

whereas the probability of choosing left (L) is obviously the complement to one.

Assuming that the error follows a logistic distribution defines the *logit* specification, used for instance by Stott (2006). The underlying mechanism is the same as in probit, the only difference being that here the probability of choosing R is evaluated using the cumulative distribution function of a logistic distribution $\Lambda(0, \mu)$. The equivalent of Equation (2) in this case has an easy closed-form solution:

$$\Pr(R) = \frac{1}{1 + e^{-(1/\mu)(EU_R - EU_L)}}. \tag{3}$$

Another error specification has been proposed by Luce (1959) and is the same used by Holt and Laury (2002) in their original contribution.[19] The probability of choosing R becomes in this case[20]

$$\Pr(R) = \frac{EU_R^{1/\mu}}{EU_L^{1/\mu} + EU_R^{1/\mu}}. \tag{4}$$

[19] Detailed instructions as well as a script to implement this procedure can be found in Harrison and Rutström (2008).

[20] The Luce choice model can also be given a Fechner representation using the logarithms of expected utilities. In fact, Equation (4) can be derived starting from $\Pr(R) = \Pr(\ln(EU_R) - \ln(EU_L) < \varepsilon)$ and $\varepsilon \sim \Lambda(0, \mu)$.

It is easily shown that in both Luce and logit models, $\Pr(R)$ converges to $\frac{1}{2}$ as $\mu \to \infty$, and, as $\mu \to 0$, it goes to 1 when $EU_R > EU_L$ and to 0 when $EU_L < EU_R$.

We assume a flexible functional form of subjects' preferences, as represented by an expo-power utility function:

$$U(x) = \frac{1 - e^{-ax^{1-r}}}{a}, \tag{5}$$

which has the advantage of encompassing the two most common utility functions, i.e., the constant absolute risk aversion (CARA) (when $r = 1$) and the CRRA (when $a \to 0$).[21]

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

$$\text{LogLik} = \begin{cases} \ln 1 - \Pr(R) & \text{if choice is option A,} \\ \ln \Pr(R) & \text{if choice is option B,} \end{cases}$$

and then separately estimate for each paper and jointly over all the data set a structural model of choice using maximum likelihood and clustering standard errors by subject.[22] The model is estimated using the amounts in dollars of the HL table, and using all the papers characterized by "full" detail.[23]

We allow for heterogeneity by gender and we include as control a dummy for hypothetical versus real payoffs and a dummy for the random order of the choices for all the parameters (r, a, and μ). We control for the money illusion possibly induced by the experimental exchange rate on r and a, and for the possible effect of the number of choices on the noise parameter μ.

Results are shown in Table 10 and are in line with what are found in the previous section using regression analysis.

The different models display a rather consistent picture. First, the expo-power estimation results suggest that preferences are well represented by a CRRA utility function. In fact, the coefficient a of the expo power is never significantly different from zero, neither for males nor females (see the test of H_o: $a + a_{\text{female}} = 0$ reported below each estimate) for all the error specifications. A CARA specification is hence rejected by the data across all error structures.[24] Nevertheless,

[21] The coefficient of relative risk aversion of the expo power is equal to $r + a(1 - r)x^{1-r}$, which depends on income as long as $a \neq 0$.

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

[23] As a robustness check we also estimate the same models excluding from the sample the three papers with the highest stakes. The reason is that extremely high stakes may have a large impact on the curvature of the estimated function, in particular in the stochastic choice models that are not invariant to the scale of the stakes such as probit and logit. Results are similar to the ones reported here and available upon request.

[24] The CARA estimates are reported for completeness in Table A.1 in the appendix.

Table 10 Maximum-Likelihood Estimation

	Probit error		Logit error		Luce error	
	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.
Expo power						
r	0.477***	0.042	0.229***	0.045	0.399***	0.036
r_{female}	0.012	0.008	0.040	0.026	0.039*	0.021
a	−0.018	0.028	−0.013	0.010	0.002	0.020
a_{female}	−0.004	0.024	0.017**	0.007	−0.001	0.036
μ	0.624***	0.089	0.217***	0.022	0.337***	0.032
μ_{female}	−0.138***	0.048	−0.004	0.008	−0.047***	0.013
Log likelihood	−21,554.863		−20,121.776		−21,135.871	
Test $a + a_{female} = 0$	0.4641		0.6261		0.9684	
CRRA						
r	0.440***	0.026	0.396***	0.023	0.186***	0.029
r_{female}	0.039*	0.020	0.050***	0.015	0.090***	0.018
μ	1.142***	0.123	0.618***	0.069	0.218***	0.021
μ_{female}	−0.103**	0.044	−0.048*	0.025	−0.009	0.007
Log likelihood	−21,643.586		−21,192.097		−20,138.628	

Notes. Number of decisions = 48,965; number of subjects = 4,899; standard error clustered by subjects. Additional controls (not reported): incentivized vs. hypothetical stakes; choices presented in random order for all parameters; exchange rate between experimental currency unit and currency for parameter r and α; number of choices for parameter μ.

*, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

directly estimating the CRRA delivers different results for our main variable of interest. Although the range of the baseline r does not differ much, as it goes from 0.23 to 0.48 in the expo power and from 0.18 to 0.44 in the CRRA, gender differences appear to be stronger and more significant in the CRRA than in the expo power.

The noise coefficient μ is rather volatile across error specifications but appears to be quite high. For the sake of comparison, Dave et al. (2010) find a μ of about 0.08 using the Luce error specification, but this discrepancy is not very surprising given the much higher heterogeneity (in terms of experimental design, shape of the HL list, and stakes) in our data set. However, all the specifications consistently show that females display a significantly lower μ, providing some evidence against females' lower numeracy as a possible explanation of gender differences. In fact, if numeracy played a role, the lower understanding of the task, ceteris paribus, should have been reflected via higher confusion and more frequent decision errors by a higher μ_{female}.

The coefficients of the controls are not reported to enhance the readability of the table. A robust and not surprising result is that measured risk aversion is consistently and significantly higher when the experiment does not entail hypothetical rewards. Subjects' behaviour reacts in a qualitatively similar way to the artificial inflation of the payoffs obtained through the exchange rate of the experimental currency unit, although to a much lower extent. Presenting lotteries in random order appears to decrease risk aversion, except in the Luce specification. The controls of the noise coefficient μ are instead characterized by a remarkably erratic behaviour across error specifications and utility functions.

Although with some difference with respect to the linear specification of §4.2.2, the structural model seems to confirm that significant gender differences are detected in the HL task when merging all the observations. The reason is to be found in the skyrocketing increase of the statistical power of the test, which drives fairly close to zero the likelihood of observing a false negative when data are merged. The magnitude of the observed differences is, however, much lower in the HL with respect to other tasks, and the next section tries to provide some possible explanations, besides arguing that the differences in the HL task might be driven by probability weighing rather than risk aversion.

5. Gender Differences and the Characteristics of the Task

The analysis carried out in §4 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.

Higher Noisiness of HL. It has been argued (Dave et al. 2010, Charness and Viceisza 2016, among others) that HL is a relatively demanding task from a cognitive point of view. Being more difficult to understand than other methods, HL might elicit noisier signals.

The noise could then blur the evidence and lead to the observed lack of significant gender differences in small to medium samples.

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

First, we compare the signal to noise ratio (SNR) of the tasks, defined as the mean choice in each task divided by its standard deviation. If HL were noisier, it should display a lower SNR than the other tasks. This is not the case. The SNR in our data set of HL replications is equal to 3.34, higher than the average of the replications of the SNR of the IG (2.06) and the EG task (2.41).[25] These results are confirmed by a replication in a homogeneous subject pool performed by Crosetto and Filippin (2016), with SNR of 3.27 for HL, 2.67 for IG and 2.16 for EG.

Second, Crosetto and Filippin (2016) simulate with virtual agents (not affected by complexity and with a known distribution of risk preferences) the effect of the mechanics of the different tasks on the measured risk attitudes. If HL induces noisier choices, one should observe a sizeable discrepancy of the variance of choices in the human relative to the virtual subjects. This is not the case. The standard deviation of measured preferences in the simulations is similar to that obtained by consistent human subjects.

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

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

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

The HL task differs from IG and EG along all three dimensions. First, lotteries are generated changing probabilities over fixed outcomes. Second, HL measures preferences both in the risk-averse and in the risk-loving domain. Third, the choice set does not include a riskless alternative. The subject must incur some risks as the degenerate lottery in row number 10 of the original HL is played with 10% probability only. It can be argued that the role of the risk-free alternative might be played by the low amount of the safe lottery, that can be secured by always choosing option A (except in row 10). Whether such an amount can be considered as a risk-free alternative is disputable, but it is definitely less focal than in the other two elicitation methods. In fact, it is not directly shown to the subjects, it requires some elaboration to be identified, and its salience is likely diluted by the existence of multiple choices, which induce row-by-row comparisons.[26]

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

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

[25] We use data from Nelson (2015b) for the IG, and our computations for EG. Note that since we do not have the microdata of the replications of IG and EG, we cannot compute the SNR of the pooled samples. However, the distribution of the SNR of the individual replications of both IG and EG is significantly different than that of the HL replications (Mann-Whitney, $p < 0.001$).

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

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

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

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

Unfortunately, such an exercise cannot be considered a valid test because the safe-option papers differ from the standard HL in several other dimensions that are confounded with the availability of a safe option itself.[27] The number of options is higher (15) and limited to the range of probability $[0.3-1]$ of the good outcome to occur. Moreover, the fixed amount is usually lower than the expected value of the 50%–50% risky lottery, it is kept constant across all rows and is therefore different from the expected value of the corresponding option A in the classic HL.

Fixed Probabilities. The role of fixed probabilities is even harder to ascertain given the existing literature. A study by Bruner (2009) tests two different HL tables, one with changing stakes and one with changing probabilities, but unfortunately no information on gender is available. Another recent paper (Andersson et al. 2016) employs an Outcome Scale method without a safe option, effectively replicating a HL method with fixed probabilities, but finds significant gender differences in one of the two experiments of the study, and not in the other.

Probability Weighting. Probability weighting might in principle play a role in explaining the different outcomes across methods. The effects of basing decisions on perceived weights that differ from objective probabilities should be higher in HL, where probabilities vary, than in other tasks in which probabilities are fixed and equal to 50%–50%. If probability weighting led females to act in a less risk-averse way, HL would detect lower gender differences. This explanation, however, fails on empirical grounds since adding probability weighting makes gender differences disappear.

Given the shape of the probability weighting function by gender, females should appear *more* risk averse than males. Therefore, probability weighting could possibly account for the small gender differences found in the HL task, but not for the heterogeneity of results across methods. If anything, under probability weighting we should observe *more* gender differences in HL than in 50–50 tasks.

We run maximum-likelihood structural estimations as done in §4.3, adding a Lattimore et al. (1992) two-parameter probability weighting function. That is, we run maximum-likelihood estimations assuming CRRA, the three different error structures explored above, and we impose that subjects weigh probabilities according to the function

$$w(p) = \frac{\delta p^{\gamma}}{\delta p^{\gamma} + (1-p)^{\gamma}}.$$

Running the exact same specification used in Table 10 results in unusual shapes of the probability weighting function, and erratic estimates across error specifications. We hence add study dummies to the μ parameter to account for heterogeneity across HL replications. The results, detailed in Table 11, show that gender differences in risk attitudes disappear, and, depending on the specification, they appear in the parameter δ of the probability weighting function. The shape of the function is nonetheless rather heterogeneous across error structures.

[27] These 15 papers come from the same group of authors employing the same design (it is the case of, among others, Sheremeta 2010, Cason et al. 2010, Price and Sheremeta 2011).

Table 11 Probability Weighting Maximum-Likelihood Estimation

	Probit error		Logit error		Luce error	
CRRA with prob. weigh.	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.
r	0.429***	0.019	0.462***	0.025	0.220***	0.006
r_{female}	0.019	0.019	0.024	0.025	0.013	0.009
μ	0.917***	0.090	0.412***	0.045	0.115***	0.013
μ_{female}	0.021	0.033	0.014	0.017	−0.000	0.004
δ	0.516***	0.085	0.278***	0.039	0.496***	0.000
δ_{female}	0.140*	0.078	0.068**	0.034	0.001	0.000
γ	0.633***	0.035	0.584***	0.041	0.158***	0.013
γ_{female}	−0.037	0.049	−0.045	0.053	−0.010	0.017
Log likelihood	−19,104.488		−18,967.302		−18,910.492	

Notes. Number of decisions = 48,965; number of subjects = 4,899; standard error clustered by subjects. Additional controls (not reported): incentivized vs. hypothetical stakes; choices presented in random order for all parameters; exchange rate between ECU and currency for parameter r and α; number of choices and study dummies (37 studies) for parameter μ.

*, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Summing up, the results in the literature indicate that when a safe option is available, and when the tasks employ 50/50 lotteries changing the amounts at stake to generate variation in the expected values (IG, EG), 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 behaviour of males and females. More research and the development of *ad hoc* tests in a controlled environment are needed to shed more light on this 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 literature with several findings.

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

Second, we provide the largest analysis of HL replications to date. Since HL is usually employed as a companion task in experiments focusing on other topics, the number of papers directly reporting gender results is small relative to the number of replications. By gathering the original data from the authors, we built a data set of 54 published papers involving about 7,000 subjects and covering more than half of all the HL published replications. We found that gender differences appear in less than 10% of the published papers. This striking difference is neither due to a different average sample size, nor to a greater noise determined by the HL task.

The creation of a comparable data set of HL replications allows us to merge the data and reach several further goals. First, we can provide a reliable estimate of the typical results obtained with the HL task. The average unconditional number of safe choices maps into an Arrow–Pratt coefficient of risk aversion equal to $r = 0.36$. Inconsistent choices are commonly found and characterize, on average, 15.8% of the subjects. We find that females are more likely to display an inconsistent behaviour than males, but the choices of inconsistent subjects do not differ by gender. Second 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. In doing so, we indeed detect significant differences. Their magnitude is, however, economically unimportant; i.e., about one-sixth of a standard deviation, which is three times lower than what is found, for instance, in the Gneezy and Potters (1997) Investment Game or in the Eckel and Grossman (2002) lottery choice task. Third, a stochastic choice model estimated with maximum likelihood finds that preferences should be represented by a CRRA utility function for both males and females. This procedure also provides evidence against numeracy as a possible explanation of gender differences.

Heterogeneity in risk preferences across tasks and domains has already been observed in the literature. The main difference between our results and the

stated view in the literature is that we link the likelihood of observing gender differences with the features of the task used to elicit risk preferences. Our results suggest that not only do subjects react differently to different tasks, but males and females may be affected by different tasks in different ways. This is an interesting and novel result *per se* because it suggests that there is a structure behind the finding of gender differences in risk attitudes. At the same time, if the measured risk preferences depend on the elicitation task, it is natural to ask why this is the case and which task gets closer to the *true* value of risk preferences.

We do not provide a final answer to this question, but we draw a map of the features of the different tasks that might trigger different behaviour by gender. We can rule out that the observed gender pattern is due to the different domain of preferences (risk averse, risk loving) investigated by the risk elicitation methods. The characteristics that correlate with the emergence of gender differences are restricted to (a) the availability of a safe option among the set of alternatives and (b) the use of 50–50 lotteries that vary only in the amounts at stake. The first determinant is likely to trigger certainty effects, and it is known in the literature that safe options increase the likelihood of observing violations of the predictions of expected utility theory. The second factor prevents misperceptions of probabilities from playing a role, and we exclude that probability weighting might account for the heterogeneity of results across task. The absence of conclusive results makes further research necessary to properly identify when and why males are more risk tolerant than females.

Acknowledgments

The authors are grateful to the Max Planck Institute of Economics (Jena) for financial and logistic support and to Janna Heider for excellent research assistance. The authors thank all the authors that kindly contributed their data, the members of the Economics Science Association mailing list for useful references, Ainhoa Aparicio Fenoll, Tore Ellingsen, Andrea Ichino for useful suggestions, together with the participants, in 2013, to International Meeting on Experimental and Behavioral Economics, Madrid; Behavioral and Experimental Economics Lab workshop, Florence; Economics Science Association World, Zurich; Società Italiana degli Economisti Conference, Bologna; Max Planck Institute Autumn Workshop, Jena; in 2014, to Association Française d'Economie Expérimentale, Besançon; Foundations of Utility and Risk, Rotterdam; European Association of Labour Economists, Ljublijana; Workshop on Gender, Alicante; and to seminars in European University Institute Florence, Milan, Oxford, Göttingen, Prague, Padua and Kiel. All remaining errors are those of the authors.

Appendix

Table A.1 Complement to Table 10: CARA Estimations

	Probit error		Logit error		Luce error	
CARA	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.
α	0.347***	0.015	0.344***	0.015	0.006***	0.002
α_{female}	−0.007	0.013	−0.004	0.013	0.007***	0.002
μ	0.646***	0.053	0.618***	0.347	0.273***	0.026
μ_{female}	−0.012	0.024	−0.011	0.015	−0.004	0.007
Log likelihood	−24,231.339		−24,219.353		−21,612.557	

Notes. Number of decisions = 48,965; number of subjects = 4,899; standard error clustered by subjects.
***Denotes statistical significance at the 1% level.

Table A.2 Maximum-Likelihood Estimation by Paper—CRRA Utility, Consistent Subjects Only

		r		r_{female}	
Article	Mann-Whitney significant	Coeff.	p-value	Coeff.	p-value
Abdellaoui et al. (2011)	No	0.159	0.187	0.050	0.777
Andersen et al. (2008)	No	0.448	0.000	0.056	0.647
Barrera and Simpson (2012)	No	0.247	0.009	−0.005	0.969
Bauernschuster et al. (2010)	No	0.310	0.000	0.097	0.237
Bellemare and Shearer (2010)	No	−0.064	0.522	0.237	0.130
Brañas-Garza and Rustichini (2011)	No	0.008	0.952	0.119	0.524
Carlsson et al. (2012)	No	0.505	0.000	−0.193	0.133
Casari (2009)	No	0.333	0.000	0.107	0.248
Chakravarty et al. (2011)	No	0.592	0.000	0.093	0.687
Chen et al. (2013)	No	0.522	0.000	0.070	0.373
Cobo-Reyes and Jimenez (2012)	No	0.071	0.649	0.290	0.153

Table A.2 (Continued)

Article	Mann-Whitney significant	r		r_{female}	
		Coeff.	p-value	Coeff.	p-value
Dave et al. (2010)	**Yes**	0.522	0.000	0.131	**0.001**
Deck et al. (2012)	No	0.591	0.000	−0.159	0.332
Drichoutis and Koundouri (2012)	No	−0.038	0.885	0.339	0.318
Fiedler and Glöckner (2012)	No	0.645	0.000	0.304	0.135
Fiore et al. (2009)	No	0.573	0.000	−0.059	0.656
Glöckner and Hilbig (2012)	No	0.477	0.000	−0.113	0.308
Glöckner and Pachur (2012)	No	0.754	0.000	−0.066	0.704
Harrison et al. (2005)	No	0.337	0.000	0.069	0.285
Harrison et al. (2007)	No	0.603	0.050	−1.089	0.084
Harrison et al. (2013)	No	0.540	0.000	−0.041	0.791
Holt and Laury (2002)	No	0.488	0.000	0.110	0.108
Jamison et al. (2008)	**Yes**	0.376	0.000	0.230	**0.004**
Levy-Garboua et al. (2012)	No	0.464	0.000	−0.173	0.904
Lusk and Coble (2005)	No	0.472	0.000	−0.307	0.260
Masclet et al. (2009)	No	0.172	0.070	0.100	0.509
Mueller and Schwieren (2012)	No	0.326	0.000	0.033	0.692
Nieken and Schmitz (2012)	No	0.396	0.000	−0.021	0.769
Ponti and Carbone (2009)	No	0.303	0.213	0.295	0.553
Ryvkin (2011)	No	0.425	0.002	−0.118	0.577
Schram and Sonnemans (2011)	No	0.468	0.000	−0.075	0.329
Shafran (2010)	No	0.144	0.148	0.150	0.278
Slonim and Guillen (2010)	No	0.316	0.000	0.274	**0.029**
Sloof and van Praag (2010)	No	0.298	0.000	0.061	0.503
Szrek et al. (2012)	**Yes**	0.335	0.000	0.106	0.269
Wakolbinger and Haigner (2009)	**Yes**	0.329	0.000	0.153	0.118
Yechiam and Hochman (2013)	No	0.312	0.079	−0.065	0.765

Notes. Along the same lines as the estimations of §4.3, we estimated a CRRA with probit error and the specification included also the noise parameter μ and μ_{female} (not reported). Significance at 5% of gender differences emphasized in bold.

References

Abdellaoui M, Driouchi A, L'Haridon O (2011) Risk aversion elicitation: Reconciling tractability and bias minimization. *Theory Decision* 71(1):63–80.

Agnew JR, Anderson LR, Gerlach JR, Szykman LR (2008) Who chooses annuities? An experimental investigation of the role of gender, framing, and defaults. *Amer. Econom. Rev.* 98(2):418–422.

Anderson L, Freeborn B (2010) Varying the intensity of competition in a multiple prize rent seeking experiment. *Public Choice* 143(1):237–254.

Andersen S, Harrison GW, Lau MI, Rutström E (2006) Elicitation using multiple price list formats. *Experiment. Econom.* 9(4):383–405.

Andersen S, Harrison GW, Lau MI, Rutström EE (2008) Eliciting risk and time preferences. *Econometrica* 76(3):583–618.

Andersen S, Harrison GW, Lau MI, Rutström EE (2010) Preference heterogeneity in experiments: Comparing the field and laboratory. *J. Econom. Behav. Organ.* 73(2):209–224.

Andersson O, Holm HJ, Tyran J-R, Wengström E (2016) Risk aversion relates to cognitive ability: Preferences or noise? *J. Eur. Econom. Assoc.* Forthcoming.

Andreoni J, Sprenger C (2012) Risk preferences are not time preferences. *Amer. Econom. Rev.* 102(7):3357–3376.

Arya S, Eckel C, Wichman C (2013) Anatomy of the credit score. *J. Econom. Behav. Organ.* 95(November):175–185.

Baker RJ, Laury SK, Williams AW (2008) Comparing small-group and individual behavior in lottery-choice experiments. *Southern Econom. J.* 75(2):367–382.

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

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

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

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

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

Binswanger HP (1981) Attitudes toward risk: Theoretical implications of an experiment in rural india. *Econom. J.* 91(364): 867–890.

Blavatskyy PR, Pogrebna G (2010) Models of stochastic choice and decision theories: Why both are important for analyzing decisions. *J. Appl. Econometrics* 25(6):963–986.

Brañas-Garza P, Rustichini A (2011) Organizing effects of testosterone and economic behavior: Not just risk taking. *PLoS ONE* 6(12):e29842.

Bruner D (2009) Changing the probability versus changing the reward. *Experiment. Econom.* 12(4):367–385.

Byrnes JP, Miller DC, Schafer WD (1999) Gender differences in risk taking: A meta-analysis. *Psych. Bull.* 125(3):367–383.

Camerer CF (1992) Recent tests of generalizations of expected utility theory. Edwards W, ed. *Utility Theories: Measurements and Applications, Studies in Risk and Uncertainty* (Kluwer Academic Publishers, Boston), 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. *J. Econom. Behav. Organ.* 84(2):525–536.

Casari M (2009) Pre-commitment and flexibility in a time decision experiment. *J. Risk Uncertainty* 38(2):117–141.

Cason TN, Masters WA, Sheremeta RM (2010) Entry into winner-take-all and proportional-prize contests: An experimental study. *J. Public Econom.* 94(9–10):604–611.

Chakravarty S, Harrison GW, Haruvy EE, Rutström EE (2011) Are you risk averse over other people's money? *Southern Econom. J.* 77(4):901–913.

Charness G, Genicot G (2009) Informal risk sharing in an infinite-horizon experiment. *Econom. J.* 119(537):796–825.

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

Charness G, Gneezy U (2012) Strong Evidence for Gender Differences in risk taking. *J. Econom. Behav. Organ.* 83(1):50–58.

Charness G, Viceisza A (2016) Comprehension and risk elicitation in the field: Evidence from rural Senegal. *Rev. Behav. Econom.* Forthcoming.

Chen Y, Katuščák P, Ozdenoren E (2013) Why can't a woman bid more like a man? *Games Econom. Behav.* 77(1):181–213.

Cleave BL, Nikiforakis N, Slonim R (2013) Is there selection bias in laboratory experiments? *Experiment. Econom.* 16(3): 372–382.

Cobo-Reyes R, Jimenez N (2012) The dark side of friendship: "Envy." *Experiment. Econom.* 15(4):547–570.

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

Cohen M, Jaffray JY, Said T (1987) Experimental comparison of individual behavior under risk and under uncertainty for gains and for losses. *Organ. Behav. Human Decision Processes* 39(1):1–22.

Crosetto P, Filippin A (2013) The "bomb" risk elicitation task. *J. Risk Uncertainty* 47(1):31–65.

Crosetto P, Filippin A (2016) A theoretical and experimental appraisal of four risk elicitation methods. *Experment. Econom.* Forthcoming, doi: 10.1007/s10683-015-9457-9.

Crosetto P, Filippin A, Heider J (2015) A study of outcome reporting bias using gender differences in risk attitudes. *CESifo Econom. Stud.* 61(1):239–262.

Croson R, Gneezy U (2009) Gender differences in preferences. *J. Econom. Literature* 47(2):448–474.

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

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

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

Delnoij J, De Jaegher K, Rosenkranz S (2014) Understanding preferences for ascending auctions, Buy-It-Now auctions and fixed prices. Discussion Paper Series 14-02, Tjalling C. Koopmans Research Institute, Utrecht University School of Economics, Utrecht, Netherlands.

Dickinson DL (2009) The effects of beliefs versus risk attitude on bargaining outcomes. *Theory Decision* 66(1):69–101.

Dohmen T, Falk A (2011) Performance pay and multidimensional sorting: Productivity, preferences, and gender. *Amer. Econom. Rev.* 101(2):556–590.

Dohmen T, Falk A, Huffman D, Sunde U (2010) Are risk aversion and impatience related to cognitive ability? *Amer. Econom. Rev.* 100(3):1238–1260.

Dohmen T, Falk A, Huffman D, Sunde U, Schupp J, Wagner GG (2011) Individual risk attitudes: Measurement, determinants, and behavioral consequences. *J. Eur. Econom. Assoc.* 9(3):522–550.

Dreber A, Rand DG, Wernerfelt N, Garcia JR, Lum JK, Zeckhauser R (2011) Dopamine and risk choices in different domains: Findings among serious tournament bridge players. *J. Risk Uncertainty* 43(1):19–38.

Drichoutis AC, Koundouri P (2012) Estimating risk attitudes in conventional and artefactual lab experiments: The importance of the underlying assumptions. *Econom.—The Open-Access, Open-Assessment E-J.* 6(38):1–15.

Duersch P, Oechssler J, Vadovic R (2012) Sick pay provision in experimental labor markets. *Eur. Econom. Rev.* 56(1):1–19.

Eckel CC, Grossman PJ (2002) Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behav.* 23(4):281–295.

Eckel CC, Grossman PJ (2008a) Men, women and risk aversion: Experimental evidence. Plott CR, Smith VL, eds. *Handbook of Experimental Economics Results*, Vol. 1 (Elservier B.V., Amsterdam), 1061–1073.

Eckel CC, Grossman PJ (2008b) Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *J. Econom. Behav. Organ.* 68(1):1–17.

Eckel CC, Wilson RK (2004) Is trust a risky decision? *J. Econom. Behav. Organ.* 55(4):447–465.

Eckel CC, Wilson RK (2006) Internet cautions: Experimental games with Internet partners. *Experiment. Econom.* 9(1): 53–66.

Eckel CC, El-Gamal MA, Wilson RK (2009) Risk loving after the storm: A Bayesian-Network study of Hurricane Katrina evacuees. *J. Econom. Behav. Organ.* 69(2):110–124.

Ehmke M, Lusk J, Tyner W (2010) Multidimensional tests for economic behavior differences across cultures. *J. Socio-Econom.* 39(1):37–45.

Eriksen KW, Kvaløy O, Olsen TE (2011) Tournaments with prize-setting agents. *Scandinavian J. Econom.* 113(3): 729–753.

Ertac S, Gurdal MY (2012) Deciding to decide: Gender, leadership and risk-taking in groups. *J. Econom. Behav. Organ.* 83(1):24–30.

Falk A, Huffman D, Sunde U (2006) Self-confidence and search. Discussion paper, Institute for the Study of Labor, Bonn, Germany.

Fellner G, Sutter M (2009) Causes, consequences, and cures of myopic loss aversion—An experimental investigation. *Econom. J.* 119(537):900–916.

Fiedler S, Glöckner A (2012) The dynamics of decision making in risky choice: An eye-tracking analysis. *Frontiers in Psych.* 3(335), doi: 10.3389/fpsyg.2012.00335.

Fiore SM, Harrison GW, Hughes CE, Rutström EE (2009) Virtual experiments and environmental policy. *J. Environ. Econom. Management* 57(1):65–86.

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

Glöckner A, Pachur T (2012) Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition* 123(1):21–32.

Gneezy U, Potters J (1997) An experiment on risk taking and evaluation periods. *Quart. J. Econom.* 112(2):631–645.

Gneezy U, Leonard KL, List JA (2009) Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica* 77(5):1637–1664.

Gong B, Yang CL (2012) Gender differences in risk attitudes: Field experiments on the matrilineal Mosuo and the patriarchal Yi. *J. Econom. Behav. Organ.* 83(1):59–65.

Grijalva T, Berrens RP, Shaw WD (2011) Species preservation versus development: An experimental investigation under uncertainty. *Ecological Econom.* 70(5):995–1005.

Grossman PJ, Eckel CC (2015) Loving the longshot: Risk taking with skewed gambles. *J. Risk Uncertainty* 51(3):195–217.

Haigh MS, List JA (2005) Do professional traders exhibit myopic loss aversion? An experimental analysis. *J. Finance* 60(1):523–534.

Harless DW, Camerer CF (1994) The predictive utility of generalized expected utility theories. *Econometrica* 62(6): 1251–1289.

Harrison GW, Rutström EE (2008) Risk aversion in the laboratory. Cox JC, Harrison GW, eds. *Risk Aversion in Experiments, Research in Experimental Economics*, Vol. 12 (Emerald Group Publishing, Bingley, UK), 41–196.

Harrison GW, List JA, Towe C (2007) Naturally occurring preferences and exogenous laboratory experiments: A case study of risk aversion. *Econometrica* 75(2):433–458.

Harrison GW, Johnson E, McInnes MM, Rutström EE (2005) Risk aversion and incentive effects: Comment. *Amer. Econom. Rev.* 95(3):897–901.

Harrison GW, Lau MI, Rutström EE, Tarazona-Gómez M (2013) Preferences over social risk. *Oxford Econom. Papers* 65(1): 25–46.

He H, Martinsson P, Sutter M (2011) Group decision making under risk: An experiment with student couples. Working Paper 2011-27, University of Innsbruck, Innsbruck, Austria.

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

Holt CA, Laury SK (2002) Risk aversion and incentive effects. *Amer. Econom. Rev.* 92(5):1644–1655.

Holt CA, Laury SK (2014) Assessment and estimation of risk preferences. Machina M, Viscusi K, eds. *Handbook of the Economics of Risk and Uncertainty*, Vol. 1 (North-Holland, Amsterdam), 135–201.

Houser D, Schunk D, Winter J (2010) Distinguishing trust from risk: An anatomy of the investment game. *J. Econom. Behav. Organ.* 74(1–2):72–81.

Jacquemet N, Rullière JL, Vialle I (2008) Monitoring optimistic agents. *J. Econom. Psych.* 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. *J. Econom. Behav. Organ.* 68(3-4): 477–488.

Kocher MG, Pahlke J, Trautmann ST (2013a) *Tempus fugit*: Time pressure in risky decisions. *Management Sci.* 59(10): 2380–2391.

Kocher MG, Pogrebna G, Sutter M (2013b) Other-regarding preferences and management styles. *J. Econom. Behav. Organ.* 88:109–132.

Koszegi B, Rabin M (2007) Reference-dependent risk attitudes. *Amer. Econom. Rev.* 97(4):1047–1073.

Lange A, List JA, Price MK (2007a) A fundraising mechanism inspired by historical tontines: Theory and experimental evidence. *J. Public Econom.* 91(9):1750–1782.

Lange A, List JA, Price MK (2007b) Using lotteries to finance public goods: Theory and experimental evidence*. *Internat. Econom. Rev.* 48(3):901–927.

Langer T, Weber M (2004) Does binding or feedback influence myopic loss aversion? An experimental analysis. Mimeo.

Lattimore PK, Baker JR, Witte AD (1992) The influence of probability on risky choice: A parametric examination. *J. Econom. Behav. Organ.* 17(3):377–400.

Laury SK (2005) Pay one or pay all: Random selection of one choice for payment. Working paper, Andrew Young School of Policy Studies, Georgia State University, Atlanta.

Levy-Garboua L, Maafi H, Masclet D, Terracol A (2012) Risk aversion and framing effects. *Experiment. Econom.* 15(1): 128–144.

Luce RD (1959) *Individual Choice Behavior: A Theoretical Analysis* (Wiley, Hoboken, NJ).

Lusk JL, Coble KH (2005) Risk perceptions, risk preference, and acceptance of risky food. *Amer. J. Agricultural Econom.* 87(2):393–405.

Masatlioglu Y, Taylor S, Uler N (2012) Behavioral mechanism design: Evidence from the modified first-price auctions. *Rev. Econom. Design* 16(2):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. *J. Econom. Behav. Organ.* 70(3):470–484.

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? *J. Econom. Psych.* 33(3):448–460.

Nelson JA (2015a) Are women really more risk-averse than men? A re-analysis of the literature using expanded methods. *J. Econom. Surveys* 29(3):566–585.

Nelson JA (2015b) Not-so-strong evidence for gender differences in risk taking. *Feminist Econom.*, ePub ahead of print July 20, doi: 10.1080/13545701.2015.1057609.

Niederle M, Vesterlund L (2010) Explaining the gender gap in math test scores: The role of competition. *J. Econom. Perspectives* 24(2):129–144.

Nieken P, Schmitz PW (2012) Repeated moral hazard and contracts with memory: A laboratory experiment. *Games Econom. Behav.* 75(2):1000–1008.

Niemeyer C, Reiss JP, Sadrieh A (2013) Reducing risk in experimental games and individual choice. Technical report, Karlsruhe Institute of Technology, Karlsruhe, Germany.

Pogrebna G, Krantz D, Schade C, Keser C (2011) Words versus actions as a means to influence cooperation in social dilemma situations. *Theory Decision* 71(4):473–502.

Ponti G, Carbone E (2009) Positional learning with noise. *Res. Econom.* 63(4):225–241.

Price CR, Sheremeta RM (2011) Endowment effects in contests. *Econom. Lett.* 111(3):217–219.

Rosaz J (2012) Biased information and effort. *Econom. Inquiry* 50(2):484–501.

Rosaz J, Villeval MC (2012) Lies and biased evaluation: A real-effort experiment. *J. Econom. Behav. Organ.* 84(2): 537–549.

Ryvkin D (2011) Fatigue in dynamic tournaments. *J. Econom. Management Strategy* 20(4):1011–1041.

Sapienza P, Zingales L, Maestripieri D (2009) Gender differences in financial risk aversion and career choices are affected by testosterone. *Proc. Natl. Acad. Sci. USA* 106(36): 15268–15273.

Schipper BC (2012) Sex hormones and choice under risk. Working Paper 2012-07, University of California at Davis, Davis.

Schram A, Sonnemans J (2011) How individuals choose health insurance: An experimental analysis. *Eur. Econom. Rev.* 55(6):799–819.

Schubert R, Brown M, Gysler M, Brachinger H (1999) Financial decision-making: Are women really more risk-averse? *Amer. Econom. Rev.* 89(2):381–385.

Schunk D (2009) Behavioral heterogeneity in dynamic search situations: Theory and experimental evidence. *J. Econom. Dynam. Control* 33(9):1719–1738.

Shafran AP (2010) Interdependent security experiments. *Econom. Bull.* 30(3):1950–1962.

Sheremeta RM (2010) Experimental comparison of multi-stage and one-stage contests. *Games Econom. Behav.* 68(2):731–747.

Slonim R, Guillen P (2010) Gender selection discrimination: Evidence from a trust game. *J. Econom. Behav. Organ.* 76(2): 385–405.

Sloof R, van Praag CM (2010) The effect of noise in a performance measure on work motivation: A real effort laboratory experiment. *Labour Econom.* 17(5):751–765.

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

Stott H (2006) Cumulative prospect theory's functional menagerie. *J. Risk Uncertainty* 32(2):101–130.

Sutter M, Kocher MG, Glätzle-Rüetzler D, Trautmann ST (2013) Impatience and uncertainty: Experimental decisions predict adolescents' field behavior. *Amer. Econom. Rev.* 103(1): 510–531.

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.

Tversky A, Kahneman D (1992) Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertainty* 5(4):297–323.

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

Wakolbinger F, Haigner SD (2009) Peer advice in a tax-evasion experiment. *Econom. Bull.* 29(3):1653–1669.

Wieland A, Sarin R (2012) Gender differences in risk aversion: A theory of when and why. Working paper, UCLA Anderson School of Management, University of California, Los Angeles.

Wik M, Kebede TA, Bergland O, Holden S (2004) On the measurement of risk aversion from experimental data. *Appl. Econom.* 36(21):2443–2451.

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