The international platform of Ludwig-Maximilians University's Center for Economic Studies and the Ifo Institute

CESifo Working Papers
www.cesifo.org/wp

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

Paolo Crosetto
Antonio Filippin
Janna Heider

CESIFO WORKING PAPER NO. 4466
CATEGORY 13: BEHAVIOURAL ECONOMICS
NOVEMBER 2013

CESifo
Center for Economic Studies & Ifo Institute

CESifo *Working Paper No. 4466*

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

Abstract

This paper exploits a large dataset of replications of the Holt and Laury (2002) risk elicitation task to study a possible outcome reporting bias using gender differences in risk attitudes. There is a strong consensus view in the experimental literature according to which women are more prudent than men in risky choices. The evidence collected in the dataset, however, does not support the consensus: only a tiny fraction of the replications displays gender differences. This striking distance between the consensus and the data gathered with this elicitation task allows us to test directly for the presence of outcome reporting bias in the risk and gender literature. We find no evidence that the likelihood of reporting about gender differences is affected by obtaining results in line or against the consensus, also controlling for authors fixed effects. The vast majority of the studies does not report gender results. The only significant determinant of the probability of reporting is the fact that the study focuses directly on the analysis of risk preferences.

JEL-Code: C810, D810, J160.

Keywords: publication bias, gender difference, risk attitude.

Paolo Crosetto
INRA, UMR 1215 GAEL
University of Grenoble
France – 38000 Grenoble
paolo.crosetto@gmail.com

Antonio Filippin
University of Milan
Department of Economics
Via Conservatorio 7
Italy – 20122 Milan
antonio.filippin@unimi.it

Janna Heider
Max Planck Institute for Economics
Kahlaische Straße 10
Germany – 07745 Jena
janna.heider@gmail.com

November 4, 2013
We are grateful to the Max Planck Institute of Economics in Jena for financial and logistic support. We would like to thank all the authors that kindly contributed their data, the members of the ESA mailing list for useful references, and the participants to the IMEBE 2013 Conference in Madrid, the 2013 BEELAB workshop in Florence, the 2013 CESifo workshop in Venice, and the 2013 SIE conference in Bologna for useful suggestions. All remaining errors are ours.

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 from 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 according to the interest or the significance of the results, the conclusions about the underlying phenomenon based on the review of literature will be biased.

The probability of (non)publication of research findings according to the nature and direction of the results can take different forms according to what exactly causes it (Higgins and Green, 2011). If the bias is introduced during the study, e.g., if 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 view in the literature, it is known as *outcome reporting bias*. If instead the bias is introduced at the moment of the peer review and before publication, e.g., if editors and referees tend to promote research that adheres to their preexisting views or tend to favor interesting, strong, counterintuitive results, then it is known as *publication bias*. Moreover, publication bias can lead to the formation of a (false) consensus that can later result in more cases of outcome reporting bias. The (false) consensus can also fuel *location bias*, i.e., the fact that results disagreeing with the consensus tend to be published in lower-ranked journals. Due to these biases, false results can persist for a long time in the literature, while studies that are not compliant with the (false) consensus are 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).

Methodologically, the research carried out on the topic has focused on proving that these biases exist. This can be done empirically, for instance by counting the number of papers in a field or subfield reporting statistically significant outcomes for the studied effects (Sterling et al., 1995), estimating the rate of false positive in economic journals (De Long and Lang, 1992), measuring how many accepted abstracts get fully published after results are known (Scherer et al., 2007), or following several scientific projects from the grant approval to eventual publication (cohort studies, for a review see Dwan et al., 2008). It has 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 exist to take into account missing studies in a meta-analysis (Duval and Tweedie, 2000) or to analyze and reduce the publication bias employing meta-regression approximations (Stanley and Doucouliagos, 2013). Through the use of these techniques, several known results in economics have been put into question, for instance the effect of a raise of the minimum wage on employment (Doucouliagos and Stanley, 2009) or the link between demand for health care and income (Costa-Font et al., 2011).

While it is easy to prove that those biases exist, it is nonetheless very difficult to disentangle the source of the bias, since it is virtually impossible to observe the counterfactual. This

paper exploits a unique dataset perfectly suited to study the influence of a well-established consensus about gender differences in risk preferences on *outcome reporting bias*, i.e., the likelihood of reporting results in favor or against such a consensus. As explained below, the dataset allows us to disregard the *publication bias*, and to focus on the mere attraction exerted by the consensus itself. Even if this study focuses solely on outcome reporting bias, we believe it to be nonetheless important: the existence of reporting bias could in fact contribute to reinforce the scientific consensus, to the point of possibly generating a false consensus based on self-confirming beliefs.

There is widespread consensus in the experimental economics literature on the existence of gender differences in risk attitudes, with women portrayed as more prudent than men when confronted with decisions under risk. The consensus is strong. It relies on surveys of laboratory studies (Bertrand, 2011; Croson and Gneezy, 2009; Eckel and Grossman, 2008) and on large scale questionnaire results (Dohmen and Falk, 2011; Dohmen et al., 2011). The result has proven to be robust along several dimensions such as the characteristics of the subject pool, the strength of the incentives, the gain *vs.* loss domain, the abstract *vs.* contextual framework. The strength of the consensus can be appreciated noting how sometimes authors underline that they might have an "atypical" subject pool since they do not find gender differences as expected (Anderson and Mellor, 2009).

This vast consensus notwithstanding, the results of the Holt and Laury (2002) risk elicitation method (henceforth HL), by far the most widely used risk elicitation procedure in experimental economics, had not yet been comprehensively analyzed from a gender perspective. Section 2 presents a comprehensive survey of the HL method in the literature, showing that significant gender differences are the exception rather than the rule. The consensus is in this case disproved by the HL data: it is in this case false. This allows us to deal with a relatively large number of papers for which the outcome reporting bias is relevant.

Crucial to our research question, in most of the papers the HL risk elicitation task is performed only as a control in 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 minor result, and reporting it is not mandatory. 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 left is the fact that presenting "strange" results, i.e., results that go against the consensus, could cast a shadow on the goodness of the sample for the entire work. Our dataset captures a situation in which *swimming upstream*, i.e., reporting results against the current consensus, implies a cost that is close to zero in terms of odds of getting published, particularly in case the HL task is used as a control. With the publication bias out of the picture, the dataset allows us to study the presence of a somewhat pure outcome reporting bias.

The dataset we exploit collects results from several dozen papers, and deals with gender differences in risk attitudes using the HL elicitation task. This dataset contains the results of a larger set of individual studies than those who directly report about gender differences in the published version, and approximates the universe of all papers replicating the HL procedure in the lab or in the field, thereby allowing to observe a good proxy of the counterfactual situation. The dataset is uniquely fit to study outcome reporting bias as distinct from publication bias because it covers a topic about which there is a widespread, possibly *false* consensus, and in which the likelihood of being published depends only marginally on

the coherence of the results with the consensus. It is hence possible to study, among published papers, if the author's attitude towards reporting or not the result is correlated with the consensus view.

We find no significant evidence of an outcome reporting bias in the literature about gender differences in risk preferences. The existence of a very strong consensus does not affect the likelihood of reporting results that are swimming upstream at least when it does not correlate much with the odds of getting published. This finding is robust to possible idiosyncratic characteristics of the authors involved in this field, as the results survive in a fixed effect specification. The only variable significantly affecting the likelihood of reporting about gender differences is the relevance of risk attitudes in the research question of the study.

The outline of the paper is as follows. Section 2 summarizes the state of the art in the literature about gender differences in risk aversion, focusing on the HL task. Section 3 describes the construction and contents of our dataset. 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 Holt and Laury elicitation method

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 (Croson and Gneezy, 2009; Eckel and Grossman, 2008) or from questionnaire studies (Dohmen et al., 2011). More recently, recognizing that the variation in the methods used to elicit preferences generates problems when comparing the results, Charness and Gneezy (2012) carried out a review of a single specific task, the investment game of Gneezy and Potters (1997), providing further supporting evidence for gender differences. The consensus is so strong that gender differences are sometimes considered as a stylized fact whose causes, rather than existence, should be investigated (Bertrand, 2011). However, the HL task has never been analyzed from a gender perspective, despite being the most popular elicitation method in economics.

The HL task uses a multiple price list to elicit the risk preference of subjects. In the HL task subjects face a series of binary choices between pairs of lotteries, with one lottery being safer (i.e., with lower variance) than the other. 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 lottery pairs is obtained by increasing the probability of the 'good' outcome (see Table 1). The subjects must make a choice for each row. At the end of the experiment, one row is randomly chosen for payment, and the chosen lottery is played to determine the payoff.

[1] Henceforth, when referring to gender differences without further specification, we mean that females are less risk tolerant than males.

	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 €

Table 1: The Original Holt and Laury (2002) task

Since the expected value of the risky lottery increases faster and beyond the one of the safe lottery, subjects will at some point 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 are not infrequent patterns; subjects displaying such behavior are regarded as inconsistent when modeling the choices without including a stochastic component.

There are only 21 papers (the original Holt and Laury (2002) and 20 replications) explicitly reporting about 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 last decade confirm that significant gender differences in HL are only rarely found. Out of all the papers quoting Holt and Laury (2002) as of January 2013, only 20 papers reported the breakdown of results by gender. Out of these, only 4 report significant differences, 2 provide mixed evidence as in the original contribution, while 14 find that males and females display a 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 on the paper. We include, when available, the details of the sample, the results by gender and their significance. We report whether the study was a laboratory or field experiment, the characteristics of the subject pool, the type of evidence reported in the paper to support the result, and the p-value of the test or of the coefficient in a multivariate regression. We see that the majority of papers enroll students as subjects and use multivariate regressions to report the significance of their results.

The four papers finding a significant gender difference are Agnew et al. (2008) and Houser et al. (2010), using an unmodified low stake HL task, Dave et al. (2010), using the

Article	n_m	n_f	safe$_m$	safe$_f$	significant gender difference	lab/field	subjects	type of evidence	p-value
Agnew et al. (2008)					yes	lab	non-student	text	
Brañas-Garza and Rustichini (2011)	72	116	4.35	5.01	yes	lab	caucasians	Mann-Whitney	0.0027
Dave et al. (2010)	347	454			yes	lab	Canadian labor force	coefficient	0.001
Houser et al. (2010)	128	76	5.75	6.07	yes	lab	students	text	
Chen et al. (2013)					mixed	lab	students	see text	see text
Menon and Perali (2010)					mixed	field	Italian high school graduates and students	see text	see text
Andersen et al. (2006)	66	24			no	lab	students	coefficient	0.38
Anderson and Freeborn (2010)					no	field	Danish population sample	coefficient	0.54
Baker et al. (2008)		11			no	lab	students	coefficient	0.891
Carlsson et al. (2009)	105	108			no	field	Chinese rural population	Wilcoxon	0.14
Chakravarty et al. (2011)					no	lab	U.S. students	coefficient	0.644
Drichoutis (2012)					no	lab	students and general population	coefficient	
Eckel and Wilson (2004)	133	99	5.30	5.50	no	lab	U.S. students	coefficient	0.586
Ehmke et al. (2010)	170	175	5.26	5.58	no	lab	Chinese, French, Nigerien and U.S. students	text	
Harrison et al. (2005)						lab	students	text	
Harrison et al. (2012)	76	32			no	lab	students	coefficient	0.78
Masclet et al. (2009)					no	lab	students, employees, self-employed workers	coefficient	0.19
Mueller and Schwieren (2012)	32	50	6.75	6.64	no	lab		text	
Ponti and Carbone (2009)					no	lab	Spanish students	test	
Viscusi et al. (2011)					no	lab	students	text	

Table 2: Results by gender reported in the HL literature

20X high stake HL treatment, and Brañas-Garza and Rustichini (2011), implementing a not incentivized version with only 9 choices.

The two contributions reporting mixed results find a significant effect only for a subsample, or only through one and not all statistical methods. In Chen et al. (2013), significant gender differences do not emerge in the unconditional distribution of choices 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). Menon and Perali (2010) on the other hand find, within the same study, females to be significantly more risk averse in one sample and significantly less risk averse in another.

The list of contributions in which the behavior of males and females does not differ significantly is longer, starting with the first replication of the original task (Harrison et al., 2005). It includes Anderson and Freeborn (2010); Carlsson et al. (2009) in the field, and Andersen et al. (2006); Baker et al. (2008); Chakravarty et al. (2011); Drichoutis (2012); Eckel and Wilson (2004); Ehmke et al. (2010); Harrison et al. (2012); Mueller and Schwieren (2012); Ponti and Carbone (2009); Viscusi et al. (2011) and Masclet et al. (2009) in the lab.

Summarizing, this branch of the experimental literature provides a unique opportunity to analyze the outcome reporting bias. In fact, there is a consensus concerning gender difference in risk preferences that turns out to be false when coming to the HL risk elicitation method. The reason is that the likelihood of observing gender differences strongly correlates with the method used to elicit preferences, but this is something that has been pointed out only very recently (Filippin and Crosetto, 2013). In the next section we show that by means of a large dataset of HL replications it is possibly to assess if and how much the presence of the consensus impacts the likelihood of reporting gender-related findings in the paper.

3. The dataset of HL replications

In this paper we use the dataset of HL replications collected by Filippin and Crosetto (2013). To build the dataset, the authors went trough the 529 papers in the Scopus bibliographic database citing Holt and Laury (2002) as of January 31st, 2013. Out of these papers, only 118 implement a version of the HL task sufficiently similar to the original to be counted as a replication. The dataset 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. The dataset 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. Out of the 118 replications, in 16 papers the authors did not record gender or have a single-gender sample, while 8 papers use the same data as another paper in the dataset and have been excluded to avoid duplication of results. The final dataset covers 52 of the 94 remaining papers (see Table 3), with a coverage of about 55% of all published HL replications.[2]

The dataset 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, given the very low reporting rate for gender findings in published articles as well as the variety of statistical approaches followed to report them when doing so. In fact, comments about gender differences are not always accompanied by quantitative results. When reported, results sometimes are expressed using non-parametric tests of the average choices of males and

[2]Since also among the remaining 42 papers some are likely to entail same-gender samples, missing gender data, or a sufficiently different version of the classic HL method, the actual coverage can safely be regarded as higher than the reported 55%.

HL published replications as of Jan 31st, 2013:	**118**	
of which:		
Not reporting gender or single gender	16	
Duplicate dataset	8	
Universe of reference	**94**	100%
of which:		
Promise of future delivery	6	6.3%
No response or not available to share the data	36	38.3%
Final dataset	**52**	55.3%
of which:		
Microdata (shared or available online)	47	
Summary statistics	5	

Table 3: Extent of the dataset of HL replications

females, sometimes take the form of coefficients in multivariate regressions. The dataset reduces to a common metric a large body of potentially heterogeneous literature, and it also allows to uniformly define and process inconsistent choices, which are another source of heterogeneity in the literature. The dataset also keeps track of differences in the implementation of the task (number of choices, probability range spanned, stakes, real or hypothetical incentives, forced or not forced consistency), and, most important for the aim of this paper, includes several studies for which no result by gender was provided in the paper.[3]

The dataset of HL replications of Filippin and Crosetto (2013) confirms that findings with the HL method go against the consensus, as in most of the studies women's behavior is not statistically different from men's. Out of the 52 included papers, males are never found to be significantly more risk averse than females, while the significant consensus gender gap significantly appears only in 6 cases.

This proportion is even lower than the already weak and mixed evidence reported when looking at reported results in Section 2. In fact, the dataset of replications is larger than the published findings, and the availability of results for a sample of studies that do not report about gender differences allows us to approximates a counterfactual situation.

4. Results

In this section we use the Filippin and Crosetto (2013) dataset about the risk and gender literature to test for the presence of an outcome reporting bias. The dataset allows us to abstract away from publication bias not only because all studies have been published, but also because the results about gender differences in risk attitudes can safely assumed not to affect the final outcome. We will show some evidence about this, too. Hence, the dataset

[3]The procedure followed to build the dataset, the reasons for exclusion and the methodological variations included in the sample are described in detail in Filippin and Crosetto (2013).

provides a clean test of the attraction exerted by a strong consensus on the likelihood of observing a reporting bias without explicit extrinsic rewards at stake.

4.1. Descriptives of the sample

As detailed in section 3, the dataset is composed of 52 studies replicating HL. These studies can be divided according to two criteria related to the outcome reporting bias. Along the *report* dimension, the papers can report a significant gender difference, report a not significant gender difference, or not report anything. Along the *result* dimension, the papers can find or not find a significant gender difference. This second dimension is computed applying a common methodology to all the papers of the dataset, and namely, results are found by means of a non-parametric test on the unconditional distributions of safe choices of consistent, i.e., not multiple-swithcing, males and females.

Table 4 includes all 52 papers of the Filippin and Crosetto (2013) dataset. Moreover, it includes a column for the 42 papers outside the dataset but within the universe of HL replications. For these studies we have no microdata. Of those, for 36 studies we could not find in the paper any gender information for the HL results; for further 6 we did find some information in the paper. Table 4 reports the results of this bidimensional breakdown of the papers.

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 only (at 10%) and 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 methodology, i.e., testing the unconditional distribution of choices of consistent subjects, we classify this paper as finding and reporting no gender differences. Menon and Perali (2010) find different results with females significantly more risk averse in one sample, significantly less risk averse in another sample, and not significantly different than males in a third one. We do not have the microdata available for this paper, and therefore we cannot classify it according to our common metric. Looking at their published figures, though, we speculate that merging the sub-samples the opposite results are quite likely to cancel out, delivering a not significant difference overall. Since Menon and Perali (2010) report all kind of results, they clearly show no reporting bias. Therefore, as a default option we classify this entry as finding and reporting no gender differences. As a robustness check we will remove it from the dataset. We follow exactly the same approach with Holt and Laury (2002), who find significant gender differences only in one of their treatments.

		Gender difference			
		Found	Not found	n.a.	**Total**
Report	significant difference	2	1	1	**4**
	not significant difference	0	13	5	**18**
	nothing	4	32	36	**72**
	Total	**6**	**46**	**42**	**94**

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

A consideration is necessary for one paper that publishes significant gender differences,

which do not emerge in our analysis of the microdata. The difference is due to the fact that we exclude inconsistent subjects from the analysis, while the authors included them. In view of our goal in this paper, we keep it in the strange position of finding no bias but reporting one, because on the one hand we want to keep a common procedure to evaluate the papers and on the other the choice of which results to publish is up to the authors and could in principle be part of the outcome reporting bias.

4.2. Testable implications

We can use the dataset to formally identify several testable implications about outcome reporting bias. We clarify the various tests with the help of Table 5.

	Gender differences	
	Found	Not Found
Significant difference reported	a	0
Not significant difference reported	0	c
Nothing reported	b	d

Table 5: Generic distribution of results

We start first with a testable implication about our assumption that the publication bias is not an issue in our framework, showing that reporting about gender differences is not an important factor in getting published. We do so in an indirect way, testing whether it is likely for papers not to report anything about gender differences, regardless of the significance of the underlying results. In terms of Table 5, this amounts to test for the existence of a low reporting rate:

$$a < b \quad \text{and} \quad c < d. \tag{1}$$

This can, though, have to do in part with the research question of the paper. Studies differ with respect to their main focus. Some studies have the exploration of risk preferences at the core of their research. They focus 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 decision under risk (for instance trying to disentangle risk aversion from loss aversion, or estimating the effect of the salience of the incentives). Another class of studies can be built to include papers that focus on other topics, like auctions, strategic games, tournaments, and use the HL task just as a control for risk preferences. This is a rather heterogeneous class, but for the goal of this paper it has in common a much looser focus on the HL task itself. We label the former category as papers having a *main* focus on HL, while the latter as using it only as a *control*.

The different importance of analyzing risk attitudes as mandated by the main research question of the papers provides a further test that the likelihood of reporting gender differences is driven by other determinants than the role it can play towards the publication outcome. The inequality sign of Equation 1 could be reversed for papers having HL as their*main* focus, but in any case we expect the report rate $(a + c)/(b + d)$ to be significantly lower for studies using HL as a *control*.

The presence of an outcome reporting bias in this context means that studies that find significant gender differences in line with the consensus should be more likely to report them. In contrast, when males and females are characterized by a similar behavior the presence of outcome reporting bias would predict that the results are less likely to be reported, as authors prefer to amend their reports rather than signalling "atypical" findings not aligned with the consensus. The testable implication is therefore that the fraction of studies finding significant gender differences should be higher among those who report rather than among those who do not. Under the reasonable assumption that the likelihood of observing significant gender differences is *ex ante* the same, 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 *report* and *result*. In particular, we formally test whether:

$$\frac{a}{a+b} > \frac{c}{c+d}. \tag{2}$$

Note that had we relied upon the literature review, even abstracting away from problems related to the different tests used to generate the results in the different papers, we would have observed a and c only. Relying on the replications dataset allows us to observe also b and d, which can be used to approximate the counterfactual situations of *not* reporting conditioning on the results observed. The counterfactual is only approximated, since we have 42 papers in the universe of HL replications that do not enter the dataset. This notwithstanding, for the 52 studies in the dataset the availability of the microdata allows us to observe the underlying latent variable about which no information has been published. For these 52 studies we can directly test the existence of outcome reporting bias, without relying upon bias reducing techniques.

4.3. Non-parametric test

Concerning the testable implication of Equation 1 it can be immediately noticed from Table 4 that about two thirds of the times gender differences are not explicitly reported. Following the discussion made in section 4.1, one could argue that the likelihood of reporting should be analyzed in a different way for *main* and *control* studies, since the importance of the HL task and therefore of gender differences is lower in the *control* papers.

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

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

We find this hypothesis to be supported by the data. Following the classification of the 94 papers between *main* and *control* detailed in Table 6, we find as expected that the likelihood of reporting about gender differences strongly correlates with the importance of risk attitudes in the paper. In fact, results are reported only in very few circumstances (about 11.5%) when the main research question of the paper does not concern risk preferences, while it is more common (about 42.5% of the times) when the paper deals with risk attitudes.

A Fisher exact test confirms that the two distributions are indeed significantly different ($p = 0.001$).

Table 4 shows that 4 out of the 22 studies reporting results do find significant gender differences. Using the 36 studies that do not report (4 with significant gender differences, 32 without) as the counterfactual, we see that the fractions are indeed different (18.2% *vs.* 11.1%), but not significantly so according to a one-sided Fisher exact test ($p = 0.351$). In this comparison we included also the five papers for which we have only published information and no microdata. Results do not change, though, if we limit the analysis to the 52 studies present in the dataset: fractions become 20% *vs.* 11.11%, not significantly different according to the one-sided Fisher exact test ($p = 0.334$).

		Gender differences					
		Main			*Control*		
		Found	Not found	n.a	Found	Not found	n.a
Report	significant difference	1	1	0	1	0	1
	not significant difference	0	10	3	0	3	2
	nothing	1	11	6	3	21	30
	Total	**2**	**22**	**9**	**4**	**24**	**33**

Table 7: Distribution of papers according to the importance of risk attitudes, detail

One could argue that the outcome reporting bias, while not detected with aggregate results, could still correlate with the relative importance of risk preferences in the research question of the paper, although it is not clear a priori in which direction. 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. An outcome reporting bias could instead exist limiting the attention to one of the two types of papers. Data, however, show that this is not the case. Table 7 replicates the distribution of Table 4 for *main* and *control*, separately. Among *main* papers reporting, 2 out 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 ($p = 0.586$). Within studies using the HL as control, the frequencies are more differentiated: significant gender differences emerge in 28.6% of the papers publishing the results, while among those that can be used as a counterfactual situation the percentage is equal to 12.5%. Also in this case a Fisher test cannot reject that the two frequencies are the same ($p = 0.312$).

In principle, the outcome reporting bias could extend also 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 (2013) dataset and therefore that the likelihood of not finding gender differences is even higher among the 36 studies about which there is no information available. Of course, there is no way we can check the distribution of significant *vs.* non-significant gender differences within this sub-sample. However, it would be enough that only one out of the 36 papers actually found significant gender differences to avoid rejecting that the two frequencies are drawn from the same distribution.

Note also that a higher likelihood of reporting significant differences could be driven by a possibly different value of the information provided. This is more likely the case among papers using HL as a control, whose focus is on other issues. Therefore, they could select the significance of the results as a screening device for the non-central results to include, without this having necessarily to do with any consensus. What we measure with the tests above is therefore an upper bound of the effect of an outcome reporting bias. Detecting null results for the upper bound makes however unnecessary to identify the two effects. We can then conclude here that there is no evidence of outcome reporting bias for gender in risk preferences, and we observe that this result is mainly due to the overwhelming rate of non reporting present across the board in the literature.

4.4. Multivariate analysis

In this sub-section we jointly analyse 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)*, once controlling for other explanatory variables. In particular, the fact that risk attitudes are the main focus of the paper or not *(main)* has already been shown to affect the probability of addressing the gender pattern. In a multivariate framework besides controlling for this additional factor we can also interact it with the observed pattern in the results. In particular the specification that we estimate is the following:

$$\text{report} = \alpha + \beta\text{find} + \gamma\text{main} + \delta\text{find*main} + \epsilon \tag{3}$$

Table 8 reports 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 the simple linear regression model detailed in Equation 3. Results show that only the fact that risk attitudes are among the main goals of the paper increases the likelihood of reporting the data. In contrast, there is no evidence of any outcome reporting bias, neither at the aggregate level, nor within each of the two subgroups.

	Dependent variable: Gender differences reported					
	Model 1			**Model 2**		
	Coeff.	St. Err.	p-value	Coeff.	St. Err.	p-value
find	.208	.228	.368	.116	.265	.665
main	.329	.134	.017	.373	.184	.050
main×find	-.063	.367	.865	-.191	.458	.679
author fixed effects	No			Yes		

Table 8: Determinants of the likelihood of reporting results

Observations are not necessarily independent, because the same authors are sometimes included more than once in our paper. Therefore, we also run a fixed effects specification (Model 2) in which we partial out the effect of author invariant observable and unobservable characteristics. Results barely change, as shown in the second column of Table 8.

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 paper focuses on one of the possible sources of this problem, namely the outcome reporting bias, i.e., the different likelihood of being reported that can characterize results in favor or against a well-established consensus.

This paper exploits a large dataset of replications of the Holt and Laury risk elicitation method, by far the most widely used risk task in experimental economics, to provide clean evidence about the outcome reporting bias. 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 in a systematic way that using the HL procedure gender differences are the exception rather than the rule. This means that there is a false consensus in this branch of the literature, and therefore that many authors had to face the choice between reporting or not results that did not conform to the consensus.

When using the HL method, reporting or not about gender differences affects only marginally, or more likely not at all, the likelihood of being published. This is true particularly for the papers in which the HL risk elicitation task is performed only as a control in experiments dealing with other topics. This means that the dataset used allows us to disregard the *publication bias*, and to focus on the mere attraction exerted by the consensus itself. We believe that such an exercise is important because the existence of a pure reporting bias could contribute to reinforce an existing consensus and even to generate a false consensus based on self-confirming beliefs.

The dataset of replications contains the results of a larger set of individual studies than those which directly report about gender differences in the published version, thereby allowing to approximate a counterfactual situation. In other words, we can observe the existence or not of significant gender differences for many papers that did not report about it in the published version.

We find no significant evidence of an outcome reporting bias in the literature about gender differences in risk preferences. The existence of a very strong consensus does not affect the likelihood of reporting results that are swimming upstream at least when it does not correlate much with the odds of getting published. This finding is robust to possible idiosyncratic characteristics of the authors involved in this field, as the results survive in a fixed effect specification. The only variable significantly affecting the likelihood of reporting about gender differences is the relevance of risk attitudes in the research question of the study.

This result is definitely good news. It should be taken with a grain of salt, though, because it refers to a very 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 dataset, gathered from dozens of studies involving altogether more than one hundred authors adopting the most widely used task in the literature. To the extent that this dataset is representative of the practices adopted in other disciplines, the insight that can be derived go beyond the specific subfield from which the data have been gathered.

References

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

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

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

Anderson, L. R., Mellor, J. M., 2009. Are risk preferences stable? comparing an experimental measure with a validated survey-based measure. Journal of Risk and Uncertainty 39 (2), 137–160.

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

Bertrand, M., 2011. New perspectives on gender. 1st Edition. Vol. 4B Handbook of Labor Economics. Elsevier, Ch. 17, pp. 1543–1590.

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

Carlsson, F., Martinsson, P., Qin, P., Sutter, M., 2009. Household decision making and the influence of spouses' income, education, and communist party membership: A field experiment in rural china. Tech. rep., IZA.

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

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

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

Costa-Font, J., Gemmill, M., Rubert, G., 2011. Biases in the healthcare luxury good hypothesis?: a meta-regression analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society) 174 (1), 95–107.

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

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

De Long, J. B., Lang, K., December 1992. Are all economic hypotheses false? Journal of Political Economy 100 (6), 1257–72.

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

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

Doucouliagos, H., Stanley, T. D., 06 2009. Publication selection bias in minimum-wage research? a meta-regression analysis. British Journal of Industrial Relations 47 (2), 406–428.

Drichoutis, Andreas C., P. K., 2012. Estimating risk attitudes in conventional and artefactual lab experiments: The importance of the underlying assumptions. Tech. rep., Social Sciences research Network.

Duval, S., Tweedie, R., 2000. A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. Journal of the American Statistical Association 95 (449), pp. 89–98.

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., Decullier, E., Easterbrook, P. J., Von Elm, E., Gamble, C., Ghersi, D., Ioannidis, J. P. A., Simes, J., Williamson, P. R., 08 2008. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS ONE 3 (8), e3081.

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

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

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

Filippin, A., Crosetto, P., 2013. Gender Differences in Risk Attitudes: a Reconsideration using a large Dataset of Holt and Laury Replications, mimeo.

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

Harrison, G., Lau, M., Rutström, E., Tarazona-Gómez, M., 2012. Preferences over social risk. Oxford Economic Papers forthcoming.

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

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

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

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

Mahoney, M., 1977. Publication prejudices: An experimental study of confirmatory bias in the peer review system. Cognitive Therapy and Research 1 (2), 161–175.

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

Menon, M., Perali, F., 2010. Eliciting risk and time preferences in field experiments: Are they related to cognitive and non-cognitive outcomes? are circumstances important? Tech. rep., Department of Economics and CHILD University of Verona.

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

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

Scherer, R. W., Langenberg, P., von Elm, E., 2007. Full publication of results initially presented in abstracts. Cochrane Database of Systematic Reviews 2.

Simmons, J. P., Nelson, L. D., Simonsohn, U., 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science 22 (11), 1359–1366.

Stanley, T. D., Doucouliagos, H., 2013. Meta-regression approximations to reduce publication selection bias. Research Synthesis Methods, n/a–n/a.

Sterling, T., 1959. Publication decision andthe possibleeffects on inferences drawnfrom testsof significance-or vice versa. Journal of the American Statistical Association 54, 30–34.

Sterling, T. D., Rosenbaum, W. L., Weinkam, J. J., Feb. 1995. Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. The American Statistician 49 (1), 108–112.

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