CESifo Economic Studies, 2014, doi:10.1093/cesifo/ifu029

A Study of Outcome Reporting Bias Using Gender Differences in Risk Attitudes

Paolo Crosetto*, Antonio Filippin† and Janna Heider‡

*INRA, UMR 1215 GAEL, University of Grenoble, 38000 Grenoble, France.
e-mail: paolo.crosetto@grenoble.inra.fr
†Department of Economics, University of Milan, Via Conservatorio 7, 20122 Milano, Italy and Institute for the Study of Labor (IZA), Schaumburg-Lippe-Str. 5-9, 53113 Bonn, Germany. e-mail antonio.filippin@unimi.it
‡Max Planck Institute for Economics, Kahlaische Straße 10, 07745 Jena, Germany.
e-mail: janna.heider@gmail.com

Abstract

There is a strong consensus in the experimental literature according to which women are more risk averse than men. However, new evidence reveals that only a tiny fraction of the replications of the Holt and Laury (2002) risk elicitation task displays gender differences. This striking distance between the consensus and the data gathered with this elicitation method provides a clean test of the presence of an outcome reporting bias in the risk and gender literature. Exploiting a large data set of replications of Holt and Laury (2002), we find no evidence that the likelihood of reporting about gender differences is affected by obtaining results in line or against the consensus. Two variables significantly increase the probability of describing results along a gender dimension: the share of women among the authors and the fact that the study focuses directly on risk preferences. Both variables, however, are orthogonal to the results being in line with the consensus or not, confirming the absence of any outcome reporting bias. (JEL codes: C81; D81; J16).

Keywords: publication bias, gender difference, risk attitude

1 Introduction

The fact that published results may not be a representative sample of all scientific studies is something that has long been debated in the literature starting with Sterling (1959). This is a relevant issue in the scientific community, because as long as some contributions have a higher likelihood of being published than others, depending on the sign of the results, conclusions based on the review of the published literature will be biased.

The probability of (non)publication of research findings depending on the nature and direction of results can take different forms (Higgins and Green 2011). If the bias is introduced during the study, for example, when the authors decide to report only a part of the results obtained or to cancel the study on the face of results disagreeing with the initial hypotheses or with a consensus in the literature, it is known as ‘outcome reporting bias’. If instead the bias is introduced in the peer review process, for example, by editors and referees who tend to promote research that adheres to their preexisting views or tend to favor interesting, strong, counterintuitive

© The Author 2014. Published by Oxford University Press
on behalf of Ifo Institute, Munich. All rights reserved.
For permissions, please email: journals.permissions@oup.com

results, then it is known as ‘publication bias’. Moreover, a publication bias can lead to the formation of a (false) consensus, which may promote further cases of outcome reporting bias. The (false) consensus can also fuel a ‘location bias’, which occurs when results that disagree with the consensus are published in lower-ranked journals. Due to these biases, a (false) consensus can persist for a long time in the literature, while studies supporting different views are possibly abandoned by the authors, rejected by referees, relegated to modest journals.

The phenomena of outcome reporting and publication bias have mainly been investigated in the medical and pharmaceutical literature, both indirectly, using meta-analyses (Dwan et al. 2008, among others) and directly, using randomized experiments (Mahoney 1977). The presence of these biases is also documented in experimental psychology (Simmons et al. 2011, mainly about reporting biases) and in macroeconomics (De Long and Lang 1992).

From a methodological point of view, the existence of these biases can be investigated in several ways. Empirical approaches include counting the number of papers in a field reporting statistically significant results (Sterling et al. 1995), estimating the rate of false positives (De Long and Lang 1992), computing how many accepted abstracts get fully published after results are known (Scherer et al. 2007), or by following several scientific projects from the grant approval to eventual publication (cohort studies, for a review see Dwan et al. 2008). Biases have also been tested experimentally, for instance, by creating fake papers differing only in the significance of results and sending them to journals (Mahoney 1977). Finally, econometric techniques allow us to take into account missing studies in meta-analyses (Duval and Tweedie 2000) and to analyze and reduce the publication bias employing meta-regression approximations (Stanley and Doucouliagos 2014). Through the use of these techniques, several established results have been put into question, like the effect of an increased minimum wage on the employment rate (Doucouliagos and Stanley 2009) or the link between demand for health care and income (Costa-Font et al. 2011).

While it is possible to show whether the balance of the published results is biased, it is difficult to identify the underlying causes. The reason is that it is virtually impossible to observe the counterfactual situation, as non-published results cannot be observed. We avoid this drawback exploiting a data set that focuses on gender differences in risk preferences using the Holt and Laury (2002) risk elicitation method (henceforth HL). The data set has been built by Filippin and Crosetto (2014), collecting data directly from the authors. The data set covers a large fraction of the population of papers replicating the HL procedure in the lab or in the field. Crucially, the data set includes a larger set of individual studies than those who directly report about gender differences in the published version. It is hence possible to study whether the author's decision to report about gender differences in risk attitudes correlates with getting findings in line with the prevailing view in the risk and gender literature.

There is widespread consensus in the experimental economics literature about the fact that women are more prudent than men when confronted with decisions under risk. The consensus is strong and relies on surveys of laboratory studies (Eckel and Grossman 2008; Croson and Gneezy 2009; Bertrand 2011) as well as on large-scale survey results (Dohmen and Falk 2011; Dohmen et al. 2011). The basic result has proven to be robust along several dimensions such as the characteristics of the subject pool, the strength of incentives, the gain vs. loss domain, and the abstract vs. contextual framework. The strength of the consensus can be illustrated by an example, in which the authors wrote that they might have an 'atypical' subject pool since the 'usual' gender differences were not found (Anderson and Freeborn 2010). Moreover, gender differences appear to be rooted in firms' belief about gender stereotypes, with effects on the labor market (Campa et al. 2011) and on entrepreneurship (Caliendo et al. 2014).

Only recently it has been suggested that gender results might be task-specific (Crosetto and Filippin 2013). At the same time, the results of the HL risk elicitation method, by far the most widely used risk elicitation procedure in experimental economics, had not been comprehensively analyzed from a gender perspective. In Section 2 we fill this gap and show that using HL significant gender differences are the exception rather than the rule. The aforementioned large set of HL replications corroborates this finding, as shown by Filippin and Crosetto (2014), who also confirm that gender differences are instead ubiquitous in the literature when risk preferences are elicited with other tasks such as the Investment Game (Gneezy and Potters 1997) or the Ordered Lottery Selection task (Eckel and Grossman 2002, 2008).

A large body of experimental literature has been developed unaware of the fact that observing females more risk averse than males is much less likely using the HL as compared with other methods. As a result, many authors had to choose whether to report, or not, results that did not conform to the consensus.

Crucial to our research question, in most of the papers, the HL risk elicitation task is performed only as a control for experiments dealing with other topics (auctions, tournaments, trust, strategic behavior in games, etc.). As such, risk preferences in general and gender differences in particular constitute a by-product that must not necessarily be reported. As a consequence, the likelihood of being published depends only marginally, or not at all, on the results about gender differences in risk attitudes. The only (indirect) link remains the fact that presenting results that do not comply with the consensus could cast a shadow on the goodness of the findings of the entire work. Therefore, 'swimming upstream', that is, reporting results against the consensus view, implies a cost that is close to zero in terms of odds of getting published. As long as the likelihood of being published does not depend on the coherence of the outcomes with the consensus, this creates a unique opportunity to test for the existence of an outcome reporting bias. Restricting our attention to the papers which focus on risk preferences allows us to test whether results are robust to the presence of a possible publication bias. Therefore, our study mainly focuses on a specific kind of bias, but we believe it to be nonetheless important. In fact, the existence of an outcome reporting bias may significantly contribute to the formation and persistence of a (possibly false) consensus.

We find no significant evidence of an outcome reporting bias in the literature about gender differences in risk preferences. The existence of a strong consensus does not affect the likelihood of reporting results that are swimming upstream. A fixed effect specification shows that this finding is robust to possible idiosyncratic characteristics of each author considered separately. The only variables which significantly affect the likelihood of reporting gender differences are the relevance of risk attitudes for the research question of the study and the fraction of females in the group of authors. Note that this evidence does not imply a reporting bias, because the effect of the two variables does not change with the sign and significance of the gender differences observed.

The outline of the article is as follows. Section 2 summarizes the state of the art in the literature on gender differences in risk aversion and provides a survey of gender results obtained using the HL task. Section 3 describes the construction and contents of our data set. Section 4 reports the results in terms of outcome reporting bias, and Section 5 concludes.

2 Published Results About Gender Differences in Risk Attitudes: The Consensus and the HL

There is a vast consensus in the experimental economics literature on the existence of a gender difference in risk attitudes, with women being generally reported as more risk averse than men.[1] This consensus stems from surveys carried out over several different tasks (Eckel and Grossman 2008; Croson and Gneezy 2009) and from surveys in the field (Dohmen et al. 2011). More recently, recognizing that results obtained with different risk

[1] Henceforth, this is what we mean when simply referring to 'gender differences'.

elicitation methods are difficult to compare, Charness and Gneezy (2012) carried out a review of a single specific task, the Investment Game of Gneezy and Potters (1997). Their findings provide further (though not conclusive, see Nelson 2013) evidence for gender differences. Despite some papers challenging the extent and strength of the consensus (Nelson 2014), gender differences are by some considered as a stylized fact whose causes, rather than existence, should be investigated (Bertrand 2011; Pan and Houser 2011).

Despite the fact that HL is the most popular elicitation method in economics, its results have never been analyzed from a gender perspective. The HL task uses a multiple price list to elicit the risk preferences of subjects. In the HL task, subjects face a series of binary choices between pairs of lotteries. One lottery is safer (that is, with lower variance) than the other. The lottery pairs are ordered by increasing expected value. The set of possible outcomes is the same for every choice, and the increase in expected value across lottery pairs is obtained by increasing the probability of the 'good' outcome (see Table 1). For each row, subjects have to pick their preferred lottery. At the end of the experiment, one row is randomly chosen for payment, and the chosen lottery is played to determine the payoff.

The expected value of the risky lottery (Option B) starts lower than the one of the safe lottery (Option A), but it increases faster ending up higher. Subjects should therefore switch from the safe to the risky option as the probability of the good outcome increases. The switching point captures their degree of risk aversion. For instance, a risk-neutral subject should start with Option A, and switch to B from the fifth choice on. The higher the number of safe choices, the stronger the degree of risk aversion. Never choosing the risky option or switching 'back' from B to A is not an infrequent pattern. Subjects displaying such a behavior are usually regarded as

Table 1 The original Holt and Laury (2002) task

	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	0.2 €

inconsistent when modeling the choices without including a stochastic component.

There are only 21 papers (the original Holt and Laury (2002) and 20 replications) of the 528 papers citing Holt and Laury (2002) as of January 2013 which explicily report gender differences in their published version. Such a low number constitutes indirect evidence of the fact that in most of the cases the HL task is just used as a control for a potential confounding factor in an unrelated experiment. What emerges immediately from the literature is that using the HL task the gender consensus is far from confirmed. For starters, in the original Holt and Laury (2002) article, gender differences appear only in the low stake but not in the high stake treatment. Several replications in the past decade confirm that significant gender differences in HL are only rarely found. Of the 20 replications, only 3 report significant differences, 2 provide mixed evidence as in the original article, while 15 find that males and females display behavior that does not significantly differ. The details of these papers are reported in Table 2.

Table 2 includes, for each study, all the information that can be gathered by reading the paper. The Table includes the numerosity and the characteristics of the subject pool, whether the study was a laboratory or field experiment, and the country in which the experiment was run. It reports, whenever available, the average results by gender together with their significance (the *p*-value of the test or of the coefficient in a multivariate regression), specifying in any case the type of evidence reported in the paper to support the results about gender differences. The majority of papers uses students as subjects and ~~use~~ multivariate regressions to report the significance of their results.

The three papers which find a significant gender difference are Agnew et al. (2008), who use an unmodified low stake HL task, Dave et al. (2010), using the high stake version of the HL task, with outcomes scaled up by a factor of 20 with respect to Table 1, and Brañas-Garza and Rustichini (2011), who implement a non-incentivized version with only nine choices.

The two papers reporting mixed results find a significant effect only for a subsample, or only through one and not all of the statistical methods employed. In Chen et al. (2013), significant gender differences do not emerge in the unconditional distribution of choices in the HL task, but choices become significantly different (at 10%) when controlling for other observable characteristics (age, race, academic major, and number of siblings) in a multivariate framework. Menon and Perali (2009) on the other hand find females to be significantly more risk averse in one subsample and significantly less risk averse in another.

The list of papers in which the behavior of men and women does not differ significantly is longer, starting with the first replication of the

Table 2 Gender results as reported in the HL literature

Article	N	n_m	n_f	safe$_m$	safe$_f$	Significant gender difference	Lab/field	Type of subjects	Country	Type	p-value
Holt and Laury (2002)	212	–	–	–	–	Mixed	Lab	Students	U.S.	Text	–
Agnew et al. (2008)	845	400	445	–	–	Yes	Lab	Non-student	U.S.	t-test	< 0.05
Brañas-Garza and Rustichini (2011)	188	72	116	4.35	5.01	Yes	Lab	Students	Spain	Mann-Whitney	0.0027
Dave et al. (2010)	801	347	454	–	–	Yes	Lab	Labor force	Canada	Coefficient	0.001
Chen et al. (2013)	80	29	51	–	–	Mixed	Lab	Students	U.S.	Various	–
Menon and Perali (2009)	6496	2898	3598	–	–	Mixed	Field	High school and university students	Italy	Various	–
Andersen et al. (2006)	90	66	24	–	–	No	Lab	Students	Denmark	Coefficient	0.38
Anderson and Freeborn (2010)	140	74	66	–	–	No	Lab	Students	U.S.	Coefficient	0.54
Baker et al. (2008)	120	–	–	–	–	No	Lab	Students	U.S.	Coefficient	0.891
Carlsson et al. (2012)	213	105	108	5.79	5.19	No	Field	Rural population	China	Wilcoxon matched pairs	0.14
Chakravarty et al. (2011)	74	–	–	–	–	No	Lab	Students	India	Coefficient	0.644
Drichoutis and Koundouri (2012)	57	20	37	–	–	No	Lab	Students and general population	Greece	Coefficient	> 0.05
Eckel and Wilson (2004)	232	133	99	5.30	5.50	No	Lab	Students	U.S.	Coefficient	0.586
Ehmke et al. (2010)	345	170	175	5.26	5.58	No	Lab	Students	China, France, Niger, U.S.	Text	–
Harrison et al. (2005)	178	–	–	–	–	No	Lab	Students	U.S.	Text	–
Harrison et al. (2013)	108	76	32	–	–	No	Lab	Students	Colombia	Coefficient	0.78
Houser et al. (2010)	204	128	76	5.75	6.07	No	Lab	Students	Germany	Text	–
Masclet et al. (2009)	144	–	–	–	–	No	Lab	Students , employees, self-employed	France	Coefficient	0.19
Mueller and Schwieren (2012)	127	57	70	–	–	No	Lab	Students	Germany	Text	–
Ponti and Carbone (2009)	48	33	15	–	–	No	Lab	Students	Spain	Correlation	–
Viscusi et al. (2011)	144	–	–	–	–	No	Lab	Students	U.S.	Text	–

original task (Harrison et al. 2005). It includes Carlsson et al. (2012) in the field, and Eckel and Wilson (2004); Andersen et al. (2006); Masclet et al. (2009); Baker et al. (2008); Ponti and Carbone (2009); Anderson and Freeborn (2010); Ehmke et al. (2010); Houser et al. (2010); Chakravarty et al. (2011); Viscusi et al. (2011); Mueller and Schwieren (2012); Drichoutis and Koundouri (2012) and Harrison et al. (2013) in the lab.

This branch of the experimental literature provides a unique opportunity to explore the existence and extent of an outcome reporting bias. The consensus about gender differences in risk preferences is in fact not reflected by the results obtained using the HL risk elicitation method. This is due to the fact that the likelihood of observing gender differences strongly correlates with the characteristics of the method used to elicit preferences, but this is something that has been pointed out only recently (Crosetto and Filippin 2013) and was not known to the authors of the HL replications. In the next section, we show that by means of a large data set of HL replications it is possible to assess if and how much the presence of such a consensus affects the likelihood of reporting gender-related findings in a paper.

3 The Data Set of HL Replications

In this article we use the data set of HL replications collected by Filippin and Crosetto (2014). The description of the data set will here be limited to what is essential for the analysis of outcome reporting bias. We direct the interested reader to the original paper for the inclusion/exclusion criteria, the procedure followed to build the data set, and the full results.

To build the data set, the authors went trough the 528 papers in the Scopus bibliographic database citing Holt and Laury (2002) as of 31 January 2013. Of these papers, only 118, including the original Holt and Laury (2002), implement a sufficiently similar version of the HL task. The data set includes versions of the HL differing in the amounts at stake, the number of binary choices (from 6 to 20), the support of the probability spanned, and the step of change in the probability of the good outcome from one row to the next. In contrast, the data set excludes multiple price lists in which the amounts at stake increase with constant probabilities, as well as versions of HL in which the less risky lottery is substituted by a safe amount. The authors of all the 118 replications were contacted and asked to share their data or to report a predetermined set of summary statistics and tests. Of all the authors who replied, in 16 papers it turned out that the authors had not recorded gender or had a single-gender sample, while further 8 papers used the same data as another study in the data set and have been excluded to avoid duplication of results. The final data set

Table 3 Extent of the data set of HL replications

Published HL replications as of 31 January 2013	118	
of which:		
Not recording gender or using single gender subjects	16	
Duplicate data set	8	
Universe of reference	94	100%
of which:		
No response or not sharing the data	40	42.5%
Final data set	54	57.5%
of which:		
Microdata (shared or available online)	48	
Summary statistics	6	

The original Holt and Laury (2002) is included in the 118 replications.

covers 54 of the 94 remaining papers (see Table 3), with a coverage of about 57.5% of all published HL replications.[2]

The data set was built in order to provide a comprehensive analysis of gender differences in the HL task. Gathering the microdata proved vastly superior to a meta-analysis of published results, given the low reporting rate for gender findings in published articles as well as the variety of statistical approaches used. In fact, as shown in Table 2, comments about gender differences are not always accompanied by quantitative results. When reported, results sometimes are expressed using nonparametric tests of the average choices of males and females, while sometimes they take the form of coefficients in multivariate regressions. The data set reduces a large body of potentially heterogeneous literature to a common metric, and it also allows us to uniformly define and process inconsistent choices, which are another source of heterogeneity in the literature. The data set also keeps track of differences in the implementation of the task (number of choices, probability range spanned, stakes, real or hypothetical incentives, forced consistency or not). Most important for the aim of this article, the data set includes several studies that did not provide results by gender.

The data set of HL replications of Filippin and Crosetto (2014) confirms that findings with the HL method are not in line with the consensus about gender differences. In most of the studies, the behavior of males

[2] Since also among the remaining 40 papers some are likely to entail same-gender samples or missing gender data, the actual coverage can safely be regarded as higher than the reported 57.5%. Considering also the authors who replied to our request but whose data could not be used because the gender variable could not be exploited in the analysis, the overall response rate amounts to 63.6%.

and females does not significantly differ. Males are never found to be significantly more risk averse than females, while females are significantly more risk averse than males in only 5 of 54 papers. This proportion is even lower than the already weak and mixed evidence reported when looking at published results in Section 2. When pooling the data from the 54 papers, results show a comeback of significant gender differences due to the boost in statistical power, but the magnitude of the effect turns out to be economically unimportant. Differences amount to one-sixth of a standard deviation, less than a third of the effect found by other elicitation methods.[3] A similar picture emerges looking at the sign of the differences. In 40 of the 54 papers, females display a more risk averse average choice, while the opposite happens in 13 cases and in 1 case they are identical. A binomial test rejects ($p < 0.001$) that such a distribution comes from a population in which the two genders make the same choice on average. Nevertheless, the frequency of studies in which females make more risk averse choices is dramatically lower than what emerges in other tasks.

These statistics agree in showing that, using the HL task, gender differences can still be observed but with a clearly different incidence and magnitude. The crucial point, however, is that all the papers included in the data set were carried out with no knowledge of these results nor of the specific survey of the HL literature presented in Section 2 above. The state of the art in the literature was that gender differences were perceived as a systematic finding even dealing with small samples, like those typically encountered in a single experiment (about 100–150 subjects). In other words, most of the times, scholars replicating the HL task had to face results that did not conform to the consensus.

We exploit such an unknown task dependence of the results to test for the existence of an outcome reporting bias. The availability of the data for a sample of studies larger than those publishing their results along a gender dimension allows us to analyze also the decision of ‘not’ reporting results by gender. We can therefore reconstruct a counterfactual situation for the survey of published contributions eliciting risk attitudes with the HL task and reporting about gender differences. This makes it possible to study whether the author’s decision of reporting gender-specific results correlates with getting findings in line with the consensus.

4 Results

In this section, we will describe in detail the sample and the variables used for the analysis, present some implications of publication and outcome

[3] For details on the effect size and on significance tests see Filippin and Crosetto (2014).

reporting bias, and provide both a nonparametric and a multivariate analysis to test those implications.

4.1 Description of the sample

As described in Section 3, the data set is composed of 54 studies implementing the HL task. For the sake of clarity, the papers can be characterized along three relevant criteria.

Report

Along this dimension, we classify papers according to whether or not they report results about gender differences, be they significant or not. We classify as 'reporting' those papers which inform the reader about the breakdown by gender of the risk elicitation results and their statistical significance. This information can take several forms: a text comment, the results of a test or a coefficient in a multivariate regression. We can identify the 'report' variable for each of the 94 HL replications, including the 40 not present in the data set. Overall, 21 papers report and 73 do not.

Find

We then classify the papers based on whether or not they 'find' significant gender differences. The significance of gender results is computed applying a common method to each of the 54 papers in the data set. For each paper, we test nonparametrically if there is a significant difference in the unconditional distribution of the number of safe choices by gender. We restrict attention to consistent subjects, that is, participants characterized by a single switching point and not making dominated choices. Overall, applying this method, we find significant differences in only 5 of the 54 papers. The 40 papers for which we do not have the microdata cannot be classified along this dimension.

Focus

We categorize the importance given to the risk elicitation measure in each paper. We distinguish papers having risk elicitation as their 'main' focus from papers using the risk elicitation task to provide background data as a 'control'. Papers are classified as 'main' if they focus on measuring risk preferences directly for different subpopulations and/or in different contexts, or study the task itself or different versions of it, or contribute from a theoretical point of view to the understanding of decisions under risk (for instance, trying to disentangle risk aversion from loss aversion, or estimating the effect of the salience of the incentives). Papers are classified as 'control' if they focus on other topics, such as auctions, strategic games, tournaments, and use the HL procedure as a companion task in the

Table 4 Distribution of the HL replications according to the information reported and the results obtained

Report	Find			Total
	Yes	No	n.a.	
Full sample				
Significant difference	1	1	1	3
Not significant difference	0	16	2	18
Nothing	4	32	37	73
Total	5	49	40	94
Focus: main				
Significant difference	1	1	0	2
Not significant difference	0	11	1	12
Nothing	1	11	7	19
Total	2	23	8	33
Focus: control				
Significant difference	0	0	1	1
Not significant difference	0	5	1	6
Nothing	3	21	30	54
Total	3	26	32	61

experimental sessions to control for risk preferences. This is a rather heterogeneous class, but for the goal of this article it is characterized by a much looser focus on the HL task itself. We can identify the 'focus' variable for each of the 94 HL replications. Thirty-three papers have a 'main' focus on risk preferences, while 61 use HL as a 'control'.

Table 4 reports the results of this three-dimensional categorization of the papers. It classifies both the 54 papers of the Filippin and Crosetto (2014) data set and the 40 replications that fall outside the data set, because for these studies we could not access the microdata. Of those 40 papers, 3 report and 37 do not; 8 have HL as their main focus, and 32 use HL as control.

In order to build Table 4, a few cases reporting mixed results had to be reconsidered (see Table 2 above). First, Chen et al. (2013) report that gender differences emerge (at 10%) only when controlling for other observable characteristics, otherwise risk attitudes do not significantly differ between males and females. Since this is also what happens applying our common method, that is, testing the unconditional distribution of choices of consistent subjects, we classify this article as finding and reporting no gender differences. Menon and Perali (2009) find different results with females significantly more risk averse in one sample, significantly less

Table 5 Generic distribution of results

Report	Gender differences	
	Found	Not found
Significant difference	*a*	·
Not significant difference	·	*c*
Nothing	*b*	*d*

risk averse in another sample, and not significantly different from males in a third one. We do not have the microdata available for this article, and therefore, we cannot classify it according to our common metric. Looking at their published figures, though, we speculate that opposite results are quite likely to cancel out when merging the subsamples and delivering a not significant difference overall. Since Menon and Perali (2009) report all sorts of results, they clearly show no reporting bias. A similar argument applies to Holt and Laury (2002), who find significant gender differences in only one of their treatments. Therefore, we classify these entries as finding and reporting no gender differences, too.

A further remark is necessary for Brañas-Garza and Rustichini (2011) who publish significant gender differences, which do not emerge in our analysis of the microdata. Such a discrepancy arises because of the different way in which inconsistent choices are treated. Our analysis excludes inconsistent subjects, who are instead included by the authors. In most of the cases, inconsistent subjects are analyzed in the literature within structural models incorporating a random component, or excluded otherwise. This is our preferred approach, too, although the alternative approach of including inconsistent subjects and counting the number of safe choices is also often used (see for instance Ponti and Carbone (2009); Kocher et al. (2011); Dorschner and Musshoff (2012) among others.) As a result, this article ends up being classified as not finding gender differences but reporting them.[4]

4.2 Testable implications

We can use the data set to formally identify several testable implications about outcome reporting bias. We introduce some shorthand notation in Table 5 to help visualize the implications.

[4] The authors underline that their results in terms of gender differences are in line with the consensus in the literature. The inconsistency issue is not discussed.

4.2.1 *Absence of publication bias*

Since we deal only with published studies, we cannot directly test for the presence of publication bias. We can nonetheless proceed in an indirect way, testing whether it is frequent for papers not to report anything about gender differences, regardless of the significance of the underlying results. A finding in this sense would imply that reporting gender differences is not an important factor in getting published, irrespective of gender results. In terms of Table 5, this amounts to a test of the existence of a low reporting rate:

$$a + c < b + d. \tag{1}$$

4.2.2 *Report rates and paper focus*

The different prominence of risk attitudes in the research question of the papers provides a further testable implication. A significantly higher report rate for studies having a 'main' direct focus on risk preferences could signal a different (and possibly non-negligible) role that reporting gender differences can have in the likelihood of getting published in the two subgroups. We will hence test if the reporting rate is higher for 'main' rather than 'control' papers.

$$\left(\frac{a+c}{b+d}\right)_{main} > \left(\frac{a+c}{b+d}\right)_{control}. \tag{2}$$

4.2.3 *Outcome reporting bias*

In the presence of an outcome reporting bias, studies finding that females are more risk averse should be more likely to report the results. In contrast, null or opposite findings should be less likely to be reported, as authors prefer to amend their reports rather than signaling 'atypical' outcomes not in line with the consensus. Under the reasonable assumption that the likelihood of observing significant gender differences is ex ante the same, the testable implication is that the fraction of studies finding significant gender differences should be higher among those who report rather than among those who do not. The presence of an outcome reporting bias can be revealed by a Fisher exact test on the joint distribution of studies across the two dimensions of 'find' and 'report'. In particular, we formally test first for the pooled sample and then separately for the 'main' and 'control' subgroups whether:

$$\frac{a}{a+b} > \frac{c}{c+d}. \tag{3}$$

Note that had we relied upon the literature review, even abstracting away from problems related to the heterogeneity of methods used to deliver the results, we would have observed *a* and *c* only. Relying on the replications data set allows us to observe also *b* and *d*, which can be used to approximate the counterfactual situations of 'not' reporting conditioning on the results obtained. The counterfactual is only approximated, since we have 40 papers in the universe of HL replications that do not enter the data set. This notwithstanding, for the 54 studies in the data set the availability of the microdata allows us to observe the underlying latent variable about which no information has been published in 36 cases. For these 54 studies we can directly test the existence of outcome reporting bias, without relying upon bias reducing techniques.

4.3 Nonparametric tests

In this section we formally test the implications outlined above.

4.3.1 *Absence of publication bias*

The testable implication of Equation (1) can be investigated using the data of Table 4. It is immediate to notice that the overall report rate is low, as gender differences are explicitly reported in only 21 of 94 cases (22.3%). We interpret these results as indirect evidence for the absence of a publication bias, since not reporting results is widespread in the literature.

4.3.2 *Report rates and paper focus*

As noted in Section 4.2, it could be argued that 'main' and 'control' studies have a different likelihood of reporting results, given the different importance of risk preferences, and should be analyzed separately. We find this hypothesis to be supported by the data (see Table 6). The reporting rates for each subgroup are different: 14 of 33 (42.4%) for 'main' and 7 of 61 (11.4%) for 'control' papers. The likelihood of reporting gender differences strongly correlates with the importance of risk attitudes in the paper, and the difference is statistically significant according to a one-sided Fisher exact test ($p = 0.001$). In light of this evidence, we will also test the presence of outcome reporting bias in the two subsamples separately.

Table 6 Distribution of papers according to the importance of risk attitudes

	Role of risk attitudes	
	Main	Control
Report about gender differences	14	7
Do not report	19	54

4.3.3 *Outcome reporting bias*

We find no evidence of outcome reporting bias in the pooled sample. Looking at the top panel of Table 4, it can be shown that 3 of the 21 studies reporting results find significant gender differences. Using as the counterfactual the 36 studies that do not report but of which we have the data, we see that 4 display significant gender differences while 32 do not. The two fractions are sightly different (14.3% versus 11.1%, respectively), but not significantly so according to a one-sided Fisher exact test ($p = 0.701$). Note that in this comparison we also included the three papers for which we have only published information and no microdata. Results do not change, though, if we limit the analysis to the 54 studies present in the data set: fractions become identical at 11.1%.

4.3.4 *Outcome reporting bias and paper focus*

While not detected with aggregate results, the outcome reporting bias could still exist, depending on the relative importance of risk preferences in the research question of the paper, although the direction is not evident a priori. On the one hand, the cost of displaying results against the consensus could be higher among 'main' papers. On the other hand, providing incomplete information could have a negative impact per se, regardless of the underlying results. In any case, the data (see Table 4, bottom panels) show no evidence of outcome reporting bias in neither of the 'main' and 'control' subgroup. Among 'main' papers reporting, 2 of 14 studies find significant gender differences. Using the 12 studies that do not report (1 with significant gender differences, 11 without) as the counterfactual, we see that both fractions are low (13.3 vs. 8.3%) and not significantly different ($p = 0.586$).

Among 'control' papers, for which publication bias can be ruled out but outcome reporting bias could be present, frequencies are more differentiated. Significant gender differences emerge in 28.6% of the reporting papers (2 of 7), while among those that can be used as a counterfactual the percentage is equal to 12.5% (3 of 24). Also in this case, however, a Fisher test cannot reject that the two frequencies are the same ($p = 0.312$).

Note that the likelihood of reporting significant differences could even be higher among papers using HL as a control without this having necessarily to do with any consensus. Studies focusing on other issues could simply set a threshold of significance as a screening device for the inclusion of noncore results. In this case, what we measure with the test would therefore be an upper bound of the effect of an outcome reporting bias. Detecting null results for the upper bound makes it unnecessary, however, to identify this additional effect.

In principle, the bias could extend to the likelihood of sharing the data. In other words, the fear of going against the consensus could imply that data are not missing at random in the Filippin and Crosetto (2014) data set and therefore that the likelihood of not finding gender differences is even higher among the 37 studies about which there is no information available. There is no way we can check the distribution of significant vs. not significant gender differences within this subsample. However, it would be enough that only one of these 37 papers actually found significant gender differences to generate results that would not significantly differ from the ones found with the data we have.

The nonparametric analysis carried out above allows us to conclude that there is no evidence of outcome reporting bias for gender differences in risk preferences. This result is mainly due to the overwhelming rate of non-reporting papers present across the board in the literature.

4.4 Multivariate analysis

In this subsection, we jointly analyze the determinants of the likelihood of reporting gender differences using a multivariate approach. In this framework, the outcome reporting bias would take the form of a significant increase of the probability of reporting driven by the fact that gender differences have been observed ('find') ceteris paribus. Among the control variables we first include the distinction between 'main' and 'control' that has already been shown to affect the probability of reporting. Second, we check whether the gender of the authors has an effect on the likelihood of reporting. To do so, we built a variable 'wshare', containing the share of women among the authors of each paper. The variable takes value 1 for studies authored by women only, value 0 for studies authored by men only, and values in-between for all other cases.[5]

We exploit the multivariate framework not only to control for these additional factors, but also to test their interaction with the results about gender differences. In this way we disentangle the pure effect on the likelihood of reporting any result from the presence of an outcome reporting bias.

Table 7 reports the results of two linear regression models in which the dependent variable is a dummy taking value 1 if results were reported, 0 otherwise. Model 1 is a simple linear regression model estimating the following equation:

$$\text{report} = \beta_0 + \beta_1\text{find} + \beta_2\text{main} + \beta_3\text{find} \cdot \text{main} + \beta_4\text{wshare} + \beta_5\text{find} \cdot \text{wshare} + \varepsilon. \quad (4)$$

[5] We also built a dummy version of the variable, taking value 1 if there is at least one woman among the coauthors, 0 otherwise, but results are unaffected.

Table 7 Determinants of the likelihood of reporting results

	Model 1		Model 2	
	Report		Report	
	Coefficient	t	Coefficient	t
Constant	0.027	(0.25)	−0.044	(−0.33)
Find	−0.034	(−0.11)	0.066	(0.19)
Main	0.343***	(2.75)	0.393**	(2.34)
main·find	0.093	(0.26)	−0.0016	(−0.00)
wshare	0.621***	(2.93)	0.918***	(3.25)
wshare·find	0.327	(0.50)	−0.227	(−0.27)
author fixed effects	No		Yes	
N	57		57	

t statistics in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

$N = 57$ includes also three studies for which enough information was reported in the article.

The results show that both the share of women among the authors and the fact that risk attitudes are among the main goals of the paper increase the likelihood of reporting the data. In contrast, there is no evidence of any outcome reporting bias. The interaction terms capture the possibility that the outcome reporting bias could concern only the studies having risk attitudes as their main focus or a high fraction of females among the authors, as argued above. The fact that the coefficients do not significantly differ from zero confirms that this is not the case. While the presence of women among the authors increases substantially the attention given to gender issues, women are not affected by outcome reporting bias. The same argument applies for papers having risk attitudes as their main focus.

Observations are not necessarily independent in our data set, because some authors contribute to more than one paper. To take this into account, we also run a fixed effects specification (Model 2) in which we partial out the effect of authors' observable and unobservable individual characteristics. Note that the share of females can be included in this specification as well, because it is a nonlinear function of the fixed effects. The results are qualitatively similar and the absence of any outcome reporting bias is confirmed.

5 Discussion and Conclusions

It has long been argued that published results may not be a representative sample of all scientific studies, as long as the likelihood of reporting the results and of being published is a function of the results obtained. This article exploits a large data set of replications of the HL to provide a clean test of the presence of an outcome reporting bias, that is, a different likelihood of reporting results in favor or against a well-established consensus.

There is a widespread and strong consensus that women are more prudent than men when dealing with risky choices. However, only recently it has been shown that this finding is task-dependent, and that using the HL procedure gender differences are the exception rather than the rule. In the meantime, many authors had to face the choice between reporting and not reporting results that did not conform to the consensus.

The data set of HL replications contains the results of a larger set of individual studies than those which directly report gender differences in the published version. Observing the gender results of a subsample of studies that do not explicitly report them allows us to study whether the author's decision to include the results about gender differences in risk attitudes correlates with getting findings in line with the well-established consensus.

We find no significant evidence of an outcome reporting bias in the literature about gender differences in risk preferences. The existence of a strong consensus does not affect the likelihood of reporting results that are swimming upstream. This result is robust to any idiosyncratic characteristic of each author considered separately, as confirmed by a fixed effect specification.

When the HL risk elicitation method is performed only as a control in experiments dealing with other topics, whether or not authors report gender differences does not affect the likelihood of being published. For this subsample of studies, it is safe to assume that a 'publication bias' plays no role and therefore that our results striclty refer to the absence of an 'outcome reporting bias'.

Two variables, however, significantly impact the likelihood of reporting gender differences: the relevance of risk attitudes in the research question of the study, and the fraction of women among the group of authors. In both cases, though, we find no evidence of outcome reporting bias. For papers aimed solely at analyzing risk preferences, being in line or not with the consensus could be slightly more relevant, possibly adding some external incentives to the mere attraction exerted by the consensus itself. However, we find that the increase in the probability of reporting is orthogonal to the results obtained in terms of gender differences and therefore no outcome reporting bias emerges. A similar argument applies to the fraction of women among the group of authors, which also positively correlates with the probability of reporting the results by gender, regardless of the direction of the findings.

The results of this article should be taken with a grain of salt, though, because they refer to a specific topic and cannot be easily generalized. The external validity of our exercise is somewhat limited and more evidence is necessary before we can extend the conclusion to the whole discipline. However, our results are based on a large and reliable data set, gathered from dozens of studies involving altogether more than 100 authors adopting the most widely used risk elicitation task in the literature. To the extent that this data set is representative of the practices adopted in experimental economics at large and in other disciplines, the insight that can be derived goes beyond the specific subfield from which the data have been gathered and is definitely good news.

References

Agnew, J. R., L. R. Anderson, J. R. Gerlach and L. R. Szykman (2008), "Who Chooses Annuities? An Experimental Investigation of the Role of Gender, Framing, and Defaults", *The American Economic Review* **98**, 418–22.

Andersen, S., G. Harrison, M. Lau and E. Rutström (2006), "Elicitation Using Multiple Price List Formats", *Experimental Economics* **9**, 383–405.

Anderson, L. and B. Freeborn (2010), "Varying the Intensity of Competition in a Multiple Prize Rent Seeking Experiment", *Public Choice* **143**, 237–54.

Baker, R. J., S. K. Laury and A. W. Williams (2008), "Comparing Small-Group and Individual Behavior in Lottery-Choice Experiments", *Southern Economic Journal* **75**, 367–82.

Bertrand, M. (2011), "New Perspectives on Gender, vol. 4B, 1st edn", in C. David and A. Orley, eds., in *Handbook of Labor Economics*, Elsevier, Ch. 17, pp. 1543–90.

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

Caliendo, M., F. M. Fossen, A. Kritikos and M. Wetter (2014), "The Gender Gap in Entrepreneurship: Not Just a Matter of Personality", *CESifo Economic Studies* **60**, http://cesifo.oxfordjournals.org/content/early/2014/06/19/cesifo.ifu023.abstract.

Campa, P., A. Casarico and P. Profeta (2011), "Gender Culture and Gender Gap in Employment", *CESifo Economic Studies* **57**, 156–82.

Carlsson, F., H. He, P. Martinsson, P. Qin and M. Sutter (2012), "Household Decision Making in Rural China: Using Experiments to

Estimate the Influences of Spouses”, *Journal of Economic Behavior and Organization* **84**, 525–36.

Chakravarty, S., G. W. Harrison, E. E. Haruvy and E. E. Rutström (2011), “Are You Risk Averse Over Other People’s Money?”, *Southern Economic Journal* **77**, 901–13.

Charness, G. and U. Gneezy (2012), “Strong Evidence for Gender Differences in Risk Taking”, *Journal of Economic Behavior and Organization* **83**, 50–8.

Chen, Y., P. Katuščák and E. Ozdenoren (2013), “Why Can’t a Woman Bid more Like a Man?”, *Games and Economic Behaviour* **77**, 181–213.

Costa-Font, J., M. Gemmill and G. Rubert (2011), “Biases in the Healthcare Luxury Good Hypothesis?: A Meta-Regression Analysis”, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **174**, 95–107.

Crosetto, P. and A. Filippin (2013, February), “A Theoretical and Experimental Appraisal of Five Risk Elicitation Methods”, Jena Economic Research Papers 2013-009, Friedrich-Schiller-University Jena, Max-Planck-Institute of Economics.

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

Dave, C., C. Eckel, C. Johnson and C. Rojas (2010), “Eliciting Risk Preferences: When is Simple Better?”, *Journal of Risk and Uncertainty* **41**, 219–43.

De Long, J. B. and K. Lang (1992, December), “Are all Economic Hypotheses False?”, *Journal of Political Economy* **100**, 1257–72.

Dohmen, T. and A. Falk (2011, September), “Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender”, *American Economic Review* **101**, 556–90.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp and G. G. Wagner (2011), “Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences”, *Journal of the European Economic Association* **9**, 522–50.

Dorschner, T. and O. Musshoff (2012), “Cost-Oriented Evaluation of Ecosystem Services Under The Consideration of Entrepreneurial Risks”, *Journal of the Austrian Society of Agricultural Economics* **127**, 249–54.

Doucouliagos, H. and T. D. Stanley (2009), “Publication Selection Bias in Minimum-Wage Research? A Meta-Regression Analysis”, *British Journal of Industrial Relations* **47**, 406–28.

Drichoutis, A. C. and P. Koundouri (2012), "Estimating Risk Attitudes in Conventional and Artefactual Lab Experiments: The Importance of the Underlying Assumptions", *Economics - The Open-Access, Open-Assessment E-Journal* **6**, 1–15.

Duval, S. and R. Tweedie (2000), "A Nonparametric "Trim and Fill" Method of Accounting for Publication Bias in Meta-Analysis", *Journal of the American Statistical Association* **95**, 89–98.

Dwan, K., D. G. Altman, J. A. Arnaiz, J. Bloom, A. W. Chan, E. Cronin, E. Decullier, P. J. Easterbrook, E. Von Elm, C. Gamble, D. Ghersi, J. P. A. Ioannidis, J. Simes and P. R. Williamson (2008), "Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias", *PLoS One* **3**, e3081.

Eckel, C. C. and P. J. Grossman (2002), "Sex Differences and Statistical Stereotyping in Attitudes Toward Financial Risk", *Evolution and Human Behavior* **23**, 281–95.

Eckel, C. C. and P. J. Grossman (2008), "Men, Women and Risk Aversion: Experimental Evidence, Vol. 1", in *Handbook of Experimental Economics Results*, Elsevier, Ch. 113, pp. 1061–73.

Eckel, C. C. and R. K. Wilson (2004), "Is Trust a Risky Decision?", *Journal of Economic Behavior and Organization* **55**, 447–65.

Ehmke, M., J. Lusk and W. W. Tyner (2010), "Multidimensional Tests for Economic Behavior Differences Across Cultures", *The Journal of Socio-Economics* **39**, 37–45.

Filippin, A. and P. Crosetto (2014), "A Reconsideration of Gender Differences in Risk Attitudes", Working Papers 2014-01, Institute for the Study of Labor (IZA). http://ideas.repec.org/p/iza/izadps/dp8184.html.

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

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

Harrison, G. W., M. I. Lau, E. E. Rutstrm and M. M. Tarazona-Gmez (2013), "Preferences Over Social Risk", *Oxford Economic Papers* **65**, 25–46.

Higgins, J. and S. Green (2011, March), "Cochrane Handbook for Systematic Reviews of Interventions", The Cochrane Collaboration, version 5.1.0, www.cochrane-handbook.org.

Holt, C. and S. Laury (2002), "Risk Aversion and Incentive Effects", *American Economic Review* **92**, 1644–55.

Houser, D., D. Schunk and J. Winter (2010), “Distinguishing Trust from Risk: An Anatomy of the Investment Game”, *Journal of Economic Behavior and Organization* **74**, 72–81.

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

Mahoney, M. (1977), “Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System”, *Cognitive Therapy and Research* **1**, 161–75.

Masclet, D., N. Colombier, L. Denant-Boemont and Y. Lohéac (2009), “Group and Individual Risk Preferences: A Lottery-Choice Experiment With Self-Employed and Salaried Workers”, *Journal of Economic Behavior and Organization* **70**, 470–84.

Menon, M. and F. Perali (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**, 593–630.

Mueller, J. and C. Schwieren (2012), “Can Personality Explain What is Underlying Women’s Unwillingness to Compete?”, *Journal of Economic Psychology* **33**, 448–60.

Nelson, J. (2014), “Are Women Really more Risk-Averse than Men? A Re-Analysis of the Literature Using Expanded Methods”, *Journal of Economic Surveys*, doi: 10.1111/joes.12069.

Nelson, J. A. (2013, November), “Not-so-strong Evidence for Gender Differences in Risk Taking”, Working Papers 19, Economics Department, University of Massachusetts Boston. http://econpapers.repec.org/paper/mabwpaper/2013_5f06.htm.

Pan, X. and D. Houser (2011), “Mating Strategies and Gender Differences in Pro-sociality: Theory and Evidence”, *CESifo Economic Studies* **57**, 653–82, http://cesifo.oxfordjournals.org/content/57/4/653.abstract.

Ponti, G. and E. Carbone (2009), “Positional Learning with Noise”, *Research in Economics* **63**, 225–41.

Scherer, R. W., P. Langenberg and E. von Elm (2007), “Full Publication of Results Initially Presented in Abstracts”, *Cochrane Database of Systematic Reviews* **2**, MR000005.

Simmons, J. P., L. D. Nelson and U. Simonsohn (2011), “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant”, *Psychological Science* **22**, 1359–66.

Stanley, T. D. and H. Doucouliagos (2014), "Meta-Regression Approximations to Reduce Publication Selection Bias", *Research Synthesis Methods* **5**, 60–78.

Sterling, T. (1959), "Publication Decision and the Possible Effects on Inferences Drawn from Tests of Significance-or vice versa", *Journal of the American Statistical Association* **54**, 30–4.

Sterling, T. D., W. L. Rosenbaum and J. J. Weinkam (1995), "Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and vice versa", *The American Statistician* **49**, 108–12.

Viscusi, W., O. Phillips and S. Kroll (2011), "Risky Investment Decisions: How are Individuals Influenced by their Groups?", *Journal of Risk and Uncertainty* **43**, 81–106.