Gender

Muriel Niederle
Stanford University and *NBER*

December 17 2014

In preparation for the *Handbook of Experimental Economics, Vol.2*
Editors: John Kagel and Alvin E. Roth

I am deeply grateful to Katherine Coffman and John Kagel for their extensive comments and to John Kagel and Alvin Roth for their patience. Foremost, I am thankful to Lise Vesterlund with whom I have most frequently worked on gender. Our long discussions have shaped my understanding of the topic tremendously. The first section of this chapter is based on our common work, and I have borrowed from it extensively. Finally, I am grateful to the NSF for its support.

I. INTRODUCTION

Gender is deeply rooted in our identity and is one of the first traits we observe about others.[1] Gender differences receive enormous attention by the popular press and the public: John Gray's (1992) "*Men are from Mars, Women are from Venus*" has "sold more than 50 million copies and in the 1990s was ranked as the third most popular book.[2] More recently, Sheryl Sandberg's (2013) "*Lean In*", a controversial and much discussed book has spent weeks on the top of bestseller lists. While the psychology literature has debated gender differences in preferences for almost 150 years (Hyde, 2005 and Shields, 1975) the discussion of gender has only recently started to gain momentum in economics. For example, most chapters in the last Handbook of Experimental Economics (Kagel and Roth, 1995) did not even mention gender differences - even Ledyard's chapter on Public Goods referencing roughly 250 papers includes only 6 papers studying gender differences.

Since the turn of the millennium the situation has changed and there has been an explosion of experimental work on gender differences in economics. There are now several surveys focusing solely on that topic (Eckel and Grossman, 2008 a,b, Croson and Gneezy, 2009). In this chapter I revisit the three traits for which gender differences have been most extensively studied: attitudes to competition, altruism or cooperative attitudes, and risk attitudes. In each section I focus on series of experiments, and also present early results from the psychology literature (though this literature does not have results for competition). One of the strengths of experimental economics is that many findings are replicated and studied in different contexts to establish whether the initial finding was a true and robust result as opposed to a false positive or rather a knife-edge result.[3] This survey focuses on experiments in which there is little or no interaction between

[1] For example, Simons and Levin (1998) when studying change blindness, the failure to detect changes when interacting with an individual, such as exchanging the clothes of that person or exchanging the person herself, confine themselves to exchanging a person with a person of the same sex. They write that "Clearly we would be quite surprised if subjects missed a switch between enormously different people (e.g., a switch from a 4 ft 9 in. female of one race to a 6 ft 5 in. male of another). The change in this case would alter not only the visual details of the person, but also their category membership. If, as suggested by other recent findings of change blindness, we retain only abstracted information and not visual details from one view to the next, changes to category membership may well be detectable." (p 648).

[2] The 50 million estimated sales were reported on Wikipedia, and the CNN article that ranks the most popular books in the 1990s is here: http://archives.cnn.com/1999/books/news/12/31/1990.sellers/index.html

[3] For the importance of replication in all fields not only experimental economics see e.g. Coffman and Niederle (2014a). Coffman and Niederle (2014b) discuss ways to promote replications and studies of robustness in experimental economics.

agents. This reduces the influence of confounds such as potential gender differences in strategic behavior or discrimination which may be present in more complex interactions such as sequential or repeated games. While there are many other areas in which gender differences have been documented, I am neglecting these areas not because they lack importance or interest, but rather to keep the chapter at a manageable length.

Why have economists not studied gender differences in psychological attributes earlier, given the interest in gender differences in economic outcomes? Attributing field evidence of gender differences in outcomes to specific traits is difficult. For example, assessing gender differences in altruism in the field often relies on observing gender differences that cannot be explained by standard economic variables such as socio-economic status, income etc.[4] Additionally, the difficulty of attributing gender differences in labor market outcomes to specific traits may contribute to labor economists focusing on two other possible sources of gender differences: discrimination and differences in human capital accumulation. The latter may either be in the form of education before labor market entry, or in the form of accumulated experience after having entered the labor market (see Altonji and Blank (1999)'s "race and gender in the labor market" in the 3rd volume of the Handbook of Labor Economics.)

In contrast to field evidence, inferring altruism as the unexplained variation of a complex choice that can be the result of many motives, the laboratory can be stripped of many confounding factors and decisions can be observed in a highly controlled environment. In doing so, we can directly measure traits such as attitudes to competition, altruism, and risk. With the rise of behavioral and experimental economics, the study of gender differences in traits has received growing attention.

As my chapter focuses on gender differences, it is worthwhile to note that these differences, while significant, are sometimes small. This has been the case for many psychological traits, and

[4] For example, there has been a long and ongoing debate on gender differences in charitable donations with, at present, no clear conclusion. An even more indirect test from the field consists of explaining voting patterns of women and men. For example, Edlund and Pande (2002) show that over time women have become more left-wing. Their paper points out that this difference may, however, be explained by an increase in divorce risk and decline in marriage. That is a preference for redistributive policies could have purely economic rather than psychological reasons. Though more recently Funk and Gathman (forthcoming) provide some evidence that gender differences in voting remain after controlling for socio-economic characteristics.

almost since its inception the literature on gender differences consisted of two "camps." One side argues for the existence and importance of gender differences, and the other side emphasizes gender similarities. As an example of the "differences are important" camp, Eckel and Grossman (1998) in their foreword cite Charles Darwin, 1874, p 586 "[w]oman seems to differ from man in mental disposition, chiefly in her greater tenderness and less selfishness…Man…delights in competition, and this leads to ambition which passes too easily into selfishness." In the "differences are small" camp, Hyde (2005) calls her review on the psychological literature "The Gender Similarities Hypothesis."

Whether statistically significant gender differences are economically significant, so that it is more appropriate to talk of gender differences rather than gender similarities depends on the question at hand. In cases where the average outcome of one decision is of interest, small gender differences may not be economically important.

However, there are cases in which even small differences can result in significant effects. When studying repeated choices, small differences might accumulate, thereby calling for policy interventions. For example, if the structure of an exam is such that there are penalties for wrong answers, small differences in risk aversion may result in women being, on average, slightly more likely to skip a question than men are. In an exam with many questions even a small difference can accumulate to generate a more sizable effect. Furthermore, small average differences in normal distributions become larger when assessing the probability of gender representations among participants with extreme versions of that trait. Indeed, there is a long and ongoing debate about gender differences in math ability and the extent to which gender differences are exacerbated among those of very high ability. Recall the heated and very ideological debate that followed Larry Summers comments on January 14, 2005, in which he suggested the underrepresentation of female scientists at top universities may be in part due to natural ability differences between men and women.

The study of gender differences has, almost since its inception, been plagued by ideologically guided interpretations. In the first review of the literature on gender differences in psychology, Woolley (1914) pointed out and deplored the gap between the predominant views on the

question including that of scientists versus the conclusions supported by data. Hyde (2005, page 581) cites Woolley (1914, page 372) as "The general discussion of the psychology of sex, whether by psychologists or by sociologists show such a wide diversity of points of view that one feels that the truest thing to be said at present is that scientific evidence plays very little part in producing convictions." When surveying gender differences in various traits I will therefore aim to provide a balanced view and provide interpretation of the magnitude of observed differences.

Summarizing the evidence presented in this chapter, I find that there are large gender differences in reaction towards competition, with women shying away from competition with men, and women underperforming when competing against men. These differences persist and are only somewhat reduced when controlling for beliefs about relative performance as well as risk aversion. The robust finding is that gender differences are particularly pronounced when performance is measured in tasks that are not stereotypically female. In addition there is some new evidence that women shy away from challenging tasks and refrain from "speaking up."

The evidence on gender differences in altruism is much more mixed. While some studies find women to be more altruistic than men, this is not always the case, and differences, when they exist, are often small. A more robust conclusion seems that women and men differ in how their utility depends on the payoffs of others. Specifically, women seem more concerned with equalizing payoffs among laboratory participants while men seem to have a higher preference for efficiency; that is, donations by women compared to donations by men respond less to the costs of giving. The behavior in more complex public good games is less amenable to a simple summary, though recent studies have provided promising inroads in understanding the interplay between donations to a public good and strategic reactions towards the way the public good is provided.

The evidence on gender differences in risk aversion is also much less clear than one might expect. Some methods of eliciting risk preferences, while showing variation across participants result in no, or very small, gender differences. Other elicitation methods produce reliable gender differences, with women being more risk averse. It is, however, somewhat disconcerting that

different elicitation methods are often not much correlated with one another, and each one seems quite valid in estimating risk preferences. This lack of a unified result could be due to the fact that risk preference in itself is complex, and is not easily reducible to the outcome of a single choice.

Once gender differences in a trait have been established through extensive replications in different laboratories, it is important to show that these differences can occur outside of the laboratory, beyond experiments with students. One way to do this is with field experiments that can bring "laboratory style" decisions to the field. Another way is to find an interesting variation that occurred naturally. To date there are at least two summaries of the literature assessing the role of experimental findings on gender differences for labor economics in field settings, Bertrand (2011) and Azmat and Petrongolo (forthcoming).

External validity shows that a result - a gender difference in behavior – can be found outside of the laboratory, and is not specific to standard student subject pools or standard laboratory tasks and decisions which are simple and short. However, non-laboratory studies often occur with subject pools that are equally (or perhaps even more) special than undergraduate students. For example, assume a specific gender difference is replicated using Austrian farmers. However, this result is not necessarily more predictive of behavior of Austrians in general, or American farmers, than the result established in the laboratory. Most importantly, finding a gender gap in a given trait among Austrian farmers does not imply that economic differences among Austrian farmers are due to this particular trait. And of course, it certainly does not inform us that gender differences in economic outcomes among Austrians in general, or American farmers, can be attributed to gender differences in this trait. External validity is exactly that: it shows that a trait is valid outside of a laboratory setting. It does not, however, necessarily tell us if that this trait is relevant for general economic outcomes.

In other words, documenting results outside of the laboratory cannot always speak to the broader importance of external relevance. Bertand (2011) makes this point in the conclusions of her Chapter on Gender in the labor market for the 4th Volume of the *Handbook of Labor Economics* (p 1583): "While the laboratory evidence shows in many cases large gender differences (say, in

attitudes towards risk, or attitudes toward competition), most of the existing attempts to measure the impact of these factors on actual outcomes fail to find large effects. This is undoubtedly a reflection of a rather new research agenda, as well as of the difficulty in finding databases that combine good measures of psychological attributes with real outcomes. More direct demonstrations of field relevance will be crucial for these new perspectives to have a lasting impact on how labor economists approach their study of gender gaps."

Each section on competition, altruism and risk will have some evidence regarding the external validity of the laboratory findings through field experiments and naturally occurring data. However, special emphasis will be placed on showing that gender differences in a given trait can account for a significant fraction of gender differences in economic decisions relevant to labor market or education outcomes of women and men. That is, I will try to emphasize the evidence for the *external relevance* of gender differences in competition, altruism and risk.

After establishing the importance of gender differences in psychological traits for education and labor market outcomes the question is what to do with this knowledge. A first natural question is whether, or how much, of these differences are due to nature or nurture. If they are due to nurture, maybe these traits can be changed, though this may require a deeper investigation into the potential benefits of doing so. A second question is whether these gender differences are indeed "true" differences in preferences, or whether they rather represent biases of women (or men), and whether awareness of those gender differences may therefore act as a way to "debias" the decisions of women and men.[5] At heart, this is one of the messages of Sandberg's (2013) "*Lean In*."

Another possible next step is to assess whether the design of the decision environment, the choice architecture or the market can affect the gender gap because they differentially activate a psychological attribute in which there are large gender differences. For example, when students decide about how much math to take in school, these choices are typically binding (once and for all choices) in continental Europe. In contrast, in the US, education choices are much more flexible. American students, upon struggling in a difficult math class, may opt to take an easier

[5] For a review on the literature on debiasing, see Soll, Milkman and Payne (2013)

one next semester in high school. Choosing the difficult path of taking hard math courses is therefore a different choice in Europe than in the US. This difference in the way decisions are made may in itself affect gender differences in choices of math intensive courses.

In his Fisher Schultz lecture Roth (2002) has emphasized the role economists can play in designing as opposed to simply studying them. More recently, the investigation of the role of choice architectures on economic choices of agents is extensively reviewed and discussed by Thaler and Sunstein (2008). It may be time to expand the debate of choice architecture to understanding their impact on gender differences in choices.

The study of gender differences in attributes as summarized in this chapter may change the way in which we interpret gender differences in labor market outcomes. We may start to attribute such differences not only to gender differences in abilities and discrimination, but to gender differences in preferences. In addition, we should start investigating which policies may be successful in ensuring that gender differences in economic outcomes reflect underlying abilities. Market design could be expanded to include institutional and education designs that help women and men make choices that reflect their underlying preferences over outcomes rather than reflecting differences in attributes which play a role due to the environment in which these decisions are made.

II. Gender Differences in Competitiveness

In this section I review the relatively new but very vibrant work on gender differences in attitudes toward competitions. Much of the work has been reviewed more extensively in an earlier survey by Niederle and Vesterlund (2011).[6] The main motivation for the work on gender differences in competition is to shed light on possible reasons for gender differences in labor market outcomes, concerning vertical as well as horizontal segregation. Historically, the main explanations for these differences are differences in preferences over jobs, differences in ability, and discrimination (see references in Niederle and Vesterlund, 2011). In this chapter I review evidence of an additional explanation, namely that there are gender differences in attitudes towards competition: women may be less likely to enter competitive and male-dominated fields, less likely to seek out promotions and their performance may suffer in competitive

[6] I am very indebted to Lise Vesterlund for numerous discussions on this topic, and for her work on our previous survey paper Niederle and Vesterlund (2011) from which I drew heavily to write the present one.

environments compared to men. This research also provides a prime example of how experimental laboratory results interplay with work in the field.

While gender differences in competitiveness have not been a topic of interest in the economics literature until the last decade, such differences have been documented in the educational and evolutionary psychology literature (see Campbell 2002): Boys spend more time at competitive games than girls, while girls often select games that have no clear end point and no winner. These differences increase through puberty, and more men than women describe themselves as competitive. However, in contrast to most of the other work described in this chapter, there has been no earlier literature in social psychology studying gender differences in competitiveness (e.g. the *Handbook of Social Psychology*, fourth edition (1998) does not have an entry on competition in the subject index).

There are two methodological issues when studying gender differences in competitiveness. The first is that the experiments in this section differ from many "standard" economic experiments in that they use real effort tasks. This has advantages and disadvantages. Real effort tasks allow measuring actual performances of women and men under both competitive and non-competitive incentive schemes. Furthermore, choices over incentive schemes indicate not only preferences over payment schemes, but also factors which are potentially important factors outside of the laboratory such as, e.g. beliefs about the ability to perform these tasks. The disadvantage of a real effort task is a loss of control as effort cannot be directly measured, only performance can be measured. Furthermore, the link between performance and effort is not always clear. Some tasks may result in performance that is very inelastic in effort and as such changes in incentive schemes may affect effort but not performance. In addition, Ariely et al (2009) have shown that higher incentives can lead to lower performance, one explanation being that participants may choke when stakes are very high. Another disadvantage of a real effort task is that because costs of effort cannot be measured easily optimal choices cannot be computed easily. However there are experimental techniques that minimize these disadvantages and allow researchers to draw sound inferences despite not knowing the costs of effort or the precise relationship between effort and performance.

The second methodological issue is that experiments on gender differences in competitive attitudes may depend on the gender of the participant as well as the gender of the other subjects. The most common solution, and the solution employed in early experiments, is to physically show subjects who they are

competing against, which allows them to determine the gender of their competitors. The main reason for this approach is that directly mentioning gender could lead to priming or to experimenter demand effects.[7]

I first describe laboratory experiments on gender differences in competition. I start with gender differences in preferences for competitive incentive schemes and discuss how this gap can be reduced. I then move to gender differences in performance under competitive incentive schemes, followed by a discussion of the field evidence of gender differences in competitiveness. The final part of this section concerns work that addresses the external relevance of gender differences in competition for education and labor market outcomes.

II.A Do Women Shy Away from Competition?

The first paper to address whether women and men differ in their choices of competitive incentive schemes is Niederle and Vesterlund (2007), henceforth NV. Participants in the experiment choose between a non-competitive piece rate scheme and a competitive tournament incentive scheme. There are several possible explanations for why a woman and a man with the same chance to win the tournament may differ in their choices. For each of those, I present the design solution in NV.

Explanation 1: Gender differences in attitudes toward competition: This will be the main hypothesis, so the experiment is designed such that other explanations can be ruled out.

Explanation 2: Gender differences in beliefs about relative performance: Men enter the tournament more than women because they are more (over)confident. Psychologists and economists often find that men tend be more optimistic about their abilities than women.

Design solution: Assess the participants' beliefs about their relative performance in the competitive tournament scheme.

Explanation 3: Gender differences in risk and feedback aversion: These are dimensions different from taste for competition that also impact the choice between a tournament and a piece rate incentive scheme. The tournament payment scheme not only is competitive, it is also more uncertain and provides more information about relative performance than the piece rate scheme. For both risk aversion as well as preferences over receiving feedback about relative performance, there may be gender differences.

[7] Keeping subjects completely in the dark about the gender of their opponent may lead to subjects each forming different beliefs about the gender of their opponent which in itself could lead to differences in the behavior of a participant.

Design solution: Instead of directly controlling for risk and feedback aversion, participants make a decision between two incentives schemes which mimic both the uncertainty in payment and the provision of feedback without any actual competition taking place.

Explanation 4: Gender differences in other regarding preferences: While a piece rate scheme has no externality on others' payments, a competitive tournament generates both winners and losers. If there are gender differences in altruism these may generate gender differences in choices.

Design solution: The tournament is designed such that choosing the tournament is an isolated individual decision that has no externality and hence no payoff consequences on any other subject.

In the experiment two women and two men were seated in rows to form groups of four participants. Participants knew they were grouped with other people in their row and could see each other, though NV never discussed gender during the experiment. Subjects perform a real effort task under various incentive schemes. The task is to add up sets of five two-digit numbers for five minutes, where the score is the number of correct answers. After each problem, participants learn the number of correct and wrong answers so far, and whether the last answer was correct or not. Participants do not receive any feedback about relative performance (e.g. whether they won a tournament) until the end of the experiment.

The experiment has four treatments, one of which was randomly chosen for payment at the end of the experiment. The first two treatments serve as a measurement of the subjects' performance at the real effort task.

Treatment 1—Piece Rate: Participants are given the five-minute addition task under a piece rate pay of 50 cents per correct answer.

Treatment 2—Tournament: Participants are given the five-minute addition task with the participant who solves the largest number of correct problems in the group receiving $2 per correct answer while the others receive no payment (in case of ties the winner is chosen randomly among the high scorers).[8]

Subjects do not receive any information about the performance of others, specifically they are not told their relative rank in either the piece rate or the tournament or whether they won the Treatment 2 tournament. Measuring performance in both competitive and non-competitive environments serves to determine the money-maximizing incentive scheme for each participant. The average performance of the

[8] By paying the tournament winner depending on the performance rather than a fixed prize, NV avoid providing information about winning performances, or distorting incentives for very high performing individuals.

40 women and 40 men under the piece rate scheme was 10.2 and 10.7 problems, respectively. Under the tournament, women solved on average 11.8 problems compared to the 12.1 of men. Neither of these differences was significant although the performance in the tournament was significantly higher than under the piece rate for both women and men. This could be due to either increased effort in the tournament which leads to increased performance, or because participants are learning how to better perform this task. The evidence points to learning.[9]

Of the 20 tournament groups 11 were won by women and 9 by men. More importantly, women and men with the same performance in Treatment 2 have the same probability of winning the tournament. This allows NV to use absolute performance rather than a computed chance of winning the tournament in their analyses of gender differences in tournament entry.

In the third treatment participants once again perform the five-minute addition task but this time select, in advance, which of the two compensation schemes to apply to their performance - piece rate or tournament.

Treatment 3 – Choice: A participant who chooses piece rate receives 50 cents for each correctly solved problem. A participant who chooses the tournament has the Treatment 3 performance compared to the Treatment 3 tournament performance of the other participants in his or her group. If the participant has the highest performance she or he receives $2 for each correct answer in Treatment 3. Otherwise she or he receives no payment.

Note that a participant's choice in Treatment 3 does not affect the payment on any other participant. This allows ruling out the possibility that women may shy away from competition because by winning the tournament they impose a negative externality. Further, the participant's beliefs about others' choices have no payoff consequences and as such should not influence choices. Finally, since the participant's competitors were all required to perform in a tournament in Treatment 2, the participant upon selecting the tournament in Treatment 3 still has to outperform a tournament performance of his or her three competitors.

[9] This is supported by the fact subjects who in Treatment 3 choose the piece rate scheme have the same change in performance between Treatment 2 and Treatment 3 than those who chose the Tournament in Treatment 3. Note that the results do not imply that participants do not provide a high effort in the tournament; rather it appears that either their baseline effort is already quite high or that the task is one in which changes in effort do not result in large changes in performance.

Given the task 2 tournament performance, 30% of women and 30% of men have substantially higher earnings from a tournament payment.[10] In fact, 35% of women and 73% of men enter the tournament (a significant difference).

Figure 1a shows for each task-2 tournament performance quartile the proportion of participants who enter the tournament, for women and men separately. Regressions confirm that men have a significantly higher propensity to enter the tournament for any performance level.[11]

Tournament entry decisions should be driven by beliefs about relative performance, not only the absolute performance of a participant. Therefore, just before the end of the experiment, after Treatment 4, participants were asked to guess their performance rank among the 4 players in their group both in the piece rate and tournament treatments (Treatments 1 and 2). Participants received $1 if their guessed rank corresponds to their actual performance rank. NV find that 30 out of 40 men (75%) believe that they were the highest performer in their group of four! Most of them were obviously wrong. Men are highly overconfident. Women are also overconfident as 17 out of 40 women (43%) believe they had the highest performance. However, men are significantly more overconfident than women given their actual rankings.

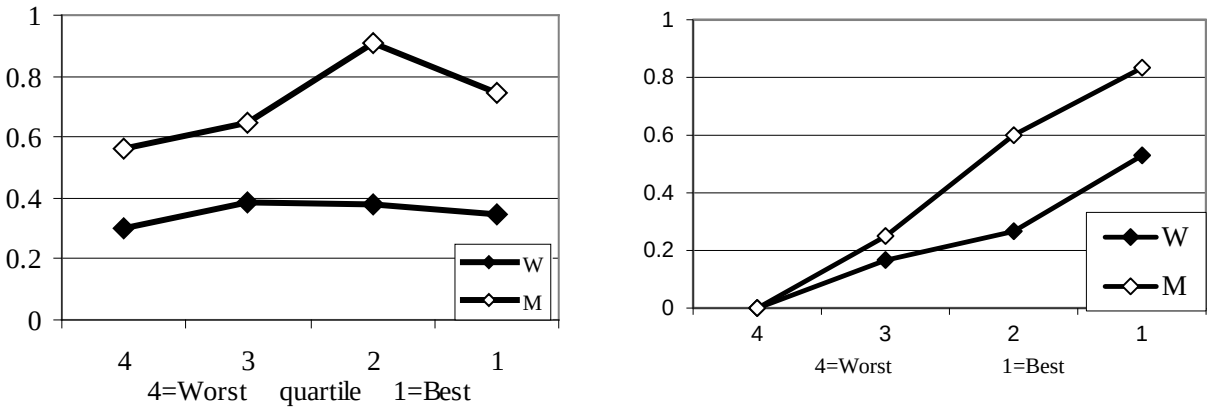

Figure 1: Proportion of participants selecting the tournament: (a) depending on performance quartile, (b) depending on believed performance rank, for women (W) and men (M), separately.

[10] If we add the subjects whose payoff would be basically identical under a piece rate and a tournament incentive scheme (because their chance of winning the tournament is roughly 25 percent), then 40% of women and 45% of men have higher or basically identical earnings from a tournament payment.
[11] Similar results are obtained when NV consider the performance after the entry decision.

Figure 1b shows for each guessed tournament rank the proportion of women and men who enter the tournament. The more confident a participant is, the more likely the participant is to enter the tournament. However, gender differences remain significant among those who guessed they had the highest, or second highest rank (both of which comprise the most common beliefs). A man with the same belief as a woman still has about a 30 percentage point higher chance of entering the tournament. Regressions show that controlling for performance, gender differences in beliefs account roughly for a third of the gender gap in tournament entry.

To account for the effects of risk and feedback aversion on the decision to enter a tournament, NV assess the impact of those factors in Treatment 4 that are as close as possible to Treatment 3 while eliminating any tournament performance.

Treatment 4—Choice of Compensation Scheme for Past Piece-Rate Performance: Participants decide between the piece rate and tournament incentive scheme for their task 1 piece rate performance, where a tournament choice results in a payment only if the participant had the highest Treatment 1 piece rate performance in their group.

There are gender differences in the propensity to prefer the piece rate incentive scheme. However, the Treatment 1 piece rate performance, and beliefs about their relative piece rate performance, largely account for choices of women and men, with the remaining gender gap being economically small and not significant. Hence, in contrast to the situation where women and men decide whether to enter the tournament and then perform, eliminating any tournament performance, leaving only the impact of beliefs, risk and feedback aversion results in no gender gap.

The results from Treatment 4 suggest that gender differences in beliefs, feedback and risk aversion do not generate a gender gap, and hence cannot account for the gender gap in tournament entry in Treatment 3. Finally, in a regression on Treatment 3 tournament entry controlling for Treatment 2 performance and beliefs about relative performance as well as controlling for the Treatment 4 decision, significant gender differences in Treatment 3 tournament entry remain. NV attribute this residual to gender differences in competitiveness.

In terms of money maximizing choices, high performing women enter the tournament too little and low performing men too much. Note however, that the losses of a high performing female not entering the

tournament are substantial while the losses of a low performing male entering the tournament are lower since even in a piece rate incentive scheme their earnings would have been low. The result is that few women enter the competition and few women win the competition.

II.B Replication and Robustness of Women Shying Away from Competition

A series of papers presents treatments that introduce minor modifications to the original Niederle-Vesterlund design and find similar results, e.g., Cason et al. (2010), Healy and Pate (2011), Balafoutas and Sutter (2012), Balafoutas, Kerschbamer and Sutter (2012), Dargnies (2012), Kamas and Preston (2012), Price (2012), Cadsby, Servátka, and Song (2013), Niederle, Segal, Vesterlund (2013), Almås et al (2014), Buser, Niederle and Oosterbeek (2014), Dreber, van Essen, Ranehill (2014), Lee, Niederle, Kang (2014), Wozniak et al. (2014) and Sutter and Glätzle-Rützler, (forthcoming).

There are two papers using the NV design whose results are not completely in line with NV. Müller and Schwieren (2012) do replicate that women enter tournaments less than men. Among participants who are expected to have higher monetary earnings from the tournament than from the piece rate incentive scheme, women are 10 percentage points less likely to enter the tournament; however, this difference is not significant. A sole failure to reproduce the basic gender gap of NV is provided by Price (2010), who fails to find gender differences in preference for competition.

Furthermore, despite the use of very different designs, a series of other papers, some of which are discussed in more detail below, also identifies circumstances in which women, conditional on performance, enter tournaments less than men, e.g., Gneezy et al. (2009), Kamas and Preston (2009), Vandegrift and Yavas (2009), Ertac and Szentes (2010), Dohmen and Falk (2011), Booth & Nolen (2012), Cardenas et al (2012), Kamas and Preston (2012), Mayr et al (2012), Shurchkov (2012), Gupta et al. (2013) and Andersen et al. (2013).

The existence of a gender gap in tournament entry has stood the test of replication. I next describe the extent to which the limited impact of gender differences in other traits, such as beliefs, risk aversion or other regarding preferences hold up.

Beliefs

NV directly elicited beliefs on relative ability and used them as controls in the tournament entry decision. Papers using this approach typically show that men are more confident than women, that beliefs help explain the gender gap in winner-take-all tournament entry, and that a significant gender difference

remains when controlling for beliefs, e.g., and Grosse and Riener, 2010, Healy & Pate (2011), Balafoutas, Kerschbamer and Sutter (2012), Balafoutas and Sutter (2012), Dargnies (2012), Shurchkov (2012), Kamas and Preston (2012), Niederle, Segal, Vesterlund (2013), Almås et al (2014), Buser, Niederle and Oosterbeek (2014), Wozniak et al (2014), Sutter and Glätzle-Rützler, (forthcoming).

A few papers find that gender differences in beliefs can account for the gender gap in tournament entry. These are Kamas & Preston (2009) who examine a ranked-order tournament in which each rank receives a different piece rate, Cadsby, Servátka, and Song (2013) and Dreber, van Essen and Ranehill (2014).

An alternative to directly measuring beliefs as in NV and related studies is to change beliefs by providing participants with information about their relative performance. There are three studies that assess the impact of such information on tournament entry. In all of them information affects tournament entry. In Cason et al (2010) a significant gender gap remains, while in Ertac and Szentes (2010) and Wozniak et al (2014) information on relative performance eliminates the gender gap in tournament entry, though for Wozniak et al (2014) a significant gender gap in tournament entry remains among participants who are expected to have higher earnings from the tournament.

Risk Attitudes

To assess the role of risk attitudes, NV employed Treatment 4 that mimicked the Treatment 3 tournament choice as much as possible while eliminating any tournament performance and used decisions in Treatment 4 as a control when estimating gender differences in the Treatment 3 tournament entry decision. The same approach has been used by Dargnies (2009a), Healy and Pate (2011) and Niederle et al. (2013) who replicate NV's result that risk has only a minor impact on gender differences in tournament entry.[12]

Other indirect approaches confirm the minor role of risk attitudes in accounting for gender differences in tournament entry. Grosse and Riener (2010) have participants choose an incentive scheme for a number to be randomly drawn, rather than a real performance. Cadsby, Servátka, and Song (2013) have subjects choose between a piece rate and chance pay. In chance pay, the subject has a 25% chance to receive the tournament payment per each correct problem (which equals four times the piece rate payment) and

[12] Specifically, Healy and Pate (2011) and Niederle et al. (2013) find that when subjects decide whether to submit their Treatment 1 Piece Rate performance to a competitive payment scheme, there is no gender difference in choices when controlling for absolute as well as beliefs about relative Treatment 1 piece rate performance. Dargnies (2009a), Healy and Pate (2011) and Niederle et al. (2013) show that gender differences in tournament entry in Treatment 3 are not very affected when controlling for the Treatment 4 choice in addition to absolute as well as beliefs about relative Treatment 2 tournament performance.

otherwise receives no payment. They found no significant gender differences in choosing the chance pay over a piece rate.[13]

A more common approach in the recent literature has been to directly elicit risk attitudes through a series of incentivized lottery choices and use those as controls. The common finding is that the risk measure has no large impact on the gender gap in tournament entry, see Cason et al (2010), Kamas and Preston (2012), Almås et al (2014), Buser, Niederle and Oosterbeek (2014), Wozniak et al (2014), Sutter and Glätzle-Rützler (forthcoming).[14]

Other Regarding Preferences

While in a naïve design gender differences in altruism could impact gender differences in tournament entry, the experiment in NV was specifically designed in such a way that concerns for altruism play no role. This is because, by design, a subject's decision had no payoff externalities on any other subject. Nonetheless, it could be that specific other-regarding preferences correlate with a preference for competitiveness and account for the gender gap in those preferences.

The Treatment 4 choice of NV mimicked the Treatment 3 tournament entry choice in all aspects, including the role of other-regarding preferences, but excluding any tournament performance. Using the Treatment 4 choice, NV found no evidence that other-regarding preferences could account for the gender gap in tournament entry.

Several papers use various measures of other-regarding preferences and in general replicate that such measures do not play a large role in accounting for the gender gap in tournament entry, see Teyssier (2008), Kamas and Preston (2009), Bartling et al (2009), Dohmen and Falk (2011) and Almås et al (2014). Only Balafoutas, Kerschbamer and Sutter (2012) found that controlling for beliefs, risk and distributional preferences eliminates the gender gap in tournament entry.

The Role of the Task

[13] Eriksson, Teyssier and Villeval (2009) have participants select an incentive scheme using a standard experimental design in which agents do not perform in a task, rather pick effort using given cost functions and corresponding performance distributions. They find no gender differences in tournament entry when controlling for risk, which is negatively correlated with tournament entry.

[14] Balafoutas, Kerschbamer and Sutter (2012), show that risk attitudes correlate with tournament entry, but it is not clear how much it affects the gender gap in tournament entry. Dreber, van Hessen and Ranehill (2014) found that adding beliefs renders the gender gap in tournament entry in a math task insignificant. Controlling for risk further reduces the gender gap.

NV selected a five-minute addition task because it requires both skill and effort and because research suggests that there are no gender differences in ability on easy math tests.[15] However, participants could perceive the task as a stereotypical male task. Changing the task to a neutral or stereotypical female task could affect the gender gap in tournament entry through many ways. It could change the gender gap in beliefs about relative performance, affect the extent to which women and men care to receive information about their relative performance or simply affect the benefits (or costs) from performing in and winning the tournament.

While most papers use the NV math task, they all use a different word task, ranging from forming words using letters out of an 8-letter word to ordering five words in a sequence so they form a sentence. Almost all papers found gender differences in tournament entry in the math but not the verbal task, see Kamas and Preston (2009), Grosse and Riener (2010), Shurchkov (2012), and Dreber, van Essen and Ranehill (2014). Only Wozniak et al (2014) find gender differences in tournament entry both in a math and a verbal task, controlling for performance and beliefs; in fact, in their paper the task has no significant impact on tournament entry.

In summary, the existence of a gender gap in tournament entry in stereotypical male tasks persists after controlling for actual performance, beliefs about the relative performance, risk attitudes and other-regarding preferences. The treatments that consistently reduce and at times eliminate the gender gap in tournament entry is providing information on relative performance or changing the task to one in which women are believed to have an advantage.

II.C. Reducing the Gender Gap in Tournament Entry

The fact that high performing women do not enter competitions and hence don't win is disconcerting not only for those women, but perhaps also from a societal point of view. How can this gender gap in competitiveness be reduced? There are two major avenues of research to address this. The first could loosely be described as trying to understand what factors, such as hormones, age, socio-economic status, culture etc. generate the gender gap in competitiveness. This line of work can help understand whether it may be possible to "fix the competitiveness of women," or whether gender differences in competitiveness are immutable. The second approach can be loosely described as "fixing the institutions." This consists of

[15] While males often score better on abstract math problems, there is no gender difference in arithmetic or algebra performance. Women tend to score better than men on computational problems (see Hyde, Fennema, and Lamon [1990] for a meta-analysis of 100 studies on gender differences in math performance).

work that studies what institutions might be the most prone to enhance or reduce the impact of gender differences in competitiveness.

What factors generate the gender gap?

Hormones and MRI studies

Results on this dimension are quite mixed. Buser (2012) finds that women (in single sex groups) are less likely to enter competitions during the phase of the menstrual cycle when the secretion of progesterone and estrogen is particularly high. Wozniak et al (2014) considering mixed sex tournaments finds just the opposite. Whereas Apicella et al. (2011) find no relationship between self-selection into a tournament and current testosterone levels, while Hoffman and Gneezy (2010) take advantage of the fact that left-handedness is thought to be an indicator of prenatal testosterone and find that left-handedness increases competitiveness.

Another possible basis for gender differences in competitiveness is if men's and women's brains are wired in such a way that their brains react differently to differences in relative income. Dohmen et al (2011) investigate this, finding that activities in brain areas related to rewards react positively to higher absolute income and negatively to lower relative income. However, they did not find any gender differences in this respect.

Clearly no obvious consensus has been reached in understanding which biological differences between women and men can directly impact the gender gap in competitiveness.

Age

One problem in assessing how competitiveness changes with age is that participants at different ages will have had different life experiences which could affect competitiveness. For example, young employees at a firm may have diverse competitive backgrounds, though older employees may perhaps be more competitive (if they have been promoted in the company) or less competitive (if they still have not received a superior position). There are also potential selection effects in terms of who signs up for the experiment; e.g., visitors to a specific locale such as a mall may be comprised of young males who are less competitive than average (since the others work) and older people of average competitiveness. Both problems can be avoided by having representative population samples or having participants who are expected to be in the same environment for a wide range of ages, like, for example, children at a typical K-12 school.

Looking at children, Sutter and Glätzle-Rützler (forthcoming) examine compensation choices of 1,000 Austrian children and teenagers ages 3 to 18. Using a math task for the older participants and a running task for the younger subjects, they find a stunningly persistent gender gap in tournament entry. Despite there being no gender gap in performance on either of these tasks, males, independent of age, are 20 percentage points more likely to enter the competition than girls. Thus the gender gap in competitiveness is already present by age three.

Several papers consider subjects older than undergraduates. Mayr et al (2012) recruit around 500 people in an indoor shopping mall who perform in an NV-style design with pairwise competition. They find that 56% of men but only 36% of women choose to compete, a difference that remains relatively stable across age (see Figure 2 below). They find that beliefs regarding relative performance ranking (they ask for percentiles) do not vary with age and only account for a small fraction of the gender gap in tournament entry.

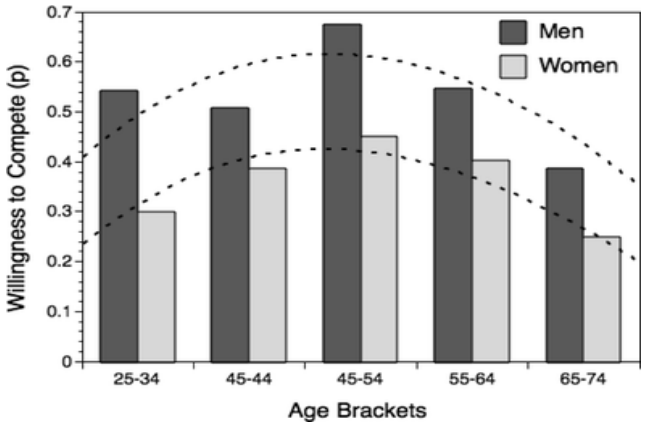

Figure 2: Competitive preferences of men and women across the life span. Dashed lines show the predicted age trajectories derived from probit regressions that model the probability of choosing competition as a function of gender, age, age-squared, gender by age and gender by age squared interactions.

Leibbrandt, Gneezy and List (2013) study villagers from a region in Brazil who either fish individually (they live close to the lake) or collectively (they live close to the sea). They find that the more experienced (or older) fishermen are, the more competitive they become in the individualistic society compared to those who live in the cooperative society. Women who live (but don't fish) in the individualistic or competitive society do not differ in their choice of tournament versus piecework pay, so overall they look more like men in the collectivist society. This suggests that the gender gap in tournament entry depends on past experience and either grows over time for men who work individually or remains constant (and perhaps shrinks a little) for men working in groups.

Charness and Villeval (2009) have two subjects decide between a piece rate and a tournament pay for an anagram task, where the person who chooses the tournament wins by default if the other chooses the piece rate. In this case choice depends not only on beliefs regarding performance but also on beliefs about others' choices, with the latter beliefs not elicited. Contrary to NV's results, they find no gender difference in tournament entry, nor any impact of age when conditioning on beliefs about relative performance. The exception to the latter is that retirees enter a little less than undergraduates.

The effects of age and work experience on competitiveness and its effect on the gender gap in competitiveness are clearly not completely resolved. However, it seems that in the Western world, gender differences in competitiveness are already present among children.

Socio-Economic Status

Almås et al (2014) consider Norwegian 9th graders and find that "children from low socioeconomic (SES) status are much less willing to compete than children from medium or high SES families, and this result holds when controlling for confidence, performance, risk- and time preferences, social preferences, and psychological traits. Second, family background is crucial for understanding the gender difference in competition preferences. .. [G]irls from well-off families are much less willing to compete than boys from well-off families, while we do not find a statistically significant gender difference in competitiveness preferences among children from low socioeconomic status." (p 1) They find that this difference in the gender gap in tournament entry conditional on SES is driven by boys with a low SES father, as they are much less willing to compete than other boys.

Bartling, Fehr and Schunk (2012) have 4-7 year old German children decide between a piece rate and a tournament pay, where each child knows they will compete against another randomly chosen child of the same sex and age. As expected (see subsection II.C.2 where I summarize the literature showing that the gender gap in tournament entry is much reduced when women compete against other women) the paper finds no gender differences in choice of compensation scheme in these single sex tournaments. However, for children from low socio-economic background, bad health implies less competitiveness. Furthermore, children who have more siblings, as well as those earlier in the birth order, are more likely to compete.

Big Five Personality Characteristics

Apart from hormonal differences and differences in risk preferences and confidence, researchers have tried to assess whether other personality traits can account for the gender gap in competitiveness. Most

prominently among these personality characteristics are the Big Five, which are openness to experience, conscientiousness, agreeableness, extroversion and neuroticism. (This reflects a growing interest among economists to understand the impact of the Big Five on labor market outcomes, see e.g. Borghans et al,2008).

Müller and Schwieren (2012) correlate an NV-style measure of competitiveness with the Big Five. They find that neuroticism significantly negatively correlates with tournament entry, a trait in which women score higher. They then show that controlling for neuroticism reduces the gender gap in tournament entry. Almås et al (2014) do not find that any of the Big Five have a significant impact on tournament entry, nor does controlling for any of the Big Five traits reduce the gender gap in tournament entry.

Clearly, this line of research is just beginning.

Priming

Cadsby, Servátka, and Song (2013) prime subjects concerning either gender/family or professional issues. Priming was administered through a questionnaire at the very beginning of the experiment. Gender or family-related concerns include questions like "what is your gender?" and "do you have children?", while professional concerns are questions like "what is your GMAT score?" and "What is your salary expectation upon the completion of your degree?"[16] They find that priming for professionalism significantly reduces the gender gap in tournament entry, whereas professional priming positively impacts women with respect to both their beliefs as well as their preference for tournament incentives.

Culture

A few papers have studied the impact of cultural differences on the gender gap in competitiveness. Gneezy, Leonard and List (2009) compare the choice of a tournament incentive scheme for patriarchal Maasai in Tanzania and the matrilineal Khasi in India. They find that among Maasai, behavior corresponds to that of the Western world: men are more likely to opt for a tournament incentive scheme than women. However, the gender gap reverses among the Khasi, where women are more competitive. Their design does not assess the impact of beliefs on these differences.

Cárdenas et al (2012) find a gender gap in tournament entry among Swedish children 9-12 years old, but not among Columbian children. "We used a similar scale to elicit how important the children consider

[16] In psychology, priming studies have recently come under scrutiny and are often hard to replicate, see e.g. Klein et al (2014).

competing against a boy and against a girl to be (0 = not at all important, 10 = very important). In both countries, boys rate competition as more important compared to girls (Colombia: $p = 0.009$, Sweden: $p < 0.001$). In Colombia, both girls and boys believe that it is more important to compete against a boy than against a girl (Girls: $p = 0.003$, Boys: $p < 0.001$). Girls in Sweden rate competing against a boy as being more important compared to competing against a girl ($p < 0.001$), whereas boys rate it as equally important ($p = 0.347$)." (p21)

II.C.2. Fixing Institutions to Reduce the Gender Gap

In addition to asking what makes men and women differ in their competitiveness, we can explore the role of institutions under which competition takes place. One possible way would be to provide participants, especially women, with feedback about their chances of winning the tournament. While this is easy to implement in the laboratory, outside the lab such information on relative tournament performance may be hard to come by. Another prominent institutional change is to make competitions more gender specific, either through single-sex competitions or a quota-style affirmative action.

Affirmative Action

Niederle, Segal and Vesterlund (2013) study whether a form of explicit affirmative action, namely a "soft" quota which basically makes the competition more gender-specific, can increase the number of high performing women that enter a tournament.

The setup is like in NV, using the same task, though now with groups of 3 men and 3 women (and mentioning the gender composition of the group explicitly). The first three treatments are just like in NV, only in the tournament the two highest performing participants receive compensation equal to three times the piece rate per correct answer (where two out of six win the tournament).[17] After the first three treatments (piece rate only, tournament only and the Choice treatment, an Affirmative Action Tournament is introduced, in which the two winners are chosen as follows. One winner is the highest performing woman; the second winner is the person with the highest performance in the remainder of the group. That is, a woman wins the tournament if either she is the highest performing woman, or if she has one of the two highest performances. A man only wins if he is both the highest performing man and among the top two performances overall.

[17] Since two out of six competitors win the tournament at a payment of three times the piece rate, a risk neutral participant who believes that all competitors are equally likely to win the tournament is, for a given performance, indifferent between the piece rate and the tournament pay, just like in NV.

While in this experiment gender was mentioned, and men outperform women, the results in terms of gender differences in the decision to enter a tournament mimic those of NV almost exactly.

Once affirmative action is introduced, women enter the affirmative action tournament at a much greater rate (both relative to the standard tournament and relative to payoff maximizing choices), while men drop out at a higher rate than predicted. Niederle, Segal and Vesterlund show that several channels are responsible for the change in the gender gap in tournament entry. First, there is an effect of purely mentioning affirmative action; women increase their tournament entry, though the effect on men is small. Second, the gender gap in beliefs vanishes when beliefs are about relative performance within gender rather than in mixed gender groups. However, even controlling for both those effects, women are more likely to enter the competition when they just have to outperform other women compared to when they compete against a mixed gender group. It seems that the gender specific competition alters the pleasure or fear of competition, or attitudes towards competition in general.

To assess the effect of affirmative action, note that the number of tournament entrants at or above a specific threshold level is almost always either the same or higher under affirmative action than under the standard tournament. The implications for this are as follows: Assume affirmative action was not announced but used "secretly", after subjects decided whether to enter the tournament. Consider the rule to hire at least one woman for every man, and being able to hire only among those participants who entered the tournament. Such an affirmative action rule would be very costly, in the following sense. When hiring participants, in order to assure having at least one woman for every man, at some early point the woman hired to fulfill the quota would be much worse than the best unemployed man. That is, many qualified men will be passed by to hire a woman to fulfill the quota. On the other hand, when affirmative action is announced, the pool of entrants to the affirmative action tournament is such that for almost all performance thresholds there are as many women as men who perform at that threshold or higher. That is, in order to hire a woman to fulfill the affirmative action requirement, the men that are passed by are "only" of the same, but not of a higher performance level than the woman is. In that sense, affirmative action is not costly when it is announced in advance, and when women had a chance to adjust their tournament entry decision accordingly.

To summarize, in an environment without discrimination, but in which the playing field is not level in that there is a gender gap in tournament entry, a quota-like affirmative action setting in which women

only have to compete against other women can reduce the gender gap in tournament entry and, in particular, increase the fraction of high performing women who enter a tournament.

This experiment and the results have been replicated (though published earlier) by Balafoutas and Sutter (2012), who in addition consider another affirmative action device, namely preferential treatments in which the performance of women is increased by one or two problems respectively. Those also reduce the gender gap in tournament entry.

Several papers consider single sex tournaments, and contrast the results with those obtained in mixed-sex tournaments. In Sutter and Rützler (2010), Austrian children ages 9-18 choose whether to enter a tournament in an NV design, when either in a group of two boys and two girls, or in a single sex group of four children. They do not find that the gender composition of the group affects choices. Likewise, Gupta, Poulsen and Villeval (2013) consider pairwise competition where each participant receives an assigned first name that corresponds to their gender, which can be observed by their opponent. If only one subject chooses the tournament, that subject wins with certainty. Otherwise, the subject with the highest performance wins. The main result is that men enter the tournament more than women. Furthermore, everyone believes that men are more likely to enter the tournament. Oddly enough, choices of tournament entry do not depend on the gender of the opponent. In contrast, Booth and Nolen (2012) consider children just under 15 and find that girls in single sex groups (of four) enter the tournament more than girls who face at least one boy. Girls are not affected by how many boys they face conditional on facing at least one, and boys do not care about the gender of their competitors.

Finally, several researchers gave participants the option to choose the gender of their competitor. In a math task, both Gupta, Poulsen and Villeval (2013) and Grosse and Riener (2010) find that men and women prefer to compete against women. In addition, Grosse and Riener (2010) find that for a verbal task, half the participants prefer to compete against women and half against men.

In summary, the majority of the evidence to date indicates that single sex tournaments, or quota-like affirmative action tournaments, reduce the gender gap in tournament entry without seriously diluting the quality of the resulting entrants. Given that some papers do not find that these alternative tournament structures do not increase entry, more work is needed to confirm that the positive effect of single-sex competitions is robust.

II.D. Performance in Tournaments

The decision to enter a tournament is not the same as the decision (or ability) to provide high effort once in a tournament. The literature discussed so far studied the decision whether to enter a tournament, which we could think of as the extensive margin, where large gender differences were found. The first economic experiment on gender differences in competitive attitudes tested for gender differences on the intensive margin;that is, whether women and men react differently when forced to perform in a competitive environment. To test for gender differences in performance once in a tournament, it is crucial to find a real effort task in which performance is not only statistically but economically significantly affected by the incentive scheme and hence, presumably, by effort.[18]

Gneezy, Niederle and Rustichini (2003), henceforth GNR, conduct an experiment at the Technion in Israel, a high profile technical university. Women and men solve mazes on the internet for 15 minutes under various incentive schemes. In each session (apart from single sex sessions) there were always exactly three female and three male participants. They could see each other and determine the gender composition of the group, though no mention of gender was explicitly made. Each treatment had 30 women and 30 men, where no one participated more than once. At the end of the experiment participants are informed only of their own earnings.

Piece Rate Treatment: In a first Piece Rate treatment participants receive about $0.5 per completed maze. Subjects have 15 minutes to solve as many mazes as they can on the internet. The task is the same for all treatments.

The average performance in the piece rate treatment for men is 11.23 mazes, while it is 9.73 for women. The difference of 1.5 mazes is not statistically significant.

Tournament Treatment: In the tournament treatment, only the highest performing participant of the 3 men and 3 women receives a payment for each solved maze that equals six times the piece rate payment, $3 per maze. In case of a tie, the winners shared the payment equally.

The average performance of men is 15, which is significantly higher than in the piece rate. The average performance of women is 10.8, not significantly different from the piece rate performance. The gender

[18] The task of adding up five two-digit numbers for five minutes does not fit that bill. Rather, this seems to be a task where changes in the incentive scheme don't lead to large changes in performance, though there may still be (hard to observe) changes in effort.

gap in tournament performance of 4.2 is significant. Most importantly, the gender gap in tournament performance of 4 is significantly higher than the gap of 1.5 in the piece rate performance.

The significant increase in the gender gap in mean performance when moving from a piece rate to a tournament scheme could be the result of two changes. First, the payment scheme became more competitive. However, the tournament payment is also more uncertain, compared to a piece rate scheme. And indeed, there is a large literature on possible gender differences in risk aversion (see Section VII).

One option to assess the potential impact of (possible) gender differences of risk aversion on changes in performance as the incentive scheme becomes more competitive is to elicit risk attitudes of women and men. One problem with this direct approach is that the magnitudes of gender differences on risk aversion could depend on the specifics of the environment used to measure them. Because the object of interest is a real effort task, it is not clear what the relevant range of lotteries should look like, since e.g., costs of effort to solve mazes cannot be easily assessed. It may prove even more difficult to extrapolate to effort choices under tournament incentives as it could be that there is additional aversion when there is uncertainty about the relevant lotteries in play, as participants may not be clear about the risk they are facing, or alternatively, the chances of winning the tournament for a given effort level.

Instead, GNR opt for a direct approach to assess whether women and men react differently to uncertainty in payments which are inherent in tournament. Subjects have to perform in a random pay treatment which is similar in terms of uncertainty to the competitive pay treatment, though without any competitive aspect.

Random Pay Treatment: Only one participant of six receives compensation equal to the tournament payment, six times the piece rate scheme, that is, $3 per maze. However, the person who receives payment – the "winner" – is chosen randomly as opposed to the number of mazes solved.

The results show that average performance of women and men is just like in the piece rate treatment. Furthermore, differences in performance between the random pay and the tournament treatment are just like those between the piece rate and the tournament treatment. GNR conclude that the significant change in the gender gap in mean performance when moving from the piece rate to the tournament can indeed be attributed to changes in the competitiveness of the incentive scheme and not to changes in the uncertainty of payment inherent to participating in a tournament.

There remain four classes of explanations for increased performance of men versus women conditional on performing in a tournament: First, it could be that women *cannot solve more mazes* without incurring very high costs.[19] Second, it could be that women simply do not perform well under competition (in general), because they do not like to, or *cannot compete*. Third, it could be that *women can compete well, but not against men*. This could be because women perform somewhat less well in this task than men, and hence they may decide not to increase their performance by not increasing their effort. In addition, this could be driven by women's beliefs that they perform less well than men.[20] Finally, the performance that needs explaining is perhaps not that of women, but that of men. It could be that *men compete too much*. This could be because men receive a direct utility boost from winning a tournament, or perhaps because men really like to compete and win especially when there are women around.

To assess the validity of each of those hypotheses, GNR run a treatment where six participants compete in single sex tournaments.

Single-Sex Tournament Treatment: In the single-sex tournament treatment groups are either comprised of 6 men or 6 women. The incentives are just like in the tournament treatment, that is, the highest performer, the winner, receives six times the piece rate payment per maze, and everyone else receives no payment.

The performance of men (average of 14.3) is not significantly different from those in mixed tournaments and significantly higher than the one in the piece rate and random pay. That is it does not appear that men compete only when they're competing against women. More importantly, women do seem to react positively to competitive incentive schemes in the single sex environment. Their performance in the single sex tournament (on average 12.6) is significantly higher than in either the piece rate or the random pay treatment.

To assess whether women respond to competitive incentive schemes in single sex groups as much as men do, GNR compare the average gender differences in performance across treatments. The gender gap in men's performance is 1.5 in both the piece rate and the random pay treatment, and is 1.7 in the single sex

[19] Other reasons include that women are not sensitive to the incentive scheme at all, and always perform similarly. It could also be that women see not increasing the performance in the tournament compared to the random pay treatment like contributing to a public good. If all participants have a 1/6 chance of winning, then they would be better off if all wouldn't increase performance.

[20] An alternative hypothesis is that women believe that there is a stereotype that they should not be able to perform well in this task, or in competition against men, and hence they may suffer from stereotype threat that provides an additional source of anxiety while performing the task and yields higher instances of "choking under pressure."

tournament. However, the gender gap in mean performance is 4.2 in mixed tournaments, significantly higher than in the single sex tournaments (p=0.08), and in all other treatments (see Figure 3a).[21] This suggests that the third explanation is the most likely: women can compete well, just not as well against men.

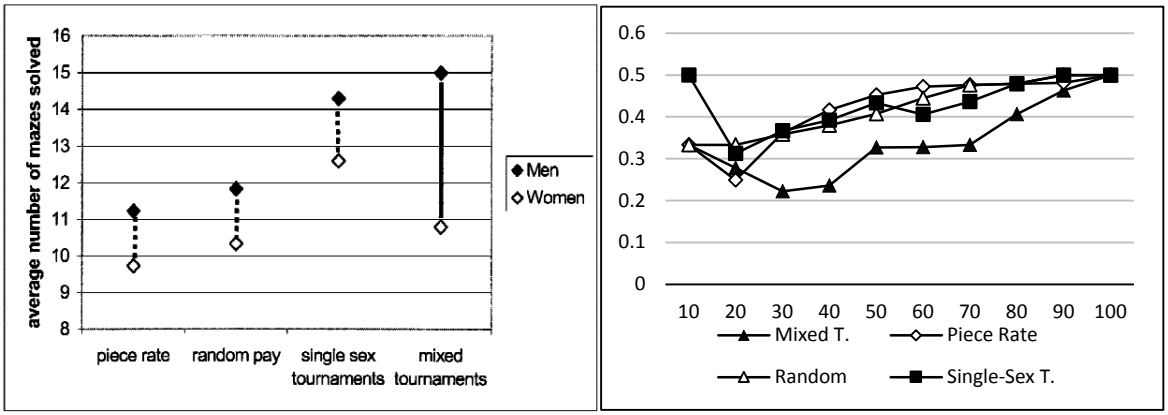

Figure 3: (a) Averages Performance of the 30 Men and 30 Women in Each of the Treatments (b), The Proportion of Women above Each Performance Decile for Each Treatment

The results summarized in Figure 3a show that there is a significantly larger gender gap in mixed competitive environments compared to both non-competitive payment schemes and a single sex competitive environment.

To describe how average experience from Figure 3a translates into individual behavior, Figure 3b shows for each decile the proportion of women above that performance decile, starting from the top 10%. The figure shows, for example, that the fraction of women among the top 40% of performers varies a lot. In the noncompetitive treatments and in the single-sex tournament, among the top 40% of performers, about 60% of them are male and about 40% are female. The fraction of women among the top performers, for any decile, is basically the same whether single-sex tournaments or non-competitive treatments are used. Hence if tournaments were run in single-sex groups, one may falsely conclude that men and women have similar responses to competition. However, running mixed-sex tournaments significantly decreases the fraction of women with a performance in the top 40% from about 40% to 24%. Women are less represented among top performers in mixed tournaments compared to any other incentive scheme when considering any performance above any specific decile but the very highest. Thus mixed-sex competitions result in a decrease in the fraction of women among top performers.

[21] A single sex piece rate treatment confirms that women perform highly when competing against women, and not merely when there are no men in the group of participants (though note that the experimenter was always male, Uri Gneezy).

There are several papers that investigate the same issues as in GNR under a variety of conditions. The first of these is Gneezy and Rustichini (2004). They consider performances of 10 year olds in competitive and non-competitive environments. Children first run 40 meters separately, and then are matched up so that the two fastest children run against each other and so on. They find no initial gender difference in speed. In competition, boys increase speed on average, while girls become slightly slower, a difference that is significant. Furthermore, a second group of children who run a second time in a non-competitive way show no significant gender difference in speed. This suggests that boys and girls did not simply become differentially tired the second time, but rather seem differentially affected by the competition. Both boys and girls improve the most when competing in mixed groups, boys competing against boys also run faster, but girls competing against girls slow down somewhat. That is, with respect to running speed boys improve more than girls when moving from non-competitive performances to a competition. However, results on running speed fail to replicate that girls compete more against female opponents than male opponents.

One thing to note from this study is that children physically ran against each other; in this setting one feedback is very salient – the winner of the race is easily determined. Therefore, one variable of interest is not how fast boys and girls were, but whether they won the competition. The initially faster child wins the competition 10 out of 17 times (59%) in boy groups, and 6 out of 12 times (50%) in girl groups. That is, the initially faster runner is equally successful at winning in homogenous groups compared to the initially slower child. In contrast, in mixed groups, 8 out of 11 times the boy won when he was initially slower (73%) and 15 out of 18 times (83%) when he was initially faster. Viewed this way, the results suggest that girls do not compete well against boys. Girls have a higher chance of winning whenever running against another girl than against a boy, independent of whether they were initially faster or slower. On the other hand, boys have an easier time winning against a girl than another boy, irrespective of whether he was initially slower or faster. In summary, running speed replicates that boys improve their performance more than girls when moving from a non-competitive to a competitive environment. This is replicated when considering who wins the competition. In addition, a specific girl has a higher chance to win the competition if she competes against another girl rather than a boy.

Several subsequent papers have employed an approach similar to Gneezy and Rustichini (2004). Cotton, McIntyre and Price (2013) have American third, fourth and sixth graders repeatedly perform in pairwise competition against a known, though in each case against a different, classmate. Using a math task, they find a gender gap in performance in the first, but not in subsequent tournaments. This is driven by low

ability males initially over-performing, and high ability females initially under-performing compared to later rounds.

Cardenas et al (2013) have 9-12 year old children first perform under a piece rate scheme, then an assortatively matched (by piece rate performance) pairwise tournament. They find no significant increases in the gender gap in performance among Columbian children, and in two out of four tasks Swedish girls actually increase their performance in competition more than boys do (in rope jumping, where girls are already much better in the piece rate, and in math, where girls increase their performance but remain worse than boys who don't change their performance). There are no controls as in Gneezy and Rustichini (2004) to assess the effects of learning or fatigue on the observed changes in performance.

Considering designs closer to GNR, Gunther at al (2010) and Shurchkov (2012) consider stereotypical male tasks – mazes and a math task, respectively. Both papers find that while men and women perform equally under a piece rate scheme, men outperform women in the tournament. On the other hand, Freeman and Gelber (2010) have participants solve mazes in groups of six, randomly formed among participants, first under a piece rate, and then various tournament schemes. While they find that participants solve more mazes in round 2, and do so differently depending on the exact round 2 payment scheme, they find no significant gender difference in round two, given round one performance. Bracha and Fershtman (2014) have participants allocate time within 10 minutes between a mindless filing task – deciding whether a number is odd or even – and a more challenging sequence task, where subjects are given three numbers and have to fill in the missing fourth number. They find that performance of women and men are not affected by the payment scheme, piece rate or tournament, and neither is the gender gap in performance.

Finally, Dato and Nieken (2014) consider pairwise competition where participants, in addition to deciding how hard to work, can also decide to sabotage the opponent. They find that men sabotage more, and hence win more often, but, because sabotage is costly, don't have higher earnings. In a treatment that controls for the gender of the opponent, they show that both men and women believe men sabotage more, though gender differences in beliefs about others' sabotage propensity cannot fully account for the gender gap in sabotage.

Effect of Different Tasks

Gunther at al (2010) and Shurchkov (2012) consider a stereotypical male task – mazes and a math task, respectively – and a stereotypical female task, a word task – forming words that start with a specific letter

and forming words out of a given set of letters, respectively. Both papers find that for the stereotypical male task, men outperform women in the tournament but perform equally under a piece rate scheme. No such gender effects are found in the verbal task. Shurchkov (2012) also considers a "low pressure" environment, where each task lasts 10 instead of two minutes, and finds that there is no significant gender gap in the math task, though women perform higher than men in the verbal task. More recently, Cotton, McIntyre and Price (2013) find a gender gap in a first tournament performance for a math, but not for a verbal, task.

The Role of Beliefs

To assess the role of beliefs and feedback on gender differences in performance, Kuhnen and Tymula (2012) have participants perform in 18 rounds in which, before each performance, participants learn whether they will receive information about their relative performance with a 0%, 50%, or 100% chance. The feedback provides information on one's ranking and the scores of all group members. Participants have no other incentives to perform. The authors find that participants have a higher performance when there is a positive chance to receive feedback that is, it appears that solely providing relative performance feedback acts may activate competitive attitudes. Furthermore, the number of men in the group affects the productivity of women. The women's expected and actual rankings are worse and their absolute performance is lower when there are more men in the groups. Men, however, are not affected by the gender composition of the group. Thus, the results mirror those of GNR if performance under the threat of feedback is akin to a competition.

Summary

Gender differences in performance increase when moving from a competitive to a non-competitive incentive scheme, a result that has been replicated several times. This implies that a woman with ability and performance in a non-competitive piece rate scheme comparable to that of a man will have an inferior performance to that man if the performance is measured in a competitive environment where women and men compete against each other. Put differently, performances under mixed-gender competitions may not equally reflect underlying abilities of women and men. To obtain this result it is crucial to consider tasks in which performances change when participants perform under a piece rate scheme or a tournament scheme. For example, the task in NV of adding up five two-digit numbers for five minutes does not fulfill this requirement.

Occasionally no changes in the gender gap in performance have been observed and in one paper, Cardenas et al (2013), females actually increase their performance more than males on one of the tasks as

the incentive scheme became more competitive. Just as in the literature on tournament entry, more research is needed to assess the extent to which differences in task characteristics can help account for variations in the change in the gender gap in performance when moving from a non-competitive to a competitive environment. Finally, there is some evidence that in cases in which mixed-sex competitions harm the relative performance of women compared to non-competitive treatments, women are as competitive as men, as long as they don't have to compete against men.[22] This would imply that affirmative action in the form of quotas may result in performances of women that are more in line with their underlying ability and comparable to that of men. Once more, this finding is still to be considered a hypothesis rather than a firmly established result.

Linking Tournament Entry and Performance in Tournaments

We can draw a parallel between gender differences in choice of incentive schemes and gender differences in performance across incentive schemes, exemplified in the papers by Niederle and Vesterlund 2005 and Gneezy Niederle and Rustichini 2003. Let "compete" mean in GNR to perform highly in a competitive environment, and in NV to enter a competitive environment. Then both papers show women do not "compete" against men. GNR also analyzed single sex tournaments and found that women compete against other women just fine. The corresponding result has been found in the literature on tournament entry, with research showing that women who do not enter tournaments against men do, in many cases, enter tournaments where they only have to compete against women.

Note that NV aimed to rule out the effects of GNR by using a task in which performance was not affected by the incentive scheme. Future research should provide a link between the two approaches and assess whether women whose performance in a competition does not reflect their piece rate performance tend to shy away from competition.

Together the results following NV and GNR show that women are less competitive than men. The last decade saw lots of work in the experimental literature on gender differences in competitiveness. Recently, this literature has also found its way into field evidence as well as field experiments.

II.E. Field Experiments on Gender Differences in Competitiveness

[22] In terms of the positive effects of affirmative action on performance, Calsamiglia, Franke and Rey-Biel (2013) find positive effects of affirmative action in the performance (effort) in tournaments that have children with prior experience in the task at hand compete against children with no prior experience. See also Schotter and Weigelt (1992) for experiments in an abstract setting.

Much of the field evidence consists of data that had no variation imposed by an experiment. I refer the interested reader to Niederle and Vesterlund (2011) for an overview of this literature. I will only detail two such field studies to provide a flavor of the existing results, and then survey field experiments on gender differences in competitiveness.

Ors, Palomino and Peyrache (2013) consider performance in a competitive entry exam to a very selective French business school (HEC), where slightly more than 10% of applicants are accepted each year. Men perform significantly better than women at this admission contest. A couple of years earlier, those same applicants took the very stressful, but non-competitive (i.e. graded on absolute performance) high school exam. In the high school exam, female HEC applicants performed significantly better than male HEC applicants. Similarly, for students who were admitted and accepted to HEC (a select sample), women performed better than men in the first year of the program though only in the nonmathematics-oriented classes. In both of these cases performance, while measured under a stressful environment, is graded more on an absolute level and not solely on a relative level. The fact that women perform worse than men, especially under the competitive entry exam, suggests that competitive exam scores reflect skills as well as responses to competition. The findings therefore corroborate the experimental results of an increased gender gap in performance when performance is measured under a competitive rather than a non-competitive incentive scheme.

Morin (forthcoming) studies grades at the University of Toronto and exploits the fact that an educational reform resulted in a "double cohort" since in one year two cohorts, namely 12^{th} (the new final high school year) and 13^{th} (the previous final high school year) graders competed for grades which are based on a curve. He finds that the gender gap in grades, controlling for background such as pre-university grades, significantly increased in the year of the double cohort, the year in which there was fiercer competition for grades. This gender gap generated by increased competition is present at all performance levels.

In a field experiment, Lavy (2013) considers teachers who are paid cash bonuses based on improvements in the test scores in their class, where payment depended on relative improvements within a specific field and school. This resulted in a variation in the gender composition of competitors. Lavy does not find gender differences in improvement, nor does he find that the gender composition of the group of competing teachers influences outcomes. Delfgaauw (2013) considers team competitions and finds that sales competitions have a large effect on sales growth, but only in stores where the store's manager and a sufficiently large fraction of the employees have the same gender.

More recently, Flory, Leibbrandt, and List (2010) conduct a field experiment on gender differences in tournament entry. They randomly offered job-seekers compensation schemes that varied in the degree of competition. They find that women are relatively less likely to apply for a job with a competitive payment scheme than men.

II.F External Relevance of Competitiveness

It is comforting that gender differences in competitiveness can be replicated using samples other than undergraduate students and using tasks that last longer than several minutes. However, the (experimental) field evidence, in general, cannot directly assess the role of gender differences in competitiveness in accounting for gender differences in education or labor market outcomes. There are two reasons for that. First, any experimental evidence, be it in the laboratory or the field cannot experiment with education and work decisions on a scale that mirrors those of a general population. For example, field evidence on Austrian farmers assessing their reaction to competitive incentive schemes may have only limited applicability to even the general population in Austria. It may not even help assess the importance of competitiveness for observed wage differences or work choices of Austrian farmers, if the experimenter was not able to manipulate those choices. Second, field evidence, while often compelling, may be hard to come by when considering existing data. There are of course always exceptions.[23]

To assess the external relevance of gender differences in competitiveness it will be useful to find, or create, data bases that combine a good measure of competitiveness with field outcomes. This is exactly the aim of Buser, Niederle and Oosterbeek (2014), henceforth BNO. They investigate gender differences in education choices of 9th graders in the Netherlands. Specifically, children who go to the pre-university school in the Netherlands share in the first three years, grades 6-9, the education experience and are somewhat randomly assigned to classes. At the end of 9th grade, children select one of four possible academic tracks, best described as Mathematics, Biology, Economics or Literature, for their last three years of high school. This is the ordering of how math and science intensive the education in each track is. It is also the ordering of how prestigious the tracks are, where the best students go, and how likely they are to actually go to university later.

BNO conduct experiments with every ninth-grader from four schools in and around Amsterdam. The in class experiment consists of a slight variation of NV: children add up sets of four two-digit numbers for

[23] For example, Ors, Palomino and Peyrache (2013) are able to exploit a natural variation in the competitiveness of various exams. They show that women perform less well compared to men in the very competitive entry exam for a selective French business school (HEC), though those same women and men performed similarly (in fact women dominated) in the non-competitive, though still very stressful, national high school graduation exam.

three minutes, first under a piece rate scheme, then a tournament scheme, where they compete against three classmates who were randomly selected by computer after the end of the experiment. Round three implements the NV choice of compensation scheme, tournament or piece rate. A participant who selected the tournament would win if her new Round 3 performance exceeded the round 2 performance of her three competitors. BNO assess the students' beliefs about their relative round 2 tournament performance and their risk attitudes. BNO also measure each student's beliefs about their subjective mathematical ability, as well as how they rank the four study tracks in terms of "Which track do the best students pick?"

The roughly 400 children in BNO exhibit behavior that mirrors those of all children of the Netherlands concerning their education choices. Children to a large extent agree that the order of prestigiousness of academic tracks corresponds to their math and science intensity, and children who chose more prestigious tracks have a higher GPA. To assess academic track choices of students, BNO control not only for objective academic ability as measured by grades, but also subjective academic ability as measured by the students' beliefs about how easy mathematics is for them as well as how good they are in math compared to their peers. Ordered probit regressions show that being female accounts for 15 percent of the distance between the most and the least prestigious tracks controlling for grades. To provide a measure of the importance of the female dummy, note that a one standard deviation in GPA only accounts for 11 percent of the gap between the most and least prestigious track.

The children also exhibit tournament entry decisions that mirror the results in NV: Controlling for performance, girls are about 23 percentage points less likely to enter the tournament. Slightly over 30 percent of this gender gap can be explained by gender differences in confidence. Risk attitudes, whether measured by a lottery choice or a simple questionnaire item significantly predict tournament entry, but reduce the gender gap in competitiveness only by a small amount once confidence is controlled for.

To assess the importance of competitiveness on academic track choice, note that the binary variable of tournament entry (controlling for performance in the experimental task, as well as grades and the subjects' beliefs about their academic ability) accounts for 18 percent of the gap between choosing the least and most prestigious track (compared to 15 percent for being female). That is, a student's competitiveness is a slightly better predictor of academic track choice than gender. When controlling for both gender and competitiveness, the gender difference drops from 15.4 to 12.3 percent, a statistically significant change. That is, 20 percent of the gender gap in choices can be accounted for by gender differences in competitiveness.

Since tournament entry is partially explained by confidence (the belief about the guessed tournament rank) and risk aversion, BNO assess the extent to which they drive the importance of tournament entry on study track choices. Figure 4 shows for each track the average net competitiveness of boys and girls that chose that track. Net competitiveness is the residual of a regression of tournament entry on the measures of performance in the experiment, the guessed rank and the risk measures. For each gender, more competitive students select more prestigious tracks. This provides a first indication that the effect of competitiveness on study track choices is not due to the impact of risk attitudes and beliefs about relative performance (or confidence) alone.

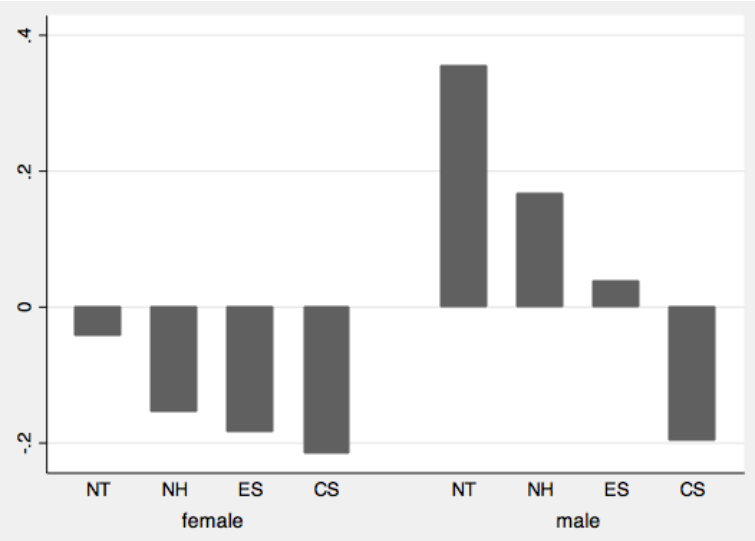

Figure 4: For each gender the average net competitive attitudes of subjects that chose a given study profile: NT: Mathematics, NH: Biology, ES: Economics and CS: Literature. The net competitive attitudes are the residual of a linear regression of tournament entry that also includes guessed rank and risk measures next to the performance measures in the experiment.

BNO provide ordered probit regressions that control for the students' risk attitudes as well as their confidence. Tournament entry in the experiment reduces the gender gap in track choices by 16 percent (compared to 20 percent without confidence and risk controls). Together, competitiveness and risk measures reduce the gender gap in track choices by 33 percent. The effect of risk only is 16 percent, and that of competitiveness only is 20 percent, hence, the combined effect (33 percent) is 92 percent of the sum of the separate competitiveness and risk effects. This suggests that competitiveness and risk attitudes have almost orthogonal effects on the gender gap in track choices. Controlling for the guessed tournament rank actually increases the gender gap in choices.Controlling for competitiveness, risk and confidence reduces the gender gap in choices by only 26 percent.

BNO then argue that tournament entry is indeed a measure of the students' competitiveness rather than an (additional) measure of the students' perceived math ability, their actual math ability or their preference for math.

Two other papers address the external relevance of competitiveness on education choices. Zhang (2012a) conducts NV style experimenters with middle school children from Ninlang County in China. Zhang also observes the students' decision to take a very competitive high school entry exam. Controlling for test scores on a previous exam, students who are more competitive are more likely to take this entry exam. However, Zhang finds no gender gap in tournament entry, or in entry rates of taking the exams. Note that Zhang (2012 b) does find a gender gap in tournament entry for ethnic minorities among high school children from the same area.

The second paper, Reuben, Wiswall and Zafar (2013), finds that among NYU students, competitiveness (as well as overconfidence, though not risk aversion) is positively correlated with earnings expectations. Furthermore, about 18% of the gender gap in earnings expectations can be accounted for by gender differences in competitiveness and overconfidence. Like in BNO, this paper finds that the experimental variables are important even when including various control variables such as test scores and family background. However, while expected earnings are related to major choices, Reuben, Wiswall and Zafar do not find that the experimental variables are related to choice of major. Note, however, that there is no obvious ranking of majors as in BNO or in Zhang, where in BNO more prestigious profiles were the more math intensive ones, and in Zhang, taking a difficult test was more prestigious than not taking it.

Overall, there is clearly more work to be done to confirm the external relevance of competitiveness as an independent trait, a trait that can account for education choices and also other labor market choices. Similarly, confirming the initial evidence that gender differences in competitiveness can account for a substantial gender gap in such choices will be important.

To conclude, gender differences in competitiveness provides a model for how new laboratory findings found their way into more mainstream economics. This has been achieved by first an important phase of experimentation with lots of replications and checks for the robustness of results, as well as trying to understand how competitiveness differs from other traits such as confidence and risk aversion. The most important step has been not only to conduct field experiments, but to create data sets that include both laboratory as well as field measures. While this last step is somewhat new, it is helpful to show that

gender differences in competitiveness can help account for gender differences in education and hence presumably labor market outcomes.

III. Gender Differences in Selecting Challenging Tasks and Speaking Up

Gender differences in competitiveness and preferences over incentive schemes have received a lot of attention. This literature has also been successful in showing that competitive attitudes predict, for example, education choices of students. Specifically, Buser, Niederle and Oosterbeek (2014) showed that students who are more competitive select more math intensive and more prestigious study profiles. However, there has been little work in assessing directly whether women and men differ in whether they select challenging or difficult tasks. For example, male and female undergraduates differ significantly in the rate with which they select to be STEM and economics majors. Therefore, it may be important to understand whether women, in general, shy away from challenging tasks, and whether or what institutional changes can affect these choices.

In this section I first review papers that tackle gender differences in task choices where tasks can be ordered in how challenging they are. I then discuss which, if any, institutional changes may affect those choices. I then present papers that address whether women may be more reluctant to speak up and put themselves forward.

III.A. Gender Differences in Task Choice

A first indication that women shy away from challenging tasks compared to men can be seen from a final treatment in Gneezy, Niederle and Rustichini (2003). Women and men could decide upon the task difficulty (mazes from level 1 to 5), where level *x* would be remunerated at *x* shekels for each completed maze (with 4 shekels about $1). All participants only saw one level 2 maze before making the decision. Men chose significantly more difficult levels than women. The average choice is 3.4 for males and 2.6 for females, a significant difference. There are, however, two limitations in interpreting this experiment. First, neither the authors nor the participants knew what the optimal choice would be for someone who has a high ability in solving mazes compared to someone of low ability. It could be that for everyone earnings are highest at task difficulty 3. Second, even if it were true that higher performing participants have on average higher earnings from choosing harder mazes, participants did not know it.

Another early work showing gender differences in task choice is by Huberman and Rubinstein (2001). Their abstract describes the setting of their experiment very well: “We asked subjects to self-select into one of two contests, "coin" or "die." The winner in each of the contests is the person with the most correct

guesses of 20 coin flips or 20 rolls of a die, respectively. The majority of subjects reported that they believed that most people would go to the "coin" group. They were correct. Although the right action under this belief is to choose "die", most people chose to be with the majority. Both men and women tended to make this mistake, but women's propensity to err in this particular experiment was stronger. This is puzzling as our overall impression (based on preliminary experiments which were not documented scientifically) does not support the existence of gender differences in other strategic situations."

The attraction of a pure "guessing" task is that neither one's own ability, nor the ability of others matters for the task. The main result of their experiment is that "Women behave less "rationally": Only 15% of the women vs. 35% of the men act optimally on their beliefs (including wrong beliefs)!" (p 6).

Bracha and Fershtman (2014) have participants allocate time within 10 minutes between a mindless filing task – deciding whether a number is odd or even – and a more challenging sequence task, where subjects are given three numbers and have to fill in the missing fourth number. They find that under a piece rate scheme, women spend less overall time on the challenging task, more time per question, though with an overall similar success rate per question. Under a tournament incentive, both women and men reduce the time spent on the challenging task, though their overall performance is the same as under the piece rate. Interestingly, the success rate of women declines significantly, and significantly more so than that of men, when performing in the tournament compared to the piece rate. This effect is mostly present in the last three minutes of the 10 minute performance.

The last paper I describe by Niederle and Yestrumskas (2008) combines two of the previous approaches. First, one task is "objectively" harder than another (as in Gneezy, Niederle, Rustichini, 2003). Second, as in Huberman and Rubinstein (2001), the environment is one where both the experimenter, as well as the subjects, know what task is payoff-maximizing for whom. Niederle and Yestrumskas (2008) studies whether women shy away from difficult and challenging tasks more than men, and what institutional changes can alleviate these gender differences in choices. The objective is to assess whether a woman and a man of the *same performance level* make different choices. To that aim, Niederle and Yestrumskas (2008) create an environment such that participants can be divided into two groups, one of which, given their ability, has higher earnings from the challenging task, while the other has higher earnings from the easier, non-challenging task.

Specifically, the task is solving mazes on paper for 10 minutes. They have an easy task (easy mazes at \$0.5 per maze) and a hard task (hard mazes, with a kinked incentive scheme: \$0.25 for each of the first

four mazes and then $3.50 for each additional one). This creates two tasks where it is the case that participants can be divided into two performance levels, high and low. The paper shows that high performance level participants have higher earnings from the hard task, while low performance level participants do so from the easy task. The reason is that low performance level participants simply do not solve sufficiently many hard mazes to reach the steep part of the piece rate incentives. Furthermore, each participant's performance level can be identified by their first performance in the easy task where approximately the top 40% performers in this first easy task are of high performance level, specifically all participants who solved 11 easy mazes or more.

In each treatment, participants first perform the easy task. This allows the experimenter to determine the performance level of each subject. Participants then choose the task difficulty for the next two tasks. When deciding about the task difficulty for the last two rounds, participants always know that a high performance participant, one who was among the top 40% of all participants in the initial easy task, has an expectation of higher earnings from a subsequent hard task, while others have higher expected earnings from a subsequent easy task. That is, participants, while not knowing their own performance level, know that the labels hard and easy task were meaningful. Participants were paid for their first performance, and one of the two subsequent performances.

A first group of subjects, who determine the validity of the high and low performance classification were asked about their relative performance, and show no gender differences in beliefs about their relative performance.

The first main treatment *Choice* has participants choose the performance level for the next two tasks, after their initial task 1 performance. While every single high performing man chose at least one hard task, only 65% of women did. On average, 86% of high performing men chose the hard task compared to only 57% of women. On the other hand, 88% of low performing men chose one hard task compared to only 58% of low performing women. Low performing men chose the hard task with 70% chance compared to the 42% for low performing women. Conditional on the performance level, women choose the hard task significantly less often than men, results that mimic those of gender differences in choices of competitive incentive schemes.

One (boring) explanation for this gender difference could be *pure task preference,* men more than women prefer the hard task. In the feedback treatment, participants receive information about whether they are of high or low performance level. If pure task preference were the major driving factor, choices of women

and men should not change much. However, once subjects learn their performance level, every single high performing subject, male or female, chose at least one hard task. Furthermore, high performing men chose the hard task with certainty, that is, every single high performing man chose the hard task twice, while women chose it with 86% chance. Low performing men chose the hard task with 25% chance, compared to 28% for low performing women. That is, when subjects learn their performance level with certainty, men and women do not differ much in their choices anymore. High performance level participants choose the hard task, and others choose the easy task.

However, as an institutional design, it may be quite unrealistic to have a perfect diagnostic exam of the performance level of a subject. The paper then describes another institutional change that could be implemented in situations where people have to choose between an easier and a more challenging option: Allowing participants to make gradual choices as opposed to choosing immediately the difficulty for the next two tasks as in the *Choice* treatment.

In the *Reduced Commitment* treatment, after the first easy task and after learning of the calibration, participants make a choice of difficulty level for their second task and perform. Only afterwards do participants decide on the difficulty level for their third and final performance. This treatment results in high performing participants to mostly choose the hard task: 88% of high performing men, and 89% of high performing women choose at least one hard task, and both high performing men as well as high performing women have an 81% chance to choose a hard task. That is, high performance level participants mostly choose the hard task, and there are no gender differences in choices among those subjects. The situation is different for low performing participants. While low performing women mostly choose the easy task (which is payoff maximizing for them) this is not the case for low performing men. Low performing men have a 72 percent chance to select the hard task, compared to 28% for low performing women.

Overall, gradual choices help high performing women to choose the hard and challenging task. Note that gradual choices do not help low performing men to avoid the hard task. One possible explanation could be that the information received from performing in the hard task is less precise, that is it is harder to learn that the hard task choice was not the money-maximizing one. The reason is that participants improve their performance in the easy task as well. Comparing earnings from the hard task in round 2 to those made when performing the easy task in round one biases results towards higher earnings for the hard task.

This final treatment also shows that the reason for gender differences in task difficulty in the initial *Choice* treatment were not due to an aversion to learning one's type. It could be that the results were driven by the fact that women may be more risk averse (and hence, for given beliefs, do not choose to enter the hard task that has higher variance in pay) or less certain in their beliefs of being a high performer, or less certain that an initial performance is indeed indicative of a high performance level.

Overall, there seem to be gender differences in task choices when one task is clearly easier than the other, with women being more likely to choose the easier task. Note that in the case of Huberman and Rubinstein (2001) the easy task would have to be the one in which a "better" performance is easier, though in this seemingly easier task the expected earnings are lower.In that sense, the easier task actually may have been the harder task.

In the earlier psychology literature there is an example, Dweck (2000), that proposes a mechanism of who may be more reluctant to engage in challenging tasks, and why females may be overrepresented in this group. Specifically, the hypothesis is that there are two extreme views of intelligence, or talent for a specific subject area. One is that intelligence is a fixed trait and tests, etc. can basically uncover how talented someone is. Another view is that intelligence is like a muscle, something that increases as it is exercised. The more someone believes that intelligence is fixed and the person has already been reinforced that he or she is intelligent, the more the person may shy away from challenges. After all, there is a chance to learn that perhaps the challenge leads to failure. On the other hand, if intelligence is like a muscle, initial failure from challenges may not be a problem, since it is understood that only by keeping at it can we improve. The gender component is that females are more likely to hear that they are smart early in their education, and so, females may shy away more from challenges than males.

An interesting avenue for future research could be to better understand the interplay between gender differences in choices and different institutional designs that either exacerbate or reduce these choices. This question came up in the last section on gender differences in competition, when I discussed the external relevance of competitiveness. Buser, Niederle and Oosterbeek (2014) showed that competitiveness predicts educational choices in the Netherlands. The Netherlands, like many continental European countries, has children make educational choices in a "once and for all" choice setting: children make one choice that determines their education for several years. These choices are not "flexible" like in the US school system, in which students are much more likely to make gradual choices. A choice of one hard mathematics class does not preclude not choosing it in the next semester. While research mentioned above hints to the fact that gradual choices may increase the chances of women selecting challenging

tasks, it clearly remains a very open question as to what extent different choice architectures affect choices of women and men.

III.B. Gender Differences in Speaking up

The final set of papers I present in this section considers whether women are as willing to speak up as men are. While there is a psychology literature on behavior in teams and whether women have less influence than men, the problem with that literature is that there are many biases that can account for potential gender differences in influence: Women may not only behave differently but also may be treated differently (see Thomas-Hunt and Phillips, 2004, and references therein). The papers I discuss share the feature that participants retain some anonymity and don't have to worry about immediate dismissal. The focus is rather on the decision to give advice or contribute an idea. Cooper and Kagel (2013) consider whether there are gender differences in the propensity to give advice, and Coffman (forthcoming) considers whether there are differences in their propensity to "speak up" when deciding which team member should answer a question.

Cooper and Kagel (2013) study advice giving in the context of a signaling game that is based on the Milgrom and Roberts (1982) entry limit pricing game. The crux of the game is that the high quality sender has to recognize that they can signal their type via a separating equilibrium. Cooper and Kagel characterize this as a "eureka" type problem in which there is a clear insight that is easily explained to others. In the first of their three main treatments, the *1x1 treatment*, one player plays against another. In the *2x2 treatment*, subjects are paired with one another and interact in two-person teams.[24] In the *advice treatment*, subjects play in two-person teams, with one subject receiving the role as advisor and the other as advisee. "Advisors and advisees played the limit pricing game separately, with no need to agree on a common action (and no mechanism for doing so). Advisors had (almost) continuous access to a messaging program which they could use to send advice to their advisee." (p 9) Advisees could not communicate with advisors, and could not observe the play of advisors. Advisors "received a bonus equal to 30% of their advisee's total payoff (along with their own payoff). These bonus payments were only reported at the end of an experimental session so that advisors could not tell what choices their advisees had made." The main question is whether there is a gender difference in both giving advice and taking advice.

[24] "Teammates must jointly agree on a choice having (almost) continuous access to a messaging program that allows for bilateral communication about possible actions. Both teammates receive the full payout from their team's outcomes." (p 9).

Earlier work (Cooper and Kagel, 2005) showed that teams play better than individuals, even when considering the truth wins norm. Specifically, forming artificial teams out of individual players and having them play strategically if one of the individuals plays strategically still leads to less strategic play than is observed in actual teams that are allowed to communicate with one another. The reason is that teams have an easier time "putting themselves in the shoes of the other" and hence understanding that the other player may make inferences from and react to one's own choices.

The present paper shows that "having an advisor significantly increases the frequency of strategic play, especially when the advisor plays strategically. But the effect is weaker than would be predicted by a truth wins model where advisees play strategically if either they or their advisor figure out strategic play." This is due partly to the behavior of advisors, where they find large gender differences, and partly due to the behavior of advisees, where there are essentially no gender differences.

First, females are less likely to play strategically than males. Therefore, to compare the behavior of advisors, the paper focuses on advisors who have played strategically. "[A]lmost half (43%) of advisors who have a history of playing strategically fail to advise their partners to play strategically. This cannot be attributed to a general unwillingness to send messages or give advice, as 93% of … advisors send at least some message and 85% send messages that include advice how to play… [a total of] 41% of …advisors who play strategically during the first half of their session never advise their partner to play strategically." (p4). This effect is to a large extent driven by female advisors. "Seventy three percent ... of ... male advisors who have played strategically also give advice to play strategically, compared to only 31% of female advisors."

To account for the aforementioned gender differences, Cooper and Kagel note that "Given that cognitive ability is basically identical for men and women in our sample, we conjecture that relatively low adoption of strategic play by women reflects lower confidence in their insights."[25]

In terms of behavior of advisees, a third (34%) of those who have received advice to play strategically fail to follow that advice. Furthermore, "[s]ubject to being advised to play strategically, the marginal effect of

[25] Cooper and Kagel note that "An odd feature of our data supports this connection: while women who have played strategically are significantly *less* likely to provide strategic advice than men, women who have *not* played strategically are significantly *more* likely to provide strategic advice than men who have *not* played strategically. It seems that men who have figured out to play strategically follow through both by playing strategically *and* by giving advice to play strategically. Women are more cautious, often only doing one or the other." (p 6).

a sound explanation as to why this advice should be followed is essentially zero." (p 2) There are, however, basically no gender differences in advisees; while women are somewhat more receptive to advice then men are, this difference is not significant.

This reluctance of women to "trust" in one's answer and put oneself forward is also found by Coffman (forthcoming). Her paper considers whether women and men differ in their propensity to be present or "speak up" and have their opinion or answer determine the answer for the whole group, depending on whether the task is a task where participants expect males to be better than females or vice versa.

Specifically, participants first answer a set of multiple choice questions in 6 categories that vary in how male or female-typed they are, as declared by the beliefs of participants whether in general, in this category, "women know more" or there are "no gender differences" or "men know more". Participants are then put in groups of two and decide how willing they are to contribute their answers to new questions in these categories to their group. Specifically, participants chose a position in line, from 1 to 4, where the participant with the lower position is the one providing the answer for the group. In case of ties the answer of one group member is chosen randomly.

For female categories, women are more likely to contribute their answers than men are, and for male categories the opposite is true. Controlling for ability, women become less and men become more likely to contribute answers to the group as the maleness of the category increases. A big part of this effect is driven by gender differences in beliefs: compared to men, women believe they have a higher chance to answer a question correctly in female categories, while the opposite is true in male categories. Controlling for beliefs and ability, there are no gender differences in contributing answers in the female categories. However, in the male categories, even after controlling for beliefs and ability, women are less likely to contribute their ideas than men are (i.e. choose a higher position in line).

Coffman (forthcoming) then addresses whether providing participants with feedback (i.e., whether or not they were the person in the group who had the highest score in a given category) can encourage high-ability members to contribute. However, "we find only weak evidence that feedback increases willingness to contribute among knowledgeable group members."

Overall, there seems to be a gender difference, with women being somewhat reluctant, compared to men, to "speak up", especially in tasks where stereotypes are that men are better. In both papers I presented this result in considerable inefficiencies.

Overall, this section is relatively short, but is one I think that should be much longer. Understanding what drives women and men to "be present and show up" for challenging and perhaps stereotypical male tasks, to "speak up" and have their opinion count is still something that needs to be better understood. It will also be important to understand what institutional changes can level the playing field.

IV. ALTRUISM AND COOPERATION

One of the traits for which gender differences are generally assumed is altruism and cooperation – with women supposedly being more altruistic and cooperative. The corresponding view is that women are more caring and nurturing and more likely to help. For example, Eagly and Crowley (1986) write that "The female gender role includes norms encouraging certain forms of helping. Many … have argued that women are expected to place the needs of others, especially those of family members, before their own. Gilligan (1982) has identified this theme as women's orientation toward caring and responsibility." Interestingly, an early meta-analysis in psychology on gender and helping behavior (Eagly and Crowley, 1986) found that in about 100 studies, men helped on average more (Cohen's d = 0.34).[26]

In this section I focus on two prominent ways economists have assessed gender differences in altruism and cooperation. The first concerns distributional preferences, over income, of women and men. These are in general studied using versions of dictator games, where one person, the dictator, decides how to distribute money among a set of participants, in general the dictator and one other person (Forsythe et al, 1994). A second way in which altruism has often been looked at is by studying cooperation in prisoner's dilemma, public good or social dilemma games. In principle, when considering behavior in games, motives besides altruism or cooperativeness may also play a role. However, these games share the feature that in the one-period version they have a dominant strategy (to defect in a prisoner's dilemma, not contribute to the public good, or consume massively, respectively). So, at least in theory, strategic motives play no role.

[26] Eagly and Crowley (1986) further argue based on social-role theory that some kinds of helping are part of the male role, such as when helping is heroic or chivalrous, and such behavior is likely to be facilitated when there are onlookers around. However, women's helping is more nurturing and caring, such as caring for children, and often occurs in private. Dividing studies into those where onlookers were present and when they were not, they found men helped much more than women did with onlookers around (d = 0.74) while there were essentially no gender differences without onlookers (d = -0.02).

Of course, in practice, there are many reasons besides altruistic or cooperative motives why behavior does not correspond to the dominant strategy. For example, the one-shot public good and dictator game share the feature that the dominant strategy is at the corner of the action set (contributing nothing). Giving positive amounts may therefore reflect confusion rather than deliberate altruistic choices. There has been some evidence that changes in the action set changes giving in the dictator game, see e.g. Bardsley (2008) and, in a very close design, List (2007).[27] Likewise, Recalde, Riedl and Vesterlund (2014) using public good games where both the dominant strategy equilibrium and the efficient outcome are in the interior show that a considerable amount of behavior is due to confusion.[28] There is, however, no evidence that women are more confused than men in such simple decisions, so, we will ignore the possibility that gender differences in altruism or cooperation in these games are due to gender differences in confusion.

Another reason for giving in those simple games could be image concerns. Andreoni and Bernheim (2009) provide evidence that participants give in a dictator game not only because participants are altruistic or fair, but because they like to be perceived as fair. Assessing the impact of image concerns in these games remains a lively and, as of now, still unresolved debate, with, most recently Exley (2014) finding that women are more image-conscious than men (see also Jones and Lenardi, 2014). We will, however, ignore this issue as well, since once repeated interactions are studied, as is often the case, strategic considerations will play a role in addition to any concerns for altruism.[29]

[27] Specifically, they consider dictator games where instead of a range of giving that is only positive, subjects can also take money away. They find that more subjects take money away compared to the number of subjects who gave 0 in a standard dictator game. See Cooper and Kagel (Chapter 4) fro further discussion of the stability of dictator game outcomes.

[28] They consider two public good games with interior dominant strategies and efficient outcomes, where in one the interior equilibrium is above 5 and below 5 in the other. They show that especially among participants who decide more quickly than others on how much to contribute to the public good, a large fraction centers around 5. These "fast" participants are "more generous" than "slow deciders" if the dominant strategy equilibrium is less than 5, and "less generous" if the dominant strategy equilibrium is higher than 5, so that the paper shows that speed of decision is not necessarily correlated with generosity. This view has gained momentum with Kahneman (2011) arguing that intuitive choices should be faster, which has been interpreted as fast choices being more intuitive (and altruistic), see Rand et (2012).

[29] The early literature on public goods has shown that some but not all giving in repeated public good games can be attributed to confusion (Andreoni, 1995).

Other games used to assess gender differences in social preferences are the ultimatum game (Güth, Schmittberger and Schwarze, 1982) and the trust game (Berg, Dickhaut and McCabe, 1995).[30] Croson and Gneezy (2009) survey on gender differences review these games in the section on differences in social preferences. Ultimatum and trust games do not have a dominant strategy, though in both the unique subgame perfect equilibrium is for the first mover to not pass any money (or, for the discrete ultimatum game, at most the smallest unit) to the second mover. However, especially for the ultimatum game, the subgame perfect strategies of the first mover result in rather low payoffs (empirically) compared to the payoff maximizing strategy. Therefore, gender differences in these games may to a much larger extent reflect gender differences in beliefs about the behavior of other player(s), or in how strategically sophisticated players are. I will therefore not review these games.[31]

A METHODOLOGICAL ASIDE: COHEN'S *d*

Much of the material in this section and in the next section on risk preferences involves work published in both the economics and psychology literature. In the latter there is frequent reference to Cohen's *d* which is used to evaluate the importance of differences between means (as opposed to just their statistical significance). It is a formalization of the common notion in economics of whether a difference in means is "economically meaningful." It is used as a measure of effect size when comparing the mean of one sample to another and is defined as the difference in population means ($\mu_1 - \mu_2$) divided by the population standard deviation σ which is supposed to be common among the two populations, that is:

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

[30] In a typical ultimatum game a proposer offers to divide a fixed amount of money between herself and the responder. The responder can accept the ultimatum proposal, in which case the division is implemented, or reject it, in which case both players receive nothing. In a typical trust game, the proposer can pass any number of tokens *x* from her endowment *m* to a responder, where tokens passed to the responder are often tripled. The responder can then give some tokens *y* back to the proposer, from zero to all the *3x* tokens, where tokens returned are not multiplied anymore. This leaves the proposer (or trustor) with *m-x+y* tokens and the responder (trustee) with *3x-y*. General outcomes in these games are also reviewed in Cooper and Kagel (Chapter 4).

[31] There also has been a lively debate whether behavior in the trust game, especially for the first mover, reflects their "altruistic" or "trusting" tendencies, versus, for example, their attitudes towards risk, see Bohnet, Herrmann and Zeckhauser (2010) who write "…differences in willingness to trust mainly came from differences in people's intolerance of betrayal, though for men differences in willingness to take risk also contributed." (p 826). Other work finds that the behavior of the second mover in the trust game is better predicted by survey questions on trust rather than trustworthiness, e.g. Glaeser et al (2000) write "In summary, to determine whether someone is trusting, ask him about specific instances of past trusting behaviors. To determine whether someone is trustworthy, ask him if he trusts others." (p 840).

In practice, Cohen's *d* is computed as the difference in sample means $\bar{X}_1 - \bar{X}_2$ divided by the pooled standard deviation, s, where s is computed as

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

where s_i^2 is the variance of the sample $i = 1,2$, each with sample size n_i for $i = 1,2$, that is

$$s_i^2 = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}(x_{i,k} - \bar{x_i})^2$$

Cohen's *d* is the standardized difference between two population means, here mostly male versus female population means. Cohen (1988) provides a guideline of how to think of effect sizes, that is used to describe many psychological effects: An effect size of 0.2 is considered small, 0.5 medium and anything of 0.8 or larger is considered a large effect.

There are several ways in which to interpret Cohen's *d*. First, assuming normal distributions, a specific effect size helps determine the minimum sample size required to get a significant result (say of 10 percent in a two-sided test) with certain power (say 0.8 that is there is an 80% chance that we correctly reject the null when it is false). Second, Cohen's *d* can give an indication of how likely it is that a person from sample 1 (male) has a higher "outcome" than a person from sample 2 (female). Third, Cohen's *d* can give an indication of how much of the variation in the "outcome" can be accounted for by gender.

With a Cohen's *d* of 0.2 there is a 56% chance that a randomly chosen man has a higher outcome than a randomly chosen woman, with 1% of the variation in "outcomes" accounted for by gender.

With a Cohen's d of 0.5 there is a 64% chance that a randomly chosen man has a higher value than a randomly chosen woman, with 6% of the variation in "outcomes" accounted for by gender.

With a Cohen's d of 0.8 there is a 71% chance that a randomly chosen man has a higher value than a randomly chosen woman, with 14% of the variation in "outcomes" accounted for by gender.

To give some examples of gender differences and their effect sizes, Hyde (2005) presents a survey over major meta-analyses of research on gender differences. She covers 44 papers with a total of 124 effect sizes analyzing cognitive variables, communication, social and personality variables, psychological well-beings, motor behavior and some other traits. Only a particular motor skill - throw velocity as well as distance – has a Cohen's *d* of about 2 (Hyde 2005). Seven more variables have a Cohen's *d* of 0.66 or higher, namely grip strength, attitudes about casual sex, masturbation, mental rotation, mechanical reasoning, agreeableness and tender-mindedness (the only one where women score higher). Almost 80 percent of effect sizes are less than 0.35.

IV.A. DICTATOR-STYLE GAMES

An early dictator game paper that examines gender differences in social preferences is Bolton and Katok (1995). They consider various dictator games over $10, pooling data from different treatments, with a total of 46 male and 31 female participants.[32] Subjects were not aware of the gender of their partner. Probably the best summary of their data is Figure 5 which shows for each amount passed to the other player (from 0 to 5), the fraction of women and men, separately, who passed that amount. They conclude that "We find no evidence for gender differences in generosity."

[32] About 50 subjects played dictator games where they could offer any dollar amount from 0 to 5 to the other player, of which half played only one dictator game whereas the others played 10 dictator games over $1, simultaneously. The remaining 25 subjects could only offer half or nothing, and also played 10 $1-dictator games simultaneously. They found no gender difference in each treatment separately.

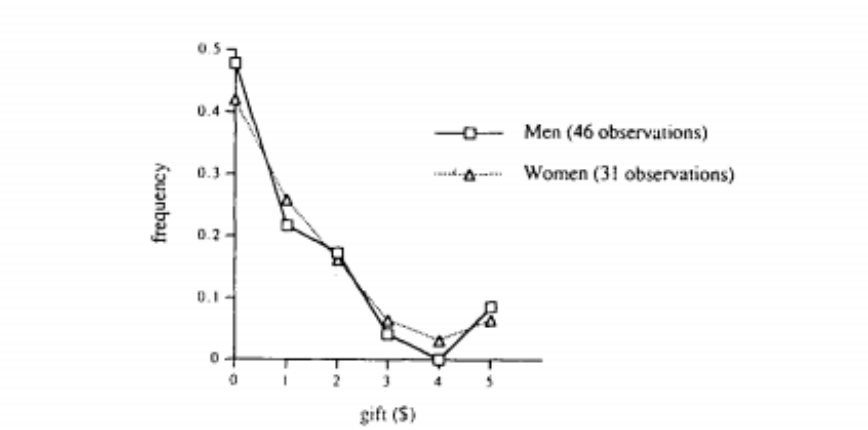

Figure 5: Dictator-giving by gender: amount left for recipients (pooled data)

The next paper in this literature, Eckel and Grossman (1998), finds a very different result. They have students divide $10 in a double blind way, in an effort to avoid potential experimenter-demand effects. One drawback to this procedure is that it requires single-sex groups, since in the double-blind they can't attribute any choice to a specific person. Table 1 below shows the distribution of outcomes. There are significant gender differences with women giving more: $1.60 versus $0.82 ($p<0.01$) for men.

Amount donated	Women	Men
$ 0	46.67	60.00
$ 1	10.00	26.67
$ 2	13.33	3.33
$ 3	11.67	5.00
$ 4	3.33	0
$ 5	15.00	3.33
$ 6	0	0
$ 7	0	0
$ 8	0	0
$ 9	0	0
$ 10	0	1.67
Average donation	$ 1.60	$ 0.82
Observations	60	60

Table 1: Percent of Decisions for each amount donated (from Eckel and Grossman,1998).

They conclude that "The double-anonymous dictator setting removes risk, the possibility of gender-related subject interactions, and the experimenter effect, leaving only underlying selflessness as an explanation for donating money. Our results indicate that women are less selfish than men when confounding factors are eliminated." (pp 732-733).

Given that there are many differences between these two studies, it would be too hasty to conclude one way or the other whether gender differences in altruism exist, and whether they are altered when participants interact in single-sex compared to mixed-sex groups. This early work perhaps also drives home the point that replications, as well of investigations of robustness, are extremely important (see Coffman and Niederle, 2014a,b).

The first paper to offer a more comprehensive study on gender differences in dictator games is Andreoni and Vesterlund (2001). They have a total of 142 subjects (95 men, 47 women) decide how to allocate a fixed amount between themselves and another person, where they vary the budget to be allocated as well as prices between own- and others-payoffs. Subjects make 8 different choices, where the budgets in tokens are 40, 60, 75 and 100, and the value of tokens for the two players range from 3:1 to 1:3. For a 3:1 value, each token is worth 3 points for the dictator but only 1 point for the recipient, meaning that giving is expensive. For each subject, one of the 8 allocations was chosen for payment, where each point was redeemed at $0.1. In three decisions tokens were worth more to the dictator, in three the exchange rate was reversed so they were worth more to the recipient and in 2 decisions the exchange rate was 1:1.

Across the eight decisions men on average passed $2.56 to the other player compared to $2.60 for the women, with the difference not significant. However, these similar averages mask important heterogeneity. The figure below shows for each price of giving in {1/3, ½, 1, 2, 3} the payoff passed to the other player as a fraction of income the dictator would have received had they kept all the tokens. For prices of giving below one, the payoff men pass to the other player is a higher fraction of their income than it is for females. In fact, for a price of 1/3 men pass more to the other player than they could have secured for themselves if they had kept everything; i.e., men are more generous than women when giving is cheap. However, when giving is expensive (for prices of 1 or more), women are more generous than men. Figure 6a nicely illustrates that the curve for men is flatter than for women, meaning that the amount of money passed as a fraction of income is more sensitive to price for men than it is for women.

Andreoni and Vesterlund (2001) then classify individual subjects as either totally Selfish, Leontief – utility is the minimum of the subjects' own payoff and the others' payoff – and Perfect Substitutes – where subjects treat own and other payoff as perfect substitutes, that is allocate all tokens to the person with the highest redemption value. Roughly 44 percent of subjects exactly fit one of these categories, with the remainder allocated to these categories using the Euclidian distance between actual behavior and the behavior predicted by one of those utility functions. Figure 6b shows the fraction of women and men for each type, where the two distributions are significantly different. Compared to women, men are more likely to be selfish or to have a utility function that views payoffs as perfect substitutes, consistent with maximizing total payoffs regardless of who receives the money. Women, on the other hand, are more likely to aim for payoff equality.

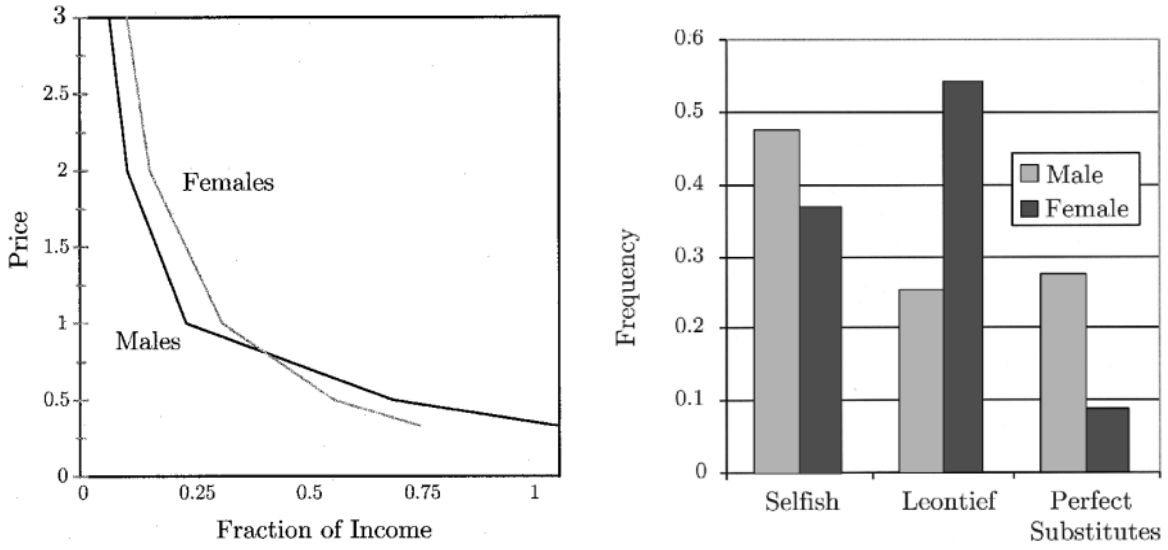

Figure 6 a: Payoff passed as a Fraction of Income, Figure 6b: Preference Distribution (Strong and Weak)

One early summary of gender differences in altruism and cooperation is Eckel and Grossman (2008). In addition to summarizing the results of the three ultimatum game experiments just discussed they look at responder behavior in the ultimatum games. Eckel and Grossman (2001) finds that women are less likely to reject offers, especially those made by other women. In contrast, Solnick, (2001) did not find this result using the strategy method version of the ultimatum game; and the behavior of one "punishment game". Given these results, Eckel and Grossman (2008) conclude: "In those settings where subjects are not exposed to risk – i.e. as respondent in ultimatum experiments employing the "game method" design and dictator games –

systematic differences are revealed. The choices women make are less individually-oriented and more socially-oriented." (section 4, conclusions).

Several other early papers explicitly study the interaction of the gender of the dictator and recipient. In Dufwenberg and Muren (2006) 352 dictators divide income between themselves and another student described as a randomly selected female (or male) student. While they find that women receive significantly more than men, donations do not differ between men and women. Ben-Ner, Kong and Putterman (2004) find quite different results: They have 154 dictators and find that information on the recipients' gender does not affect giving for men, but does for women (who give less to women). Largely as a result of this they find that overall women give less than men, though the difference fails to be statistically significant.

Houser and Schunk (2009) have German children between 8 and 10 divide M&M's between themselves and a child at another school. In all the three treatments the name of the child who ends up with the most M&M's is announced to the whole class by writing it on the blackboard. In addition in treatments 2 and 3, this child also receives his/her favorite hand-stamp. In treatment 3 (but not in treatments 1 and 2) dictators are told the gender of the child with whom they are paired. While in treatment 1 boys and girls send the same amount, around 8.5 (out of 20), the amount of M&M's given is reduced in treatment 2 for boys, who now only give 5.2 M&M's, which is significantly less than before and significantly less than girls give (around 9 M&Ms). In treatment 3, boys give around 5.6 M&M's while girls give 8.9, replicating the effect of treatment 2. Furthermore, boys send on average 2.1 fewer M&M's to girls than to boys (though the difference is not significant, $p > 0.10$), with girls also sending fewer M&M's to girls, a smaller average effect (1.9 fewer M&Ms) but one that is significant ($p < 0.05$). Although this difference in giving between boys and girls to girls is very unlikely to be statistically significant Houser and Schunk (2009) write: "While not comparable due to the influence of competitive pressure in our third treatment, this latter finding seems to contrast with findings from adult samples. In particular, both Holm and Engseld (2005) as well as Dufwenberg and Muren (2006) report men generally receive less than women when information regarding one's receiver's gender is available. The main finding from the three treatments is that competition decreases fairness for boys, but not for girls." (p 638).

Croson and Gneezy (2009) in their section on Dictator Games summarize nine studies. These include the three dictator games summarized by Eckel and Grossman (2008) described above, three games that are somewhat different from dictator games (e.g. one has a disinterested third party make allocations), and the three dictator games that explicitly study the interaction of the gender of the dictator and recipient just described. They summarize this literature as "… men choose efficient allocations while women are more inequality averse." Their paper then addresses whether donations by women are more dependent on the sex of the recipient than donations by men. "However, comparisons between the first two studies (Eckel and Grossman 1998 and Bolton and Katok 1995), and within the final two studies (Ben-Ner et al 2004 and Houser and Schunk [2009]), suggests that women's decisions are more context specific than men's."

General conclusions of this sort would appear to be premature. Recall that Dufwenberg and Muren (2006) do not find gender differences in giving. Furthermore, while Ben-Ner et al (2004) and Houser and Schunk (2009) find that women react to the recipients gender, and men do not, neither paper finds that women react to the recipients gender significantly more than men do. In fact, one could argue the opposite, namely that the behavior of men is more context dependent than women, since, (i) giving for men depends more on the price than forwomen (Andreoni and Vesterlund, 2001), and (ii) Houser and Schunk (2009) found that boys are much more inclined to become less prosocial than girls when competitive pressure is induced. Of course, arguing that the behavior of men is more context dependent than that of women would be equally premature given the small number of papers surveyed up to this point.

A more recent summary of the dictator game literature is provided by Engel (2011), a meta-analysis covering 129 dictator papers published between 1992 and the end of 2009.[33] Summarizing demographic variables, he writes that (p 597) "Since in ordinary papers on dictator games gender is not reported, meta-regression with all data would not be meaningful. If one confines the sample to those papers that have explicitly tested gender, it turns out that women

[33] He also includes "4 papers to come out in 2010 but already available through advance access. 4 papers do not report sample size. The remaining papers cover a total of 41,433 observations."

give significantly more…" since men give on average 21% of the pie to the other participant, while women give on average 27% (difference that is significant at the 10% level). Note, however, that only 12 papers, 10% of the surveyed dictator games, explicitly test for gender effects of the dictator. Unfortunately, Engel makes no attempt to assess if there is any potential bias on what papers report gender results. Clearly, if gender results are reported more often when they conform to the view of existing summaries of the literature, the result that women are more generous may be derived from a biased sample. While Engel includes Andreoni and Vesterlund (2001), it is not clear how that paper is coded in this meta-analysis.

Engel (2011) also discusses whether women receive more than men do: "Women do not only give more in dictator games, they also get more as recipients. In a meta-regression confined to those experiments that have explicitly tested recipient gender, this factor alone explains 73.2% of the observed variance, …" with male recipients receiving on average only 5% of the total pie compared to women who receive significantly about 20% of the total pie, a difference that is significant at the 1% level. He goes on to note that "If one controls for recipient gender, dictator gender is insignificant…"Given that the number of papers covered in this subsection (over 30) is larger than in the previous subsection (12), and Engel (2011) does not provide a table on how each paper is coded, it seems that in the present regressions all papers are included that have a gender variable, even if they have single-sex groups.

One summary of Engel (2011) may be that women are more generous than men are. However, a more careful analysis reveals that, conditional on the recipient of the gender, there is no significant difference in giving between women and men.

While an overall statement of whether males or females are more generous seems not to be clearly supported by the data on dictator games, the result that giving by men depends more on price than women might be.

Papers that replicate the Andreoni and Vesterlund (2001) result that males are more efficiency oriented while females are more focused on equity – that is, giving by men is more price-elastic – include Visser and Roelofs (2011), Boschini, Muren, Persson (2012), Fisman, Jakiela and

Kariv (2014). Papers that find directional, but non-significant results include Leider et al (2009) and Balafoutas, Kerschbamer and Sutter (2012). Cox and Deck (2006) often cited as refuting the Andreoni and Vesterlund (2001) result, does not, in fact, vary the price of giving in their experiments.[34]

IV.B Field Evidence and External Relevance of Gender Differences in giving

I confine myself to discussing the field evidence with respect to the gender gap in the elasticity of giving.

Andreoni and Vesterlund (2001) report that Conlin, Lynn, O'Donoghue (2003) "interviewed customers leaving over 40 restaurants in Houston, Texas. The results indicate that people tend to view 15 percent of the bill as the appropriate tip for a server who performs well. As the bill size gets larger, however, meeting this social norm becomes more expensive. What Conlin, O'Donoghue, and Lynn's data reveal is that, in fact, the percent-tip is a decreasing function of the bill size for both men and women and that men's percent-tip is more responsive to the bill size than women's." (p 306)

Andreoni, Brown and Rischall (2003) write "An important aspect of our results is that they provide direct evidence to support the growing feeling among fundraisers that men and women behave very differently with respect to charitable giving. Men are more sensitive to both price and income, for instance, and tend to concentrate their giving among fewer kinds of charities. And when the price of giving is low, men tend to give more to charity than women, but when the price is high the opposite is true." (p 128).

Craig et al (2014) consider the effect of an increase in time cost on the return behavior of over 900 blood donors in Australia, using both post donation questionnaires and blood donations up to 3.5 years later. Exploiting the natural variation in time cost involved with donating blood, a one standard deviation increase in the average wait time (an additional 20 minutes relative to the

[34] They write in their abstract that "women are more sensitive than men to the costs of generous actions when deciding whether to be generous." While this reads as if they do not replicate Andreoni and Vesterlund (2001), a more careful reading of their paper reveals that Cox and Deck do not consider variation in the price of giving, but rather variation in total payoff – or the pie to be distributed (holding the price of giving constant at a 1:1 rate). They find that women respond more to a reduction in their budget than men do.

average wait time of about 45 minutes) would result in an 11% decrease in blood donations per year. For men, longer wait time is associated with less satisfaction from donating, lower intent to return and longer delay before actually returning. While women also report less satisfaction and indicated they were less likely to return if they had to wait longer, longer wait times had no impact on the time until women returned to donate blood once more. That is, the return behavior of males is more elastic than that of females.

Fisman, Jakiela and Kariv (2014) study the external relevance of differences in distributive preferences on voting behavior. They conduct an incentivized experiment using the American Life Panel (ALP) with about 1000 participants dividing money between themselves and another (not sampled) random ALP participant. Each participant makes about 50 decisions, in which both the budget, as well as the price, of giving was varied. One of these decisions was chosen for payment. They find that women are less likely to be efficiency oriented than men. About 750 of those participants also indicated who they voted for in the 2012 presidential election. A binary indicator for efficiency-focused distributional preferences is negatively correlated with the likelihood of voting for Obama (in 2012), as well as belonging to the Democratic Party. In private communication, the authors note that the indicator for efficiency-focused preferences is larger than the coefficient for the female gender dummy with females more likely to vote for Obama and to belong to the Democratic Party.

IV.C. Prisoner's Dilemma and Public Good Games

Rapoport and Chammah (1965), published in a psychology journal, is perhaps the earliest study of gender differences in incentivized prisoner's dilemma games. They have pairs playing a 300 period repeated prisoner's dilemma game.[35] They find that male pairings exhibited the greatest rate of cooperation, followed by mixed pairings, with female pairings cooperating the least. Following this, there have been a number of papers considering gender differences in prisoner's dilemma and public good games in the psychology literature.

The literature on gender differences in public good and prisoner's dilemma games in economics started later and was slower taking off than in psychology. For example, Ledyard (1995) surveys

[35] The presented results are over 7 different payoff matrices.

6 papers on gender differences in public good games. He asks whether and how gender affects the rate of contribution, He concludes that "I think the question remains open." Eckel and Grossman in their 2008 survey (with the most recent referenced paper published in 2001) included 8 public good games, concluding that: "In those settings where subjects are exposed to risk – i.e. public good experiments […] – there is no significant evidence of systematic differences in the play of women and men."

The survey by Croson and Gneezy (2009) has 18 prisoner's dilemma, social dilemma and public good games with gender differences in behavior. Their summary is that "[a] large body of work identifies gender differences in other-regarding preferences. However, many of the results are contradictory. In some experiments, women are more altruistic, inequality averse, reciprocal, and cooperative than men, and in other they are less so. We believe that the cause of these conflicting results is that women are more sensitive to cues in the experimental context than men."

Balliet, Li, Macfarlan and Van Vugt (2011) published "Sex Differences in Cooperation: A Meta-Analytic Review of Social Dilemmas" The studies they included had to have either adolescent or adult participants along with reporting participants gender. Further, "only studies using pure social dilemma paradigms were included (i.e. prisoner's dilemma, public good, and resource dilemma). […] We coded effect sizes for studies that either involved participants interacting with a confederate, a preprogrammed strategy, or another participant. Importantly, in all studies, participants believed they were interacting with other participants." The latter makes it clear that these studies are not all published in economics journals. The meta-analysis contains 272 effect sizes. For each paper they use Cohen's *d* value as the measure of effect size which is the difference in means divided by the pooled standard deviation.[36] The Table below shows the distribution of effect sizes for sex differences in cooperation in a stem-and-leaf plot, for those studies that did not have null effects.[37] To read this figure, note that Rapoport and Chammah (1965) is one of two studies with $d = 0.57$, that is men cooperating significantly more than

[36] See below for discussion of Cohen's *d* value rro effect size.

[37] "Several articles reported a null relationship between sex and cooperation, but failed to provide the statistics necessary to calculate the effect size. We estimated that these studies had an effect size of zero. This is a very conservative estimate, as several of these articles observed a mean difference between men and women, but lacked the statistical power to detect a small effect size. Therefore, for all analyses, we first report the results excluding the null findings coded as zero effect size, followed by an additional analysis including these estimated null findings." (p 887). In the stem-and-leaf plot these studies are omitted.

women. They conclude that the relationship between sex and cooperation in social dilemmas is not statistically different from zero.[38] Further, when considering gender differences in cooperativeness based on the gender of one's group members, Balliet et al (2011) conclude that during mixed-sex interactions women were more cooperative than men, though during same-sex interaction men were more cooperative than women, a result that is also borne out in pairwise interactions.

d value	0.1 units of the *d* value
1.5	3
1.4	
1.3	0
1.2	
1.1	4 8
1.0	
0.9	8
0.8	0 6
0.7	0 0 0 3 5 5 7 8
0.6	0 5 7 9
0.5	0 2 3 3 4 5 7 7
0.4	0 0 7 7 7 9 9
0.3	0 1 3 5 5 6 6 8
0.2	0 0 0 2 3 5 5 6 6 6 6 7 8 9
0.1	0 0 1 1 2 3 4 5 6 6
0.0	0 0 0 0 2 2 3 3 4 5 5 5 6 6 7 7 7 7 8 8 8 8 8 9 9
−0.1	0 0 0 0 1 1 2 2 3 4 4 4 5 5 6 6 6 6 7 9 9
−0.2	0 0 0 1 1 1 1 1 3 3 3 4 6 7 8 8
−0.3	0 0 1 1 2 4 5 6 6 6 7 7
−0.4	0 3 3 3 5 6 6 6 7 7
−0.5	0 0 1 2 3 5 6 6 7 7
−0.6	0 0 1 2 3 5 8
−0.7	2 5 6 6
−0.8	5 7
−0.9	2 4 4 8
−1.0	4
−1.1	7
−1.2	
−1.3	
−1.4	
−1.5	3

Note. This plot omits three outliers: 1.65, −1.76, and −1.90. This plot only includes the 176 effect sizes that were coded and does not include the null effect studies that were estimated to have zero effect size.

Table 2: Stem-and-Leaf Diagram of the Overall Distribution of Effect Sizes for Sex Differences in Cooperation

[38] They note, however, that the conclusions differ when using a fixed-effects rather than random-effects analysis. In which case women are significantly more cooperative than men, but with an "exceptionally small effect size (*d*= -0.04).

One of Balliet et al's (2011) most intriguing results concerns what happens as interactions are repeated as opposed to one-shot. They find that "men, compared to women tended to become more cooperative as iterations continued". This result holds when excluding one-shot interactions, and indeed as the number of repeated interactions increases, men become significantly more cooperative compared to women. While there are obvious caveats with respect to meta-analyses, one that is particularly worrisome with respect to this result is the coding of repeated interactions. For example, Fudenberg, Rand and Dreber (2012) conduct an infinitely repeated prisoners' dilemma game with a continuation probability of 7/8, meaning that the average length of the game is 8, with Balliet et al (2011) recording this as 8 repetitions.

Most recently, Gäechter and Poen (2013) survey data from 17 papers on linear public good games with 6037 subjects in 274 sessions, all papers have Gáchter as a co-author and used similar parameters. In one-shot public good games they find that women gave 0.77 more than men (out of 20 with average contributions around 9). While the effect is significant, the difference is small. Iterated one-shot public good games – where subjects are randomly re-matched after each round – reveal no gender differences in average contributions, though males are more likely to choose extreme contributions (0 or 20 out of 20). The lack of gender differences in one-shot public good games is confirmed when preferences for contribution are elicited by the strategy method (dependent on the average contribution of others). Slightly over 50 percent of subjects prefer to be conditional cooperators – that is match the average contribution of the other players - and about 20 percent are perfect egoists – contribute zero – which is slightly more common among men than women (25 versus 18 percent). The remaining strategies are unclassified, though women give slightly less than man for large contributions of others', making up the lower fraction of participants who always donated nothing.

Gäechter and Poen also have data for women and men playing ten-time finitely repeated public good games with the same group members. While on average, the contributions of women and men are not significantly different, men are slightly more responsive than women are to the average contribution of others' in the previous round. More importantly, men are much more strategic: Their contributions decline more steeply than those of women over time and they contribute significantly less in the pen-ultimate and the ultimate round, with a positive

correlation between how sensitive men are to contributions of others and how much they reduce their end game contributions. Women, on the other hand, basically do not change their contribution over time, and there is no correlation between how sensitive they are to the contribution of others and how much they reduce their contributions over time.

This higher level of strategic behavior can also be seen in how women and men react to punishment opportunities in the public good game. The standard result is that the opportunity to punish after a contribution round increases the donation to the public good in societies in which mostly free-riders and not cooperators are punished, though no such increase is found in societies in which cooperators are heavily punished as well (Herrmann, Thoni and Gäechter, 2008). Gäechter and Poen find that in societies where mostly free-riders are punished, men contribute significantly more to a public good with punishment than women do. No such difference is found in societies where cooperators are punished as well.

To summarize, there do not seem to be large differences in average contributions in public good games between men and women. However, these similar average contributions may mask important strategic differences. More work is needed to robustly understand the extent to which women and men differ in their strategic behavior in cooperative games.[39]

Since there are no reliable gender differences in average behavior, and the investigation of gender differences in strategic behavior has only began, I do not cover (and did not find many papers on) the external validity, or external gender similarities for behavior in public good games or public good contributions, and volunteering, outside of the laboratory. For a paper that confirms the external relevance of behavior in public good games (though without addressing gender differences), see Rustagi, Engel and Kosfeld (2010).

IV.D. NEW DIRECTIONS

The vast majority of papers on gender differences in altruism fall into a somewhat narrow band of games. It would be helpful to expand the set of games where altruism, or other regarding

[39] For an early paper on gender differences in strategic environments see Casari, Ham and Kagel (2007). They study bidding in common value auctions and find that women suffer from a stronger winner's curse than men to begin with, but eventually catch up as they learn faster.

preferences, are studied. Vesterlund, Babcock, and Weingart (2014) have taken a step in that direction addressing the question of whether women volunteer more often than men to perform non-promotable tasks. This, in itself, may result in women falling behind in the workplace, being less likely to be promoted. Using data on volunteering for Senate committee duties at a large university corroborates the fact that women are more likely to perform such undesirable and under-valued duties. Looking at this relationship more closely, they conduct a laboratory experiment in which groups of three people are randomly re-matched across rounds. Each person has to decide within a two minute interval whether to make an investment. There is no cost to delay but if no one makes the investment, everyone receives $1. If at some point one player in the group makes the investment, the round ends, the investor receives $1.25 and the other 2 group members receive $2. Under mild assumptions there are three kinds of equilibria: a pure strategy asymmetric equilibrium where one individual invests, a symmetric mixed strategy equilibrium where each player invests 23.3% of the time, and an asymmetric mixed strategy equilibrium where one person does not invest and the other two invest 40% of the time. Data from 132 participants (72 males and 60 females) show that in about 82% of the rounds the investment was made, and in roughly 63% of these it was made with one second or less to go. The figure below shows for each round the average investment rate by women and men when participants played in mixed-sex sessions (mixed_w and mixed_m, respectively). Women are significantly more likely to make the investment compared to men (35 percent compared to 21 percent). The distribution of investments by men is consistently lower than women across all ten rounds. Gender differences in investment are only mildly attenuated when controlling for (gender differences in) risk, conformity and other psychological variables (none of which are significant predictors of investment).

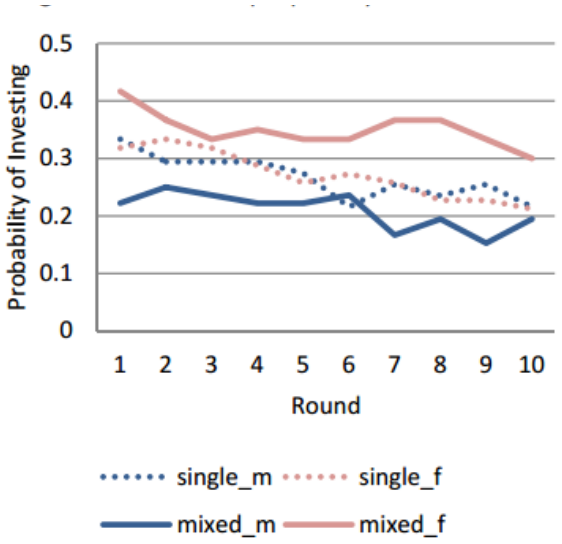

Figure 7: Probability of Investing.
(from Vesterlund et al., 2014)

To assess the extent to which the results are driven by inherent gender differences in preferences they repeat the experiment in single sex groups. Figure 7 shows that the investment rates of women and men in single sex groups are indistinguishable from one another (single_w and single_m, respectively). Women reduce and men increase their investment rate in single sex compared to mixed groups. Note, however, that the overall chance that a group makes an investment is the same whether in mixed-sex or in single- sex male or female groups. While the average rate of investment for men and women is essentially the same (occurring in 2.7 rounds on average) men are more likely to invest either seldom or very often, while women seem to be more concentrated around the "fair" investment rate of about one third of the time.

IV.E. Conclusions

Results for gender differences in altruism and cooperation are much more mixed than one might have expected. Considering average behavior, it seems that there are no reliable gender differences in average giving in the dictator game once the gender of the recipient if controlled for. In one-shot public good games women are found to be slightly more cooperative than men. However, this result did not replicate when using a sequence of one-shot public good games (though men were found to be more likely to contribute extreme amounts – all or nothing).

Likewise, repeated public good games show no significant gender differences in average contributions. However, both for dictator game giving as well as for repeated public good games, average similarities mask important differences. In dictator games, women are found to be less sensitive to the cost of giving (i.e. they are more equity than efficiency concerned compared to men). In finitely repeated public good games, men are found to be more strategic than women, especially in terms of adjusting their contributions downward as the end game approaches. More work is needed to understand the interplay between strategic sophistication, potential gender differences therein, and behavior in cooperative games.

The literature on gender differences in altruism and cooperation has only considered a small spectrum of games and could benefit from breaking out of this confinement. A promising avenue is the work by Vesterlund, Babcock, and Weingart (2014), which combines preferences for contributions to a public good, as well as concerns about discrimination and gender differences in beliefs about the behavior of others'.

V. Risk

After altruism and cooperation, the second strand of the literature on gender differences that has received an enormous amount of attention is risk attitudes. Since this topic attracted the attention of economists as well as psychologists, I'll review the evidence in both. This literature, perhaps as much as the literature on gender differences in altruism, seems to potentially suffer not only from a publication bias, but also the fact that many people seem to have a clear idea on what the "correct" finding is.

There are two main points and a piece of advice I want to convey in this section. The first point is that while gender differences in risk taking seem to exist, the evidence is far from persuasive that this gender difference is substantial in all environments. Because of the large heterogeneity in results, many surveys of the literature arrive at different conclusions, and at times some reach much stronger conclusions than seem warranted by the evidence. The second point is that the heterogeneity in results of gender differences on risk preferences stems from the fact that under some elicitation techniques gender differences in risk taking are very small (about 16% of a standard deviation, or alternatively, assuming normal distributions of risk preferences, if a

random man and a random woman are compared, there would be a 55% chance of being correct when saying that the more risk averse of the two is a woman). In fact, such an effect would often appear as a null result with sample sizes of a couple of hundred as is common in many experiments. Other elicitation techniques yield somewhat larger gender differences, around 55% of a standard deviation (if a random man and a random woman are compared, there would be a 65% chance of being correct when saying that the more risk averse of the two is a woman). It is this variance in results that leads to very different conclusions when covering only a small fraction of the literature. This point really becomes evident when looking at the conclusions reached by various surveys of experimental studies in risk aversion. In principle such heterogeneity in results is not so surprising given that risk itself does not seem to be a simple and stable factor. As a consequence, perhaps even more than for any other literature on gender differences, it will be important to show the extent to which (small) gender differences in the laboratory translate to externally relevant gender differences in observed behavior of economic interest, both in the lab and the field. Another reason for this special need stems from the enthusiastic adoption of (experimentally documented) gender differences in risk aversion as a plausible explanation for a given finding by many economists. I obviously applaud that economists embrace experimental results. However, experimental economists should be careful in their studies and conclusions, and should not produce papers with biased results solely to pander to the taste of other economists. The goal of experimental economics should not be to produce evidence for any hypothesis.

The main result of my survey of the literature on gender differences in risk aversion is that those differences do exist. However, there is substantial heterogeneity of that gender gap across situations and elicitation methods. The gender gap is very small to the point of being almost non-existent in some areas, though the gap is more pronounced in other environments. This leads my main advice. An experiment that investigates a hypothesis that could rely on gender differences in risk aversion should plan to implement a risk elicitation procedure that is germane to the question at hand. I will demonstrate this using the last experiment I describe in this subsection.

There are several methodological issues when considering experiments on risk aversion. First, is the question already addressed above, what is a good measure of risk aversion, and what are the

correlations when using several measures? Such a debate has been present almost since the advent of studying risk aversion. An example is preference reversals by Lichtenstein and Slovic (1971). In a typical experiment, subjects are given the choice between a lottery with a high probability of winning a small amount (the P bet) or one with a small probability of winning a large amount of money (the $ bet). Subjects are also asked to state their willingness to sell each of these lotteries, or state their certainty equivalent. The common finding is that subjects choose the P bet over the $ bet, but have a higher selling price for the $ bet than the P bet. This robust preference reversal is, of course, not reconcilable with expected utility theory. In the last Handbook, Camerer (1995) described the already then very impressive bulk of evidence on different ways in which subjects deviate from expected utility. It is clearly problematic to map all those deviations back to a single measure of risk preferences and study gender differences in that measure.

Perhaps as a result, the vast majority of the evidence on gender differences in risk aversion considers very simple choices over gambles that are not able to assess all the intricate ways in which behavior deviates from expected utility theory. Different gamble choices and different elicitation methods can capture different deviations from expected utility. They may, however, also result in differences in the extent to which there are gender differences in behavior. Section V.D addresses this concern and presents evidence that the extent of gender differences in risk aversion may strongly depend on the elicitation method. This suggests perhaps a more complex view than a simple "Women are more risk averse than men."

A further complication when studying risk aversion in the laboratory through small lotteries comes from, what is now known as the Rabin critique (Rabin 2000). He posits that perhaps deviations from risk neutrality observed in low stakes lab experiments should not be interpreted as risk aversion. This is because such risk aversion, if scaled to larger amounts, would lead to implausible choices. Despite that, I'll keep using the term "risk aversion" to describe behavior observed with small scale experiments. There has been only very little effort, so far, to assess changes in gender differences in risk aversion when changing the stakes, though see Holt and Laury (2002) for a prominent exception.

Another issue in studying risk aversion in the lab, concerns how to pay participants if they make multiple decisions over lotteries and whether subjects should even make several choices or only one. Azrieli, Chambers and Healy (2014) present a theoretical analysis of the issues that arise when subjects are given multiple decisions, and in what instance choices in one decision may be distorted by choices in other decisions. They show that under mild assumptions paying for one randomly chosen problem is essentially the only incentive compatible mechanism. When the decisions by subjects are choices over lotteries, and subjects are paid for one decision randomly with known probabilities, this generates a two-stage compound lottery. They claim that "if we assume the reduction of compound lotteries, then the RPS [Random Problem Selection] mechanism is incentive compatible only if subjects are expected utility maximizers." However, there is abundant evidence that subjects are not, in fact, expected utility maximizers.[40] In terms of direct experimental evidence of the validity of the RPS mechanism, careful experimental comparisons do not reveal behavioral differences due to paying for one decision randomly chosen compared to paying for that same decision when subjects know in advance that only this decision is chosen for payment.[41]

Putting aside all the complications of studying (gender differences in) risk aversion, I will first cover early work and survey papers by psychologists. I then describe early economic papers and common elicitation methods in economics used to assess gender differences in risk aversion. I next discuss early economic surveys, followed by surveys using only a single elicitation method. I then discuss some recent results and I close with an example of how to design a risk elicitation method germane to the problem at hand.

[40] "This explains why Holt (1986) and Karni and Safra (1987) found the RPS mechanism not to be incentive compatible: they both assume reduction of compound lotteries and non-expected utility preferences" (p 13).

[41] Azrieli et al (2014) compare papers that study whether paying for one decision randomly in a long multiple price list to paying for only one decision. They write: "In all of these direct-comparison studies, subjects who are given a single choice do not see the other decision problems. Thus, behavior differences may be attributed to framing effects (causing a change in underlying preferences) rather than monotonicity violations. Disentangling these confounded explanations is clearly important. In our view, Cubitt et al (1998, Experiment 3) provides the cleanest test of incentive compatibility of the RPS mechanism because theconfound with framing is eliminated. Subjects are randomly assigned to one of three groups. All groups are given the same 20 decision problems. The first group is paid only for D1; the second group only for D2; and the third group is paid for one randomly selected problem out of the twenty, each selected with equal probability. Choice frequencies in D1 do not significantly differ between groups one and three (Chi square p – value of 0.355) and choice frequencies in D2 do not differ between groups two and three (Chi square p – value of 0.285). Thus incentive compatibility of the RPS mechanism holds for that experiment" (p 16).

V.A Early Work and Surveys by Psychologists

The very first experimental test on gender differences in risk aversion I could find is Swineford (1941). "A previous article [Swineford (1938)] introduced a formula for measuring a personality trait by means of any objective test. The trait was defined as the tendency to gamble, and was found to be independent of the achievement score on the same test". Specifically, subjects take a multiple choice test. "The pupil is permitted to ask for credit of two, three, or four points for each question, with the understanding that twice the requested credit will be deducted from his score if the answer is wrong. It may be assumed that the pupil is gambling on his score against odds of two to one to the extent that he asks for extra credit for those items on which he is guessing. There being no way to separate the items guessed correctly from those representing correct knowledge, the gambling score must be based upon the incorrect items, all of which may be regarded as guesses." […] "The formula adopted to measure gambling, or G, was based on only the items marked "4," as follows: G = 100 x (Errors marked "4")/(Total errors + ½ omissions)", where omissions "are the items which were skipped within the test—not those omitted at the end of the test for lack of time."[42] The paper provides outcomes of 344 students taking each of four tests.[43] The result is that "(1) Boys have a significantly greater tendency to gamble on their test scores than do girls, particularly on an unfamiliar type of test. (2) Both boys and girls have a significantly greater tendency to gamble on unfamiliar material than on familiar material." (4) […] the G scores are independent of the scores on the tests from which they were computed […]." (pp 443-4).

One reason to describe this paper, apart from its being the first, is that the last paper discussed in this section provides a nice example of a modern look at gender differences in test taking. Specifically, the paper disentangles the extent to which gender differences in test taking may be due to gender differences in risk aversion or gender differences in other domains such as test scores or beliefs about the chances the question was answered correctly.

[42] The reason to use only those questions for which the pupil asked for 4 points is that Swineford (1938) showed that the number of 4's asked for were less likely to be 0's than for any other number of points, and that "[T]the correlation between the number of "4's" on the odd and even items is .911 and that for the "3's" is .788" (p 299).

[43] An additional 74 boys and 39 girls were eliminated from the study "either because on one or more of these tests no extra credits were requested, or because on one or more tests no errors were made among the items attempted." (p 439). For these children, Swineford's gambling measure is either not defined, or zero. It is not clear how those are distributed between girls and boys.

The first paper I found that uses incentivized choices over lotteries to study gender differences towards risk is Kass (1964). He has 52 children aged 6 to 10 choosing between three slot machines.[44] The payoffs of each machine were illustrated by pulling the lever on each machine 5 consecutive times (the returns were not described). Each slot machine had an expected return of 0, one gave a penny back for each penny put in, one gave 3 pennies back with probability 1/3 (in fact: "On the 1/3 machine S[ubject] won three pennies at a time, dispensed once in random position, within a block of three trials" p 579), and the last gave 8 pennies back with probability 1/8.[45] In the experiment the "S[ubject] was stopped after 210 trials. At this time S was told he could now play only the machine he liked best and could not play the other two. S's preference was recorded, and he was then told that he could choose the prize he wanted to buy with his pennies." The first result concerns the last 30 trials before this forced choice, where a response is a choice to put money in a slot machine: "boys made the greatest number of responses on the machines with the intermediate and low probabilities of payoff and the least number of responses on the high probability machine. For the girls, the opposite effect is apparent." The difference is significant. "At the end of the experimental session, each S was told to pick the machine he liked best and to play only that one. On these forced choice trials, 61.1 per cent of the boys and 38.0 per cent of the girls did not pick the machine they had played most frequently during the previous 30 trials." There is no more discussion about possible gender differences in those choices. Indeed, the summary conclusion suggests that only in the free choice part were there any gender differences in choices.[46] Therefore, the first paper with incentivized choices over lotteries reports two effects, one with a gender difference in choices, and one where there is no gender difference.

However, both of those previous papers are seen as evidence of gender differences in risk aversion. For example, the second oldest paper I am aware of that uses incentivized gambles,

[44] "The S's were 52 preschool and elementary school children evenly divided by sex" (p 578).

[45] "You see how each machine works? Now I am going to give you 14 pennies. You can use these pennies and the pennies you get out of the machine to play with. At the end of the game you can use the pennies you have won to buy prizes. Now you can play the game. You can play any machine you want or you can play all the machines. I'll tell you when to stop. Remember, the more pennies you have at the end of the game, the better the prize you can buy with your pennies." (pp 579, 580).

[46] Note, though, that all machines had an expected return of 0. Perhaps a summary of this paper (given after the first result concerning the free choice in the last 30 trials) is: "The significant findings of this study are related to sex differences in probability preference. In a free choice, repetitive play situation, boys preferred probabilities of winning involving greater risk than did girls" (pp 580-1).

Slovic (1966), announces that: "At present, evidence indicating that boys are willing to take greater risks than girls is scarce" (p 170). The exceptions are the two papers described earlier.[47]

A first summary of the experimental literature on gender differences in risk aversion is from psychologists. Byrnes, Miller, Schafer (1999) conduct a meta-analysis of 150 studies comparing the risk taking behavior of women and men.[48] They have a total of 322 effects. The mean of Cohen's *d* (weighted by the inverse of each *d*'s variance) was found to be *d*=.13 with a 95% confidence interval of .12 to .14.[49] Assuming normal distributions, a Cohen's *d* of 0.13 implies that when comparing a random man to a random woman, there is a 54% chance that the woman is more risk averse.[50] The paper states that "nearly half (48%) [the effects] were larger than .20 (the conventional cutoff point for small effects)." Byrnes, Miller and Shafer (1999) conclude that while the overall mean effect size of $d = 0.13$ would be labeled as small in some statistical circles, such differences should still be "a matter of concern" (p 378) since small differences can accumulate across behaviors and time to produce substantial gender differences in various outcomes, such as driving injuries or deaths.

To see that in psychology the study of gender differences in risk aversion has been a growing field, note that in the first 17 years covered by the study (1964-1980) there were 83 effects compared to 235 effects in papers published from 1981 to 1997.[51] The mean and confidence interval for these two periods were d=.20 (.17 to .23) and d=.13 (.12 to .14), so, "the gender gap seems to be growing smaller over time" (p 366). It is not clear that such a conclusion is really related to changes in how women behave differently from men compared to a conclusion about changes in the discipline. While publication date is the obvious difference between those two

[47] Slovic (1966) cites two other papers that provide evidence that boys are willing to take greater risks than girls. Of those one observes children in the playground or home and the other finds that "boys were less ready to withdraw from threat of failure on an intellectual-achievement task than girls were." (p 170).

[48] The authors use several steps to aim to retrieve all papers on gender differences in risk aversion published between 1967 and 1997 using PsycINFO and PsycLIT by searching for "risk" or "risk taking" and "gender differences" or "sex differences", and then searching in MEDLINE using terms associated with specific risks, such as smoking, driving and framing effects, and finally searching Dissertation Abstracts.

[49] When considering individual effect sizes, note that the first quintile of effect sizes is -1.23 to -.09, and the second -.08 to .07 "(indicating essentially no difference)". The authors conclude hence that "a sizable minority (i.e. 40%) were either negative [that is, men are more risk averse] or close to zero." (all p 372). The intervals for the third, fourth and fifth quintiles were .08 to .27, .28 to .49, and .50 to 1.45, respectively.

[50] Put differently, to find that women are more risk averse than men with a p-value of 0.1 using a two-sided t-test and a power level of 0.8, one would need roughly 700 women and 700 men in one's sample.

[51] The oldest two studies in the survey, and also the oldest using "gambling tasks" were published in 1964, a book by Kogan and Wallach and the paper by Kass (1964) described earlier.

sets of papers, it could of course also be that different risk elicitation methods became more or less fashionable over time, and different elicitation methods may yield different gender differences in risk aversion. Interestingly, the papers in top tier journals (containing 14 studies) have the lowest effect size, d=.03 (.01 to .05).

Many of the 150 papers analyzed in Byrnes, Miller and Schafer (1999) are from psychology. In fact, not a single one was published in an economics journal, and in only 48 were "participants observed by researchers as they engaged in various activities that were judged by the researchers to have some degree of risk (e.g. making a left turn in front of traffic)." (p 370). Of those the total Cohen's d was d=.19 (95% confidence interval .16 to .22). Of those 48 papers (and dissertations), only 17 use "gambling tasks" (see their Table 1).[52]

While I counted 17 papers with 35 effects from their Table 1, Byrnes, Miller and Shafer, (1999) report that the mean effect size for the 33 effects in the gambling task is .21 (95% confidence interval of .14 to .28, see Table 2 on page 377.) They then divide those effects by age of participants, and find a d = .03 for children aged 9 and younger, d=0.27 for children aged 10-13 as well as those aged 14-17, and d=.31for college students. The total distribution of the 35 effects I counted can be seen in Figure 8 below.[53] Note, however, that the Kass (1964) study described

[52] The others use tasks such as: "*informed guessing*" where "participants could earn points or money for correct guesses but could also lose points or money for incorrect guesses (e.g. standardized achievement tests that have penalties for incorrect guesses); "*physical activity*" included behavior such as "climbing a steep embankment, playing in the street, trying out gymnastics equipment [..] and taking a ride on an animal (e.g. a donkey)"; "*driving*" includes "taking a left turn in front of incoming traffic, gliding through a stop sign rather than coming to a complete stop and engaging in simulated driving tasks"; "*physical skills*" described "playing shuffleboard or tossing rings onto pegs. In most cases, options differed in terms of their probability of success [..] and the number of points that could be won or lost; "*risky experiments*" "involved an individual's willingness to participate in an experiment that was described to them as involving the chance of physical or psychological harm"; "*intellectual risk taking*" "involved tasks that required mathematical or spatial reasoning skills. Participants were presented with items of various levels of difficulty and asked to indicate their preferred level of choice. Unlike the tasks in the informed guessing category, points were not subtracted for incorrect answers on the intellectual tasks. Thus participants were mainly concerned about getting stuck on items or exposing their lack of skill when they fail." The final category is all the rest and includes "lying about finding someone else's money, cheating during a computerized game, [..] and administering an electric shock to a confederate to increase his learning rate." The category "*gambling tasks*" is the category closest to experimental economics tasks, and is described as "similar to the category of physical skills in terms of the varied risks/reward options." However, "a person's skill level had no bearing on the likelihood of success" (all p 371).

[53] Note that the Kass (1964) study is coded as (d = .80, n_{Male} = 21, n_{Female} = 21). (While Kass (1964) mentions in the paper 52 participants, the abstract describes the study as including 21 boys and 21 girls.) While d = 0.80 represents the choices in the last 30 trials, this is only analysis where the data were described, so, the fact that the (presumable)

earlier is coded as d = .80. While this represents the choices in the last 30 trials before the forced choice it also suggests that the (presumable) lack of a gender gap in the forced choice part of the experiment is not represented.

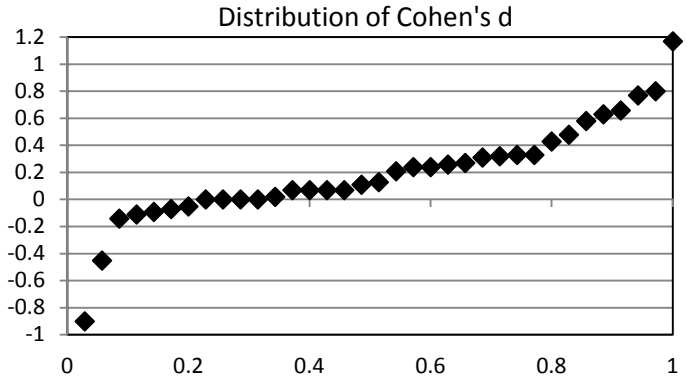

Figure 8: Distribution of the 35 effects sizes (women more risk averse than men) from 17 papers that use incentivized gambles and are analyzed by Byrnes, Miller and Shafer (1999)

So, while there seems to be a gender difference in risk aversion, about 20% of the effects show men to be more risk averse than women, and only about 50% of the studies have an effect size bigger than 0.2, what is often considered the cutoff for a small effect. Recall that, assuming normal distributions, a Cohen's d of 0.2 still only implies a 56% chance that a random man is less risk averse than a random woman. Analogously, about 600 subjects would be needed to get a significant gender effect at 10 percent assuming a power level of 0.8.

This first summary of the psychology literature already suggests a very moderate message: While gender differences in risk taking exist, and women are more risk averse than men, those differences are small. In fact many studies with sample sizes of a few hundred will not find significant gender differences. Furthermore, because the effect is small, gender differences in risk aversion will probably not account for gender differences in many experimental findings nor for large differences in many economic decisions.

V.B Early and most commonly used elicitation methods in economics

The oldest published paper in economics on gender differences in risk aversion that I found is by Schubert, Gysler, Brown and Brachinger (1999), in the American Economic Review, Papers &

lack of a gender gap in the forced choice part of the experiment is not represented is by no fault of Byrnes, Miller and Shafer (1999).

Proceedings. In their experiments 76 men and 65 women provide certainty equivalents for four gambles. “Payoffs in Swiss francs (1 SFr = $ 0.60) and their probabilities were (30 SFr, 1/6; 10 SFr, 5/6), (30 SFr, 1/2; 10 SFr, 1/2), (30 SFr, 5/6; 10 SFr, 1/6) and (50 SFr, 1/2; 20 SFr, 1/2), respectively” (p 382). Subjects were either in a “context” treatment, in which case they first made four decisions in an investment/gain frame, and then the same four in an insurance/loss frame (where losses were relative to an initial endowment). In the “abstract” treatment, participants made the same choices, though choices were now framed as abstract gambling decisions. While there are no gender differences in the context treatment, in the abstract gambling treatment women are more risk averse in the gain domain - their certainty equivalent is almost lower by one - but more risk seeking in the loss domain where the female certainty equivalent is higher by 1.3, when controlling for disposable income per month in thousands of Swiss francs.

This first paper suggests that economic style experiments may lead to a similar conclusion more so than psychology experiments. Small samples may often not find gender differences in risk aversion, which suggests that gender differences in risk aversion, even if they exist, are probably rather small. This is only the beginning of a large literature in experimental economics on measuring risk aversion and checking for gender differences therein. While many different methods have been used to study gender differences in risk aversion, two stand out in their adoption by other researchers.

Probably the most popular method is by Holt and Laury (2002). They had students make a series of binary choices between two lotteries, Option A or Option B, where the variance of Option A is lower. In each row, the set of possible outcomes of Option A (and B, respectively) are held constant; what varies is only the probability to receive the higher outcome. The choices are presented in Table XX below (where subjects, however, did not see the third column that computed the expected payoff difference between Option A and Option B). A risk neutral subject would choose Option A in the first 4 choices, and then Option B thereafter. A risk averse person may switch to Option B only later, as the probability of the best outcome increases, though everyone should choose Option B in their last choice.

The standard Holt-Laury task is to give subjects those ten choices (or sometimes less, eliminating the final rows). Then one row is randomly chosen, and subjects are paid the outcome of their chosen lottery in that row.

Option A	Option B	Expected Payoff Difference
1/10 of $2.00, 9/10 of $1.60	1/10 of $3.85, 9/10 of $0.10	$1.17
2/10 of $2.00, 8/10 of $1.60	2/10 of $3.85, 8/10 of $0.10	$0.83
3/10 of $2.00, 7/10 of $1.60	3/10 of $3.85, 7/10 of $0.10	$0.50
4/10 of $2.00, 6/10 of $1.60	4/10 of $3.85, 6/10 of $0.10	$0.16
5/10 of $2.00, 5/10 of $1.60	5/10 of $3.85, 5/10 of $0.10	-$0.18
6/10 of $2.00, 4/10 of $1.60	6/10 of $3.85, 4/10 of $0.10	-$0.51
7/10 of $2.00, 3/10 of $1.60	7/10 of $3.85, 3/10 of $0.10	-$0.85
8/10 of $2.00, 2/10 of $1.60	8/10 of $3.85, 2/10 of $0.10	-$1.18
9/10 of $2.00, 1/10 of $1.60	9/10 of $3.85, 1/10 of $0.10	-$1.52
10/10 of $2.00, 0/10 of $1.60	10/10 of $3.85, 0/10 of $0.10	-$1.85

Table 3: The Ten Paired Lottery-Choice Decisions with Low Payoffs (Holt and Laury, 2002)

In their baseline treatment Holt and Laury (2002) have subjects first make choices for each line in Table 3 above with the understanding that one of those choices would be paid. In round 2, subjects made the same choices again, with hypothetical payoffs at 20 times the level of Table 3. In the third round, subjects once more choose with payoffs 20 times higher than the payoff in round one, with one round chosen for payment. "To control for wealth effects between the high and low real-payoff treatments, subjects were required to give up what they had earned in the first low-payoff task in order to participate in the high-payoff decision. […] Nobody declined to participate so there is no selection bias." (p 1646). The final round 4 was that subjects made real choices once more with the initial low payoffs from Table 3. 93 subjects made those four choices, while 25 subjects had rounds 1, 2 and 4 only and 57 rounds 1, 3 and 4 only. Finally, 19 subjects had a treatment where instead of multiplying payoffs by 20 they were multiplied by 50 in rounds 2 and 3, and a final 18 subjects had payoffs in rounds 2 and 3 multiplied by 90. The choices of subjects are summarized in the Figure 9 below, which clearly shows that as stakes increase, subjects become more risk averse.

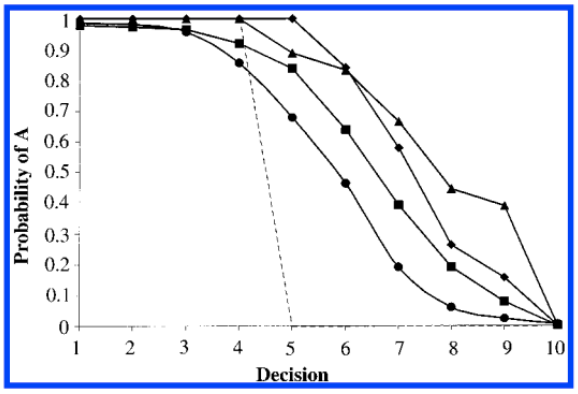

Figure 9: Proportion of Safe Choices in each decision: Data averages and predictions

Data averages for low real payoffs [solid line with dots], 20x real [squares], 50x real [diamonds], 90x real [triangles], and risk-neutral prediction [dashed line]

At the end of their Section II (p 1651) Holt and Laury have the following sentences. "Using any of the real-payoff decisions to measure risk aversion, income has a mildly negative effect on risk aversion ($p < 0.06$). Other variables (major, MBA, faculty, age, etc.) were not significant. Using low-payoff decisions only, we find that men are slightly less risk averse ($p<0.05$), making about 0.5 fewer safe choices. […] The surprising result for our data is that this gender effect disappears in the three high-payoff treatments."

A second common way to assess gender differences in risk aversion is what became known as the Eckel-Grossman (EG) task, and was used by Eckel and Grossman (2002). It is based on a method originally used by Binswanger (1980). Eckel and Grossman (2002) study gender differences in choices over lotteries, as well as gender differences in beliefs about each other's risk behavior. Subjects were shown a sheet with five possible gambles and choose one, see the Table 4 below.[54] One option is a risk-less sure payoff, and the other options are 50:50 gambles where both the variance and the mean in payoffs increase.

Option	Event	Probability	Outcome
1	A	50	16
	B	50	16
2	A	50	24
	B	50	12
3	A	50	32
	B	50	8
4	A	50	40
	B	50	4
5	A	50	48
	B	50	0

Table 4: Five option of the Eckel-Grossman task, Eckel and Grossman (2002)

[54] Eckel and Grossman (2002) had both a gain and a loss treatment. In the "loss aversion" treatment, subjects first received $6, and all payoffs from the Table were reduced by 6. "Subjects in the Loss treatment were informed that if they selected "…either Gamble 4 or 5 and Event B occurs, your losses will be deducted from your $6 fee for completing the survey…"."(p 286). 149 students participated in the Loss treatment (eight sessions) and 55 in the No-Loss treatment (five sessions). They found no difference between the loss and the gain frame for either men and women, and hence pool the data across treatments.

“Comparing men's and women's gamble choices, we found that women were significantly more risk averse than men. For example, less than 2% of the men, but over 8% of the women, chose the least risky gamble, whereas over one-third of the men, but only 13% of the women, selected the riskiest gamble. The median gamble choice was 4 and 3 for men and women, respectively. Men's mean gamble choice was 3.72 (95% confidence intervals: 3.49–3.95) versus 3.10 (2.87–3.33) for women, a significant difference [t(198)=3.83, P<.001].” (p 287). For the distribution of male and female choices see Table 5 below.

Choices	Men	Women
1	2	8
2	17	18
3	25	40
4	24	17
5	36	13
Average	3.72	3.10

Table 5: The choices of women and men in Eckel and Grossman (2002)

To summarize, in the very first experimental study by Schubert et al (1999) with a total of four effects (or treatments), women were found to be more risk averse in one, more risk seeking in one, and not significantly different from men in two treatments. However, in the environment that received the most attention by experimental economists, abstract gambles in the gains domain, women were found to be more risk averse than men. Of the two most common risk elicitation methods used to study gender differences in risk aversion, Holt and Laury (2002) found women to be more risk averse than men when stakes were around a few dollars. However, there were no gender differences in risk aversion in the three treatments where payoffs were around ten dollars or more. Finally, Eckel and Grossman (2002) found gender differences in risk aversion with their elicitation method. Overall, it would clearly be quite heroic to make strong claims that women are statistically and economically significantly more risk averse than men in the vast majority of settings and to an extent that gender differences in risk aversion could account for many differences in economic experiments and other economic outcomes.

VII.C Early Economic Surveys

There are two (early) surveys in experimental economics on gender differences in risk aversion; Eckel and Grossman (2008c), which seems to have been written quite a bit earlier, and Croson and Gneezy (2009). These two surveys, as well as the survey by Byrnes, Miller and Shafer (1999) described earlier all reach the same overall conclusion, namely that women are more risk averse than men. However this overall message is delivered with quite different forcefulness. The most moderate is perhaps Byrnes, Miller and Shafer (1999) who write that "the majority (i.e. 60%) of the effects support the idea of greater risk taking on the part of males." In fact, nearly half the studies (48%) had Cohen's d effect sizes larger than 0.2 (the conventional cutoff point for small effects). However, in a sizable minority (i.e. 40%) was the effect size either negative – that is, males were found to be more risk averse than females - or close to zero. Eckel and Grossman (2008) in their conclusions write: "The findings from field studies conclude that women are more risk averse than men. The findings of laboratory experiments are, however, somewhat less conclusive. While the preponderance of laboratory evidence is consistent with field evidence, there is enough counter-evidence to warrant caution." This message is somewhat less moderate in their introduction: "In most studies, women are found to be more averse to risk than men. Studies with contextual frames show less consistent results." And the strongest conclusions are reached by (Croson and Gneezy, 2009) on page 449: "The robust finding is that men are more risk-prone than are women."

While Byrnes, Miller and Shafer (1999) suggested that "the gender gap seems to be growing smaller over time", the opposite seems to have happened in the experimental economics literature. However, it could also be that different authors interpret the existing evidence differences. For example, Croson and Gneezy (2009) summarizing Eckel and Grosman (2008c) and Byrnes, Miller and Schafer (1999) write that "Previous surveys of economics report the same conclusions: women are more risk averse than men in the vast majority of environments and tasks" (p.449).

To address whether the findings on gender differences in risk aversion have changed in the experimental economics literature, Table 6 considers the papers summarized by each survey, focusing on experiments using objective lotteries. Eckel and Grossman (2008c) review 14

papers, and seem to have aimed to provide a review of the existing literature at the time of writing the survey, which was quite before 2008 (see their Table 1 in the paper, which however fails to include one paper I take the liberty to add in the Table 6 below).[55] Croson and Gneezy (2009) review 10 papers in their survey (see their Table 1) without providing any obvious selection criterion. They have two papers by Eckel and Grossman published in 2008, though the two references refer to the two survey chapters in the Handbook of Experimental Economics Results (one on risk and one altruism and cooperation, cited here as Eckel and Grossman 2008c and 2008b respectively). I will treat the two as Eckel and Grossman (2008). Croson and Gneezy also mention in a footnote an 11th study they dismiss.[56] Table 6shows for each paper covered in either Eckel and Grossman (2008c) or Croson and Gneezy (2009) whether only one of them or both surveyed it, as well as which, if any, gender was found to be more risk averse. Some papers have multiple results and are hence in several columns.

Table 6: Experiments on objective lotteries surveyed by Eckel & Grossman and Croson & Gneezy.

More risk averse	Women	Neither	Men
	Moore & Eckel 2003 Brinig 1995 Levy et al 1999	Moore & Eckel 2003	Moore & Eckel 2003
Eckel and Grossman Only		Schubert et al 2000 Gysler et al 2002 Harbaugh et al 2002 Kruse & Thompson 2003 Harrison et al 2005	
Eckel and Grossman as well as Croson and Gneezy	Schubert et al 1999 Holt & Laury 2002 Levin et al 1988 Powell & Ansic 1997 Hartog et al 2002 Eckel & Grossman 2008	Schubert et al 1999 Holt & Laury 2002	Schubert et al 1999
Croson and Gneezy	Finucane et al 2000		

[55] Interestingly, Eckel and Grossman (2008) do not cite the survey by Byrnes, Miller and Shafer (1999). Only one of the studies included in Byrnes, Miller and Shafer (1999) was summarized by Eckel and Grossman (2008), who in turn include two studies published in '95 and '97 (one in William and Mary Journal of Women and the Law, and the other in the Journal of Economic Psychology) that were not included in Byrnes, Miller and Shafer (1999). Furthermore, Eckel and Grossman have two papers, Eckel and Grossman (2002) and Eckel and Grossman (2008) that are always mentioned together. It seems that the data from Eckel and Grossman 2008 are (almost) the same as those of Eckel and Grossman (2002), the former has 261 subjects and the latter 2004, though no paper cites the other.

[56] They write that the study "finds no significant risk differences in estimations of prospect-theory preferences (no gender differences in loss aversion or in the curvature of the value function). However, they do not report gender differences in risk aversion parameters from traditional expected utility models." (p 449).

Only	Dohmen et al 2005 Fehr-Duda et al 2006

The papers are ordered such that papers with multiple but different effects are listed first, otherwise papers are listed in order of the year they were published (be it working paper or publication).

Excluding papers that find both evidence of no gender differences and of women being more risk averse, Eckel and Grossman (2008c) cite six papers that found women to be more risk averse and five that found no gender differences. It is easy to see how they reached the conclusion that while there is evidence that women are more risk averse than men, there is enough counter-evidence to warrant caution. Table 6 also makes it clear why Croson and Gneezy (2009) reached a much stronger conclusion. Apart from Schubert et al (1999) - the first experimental economics paper on gender differences in risk aversion in a top economics journal – and Holt & Laury (2002) all the papers they surveyed found women to be more risk averse than men.[57]

One reason different surveys reach different conclusions is the heterogeneity of experimental results. This could have two potential reasons. First, it could be that there is a wide range of results because risk preferences are very malleable and subject to framing. It could also be that gender differences in risk aversion, while perhaps statistically significant, are economically small such that samples of a several hundred do not yield reliable results. In that case the existing pattern of published papers would suggest a publication bias favoring papers in which women are more risk averse and penalizing papers that find men to be more risk averse. A second possible explanation for the heterogeneity of experimental results is that different elicitation methods measure different aspects of risk preferences and the gender gap in risk preferences is dependent on the specific way in which attitudes towards risk are measured.

Different explanations suggest different advice concerning control treatments designed to assess the role of risk aversion in the main result of any given experiment. If gender differences in risk preferences are economically small but the results are sufficiently noisy that small samples may not reflect the general finding, then experiments should assess the risk preferences of their sample rather than assuming specific risk distributions. If gender differences in risk depend on

[57] Note that of the six papers cited by Eckel and Grossman (2008c) that found no gender difference in risk aversion, three were at the time already published in economics journals.

the specific risk measure, then each experiment that attributes some portion of gender differences in a given task to gender differences in risk aversion should be careful to choose an elicitation method germane to the task at hand.

One way to address which of the two possible explanations is more responsible for the variety of results is to assess the extent of gender differences in risk attitudes using a single experimental method and capturing all (or at least many) papers that use that method. This point has been made both by Eckel and Grossman (2008c) as well as Byrnes, Miller and Shafer (1999).

VII.D Recent Economic Surveys and Meta-analyses on specific elicitation tasks

The first paper in economics that I'm aware of that summarizes risk preference experiments employing the same method involves a new task to measure risk aversion. It was originally developed by Gneezy and Potters (1997) to assess whether there are differences in risk aversion due to framing.[58] In their investment game, agents receive a fixed amount of money, $\$X$ and can decide to invest any part x of X in an investment. The investment yields dividends of kx with probability p and nothing otherwise. The several papers using this game were designed to study questions different from gender differences in investment behavior.

Charness and Gneezy (2012) summarize a series of papers that use the Gneezy and Potters (1997) investment game, where the values of k and p are such that $pk > 1$, meaning a risk neutral agent would invest everything, i.e. $x = X$. Charness and Gneezy (2012) "report data from all studies (of which we are aware) using this method for testing risk aversion", which turns out to be 14. They find that in all but one study women invest less than men. This leads to the following summary statement (made in their introduction) that "The striking and consistent result is that despite the large environmental differences among the sets of experiments, a consistent gender difference is reported: Men choose a higher x than women do."[59]

[58] Interestingly, in the Croson and Gneezy (2009) survey, it was not used as an example of gender differences in risk aversion.

[59] Nelson (2013) criticizes Charness and Gneezy (2012) by pointing out that average differences may not necessarily translate into significant differences, and those in turn may not necessarily translate to large economic differences on the individual level. She computes Cohen's d for studies found in Charness and Gneezy (2012), for the results see below. The paper also provides tests pertaining to the significance of gender differences and shows that in many papers in Charness and Gneezy (2012), while women invest less than men, those differences are not significant.

The second recent survey I am aware of that focuses on a specific elicitation method is Filippin and Crosetto (2014). They note that while many papers replicate the Holt and Laury (HL) task, not all of them report gender effects. They set out to do a meta-analysis of published results, and tried to get the data from all these papers to present a unified analysis. They ended up with the data from 63 papers for a total of 8713 subjects.[60] To ensure comparability across papers, they code the number of safe choices as the last probability of the high outcome at which Option A, the gamble with lower variance, was chosen over Option B, the gamble with the higher variance (recall Table 3 from Section V.B).[61]

In a first analysis, they provide for each paper for which they have detailed data (54), statistics of subjects who made consistent choices (i.e switched once and did not make dominated choices).[62] They compute "the mean number of safe choices by gender, as well as the results of a non-parametric Mann-Whitney test" (p 12). In 40 papers women are found to be more risk averse, in 7 of which differences are significant at a 5% level. For the remaining 14 papers, men are found to be more risk averse, though not significantly so. Such a mixed message leaning towards a finding there are either no gender differences or that women are more risk averse but not by much, is also found when considering Cohen's *d* for each paper. A total of 3 papers find a medium effect (*d* of 0.5 to 0.8), 23 find a small effect (d of 0.2 to 0.5), 22 a null effect (d less than 0.2, in both directions, i.e. women or men being more risk averse). At the same time 5

[60] A query on 31.1.2013 on Scopus bibliographic database revealed that Holt and Laury (2002) had been cited 528 times, and they found another 26 working papers through conferences and the Economics Science Association discussion group. "We regard as comparable the multiple choice lists in which the amount at stake is held constant while the increase in the expected value of the lotteries is obtained through a higher probability of the good outcome." While 48 papers were not accessible to them, 118 publications (and 17 working papers) replicate Holt and Laury's method, of which 94 publication have their own dataset and have data for both male and female subjects. Of the 94 papers and 17 working papers, they were able to obtain the data from 54 publications and 9 working papers, of which 48 and 6, respectively, shared micro-data and not just summary statistics. When a subject in an experiment made multiple Holt and Laury choices, then, for each subject, only the first such test was chosen.

[61] So, in the usual HL task, as in Table 1 of Section V.B a subject makes 6 safe choices if in the 6th row she chooses Option A, but chooses Option B when the chance of the high outcome is 7/10.

[62] "Females are significantly more likely to be inconsistent" (p 17). About 14 percent of subjects switch from Option B to Option A, where this is done by 12.1 percent of males but 15.8 percent of females. "Inconsistent subjects make on average 5.15 safe choices, without significant gender differences (Mann Whitney test, $p = 0.67$). This number is lower than that of consistent subjects (5.63), and significantly so (Mann Whitney test, $p < 0.001$). At first glance this seems to suggest that inconsistent subjects tend to systematically bias downward the number of safe choices. However a more careful interpretation suggests that inconsistent subjects simply tend to make choices that are closer to a random decision, which in the framework of the HL [Holt and Laury] task coincides with choosing each option half of the times." (p 18).

papers find a small effect, and 1 paper a medium effect, in the opposite direction (i.e., males more risk averse than females).

Filippin and Crosetto (2014) then merge the data sets. "Microdata" consists of all data sets that include every binary choice of participants, while "whole sample" includes data sets that only report the number of safe options a subject chose and whether the subject made consistent choices. Table 7 below shows the mean number of safe choices for both women and men, as well as the standard deviation.

Table 7 Summary statistics of safe choices of consistent subjects

	Mean	St.Dev	N
Whole Sample	**5.63**	**1.91**	**5935**
Males	5.47	1.89	2998
Females	5.78	1.91	2937
Microdata	**5.73**	**1.96**	**4324**
Males	5.59	1.94	2119
Females	5.87	1.97	2205

Notes: Data from papers using the Holt-Laury method that report: Microdata: Every binary choice of all subjects, Whole Sample: The number of safe choices as well as whether the subject made consistent choices. (Filippin and Crosetto, 2014)

On average males seek more risk (make fewer safe choices), significantly so in both samples, though the variance is similar. "The Cohen's *d* on the pooled sample is *d*=0.163, a tiny 16% of a standard deviation, even below the threshold of 0.2 used to identify a small effect. To give an example of how small this is, consider that if we compare two random persons, and assume normal distribution of risk preferences, we would have a [55]% chance of being correct when saying that the more risk averse of the two is a woman, against a 50% rate if we just randomized our answer." (p 18). Put differently, the minimum sample size to get a 5% significant result using a two-tailed t-test study with statistical power level of 0.8 would be about 600 subjects per gender.

Filippin and Crosetto (2014) also discuss other elicitation methods. They report that the sizeable gender gap in choices observed by Eckel and Grossman (2002) in the Eckel-Grossman (EG) task also appear in replications of this task. They cite six papers coauthored by Catherine Eckel, as well as by Crosetto and Filippin (2013b) and Wik et al (2004). Buser, Niederle and Oosterbeek (2014), described in Section II, administer an EG task to almost 400 fifteen year old Dutch

school children and also find significant gender differences in risk taking. The only paper Filippin and Crosetto (2014) are aware of that does not find significant gender differences is Cleave et al (2010). While they replicate the gender gap in a wide sample, a specific subsample of subjects who also participated in later experiments does not find that women are more risk averse.

Filippin and Crosetto (2014) then provide a Cohen's *d* both for the investment game from Gneezy and Potters (1997) and the Eckel Grossman (EG) task from Eckel and Grossman (2002).[63] The average effect size coincides for the two elicitation methods and it is equal to $d = 0.55$, with women being more risk averse than men. To compare the effect size of the EG task (and the investment game) vis-à-vis the HL task, note that the effect size is more than three times as high in the EG than in the HL task. This means if we were to compare two random persons, and assume normal distribution of risk preferences, we would have a 65% chance of being correct when saying that the more risk averse of the two is a woman in the EG task, against a 55% rate in the HL task. Put differently, the minimum sample size to get a 5% significant result using a two-tailed t-test study with statistical power level of 0.8 in the EG task would be about 55 subjects per gender compared to 600 in the HL task.

Filippin and Crosetto (2014) speculate as to what determines the extent to which women are more risk averse than men. They say that it has been argued that HL is more difficult to understand than other methods, making differences harder to detect (e.g. Dave et al, 2010). However, "[t]he SNR [signal to noise ratio (mean/standard deviation)] in our dataset of HL replications is equal to 3.34, higher than the average of the replications of the SNR of the Investment Game [of Gneezy-Potters] (2.06) and the EG [Eckel-Grossman] task (2.41)." (p22).

Given that differences between methods do not seem to stem from a different precision in measuring risk attitudes, Filippin and Crosetto (2014) offer three dimensions that differ between

[63] "For the investment game we use Cohen's d computed by Nelson (2013) for all studies included in the survey paper by Charness and Gneezy (2012). For the Eckel and Grossman task we use the data provided by the papers replicating the task, when available. In both cases we add the Cohen's d computed from our own data presented in Crosetto and Filippin (2013)" (p19).

the Holt-Laury task on the one hand, and the Investment Game and the Eckel-Grossman task on the other hand:

1. Is the menu of lotteries generated by
    - Changes in probabilities (HL)
    - Changes in outcomes (EG and investment game)?
2. What domains of risk preferences are considered? HL measures preferences both in the risk averse and risk loving domain, while other methods do not.
3. Is there a safe option available? A safe option is present both in the Investment Game and EG task, but not in the HL task.

First results of Filippin and Crosetto (2014) indicate that while adding a safe option to HL increases the gender gap in choices, removing it from EG does not seem to reduce the gender gap.

To summarize, the message from the experimental literature is complex. While the overall evidence points to women being more risk averse than men, there is large heterogeneity in the extent of this gender gap. The Holt-Laury task, which is the experimental method for which Filippin and Crosetto (2014) found the most papers employing it, generates a gender gap in risk aversion small enough that experiments using several hundreds of subjects will in general not find significant gender differences. On the other hand, elicitation methods such as the Eckel-Grossman task or the investment game generate a larger gap in risk aversion, 0.55 of a standard deviation. However, in psychology this is considered a medium effect that can be achieved quite reliably with just over 100 subjects. Clearly, understanding when gender differences in risk aversion are present and when they are rather small to almost non-existent remains an open question.

The heterogeneity in results suggests that any given experiment should not presume a specific distribution of risk preferences. Rather, an experiment that aims to assess the impact of risk aversion on the main result should aim to generate a risk measure of the kinds of risk subjects are exposed to in the experiment. That is, the risk measure may have to be very germane to the task at hand.

VII.E Stability of Risk Preferences and their External Relevance

The heterogeneity of results on gender differences in attitudes towards risk suggests a concern whether elicited risk preferences are a reliable measure of a subject's risk attitude. Put differently, to what extent is there a stable risk preference, and which elicitation method comes closest to capturing it? And more importantly, is there a risk measure that captures sufficiently broad risk attitudes and reliably correlates with behavior we expect to depend on the subjects' risk preferences? While the first question concerns internal validity, the second concerns external validity, both of which are important. Eventually, however, the question is whether gender differences in risk attitudes have external relevance? That is, do experimental risk measures correlate with economically relevant behavior or outcomes? Furthermore, given the focus of this chapter, can gender differences in the experimental risk measure account for gender differences observed in economic behavior or outcomes?

I start by describing three ways in which we can assess the external validity of risk measures. These can also be seen as three hurdles to using experimentally elicited risk measures to predict choices outside the lab. The first way or hurdle concerns the stability of risk preferences across elicitation methods but within a domain, or more precisely, for the same lottery choices. Second, is there stability in risk preferences across domains? For example, will risk preferences measured using the EG task correlate with those using the HL task? Or, will risk preferences measured in, say, choices over different car insurances match those of choices over different homeowner or health insurances? Finally, is there stability of risk preferences using the same elicitation method and the same domain? That is, if we ask participants at two separate times the same questions in the same ways, how correlated will their choices be? For each of those three ways to assess external validity of risk measurements I'll mention a few relevant papers, covering early work and some selected more current work. I will then discuss work relating risk preferences to choices outside the lab, focusing on work that uses risk measures to account for gender differences in economic outcomes (given the focus of this chapter).

The Stability of Risk Preferences

A first problem when considering the external relevance of experimentally elicited risk preferences is the considerable heterogeneity of results across elicitation methods even for the

same decisions over uncertain outcomes. For economic-style experiments that show that the same individual may have different risk attitudes depending on the elicitation method see Slovic (1972).[64] The paper uses the two elicitation methods from the preference reversal literature (see Lichtenstein and Slovic, 1971). Specifically, subjects chose between a lottery with a high probability of winning a small amount (the P-bet), for example [30/36 chance to win 250 points and 6/36 chance to lose 230 points] or one with a small probability of winning a large amount of money (the $-bet) for example [9/36 chance to win 980 points and 27/36 chance to lose 100 points]. Subjects are also asked to state their willingness to sell each of these lotteries, or state their certainty equivalent. The common finding is that subjects choose the P bet over the $ bet, but have a higher selling price for the $-bet than the P-bet. Some economists have been intrigued, but skeptical, of this result. Grether and Plott (1979) replicated the preference reversal result and state (p 634): "Needless to say the results we obtained were not those expected when we initiated this study. Our design controlled for all the economic-theoretic explanations of the phenomenon which we could find. The preference reversal phenomenon which is inconsistent with the traditional statement of preference theory remains."

Slovic (1972) asks whether the two methods not only result in a different magnitude of estimated risk preferences, but also in a different ordering of which subjects are more risk averse than others. Specifically, when subjects choose between lotteries they indicated whether the preference was from (1) "slight" to (4) "very strong". In Figure XX each subject is represented as a point, where the x-coordinate is the mean preference for the $-bet using the augmented choice index, and their y-coordinate is the mean difference in selling price between the $-bet and the P-bet. Figure 10 confirms the preference reversal: While the preference of the $-bet using the choice index is negative in general, most subjects have a higher selling price for the $-bet than the P-bet. The correlation between these two measures 0.46. This correlation is similar for women and men, separately (0.4 and 0.55, respectively). Slovic concludes: "The fact that a simple change in response mode can create so much inconsistency among individuals' relative standings in the group implies that high correlations between risk-taking measures in structurally

[64] For early economic papers check out e.g. Harrison (1990) who used different elicitation methods across similar subjects and found different elicited risk measures, and Isaac and James (2000) who estimated individual risk preferences based off two games (and many assumptions) and found that different methods yielded not only different risk estimates, but also different rankings of which subjects are the most risk averse.

different settings and other behaviors are unlikely to be found" (p 133). Very similar results have been obtained more recently by Harbaugh, Krause and Vesterlund (2010).

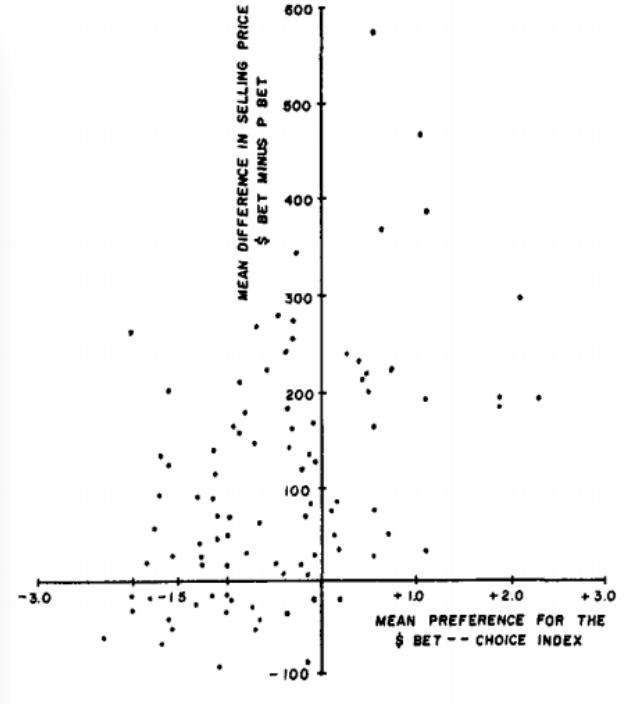

Figure 10: Relationship between choice and selling-price indexes across the total sample of subjects (r = 0.46).

A second issue in determining the external relevance of experimentally elicited risk preferences is whether people even have a unique risk attitude or risk parameter that guides all their decisions. That is, what is the variance across different elicitation methods or across different domains, such as attitudes towards risk in a health or a car insurance domain? Given that gender differences in risk aversion depend on the elicitation method, it is clear that even rankings of ordinal risk attitudes may be different across measurements when the sample contains both women and men. The problem could however also arise within gender.

Slovic (1962) presents one of the first investigations on the correlation of nine different hypothetical risk measures (8 of which can be ordered) that have been previously used by psychologists (see also Slovic 1964).[65] "The results show that only 5 correlations out of 28 reach

[65] For two earlier but in scope more limited investigations see Kogan and Wallach (1960) and Wallach and Kogan (1961).

significance in the predicted direction and none of these correlations exceed 0.34. Another 10 of these correlations are negative, 2 significantly so." (p 69). The paper concludes that "[t]he implications of the present study for the existence and measurement of a general risk taking trait are (a) none or only a few of the variables analyzed actually measure the trait; or (b) willingness to take risks may not be a general trait at all but rather one which varies from situation to situation within the same individual." (p 70).[66]

Blais and Weber (2006) propose a "Domain-Specific Risk-Taking (DOSPERT)" scale for adult populations, see also Weber, Blais and Betz (2002). "The risk-taking scale of the 30-item version of the revised DOSPERT Scale evaluates behavioral intentions, that is, the likelihood with which respondents might engage in risky behaviors originating from five domains of life (ethical, financial, health/safety, social, and recreational risks) using a 7-point rating scale ranging from 1 (Extremely Unlikely) to 7 (Extremely Likely). Sample items include "Having an affair with a married man/woman" (Ethical), "Investing 10% of your annual income in a new business venture" (Financial), "Engaging in unprotected sex" (Health/Safety), "Disagreeing with an authority figure on a major issue" (Social), and "Taking a weekend sky-diving class" (Recreational)." (p 36). Weber, Blais and Betz (2002) find that the degree of risk-taking was highly domain-specific, subjects were not consistently risk averse or risk seeking across all 5 domains. However, women were found to be more risk-averse in all domains apart from social risk. Weber, Blais and Betz (2002) also ask about the perceived benefits and the perceived risk of an action. "The risk-perception scale evaluates the respondents' gut level assessment of how risky each behavior is on a 7-point rating scale ranging from 1 (Not at all) to 7 (Extremely Risky)." These have a large influence on risk taking. "A regression of risk taking…on expected benefits and perceived risks suggests that gender and content domain differences in apparent risk taking are associated with differences in the perception of the activities' benefits and risk, rather than with differences in attitude towards perceived risk." The importance of perceptions of risk is often neglected in economics inquiries where we prefer to control for the risk at hand. See Erev and Haruvy (2015) and Erev and Roth (2014) for a discussion on how focusing on objective risk may provide a distorted view of the importance of various biases in decision making.

[66] For evidence using a between subject design see Harrison, List and Towe (2007).

A hypothetical risk question that is more common in economics is the one used by the German Socio Economic Panel (SOEP), a representative sample of the adult population living in Germany. In one of their waves they introduced the following risk question: "How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks? Please tick the box on the scale, where the value 0 means: 'not at all willing to take risks' and the value 10 means: 'very willing to take risks'. To assess the correlation of this risk question to choices over gambles, Dohmen et al (2011) recruit 450 subjects following the SOEP sampling procedure. Subjects are asked the SOEP risk question from above and, after completing a questionnaire similar to the standard SOEP questionnaire participate in a paid lottery experiment. "[P]articipants were shown a table with 20 rows. In each row they had to decide whether they preferred a safe option or playing a lottery. In the lottery they could win either €300 or €0 with 50% probability (1 ~ US$ 1.2 at the time of the experiment). In each row the lottery was exactly the same, but the safe option increased from row to row. In the first row the safe option was 0, in the second it was 10, and so on up to 190 in row 20" (p 532). To ensure incentive compatibility, 1/7 participants had one of their rows randomly chosen for payment. A regression of the value of the safe option at the switching point on the general risk question shows a significant (and positive) relationship.

More recently, Crosetto and Filippin (2013b) consider a between subject design where subjects do one of the following five incentivized risk measures: The Holt-Laury multiple price list, the Eckel Grossman task, the Gneezy-Potters investment game, the Balloon Analogue Risk Task (Lejuez et al, 2002) or the Bomb Risk Elicitation task (Crosetto and Filippin, 2013a).[67] Every subject also answers the SOEP risk question described above, and the DOSPERT (Blais and Weber 2006). They find that the SOEP question and the DOSPERT score are highly (and significantly) correlated (around 0.57). Comparing questionnaires with experimental outcomes,

[67] The Balloon Risk Analog Task technically has balls drawn from an urn where n-1 balls are safe and one is not. The task is visualized in that participants pump air into a ball and receive money each time they do so. Subjects can decide to stop pumping and collect their earnings. If they continue and the ball explodes (that is, when the unsafe ball is drawn from the urn), subjects earn 0. Subjects are in general not informed what n is (i.e. how many safe balls there are in the urn. The Bomb Risk Elicitation Task is similar, only now subjects are confronted with 100 boxes, know that one of them contains "the bomb" and basically decide how many boxes to open. That is, it allows for a "strategy method" implementation (since the bomb is only detected after subjects decided how many boxes to open) though subjects, through waiting, draw more boxes to open. Second, subjects are informed at any moment how many boxes they already have chosen to be opened later.

the SOEP correlates significantly with HL, EG and the Balloon task (though only with correlations of .23, .30 and .37, respectively.) For the investment game the correlation is 0.13 (not significant) and it is only 0.03 for the bomb task. Furthermore, "after running a linear regression of each choice on the observed demographics (age and gender) as a benchmark, we include each questionnaire separately in the regression, measuring the contribution of the last measure added to the adjusted R^2." (p 21) Only for the EG and the Balloon task is the percentage point change in the adjusted R^2 positive (around 3 and 10, respectively). The results are similar when considering correlation with DOSPERT (though the Balloon task is not significantly correlated with DORPERT).[68]

The fact that there are important domain-specific components in risk preferences is also evident in (non-experimental) empirical work. For example, Einav et al (2012) consider individuals' choices over five employer-provided insurance coverage decisions and one 401(k) investment decision. They consider ordinal rankings in the riskiness of choices of individuals, and find that the average spearman rank correlation is 0.19. This is in large part due to the fact that the correlation between the 401(k) choice and any insurance decision is in general lower than 0.061. However, within insurance choices, there is a domain-general component to preferences that seems substantively important. "For example, we find that one's choices in other insurance domains have about four times more predictive power for one's choice in a given insurance domain than does a rich set of demographic variables." (p 2636).

A third issue when considering the robustness of a trait is whether within a domain and within an elicitation method there is stability over time. Specifically, will a subject show similar responses if asked the exact same risk decision later? Andersen et al (2008) using the Danish population "find some variation in risk attitudes over time, but we do not detect a general tendency for risk attitudes to increase or decrease over a 17-month span." There are not many studies that use a time-frame of a year or more. One exception that in addition uses incentivized choices is Levin et al (2006). Parents and their 6-8 year old children complete a first set of risk experiments and then a follow up roughly three years later. There are significant correlations in choices of both

[68] For other papers that find week correlations of risk preferences across tasks see e.g. also Bruner (2009), Reynaud and Couture (2012) and Andreoni and Sprenger (2011).

children and parents, though children's choices become less correlated with those of their parents.

Can Gender Differences in Risk account for Gender Differences in Economic Outcomes?

Despite all those hurdles, there have been some attempts to correlate risk measures estimated through experiments or questionnaires with choices outside the lab. The first paper I could find is by Ziller (1957). He correlates a risk measure using Swineford (1938) (see Section V.A) with (expected) vocational choices of 182 sophomores from the University of Delaware Army ROTC program. Subjects who expect to work in sales showed the highest index for risk preference, followed by Mechanical Engineering and Education, and those choosing Engineering showed the least tolerance for risk. However, it is not clear per se how different jobs differ in their inherent amount of risk. An early economic paper relating risk measures to economic outcomes is Barsky, Juster, Kimball and Shapiro (1997), that finds that an nonincentivized risk questions can predict various health risks (such as smoking and drinking), immigration status, self-employment and whether the person holds stock. While they find gender differences in the risk question, they do not use them to ask whether gender differences in risk can help account for gender differences in economic outcomes.

A notable line of work was started by Dohmen et al (2011). They consider the hypothetical risk question asked on the 2004 SOEP "How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks? Please tick the box on the scale, where the value 0 means: 'not at all willing to take risks' and the value 10 means: 'very willing to take risks'. A large fraction of the 22, 019 individuals in 11,803 households of the 2004 SOEP answered the risk question, as well as risk questions that asked about willingness to take risks in a specific context: car driving, financial matters, sports/leisure, career, and health. They find that women see themselves as less willing to take risk (by 0.6). Likewise, older people (in years) and shorter people (in cm) are also less willing to take risk, with coefficients about 5% of the gender coefficient.[69] Most importantly "[a]ll of the survey measures are shown to explain various risky behaviors, including holding stocks, smoking, self-employment, and participation in active sports. The best all-round predictor is the general risk question. On the other hand, asking about

[69] These results are robust when including control variables for income and wealth.

risk attitudes in a more specific context gives a stronger measure for the corresponding context." (p 542). Unfortunately, there is no direct analysis about the extent with which gender differences in behavior such as stock holding are accounted for by gender differences in the risk measure.

Dohmen and Falk (2011) exploit the fact that the 2004 SOEP also asked "whether the performance of a respondent is regularly evaluated in a formal procedure, a requisite element of performance contingent remuneration schemes." (p 585). They find that more risk tolerant workers as measured by the SOEP question are more likely to work in jobs with performance evaluation and that women are less likely to work for variable pay than men. Unfortunately, no direct link is given as to how much of the gender gap in work for variable pay could be accounted for by gender differences in risk tolerance.

Dohmen et al (2011) correlate the SOEP risk question with incentivized lottery choices of one set of participants, as well as with economic outcomes of participants in the 2004 SOEP. Of course, in and of itself this does not imply that the reason the risk question predicts economic outcomes is due to the component that correlates with incentivized lottery choices. It could be that the risk question captures other behavioral attitudes of participants that, while correlated with behavior and economic outcomes, do not correlate with risk preferences.

To assess whether the risk question correlates due to its capturing risk preferences, one could, for example, consider the behavior of subjects who both answer a risk question as well as provide data on an incentivized lottery choice. Lonnqvist et al (2011) have participants play a trust game (see Section IV), make choices in an incentivized HL task, and answer the SOEP risk question, among others. The paper finds that the two measures of risk-attitudes are uncorrelated, though both correlate with the decision of a trustor in a trust game. However, the coefficient of either risk measure on behavior of the trustor (the first mover in the trust game), as well as the impact of including a risk measure on the adjusted R^2 of the behavior of the trustor, are virtually unaffected whether the other risk measure is already controlled for. This suggests that the two risk measures are almost orthogonal to each other in terms of accounting for behavior in the trust game, a fact that is already suggested by the lack of correlation between the two risk measures.

Buser, Niederle and Oosterbeek (2014), discussed in more detail in Section II, measured not only the competitiveness of almost 400 children in four schools in the Netherlands using a Niederle-Vesterlund elicitation method, but also their risk attitudes. BNO used both an Eckel-Grossman task, where subject could choose one of five gambles (a sure payoff of €2 and four 50/50 lotteries with increasing riskiness and expected payoffs: 3 or 1.5; 4 or 1; 5 or 0.5; 6 or 0) and the non-incentivized risk question used in the SOEP studied by Dohmen et al (2011). BNO found that both risk measures are correlated with tournament entry choices. However, when assessing whether education choices are correlated with risk preferences, only the EG lottery measure was significantly correlated, while the nonincentivized risk question was not. Furthermore, including the EG risk question significantly reduced the gender gap in education choices, while the unincentivized risk question did not. That is, while risk preferences as measured by EG accounted for a significant fraction of the gender gap in education choices, this was not the case for the answer to the question "'How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?''

Clearly this line of work needs to be expanded and needs to confirm that experimentally measured gender differences in risk preferences are able to account for gender differences in economic outcomes or choices.

VII.F An Example of a careful control for risk aversion

Given that the first paper on gender differences in risk aversion discussed concerned decisions in a multiple choice exam, the last paper discussed will close this loop and provide a modern view on the possible effect of gender differences in risk aversion in accounting for gender differences in exam grades. Multiple choice exams are, in the US, ubiquitous and important, a prime example is the SAT which is required by many colleges. While SAT scores seem to predict college grades, women perform relatively worse on multiple choice tests compared to essay questions, and their SAT scores under-predict their college performance (see references in Coffman, forthcoming).

Coffman (forthcoming) directly tests the extent to which gender differences in test taking (and scores) can be attributed to gender differences in risk aversion, as opposed to other effects, such

as gender differences in ability or confidence. Subjects in her experiment first answered SATII U.S. and World history practice questions, where each question had four (instead of 5, as the real SAT) possible answers. Subjects received 1 point for every correct answer. There are two treatments: In the no-penalty treatment, subjects were not penalized for wrong answers; in the penalty treatment, subjects were penalized by ¼ of a point for each wrong answer. In both treatments subjects earned 0 points for each skipped question. Therefore, in both treatments, a risk neutral subject should answer all questions.

In part II of the experiment, subjects were offered 20 gambles, aimed to assess their attitudes towards lotteries that mimic the risk from answering questions on the part I test. Specifically, subjects answered gambles where they had a chance of winning 1 point (ranging from 25% to 100% chance of winning) and in the other event lost X points, where X= 0 for subjects in the no penalty condition and X=¼ in the penalty condition. Subjects could also decline the gamble, and get 0 for sure. Accepting such a gamble with a 75% chance to win is akin to answering the SAT question where subjects are 75% confident that they know the correct answer. To assess knowledge, part III of the experiment had subjects answer the same questions from Part I, but subjects were not allowed to skip questions. In addition, an incentivized measure of confidence was elicited using a belief elicitation procedure that mimics a BDM as employed by Mobius et al (2014). That is, for each question, participants provided the probability with which they thought their (forced) answer was correct.[70] Subjects were paid for one of the three parts of the experiment. Beyond the difference in penalties for a wrong question, there were two additional treatment designs. One was the no frame treatment, while the frame treatment emphasized that the questions were SATII practice questions. This could make a difference since participants were US college students who presumably received a lot of coaching on how to approach SAT's, and as such may have internalized not to skip questions.

Coffman shows that in the no penalty condition, basically no one skips questions. In the penalty condition without a frame, men skip 2 questions (1 with SAT frame), compared to the 3.7 (2 with

[70] After eliciting the chance Z with which the answer was correct, a random number R was drawn. If R<Z or R=Z, then the "final" answer was the subjects answer. If R>Z, then the final answer was the correct answer with probability R. Subjects received 1 point if the final answer was correct. If the final answer was wrong, they lost X points, where X=0 in the no penalty condition and X=1/4 in the penalty condition. For a discussion on the incentive compatibility of such a mechanism.

SAT frame) questions skipped by women, a significant difference in both cases. Women are about 6.5 percentage points more likely to skip a question, where about 10% of this gap can be accounted for by gender differences in knowledge as measured by Part III. Therefore, women are still about 6 percentage points more likely to skip a question compared to men.[71] While beliefs are found to predict whether a subject answered correctly, there are no gender differences in beliefs conditional on measured knowledge of the material. Conditional on beliefs of knowing the answer in Part III, women are more likely to skip that same question in Part I then men are. In fact, the gender gap on question skipping is hardly affected by controlling for beliefs and remains significant at 5.9 percentage points.

Concerning risk, women are significantly more risk averse then men in the penalty treatment. The mean probability of success of the riskiest bet taken by men is 39.46% chance of winning, compared to 43.44% for women.[72] Regressions confirm that risk accounts for roughly one-third of the gender gap in skipping questions; however, the remaining gap of 4 percentage points is still significant. So, while gender differences in risk account for a significant portion, about one third of the gender gap in SAT test taking, the results suggest that it may be far too hasty to attribute the whole gender gap to risk aversion.

Coffman (forthcoming) then studies the impact of the gender difference of question skipping on the final score and shows that as a result women receive lower test scores than men with the same knowledge of material. Specifically, women score about four-tenths of a point worse on the 20 point test, which corresponds to approximately 1/12 of a standard deviation in Part I scores. This result is perhaps a great example how a small gender difference in question skipping – 6 percentage points of which about 2 can be attributed to gender differences in risk aversion - can accumulate over a longer test to much larger total effect. Recently, Tannenbaum (2012) analyzed data from a subsample of the Fall 2001 mathematics SAT and found that women skip significantly more questions than men. The paper exploits variations in the penalties for

[71] The low impact of knowledge isn't surprising, since on average, when subjects aren't allowed to skip questions, women have 11.9 correct answers compared to 12.7 for men.

[72] Note that the gender differences in question skipping seem not to be driven by gender differences in ambiguity aversion. Clearly, deciding to answer a question is a more ambiguous gamble than the gambles faced by subjects in Part 2 of the experiment. Ambiguity aversion would suggest that subjects would be more likely to decline the ambiguous gamble (i.e. not answer a question) than the objective gamble in Part II. However, Coffman (forthcoming) finds that subjects are in general more willing to accept the ambiguous gambles.

answering a question wrongly, and confirms the conclusion of Coffman (forthcoming) that roughly 40% of the gender gap in test scores can be attributed to gender differences in risk aversion.

The risk assessment used by Coffman can serve as guidance on how to handle the question whether gender differences in risk aversion can account for gender differences in the question or task at hand. Specifically, the risk assessment very nicely complemented the belief question on chances of answering a question correctly. It would, of course, not be justified to consider the exact portion of the gender gap in question skipping attributable to gender differences in risk aversion as a fixed constant for all environments and subjects. However, the paper provides a strong piece of evidence that in an environment in which gender differences in performance in risky environments are important, gender differences in risk aversion may be far from accounting for the total difference, or perhaps even the majority of the gender gap.

VII.G Conclusions

Throughout this chapter, I have provided examples of papers where the control for risk aversion was germane to the task at hand, which, however, does not allow to compute specific parameters of risk aversion, see e.g. Gneezy, Niederle and Rustichini (2003) and Niederle and Vesterlund (2007) in Section II.

Overall, gender differences in preferences towards risky prospects seem to exist, though they vary considerably depending on the elicitation method. Some methods such as the Eckel-Grossman task (Eckel and Grossman, 2002) quite reliably produce results where women behave as if they are more risk averse. Others, such as the Holt-Laury method (Holt and Laury, 2002), in general do not find that women are significantly more risk averse than men. A meta-study by Filippin and Crosetto (2014) including several thousands of women and men found a statistically significant gender difference, with women being more risk averse. However, the gender gap is only about 16% of a standard deviation, and assuming a normal distributions of risk preferences, if a random man and a random woman are compared, there would be a 53% chance of being correct when saying that the more risk averse of the two is a woman. Therefore, experiments with several hundred subjects may not reliably find gender differences in risk aversion.

More work is needed to understand the exact nature of gender differences in risk aversion. This heterogeneity of results on gender differences in risk aversion is also present when considering whether a risk parameter has external relevance, and which elicitation method is most likely to capture a risk parameter that can account for various economic outcomes.

Finally, the heterogeneity of results on gender differences in risk aversion suggests that an experiment should probably employ an elicitation method that is germane to the task in question. At the very least it suggests that gender differences in various experiments cannot automatically be attributed to gender differences in risk aversion.

VI. CONCLUSIONS

The last decade has seen an explosion of experiments documenting gender differences in behavior. In this chapter I focused on only three such traits: attitudes to competition, altruism or cooperation and risk. While gender differences are large and robust in attitudes to competition, they are at times small for cooperative and altruistic attitudes. This seems to be at odds with the "common wisdom" and with some of the perhaps too strongly formulated conclusions of previous summaries of the literature.

One insight into the causes for the discrepancy between documented results and beliefs can be gained from Eckel and Grossman (2002). They had subjects choose one out of five gambles, where choice 1 was a certain payoff of $16, while choices 2-5 were 50-50 gambles of dollar amounts (24,12), (32,8), (40,4) and (48,0) respectively. They had subjects not only pick a choice for themselves, but also guess what choices others made. "For the forecasting task, each subject stood in turn and was visible to all others in the room. The other subjects indicated on their prediction forms which of the five choices they thought the standing person had chosen. For every correct prediction, they received a $1 bonus. Forms were collected and matched with decisions, and payoffs for this task were calculated." (p 286).

"[T]here was consensus between the sexes regarding men's risk aversion but not women's. The mean predictions for men did not differ significantly by sex (3.33 by men vs. 3.26 by women,

t=1.06, P=ns), but men under-predicted women's risk acceptance even more than did women (2.48 and 2.61, respectively, t=2.12, P<.02)." (p 289).

The results suggest that men believe the gender gap in risk aversion to be larger than females do, who, in turn, do not overestimate the gender gap. The literature on belief differences between different groups of subjects is still in its early stages, though for notable early work see Fershtman and Gneezy (2001), Mobius and Rosenblatt (2006), Bohnet, van Green and Bazerman (forthcoming). It may very well be that for many traits both women and men have beliefs that exacerbate the existing gender gap.[73]

While clear results on gender differences in preferences start to emerge, there is still work to do concerning the external validity of findings. For example, which measures of risk aversion are correlated with choices in other tasks, and which are better able to predict behavior out of sample? The biggest gap in the literature, however, concerns the external relevance of laboratory findings. To date there have been few datasets and papers combining traits with behavior outside the laboratory, and even fewer assessing whether gender differences in a trait that can account for gender differences in economic outcomes.

One way to facilitate such endeavors is to provide more research linking easy-to-use hypothetical measures with incentivized experimental measures. That is, to what extent is a non-incentivized hypothetical choice such as "Do you think of yourself as someone who is eager to participate in competitions?" or a non-choice measure such as "Do you enjoy being in a competition" correlated with incentivized tournament entry decisions à la Niederle and Vesterlund (2007)? And, more importantly, when predicting behavior out of sample, or in a different environment, how much is lost when using hypothetical or non-choice measures compared to incentivized measures? Note that for such non-incentivized measures to be useful three criteria have to be fulfilled. First, non-incentivized measures as well as incentivized measures have to correlate with choices outside of the laboratory. Second non-incentivized measures have to correlate with incentivized measures. Third, and perhaps most importantly there has to be evidence that the

[73] Bordalo, Gennaioli and Shleifer (2014) suggest a simple mechanism for how stereotyping can exacerbate existing differences.

non-incentivized measure captures some variation of behavior associated with the incentivized measure and ideally not much more. Specifically, it should certainly not be the case that when including both the incentivized and the non-incentivized measure, they act as if they were two orthogonal measures of economic behavior.

Evidence linking behavioral traits with behavior outside of the laboratory is crucial to demonstrate the value of behavioral corresponding laboratory experiments. I hope that the next *Handbook of Experimental Economics* will have sufficient work that there could be a chapter covering the external relevance of behavioral traits.

REFERENCES

Almås, Ingvild, Alexander W. Cappelen, Kjell G. Salvanes, Erik Ø. Sørensen, and Bertil Tungodden (2014). "Willingness to compete: Family matters", NHH Discussion Papers SAM 3/2014

Altonji, Joseph G. and Rebecca M. Blank, 1999, "Race and Gender in the Labor Market," in: Orley C. Ashenfelter and David C. Card (Eds) Handbook of Labor Economics, Vol 3, Elsevier.

Andersen, Steffen, Seda Ertac, Uri Gneezy, John A. List, and Sandra Maximiano, "Gender, Competitiveness, and Socialization at a Young Age: Evidence From a Matrilineal and a Patriarchal Society," *Review of Economics and Statistics* 2013 95:4, 1438-1443.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutstrom. 2008. "Lost in State Space: Are Preferences Stable?" *International Economic Review* 49 (3): 1091–1112

Andreoni, James, "Cooperation in Public Goods Experiments: Kindness or Confusion?" *American Economic Review*, v.85, no.4, September 1995, 891-904.

Andreoni, James and B. Douglas Bernheim, "Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects," Econometrica, Vol. 77, No. 5 (September, 2009), 1607–1636.

Andreoni, James, Eleanor Brown and Isaac Rischall, "Charitable Giving by Married Couples: Who Decides and Why Does It Matter?,"The Journal of Human Resources, Vol. 38, No. 1 (Winter, 2003), pp. 111-133

Andreoni, James and Charles Sprenger, 'Uncertainty Equivalents: Testing the Limits of the Independence Axiom," 2011 WP NBER w17342.

Andreoni, James, and Lise Vesterlund, "Which is the Fair Sex: Gender Differences in Altruism," *Quarterly Journal of Economics*, CXVI (2001), 293–312.

Apicella, Coren L., Anna Dreber email the author, Peter B. Gray, Moshe Hoffman, Anthony C. Little, Benjamin C. Campbell (2011). Androgens and competitiveness in men. Journal of Neuroscience, Psychology, and Economics, 4(1), 54-62.

Ariely, Dan, Uri Gneezy, George Lowenstein, and Nina Mazar (2009), "Large Stakes and Big Mistakes." *Review of Economic Studies* 76, 451–469.

Azrieli, Yaron, Christopher P. Chambers, and Paul J. Healy, "Incentives in Experiments: A theoretical Analysis," working paper, 2014.

Balafoutas, Loukas, Rudolf Kerschbamer, Matthias Sutter, "Distributional preferences and competitive behavior," *Journal of Economic Behavior & Organization*, Volume 83, Issue 1, June 2012, Pages 125-135.

Balafoutas Loukas and Matthias Sutter, 2012, "Affirmative Action Policies Promote Women and Do Not Harm Efficiency in the Laboratory," *Science*, 335, 579-582.

Balliet, Daniel, Norman P. Li, Shane J. Macfarlan and Mark Van Vugt, "Sex Differences in Cooperation: A Meta-Analytic Review of Social Dilemmas," Psychological Bulletin, 2011, vol 137, No6, 881-909.

Bardsley, Nicholas, "Dictator game giving: altruism or artefact?," *Experimental Economics*, June 2008, Volume 11, Issue 2, pp 122-133.

Barsky, Robert B., F. Thomas Juster, Miles S. Kimball and Matthew D. Shapiro, "Preference Parameters and Behavioral Heterogeneity: An Experimental Approach in the Health and Retirement Study," *Quarterly Journal of Economics*, 1997, May, 537-579.

Bartling, Bjorn, Ernst Fehr, Michel Andre Marechal, and Daniel Schunk. 2009. "Egalitarianism and Competitiveness." *American Economic Review, Papers & Proceedings*, 99(2): 93-98.

Bartling, Björn, Ernst Fehr and Daniel Schunk, "Health effects on children's willingness to compete," *Experimental Economics* 2012 58-70.

Ben-Ner, Avner, Fanmin Kong, Louis Putterman, Share and share alike? Gender-pairing, personality, and cognitive ability as determinants of giving, *Journal of Economic Psychology*, Volume 25, Issue 5, October 2004, Pages 581-589.

Berg, Joyce, John Dickhaut, Kevin McCabe, "Trust, Reciprocity, and Social History," Games and Economic Behavior Volume 10, Issue 1, July 1995, Pages 122–142.

Bertrand, Marianne, 2010, "New Perspectives on Gender," Handbook of Labor Economics, Vol 4, Elsevier.

Binswanger, Hans P., "Attitudes toward Risk: Experimental Measurement in Rural India," *American Journal of Agricultural Economics*, Vol. 62, No. 3 (Aug., 1980), pp. 395-407.

Blais, Ann-Renée and Elke U. Weber "A Domain-Specific Risk-Taking (DOSPERT) scale for adult populations," *Judgment and Decision Making*, Volume 1, Number 1, July 2006 pp. 33-47.

Bohnet, Iris, Benedikt Herrmann and Richard Zeckhauser (2010). "Trust and the Reference Points for Trustworthiness in Gulf and Western Countries." Quarterly Journal of Economics, CXXV(2), May 2010: 811–828.

Bohnet, Iris, Alexandra van Geen and Max Bazerman. When Performance Trumps Gender Bias. Joint Versus Separate Evaluation. *Management Science*, forthcoming.

Bolton, Gary E. and Elena Katok, "An experimental test for gender differences in beneficent behavior", *Economics Letters* 48 (1995) 287-292.

Booth, Alison and Patrick Nolen, "Choosing to compete: How different are girls and boys?," *Journal of Economic Behavior & Organization*, Volume 81, Issue 2, February 2012, Pages 542-555.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. Working Paper. "Stereotypes," 2014

Borghans, Lex, Angela Lee Duckworth, James J. Heckman and Bas ter Weel, 2008. "The Economics and Psychology of Personality Traits," *Journal of Human Resources*, Vol. 43, pp. 972-1059.

Boschini, Anne, Anna Dreber, Emma von Essen, Astri Muren and Eva Ranehill, "Gender and Economic Preferences in a Large Random and Representative Sample," working paper, http://ssrn.com/abstract=2443315.

Boschini, Anne, Astri Muren, Mats Persson, Constructing gender differences in the economics lab, *Journal of Economic Behavior & Organization*, Volume 84, Issue 3, December 2012, Pages 741-752

Bracha, Anat and Chaim Fershtman, (2013) Competitive Incentives: Working Harder or Working Smarter?. *Management Science* 59(4):771-781.

Brinig, M.F. (1995). "Does mediation systematically disadvantage women?" William and Mary Journal of Women and the Law 2, 1–34.

Bruner, David M., "Changing the probability versus changing the reward," *Experimental Economics* December 2009, Volume 12, Issue 4, pp 367-385.

Buser, Thomas, "The impact of the menstrual cycle and hormonal contraceptives on competitiveness," *Journal of Economic Behavior & Organization*, Volume 83, 2012, Pages 1-10.

Buser, Thomas, Muriel Niederle and Hessel Oosterbeek, "Gender, Competitiveness and Career Choices," forthcoming, *Quarterly Journal of Economics*, 2014.

Byrnes, James P., David C. Miller, and William D. Schafer, "Gender Differences in Risk Taking: A Meta-Analysis," *Psychological Bulletin*, LXXV (1999), 367–383.

Cadsby, C. Bram, Maroš Servátka, Fei Song, How competitive are female professionals? A tale of identity conflict, *Journal of Economic Behavior & Organization*, Volume 92, August 2013, Pages 284-303.

Calsamiglia, Caterina, Jörg Franke and Pedro Rey-Biel "The incentive effects of affirmative action in a real-effort tournament," Journal of Public Economics, 2013, vol 98, 15-31.

Camerer, Colin, Individual Decision Making, Handbook of Experimental Economics, Kagel and Roth.

Cárdenas, Juan-Camilo, Anna Dreber, Emma von Essen, Eva Ranehill, "Gender differences in competitiveness and risk taking: Comparing children in Colombia and Sweden," *Journal of Economic Behavior & Organization*, Volume 83, Issue 1, June 2012, Pages 11-23.

Casari, Marco, John C. Ham and John H. Kagel, "Selection Bias, Demographic Effects abd Ability Effects in Common Value Auction Experiments," American Econommic Review, September 2007, vol 97 (4) 1278 – 1304.

Cason Timothy N.; Masters, William A. and Sheremeta, Roman M. 2010. "Entry into winner-take-all and proportional-prize contests: an experimental study," *Journal of Public Economics*, 94, 604–11.

Charness, Gary and Uri Gneezy, Strong Evidence for Gender Differences in Risk Taking, *Journal of Economic Behavior & Organization*, Volume 83, Issue 1, June 2012, Pages 50-58.

Charness, Gary, and Marie-Claire Villeval. 2009. "Cooperation and Competition in Intergenerational Experiments in the Field and the Laboratory." *American Economic Review*, 99(3): 956-78.

Cleave, B.L., Nikiforakis, N, Slonim, Robert, 2010. "Is there Selection Bias in Laboratory Experiments?" Department of Economics – Working Paper Series 1106, The University of Melbourne.

Cauffman, Katherine, "Gender Differences in Willingness to Guess," forthcoming, *Management Science*.

Coffman, Katherine B., "Evidence on Self-Stereotyping and the Contribution of Ideas," forthcoming, *Quarterly Journal of Economics*.

Coffman, Lucas C. and Muriel Niederle, "Pre-Analysis Plans are not the Solution, Replications Might Be," working paper 2014a.

Coffman, Lucas C. and Muriel Niederle. "A Proposal for Promoting Replications: The Case of Experimental Economics," working paper 2014b.

Cohen, Jacob (1988). Statistical power analysis for the behavioral sciences. (2nd ed.). Hillsdale, NJ: Erlbaum

Conlin, Michael Michael Lynn, and Ted O'Donoghue "The norm of restaurant tipping," Journal of Economic Behavior & Organization, Volume 52, Issue 3, November 2003, Pages 297–321

Cooper, David and John H. Kagel, "Are Two Heads Better Than One? Team Versus Individual Play in Signaling Games," American Economic Review, 2007, 97, 1278-1304.

Cooper, David and John H. Kagel, "A failure to communicate: An experimental investigation," working paper, 2013.

Cooper, David and John H. Kagel, "Other-Regarding Preferences: A Selective Survey of Experimental Results" Handbook of Experimental Economics, vol 2

Cotton, Christopher, Frank McIntyre, Joseph Price, "Gender differences in repeated competition: Evidence from school math contests," Journal of Economic Behavior & Organization, Volume 86, February 2013, Pages 52-66.

Craig, Ashley, Ellen Garbarino, Stephanie A. Heger and Robert Slonim, "Waiting to Give," working paper, 2014.

Crosetto, Paolo and Antonio Filippin, 2013b "A theoretical and experimental appraisal of five risk elicitation methods," Jena Economic Research Papers 2013-009, Friedrich-Schiller_University Jena, Max-Planck-Institute of Economics, 2013. Competition. *Management Science*, 58(11):1982-2000.

Croson, Rachel and Uri Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature* 47 (2):448–474.

Cubitt, Robin P., Chris Starmer, and Robert Sugden, "On the validity of the random lottery incentive system" *Experimental Economics*, September 1998, Volume 1, Issue 2, pp 115-131.

Dargnies, Marie-Pierre (2012) Men Too Sometimes Shy Away from Competition: The Case of Team Crosetto, Paolo and Antonio Filippin, 2013a, "The "bomb" risk elicitation task," *Journal of Risk and Uncertainty* August 2013, Volume 47, Issue 1, pp 31-65.

Darwin, Charles, 1874, The Descent of Man, and Selection in Relation to Sex, New York, NY, Hurst and Company.

Dato, Simon and Petra Nieken, Gender differences in competition and sabotage, *Journal of Economic Behavior & Organization*, Volume 100, April 2014, Pages 64-80.

Dave, C., Eckel, C.,Johnson, C., and Rojas, C., 2010, "Eliciting Risk Preferences: When is simple better?", *Journal of Risk and Uncertainty*, 41(3)219-243.

Delfgaauw, Josse, Robert Dur, Joeri Sol, and Willem Verbeke, "Tournament Incentives in the Field: Gender Differences in the Workplace", *Journal of Labor Economics*, Vol. 31, No. 2 (April 2013), pp. 305-326.

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

Dohmen, Thomas, Armin Falk, Klaus Fliessbach, Uwe Sunde, Bernd Weber, Relative versus absolute income, joy of winning, and gender: Brain imaging evidence, *Journal of Public Economics*, Volume 95, Issues 3–4, April 2011, Pages 279-285.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2005. "Individual Risk Attitudes: New Evidence from a Large, Representative, Experimentally-Validated Survey." Institute for the Study of Labor Discussion Paper 1730.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp and Gert G. Wagner, 2011 "Individual risk attitudes: Measurement, determinants, and Behavioral consequences," *Journal of the European Economic Association* Volume 9, Issue 3, pages 522–550.

Dufwenberg, Martin and Astri Muren, Generosity, anonymity, gender, *Journal of Economic Behavior & Organization*, Volume 61, Issue 1, September 2006, Pages 42-49.

Dweck Carol, Self-theories: Their Role in Motivation, Personality, and Development in the series: Essays in Social Psychology, Psychology Press; 1 edition (January 1, 2000).

Dreber, Anna, Emma von Essen, Eva Ranehill, "Gender and competition in adolescence: task matters," *Experimental Economics* March 2014, Volume 17, Issue 1, pp 154-172.

Eagly, Alice H. and Maureen Crowley, "Gender and Helping Behavior: A Meta-Analytic Review of the Social Psychological Literature," *Psychological Bulletin*, 1986, №1. 100, No. 3,283-308.

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

Eckel, Catherine C. and Philip J. Grossman. 1998. Are Women less Selfish than Men?: Evidence from Dictator Experiments. *The Economic Journal*, 108, May, 726-735.

Eckel, Catherine C. and Philip J. Grossman. 2001. "Chivalry versus solidarity in ultimatum games," *Economic Inquiry*, 39, 171-188.

Eckel, Catherine C., and Philip J. Grossman. 2008. "Forecasting Risk Attitudes: An Experimental Study Using Actual and Forecast Gamble Choices." *Journal of Economic Behavior and Organization*, 68(1): 1–17.

Eckel, Catherine C. and Philip J. Grossman, Chapter 57 Differences in the Economic Decisions of Men and Women: Experimental Evidence, In: Charles R. Plott and Vernon L. Smith, Editor(s), Handbook of Experimental Economics Results, Elsevier, 2008b, Volume 1, Pages 509-519, ISSN 1574-0722,

Eckel, Catherine C. and Philip J. Grossman, (2008c) "Men, Women and Risk Aversion: Experimental Evidence", in the Handbook of Experimental Economics Results.

Edlund, Lena and Rohini Pande, "Why have women become left-wing? The political gap and the decline in marriage," Quarterly Journal of Economics, 117, 2002, 917 – 961.

Einav, Liran, Amy Finkelstein, Iuliana Pascu, and Mark Cullen, "How General Are Risk Preferences? Choices under Uncertainty in Different Domains," *American Economic Review*, 102(6), October 2012, 2606-2638.

Engel, Christoph, "Dictator games: a meta study," *Experimental Economics* (2011) 14:583–610.

Erev, Ido and Ernan Haruvy, Learning and the Economics of Small Decisions, This HANDBOOK

Erev, Ido and Alvin E. Roth, "Maximization, Learning and Economic Behavior," *Proceedings of the National Academy of Sciences (PNAS)*, July 22, 2014, vol. 111,suppl. 3, 10818–10825

Eriksson, Tor, Sabrina Teyssier and Marie-Claire Villeval, "Self-Selection and the Efficiency of Tournaments," *Economic Inquiry*, Volume 47, Issue 3, pages 530–548, July 2009

Ertac Seda and Balazs Szentes. 2010. The effect of performance feedback on gender differences in competitiveness: experimental evidence. Working Paper, Koc Univ., Turkey.

Exley, Christine L. "Incentives for Prosocial Behavior: The Role of Reputations," working paper, 2014.

Fehr-Duda, Helga, Manuele de Gennaro, and Renate Schubert. 2006. "Genders, Financial Risk, and Probability Weights." *Theory and Decision*, 60(2–3): 283–313.

Fershtman, Chaim, and Uri Gneezy, (2001) "Discrimination in a Segmented Society: An Experimental Approach," *Quarterly Journal of Economics*, 116(1), 351-377.

Filippin, Antonio and Paolo Crosetto, A Reconsideration of Gender Differences in Risk Attitudes, working paper, 2014.

Finucane, Melissa L., Paul Slovic, C. K. Mertz, James Flynn, and Theresa A. Satterfield. 2000. "Gender, Race, and Perceived Risk: The 'White Male' Effect." *Health, Risk and Society*, 2(2): 159–72.

Fisman, Raymond, Pamela Jakiela and Shachar Kariv, "The Distributional Preferences of Americans," working paper, 2014.

Flory, Jeffrey A., Andreas Leibbrandt, and John A. List. 2010. "Do Competitive Work Places Deter Female Workers? A Large-Scale Natural Field Experiment on Gender Differences in Job-Entry Decisions." Tech. rep., NBER Working Paper No. w16546.

Forsythe, Robert, Joel L. Horowitz, N.E. Savin and Martin Sefton, "Fairness in Simple Bargaining Experiments," *Games and Economic Behavior*, 1994, 6, 347 – 369.

Fudenberg, Drew, David G. Rand and Anna Dreber, "Slow to Anger and Fast to Forgive: Cooperation in an Uncertain World", *American Economic Review*, 2012, 102(2), 720 – 749.

Funk, Patricia and Christina Gathmann, "Gender Gaps in Policy Making: Evidence from Direct Democracy," Economic Policy, forthcoming.

Gilligan, Carol (1982). *In a different voice: Psychological theory and women's development*. Cambridge, MA: Harvard University Press.

Glaeser Edward L, David I. Laibson, Jose A. Scheinkman and Christine L. Soutter. "Measuring Trust . Quarterly Journal of Economics 2000;115(3):811-846.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini, "Performance in Competitive Environments: Gender Differences," *Quarterly Journal of Economics*, CXVIII, August 2003, 1049 – 1074.

Gneezy, Uri and Jan Potters, 1997. "An experiment on risk taking and evaluation periods." *Quarterly Journal of Economics* 112, 631–645.

Gneezy, Uri, and Aldo Rustichini "Gender and competition at a young age," *American Economic Review Papers and Proceedings*, May 2004, 377-381.

Gneezy, Uri; Leonard, Kenneth L. and List, John A. "Gender Differences in Competition: Evidence from a Matrilineal and a Patriarchal Society," *Econometrica*, Vol. 77, No. 5 September 2009, 1637–1664.

Grether, David M. and Charles R. Plott, "Economic Theory of Choice and the Preference Reversal Phenomenon," *The American Economic Review*, Vol. 69, No. 4 (Sep., 1979), pp. 623-638

Grosse Niels D and Gerhard Riener. 2010. Explaining gender differences in competitiveness: gender-task stereotypes. Working Paper, Friedrich Schiller University, Jena, Germany.

Günther, Christina, Neslihan Arslan Ekinci, Christiane Schwieren, Martin Strobel, "Women can't jump?—An experiment on competitive attitudes and stereotype threat," *Journal of Economic Behavior & Organization*, Volume 75, Issue 3, September 2010, Pages 395-401.

Güth, Werner, Rolf Schmittberger and Bernd Schwarze, "An experimental analysis of ultimatum bargaining, Journal of Economic Behavior & Organization Volume 3, Issue 4, December 1982, Pages 367–388.

Gupta, Nabanita Datta, Anders Poulsen and Marie-Claire Villeval, "Gender Matching and Competitiveness: Experimental Evidence", *Economic Inquiry*, Vol 51(1) 816-835, 2013.

Gysler, M., Kruse, J.B., Schubert, R. (2002). "Ambiguity and gender differences in financial decision making: An experimental examination of competence and confidence effects". Working paper, Center for Economic Research, Swiss Federal Institute of Technology.

Harbaugh,W.T., Krause, K., Vesterlund, L. (2002). "Risk attitudes of children and adults: Choices over small and large probability gains and losses". *Experimental Economics* 5 (1), 53–84.

Harbaugh, William, Kate Krause and Lise Vesterlund, "The Fourfold Pattern of Risk Attitudes in Choice and Pricing Tasks," *The Economic Journal*, June 2010, vol. 120(545), 569-611.
Harrison, Glenn W. 1990. "Risk Attitudes in First-Price Auction Experiments: A Bayesian Analysis," *Review of Economics and Statistics* 72, 541-546.
Harrison, Glenn W., Eric Johnson, Melayne M. Mcinnes, E. Elisabet Rutström. 2005. Risk Aversion and Incentive Effects: Comment. American Economic Review 95:3, 897-901.
Harrison, Glenn W., John A. List and Charles Towe, "Naturally Occurring Preferences and Exogenous Laboratory Experiments: A Case Study of Risk Aversion", *Econometrica*, Vol 75(2), March 2007, 433-458.
Hartog, Joop, Ada Ferrer-i-Carbonell, and Nicole Jonker. 2002. "Linking Measured Risk Aversion to Individual Characteristics." *Kyklos*, 55(1): 3–26.
Healy, Andrew and Jennifer Pate (2011), Can Teams Help to Close the Gender Competition Gap?. *The Economic Journal*, 121: 1192–1204.
Herrmann, Benedikt Christian Thöni and Simon Gächter. 2008. "Antisocial punishment across societies," Science, 319 (5868), 1362-1367
Hoffman Moshe, Gneezy Uri. 2010. Left-handed women are more competitive than right-handed men: on the biological basis of gender differences in competitiveness. Work. Pap., Univ. Calif., San Diego.
Holt, Charles A., 1986. Preference reversals and the independence axiom. *American Economic Review* 76, 508–515.
Holt, Charles A. and Laury, Susan K. "Risk Aversion and Incentive Effects." *American Economic Review*, 2002, 92(5), pp. 1644-55.
Huberman, Gur and Ariel Rubinstein, "Correct Belief, Wrong Action and a Puzzling Gender Difference", SSRN working paper 2001.Hyde, Janet S., Elizabeth Fennema, and Susan J. Lamon, "Gender Differences in Mathematics Performance: A Meta-Analysis," *Psychological Bulletin*, CVII (1990), 139 –155.
Houser, Daniel and Daniel Schunk, "Social environments with competitive pressure: Gender effects in the decisions of German schoolchildren," *Journal of Economic Psychology*, Volume 30, Issue 4, August 2009, Pages 634-641.
Hyde, Janet S., "The Gender Similarities Hypothesis," American Psychologist, 2005, Vol 60, 581-592.
Isaac R. Mark, and Duncan James, "Just Who Are You Calling Risk Averse?", *Journal of Risk and Uncertainty*, 20:2; 177-187, 2000.
Jones, Daniel and Sera Linardi, "Wallflowers: Experimental Evidence of an Aversion to Standing Out," *Management Science*, 2014, 60(7):1757-1771.
Kahneman, Daniel Thinking, Fast and Slow (Straus and Giroux, 2011)
Kamas Linda and Anne Preston 2009. Social preferences, competitiveness and compensation: Are there gender differences? Working Paper, Santa Clara University.
Kamas, Linda and Anne Preston, The importance of being confident; gender, career choice, and willingness to compete, Journal of Economic Behavior & Organization, Volume 83, Issue 1, June 2012, Pages 82-97.
Kamas, Linda and Anne Preston, "Gender and Social Preferences in the US: An Experimental Study", Feminist Economics, Volume 18, Issue 1, 2012, 135-160.
Karni, Edi, Safra, Zvi, 1987. "Preference Reversal" and the Observability of Preferences by Experimental Methods," *Econometrica* 55 (3), 675–685.

Kass, Norman, "Risk in Decision Making as a Function of Age, Sex, and Probability Preference," *Child Development*, Vol. 35, No. 2 (Jun., 1964), pp. 577-582.

Klein Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams Jr., Štepán Bahník, Michael J. Bernstein, Konrad Bocian, Mark J. Brandt, Beach Brooks, Claudia Chloe Brumbaugh, Zeynep Cemalcilar, Jesse Chandler, Winnee Cheong, William E. Davis, Thierry Devos, Matthew Eisner, Natalia Frankowska, David Furrow, Elisa Maria Galliani, Fred Hasselman, Joshua A. Hicks, James F. Hovermale, S. Jane Hunt, Jeffrey R. Huntsinger, Hans IJzerman, Melissa-Sue John, Jennifer A. Joy-Gaba, Heather Barry Kappes, Lacy E. Krueger, Jaime Kurtz, Carmel A. Levitan, Robyn K. Mallett, Wendy L. Morris, Anthony J. Nelson, Jason A. Nier, Grant Packard, Ronaldo Pilati, Abraham M. Rutchick, Kathleen Schmidt, Jeanine L. Skorinko, Robert Smith, Troy G. Steiner, Justin Storbeck, Lyn M. Van Swol, Donna Thompson, A. E. van 't Veer, Leigh Ann Vaughn, Marek Vranka, Aaron L. Wichman, Julie A. Woodzicka, and Brian A. Nosek, "Investigating Variation in Replicability: A ''Many Labs'' Replication Project," *Social Psychology* 2014; Vol. 45(3):142–152.

Kimball, Miles S., Claudia R. Sahm, and Mat thew D. Shapiro. 2008. "Imputing Risk Tolerance from Survey Responses." *Journal of the American Statistical Association*, 103(483): 1028–38.

Kogan, Nathan and Michael A. Wallach, *Risk Taking: A study in cognition and personality*. New York. Holt, Rinehart, and Winston, 1964.

Kogan, Nathan and Michael A. Wallach (1960) "Certainty of Judgment and the evaluation of risk." Psychological Reports: Volume 6, Issue 2, pp. 207-213.

Kruse, J.B., Thompson, M.A. (2003). "Valuing low probability risk: Survey and experimental evidence". *Journal of Economic Behavior and Organization* 50, 495–505.

Kuhnen Camelia M. and Agnieszka Tymula, "Feedback, Self-Esteem, and Performance in Organizations," *Management Science* 2012 58:1, 94-113.

Lavy, Victor, "Gender Differences in Market Competitiveness in a Real Workplace: Evidence from Performance-Based Pay Tournaments Among Teachers," *The Economic Journal*, Vol. 123, Issue 569, pp. 540-573, 2013.

Ledyard, John, "Public Goods: A Survey of Experimental Research," in *Handbook of Experimental Economics*, edited by J. Kagel and A. Roth, Princeton University Press, 1995.

Lee, Soohyung, Muriel Niederle and Namwook Kang, "Do Single-Sex Schools make Girls more Competitive?", *Economics Letters* Volume 124, Issue 3, September 2014, Pages 474–477.

Leibbrandt, Andreas, Uri Gneezy, and John A. List, "Rise and fall of competitiveness in individualistic and collectivistic societies," *PNAS* 2013 110 (23) 9305-9308.

Leider, Stephen, Markus M. Möbius, Tanya Rosenblat and Quoc-Anh Do, "Directed Altruism and Enforced Reciprocity in Social Networks, *Quarterly Journal of Economics* (2009) 124 (4): 1815-1851.

Lejuez, C. W., Jennifer P. Read, Christopher W. Kahler, Jerry B. Richards, Susan E. Ramsey, Gregory L. Stuart, David R. Strong, Richard A. Brown, "Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART)," *Journal of Experimental Psychology*: Applied, Vol 8(2), Jun 2002, 75-84.

Levin, I. P., M. A. Snyder, and D. P. Chapman. 1988. "The Interaction of Experimental and Situational Factors and Gender in a Simulated Ricky Decision-Making Task." *Journal of Psychology*, 122(2): 173–81.

Levin, Irwin P., Stephanie S. Hart, Joshua A. Weller and Lyndsay A. Harshman, "Stability of choices in a risky decision-making task: a 3-year longitudinal study with children and adults," *Journal of Behavioral Decision Making* Volume 20, Issue 3, pages 241–252, July 2007

Levy, H., Elron, E., Cohen, A. (1999). "Gender differences in risk taking and investment behavior: An experimental analysis". Unpublished manuscript, The Hebrew University.

Lichtenstein, Sarah, and Paul Slovic, "Reversal of Preferences Between Bids and Choices in Gambling Decisions," *Journal of Experimental Psychology*, July 1971, 89, 46-55.

List, John A., "On the Interpretation of Giving in Dictator Games," *Journal of Political Economy*, 2007, vol. 115, no. 3, 482-493.

Lonnqvist, Jan-Erik, Markku Verkasalo, Gari Walkowitz and Philipp C. Wichardt, "Measuring Individual Risk Attitudes in the Lab: Task or Ask? An Empirical Question," working paper, 2011.

Maccoby, E.E. and C.N. Jacklin, 1974, The psychology of sex differences, Stanford, CA, Stanford University Press.

Mayr, Ulrich; Dave Wozniak, Casey Davidson, David Kuhns, William T. Harbaugh, "Competitiveness across the life span: The feisty fifties," *Psychology and Aging*, Vol 27(2), Jun 2012, 278-285.

Milgrom, Paul and John Roberts (1982), "Limit Pricing and Entry under Incomplete Information: An Equilibrium Analysis," *Econometrica*, 50(2) 443-459.

Mobius, Markus M. and Tanya S. Rosenblat. 2006. Why beauty matters. *American Economic Review* 96, no. 1: 222-235.

Moore, E., Eckel, C.C. (2003). "Measuring ambiguity aversion". Unpublished manuscript, Department of Economics, Virginia Tech.

Morin, Louis-Philippe, "Do Men and Women Respond Differently to Competition? Evidence from a Major Education Reform," forthcoming, Journal of Labor Economics.

Müller, Julia and Christiane Schwieren, "Can personality explain what is underlying women's unwillingness to compete?", *Journal of Economic Psychology*, Volume 33, Issue 3, June 2012, Pages 448-460.

Nelson, Julie A., "Not-So-Strong Evidence for Gender Differences in Risk Taking," working paper 2013-06, Department of Economics, University of Massachusetts Boston.

Nelson, Julie A., "Are Women Really More Risk Averse than Men? A Re-analysis of the Literature using expanded methods," *Journal of Economic Surveys*, forthcoming, 2014.

Niederle, Muriel, Carmit Segal, and Lise Vesterlund, "How Costly is Diversity? Affirmative Action in Light of Gender Differences in Competitiveness" *Management Science*, 2013, Vol 59, No. 1, 1-16.

Niederle, Muriel, and Lise Vesterlund, "Do Women Shy Away from Competition? Do Men Compete too Much?," *Quarterly Journal of Economics*, August 2007, Vol. 122, No. 3, 1067-1101.

Niederle, Muriel and Lise Vesterlund, "Gender and Competition", *Annual Review in Economics*, 2011, 3, 601–30.

Niederle, Muriel and Alexandra H. Yestrumskas, "Gender Differences in Seeking Challenges: The Role of Institutions", NBER working paper, 2008

Ors, Evren, Frédéric Palomino, and Eloïc Peyrache, "Performance Gender Gap: Does Competition Matter?," *Journal of Labor Economics*, Vol. 31, No. 3 (July 2013), pp. 443-499.

Powell, Melanie, and David Ansic. 1997. "Gender Differences in Risk Behaviour in Financial Decision-Making: An Experimental Analysis." *Journal of Economic Psychology*, 18(6): 605–28.

Price, Curtis R. (2012) Gender, Competition, and Managerial Decisions. *Management Science* 58(1):114-122.

Rabin, Matthew, "Risk Aversion and Expected-Utility Theory: A Calibration Theorem," *Econometrica*, LXVIII (2000), 1281–1292.

Rand David G., Joshua D. Greene and Martin A. Nowak, "Spontaneous giving and calculated greed," Nature, 20 September 2012, vol 489, 427-430.

Rapoport, Anatol and Albert M. Chammah, "Sex Differences in Factors Contributing to the Level of Cooperation in the Prisoner's Dilemma Game", *Journal of Personality and Social Psychology*, 1965, Vol 2, No. 6, 831-838.

Recalde, María P., Arno Riedl and Lise Vesterlund, "Error prone inference from response time: The case of intuitive generosity," working paper, 2014.

Reuben, Ernesto, Matthew Wiswall, and Basit Zafar. 2013. "Preferences and biases in educational choices and labor market expectations: shrinking the black box of gender." Staff Report, Federal Reserve Bank of New York.

Reynaud, Arnaud and Stéphane Couture, "Stability of risk preference measures: results from a field experiment on French farmers", *Theory and Decision*, 2012, 73(2), 203-221.

Roth, Alvin E., "The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics," Fisher-Schultz Lecture, Econometrica, 70,4, July 2002, 1341-1378.

Rustagi, Devesh, Stefanie Engel and Michael Kosfeld, "Conditional Cooperation and Costly Monitoring Explain Success in Forest Common Management," Science, vol 330, 12 November 2010, 961-965.

Sandberg, Sheryl, "Lean In: Women, Work, and the Will to Lead" 2013, Knopf.

Schotter, Andrew and Keith Weigelt, "Asymmetric Tournaments, Equal Opportunity Laws, and Affirmative Action: Some Experimental Results," The Quarterly Journal of Economics (1992) 107 (2): 511-539.

Schubert, Renate, Martin Brown, Matthias Gysler and Hans Wolfgang Brachinger (1999). "Financial decision-making: Are women really more risk averse?" *American Economic Review Papers and Proceedings* 89, 381–385.

Schubert, R., Gysler, M., Brown, M., Brachinger, H.W. (2000). "Gender specific attitudes towards risk and ambiguity: An experimental investigation".Working paper, Center for Economic Research, Swiss Federal Institute of Technology.

Shields S. A. 1975, "Functionalism, Darwinism and the psychology of women: A study in social myth. American Psychologist, 30, 739-754.

Shurchkov, Olga "Under pressure: gender differences in output quality and quantity under competition and time constraints," *Journal of the European Economic Association*, Volume 10, Issue 5, pages 1189–1213, October 2012.

Simons, Daniel J. and Daniel T. Levin, "Failure to detect changes to people during a real-world interaction," Psychonomic Bulletin & Review, December 1998, Volume 5, Issue 4, pp 644-649.

Slovic, Paul, "Convergent validation of risk taking measures," *The Journal of Abnormal and Social Psychology*, Vol 65(1), Jul 1962, 68-71.
Slovic, Paul. (1964). "Assessment of risk taking behavior," *Psychological Bulletin*, 61, 330–333.
Slovic, Paul, "Risk-Taking in Children: age and Sex Differences," *Child Development*, Vol 37, No 1, March 1966, 169-176.
Slovic, Paul, "Information Processing, Situation Specificity, and the Generality of Risk-Taking Behavior", *Journal of Personality and Social Psychology*, 1972, Vol 22, No.1, 128-134.
Solnick, Sara J., "Gender differences in the ultimatum game," *Economic Inquiry*, Volume 39, Issue 2, pages 189–200, April 2001.
Soll, Jack B., Katherine L. Milkman, and John W. Payne. "A user's guide to debiasing," in preparation for K. Gideon and G. Wu (eds.) Wiley-Blackwell Handbook of Judgment and Decision Making.
Steele, Claude M., "A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance," *American Psychologist*, LII (1997), 613–629.
Steele, Claude M., and Joshua Aronson, "Stereotype Vulnerability and the Intellectual Test Performance of African Americans," *Journal of Personality and Social Psychology*, LXIX (1995), 797–811.
Sutter, Matthias and Daniela Glätzle-Rützler, "Gender differences in the willingness to compete emerge early in life and persist," forthcoming, *Management Science*.
Swineford, Frances: "The Measurement of a Personality Trait." *Journal of Educational Psychology*, Vol. XXIX, April, 1938, pp. 295-300.
Swineford, Frances, "Analysis of a personality trait", *Journal of Educational Psychology*, Vol 32(6), Sep 1941, 438-444.
Tannenbaum, Daniel I., "Do gender differences in risk aversion explain the gender gap in SAT scores? Uncovering risk attitudes and the test score gap," working paper 2012.
Teyssier Sabrina. 2008. Experimental evidence on inequity aversion and self-selection between incentive contracts. GATE Work. Pap. 08-21, Ecully, France.
Thaler, Richard H., and Cass R. Sunstein, (2008). Nudge: Improving decisions about health, wealth, and happiness. New Haven, CT: Yale University Press.
Thomas-Hunt, Melissa C. and Katherine W. Phillips, "When What You Know Is Not Enough: Expertise and Gender Dynamics in Task Groups," *Personality and Social Psychology Bulletin*, 2004, 30, 1585-1598.
Vandegrift Donald and Abdullah Yavas. 2009. "Men, women, and competition: an experimental test of behavior." *Journal of Economic Behavior & Organization*, 72:554–70.
Vesterlund, Lise, Linda Babcock and Laurie Weingart, "Breaking the Glass Ceiling with "No": Gender Differences in Declining Requests for Non Promotable Tasks," working paper, 2014.
Visser, Michael S. and Matthew R. Roelofs, "Heterogeneous preferences for altruism: gender and personality, social status, giving and taking," *Experimental Economics*, (2011) 14:490–506.
Wallach, Michael A. and Nathan Kogan, "Aspects of judgment and decision making: Interrelationships and changes with age" *Behavioral Science*, Volume 6, Issue 1, pages 23–36, January 1961.
Weber, Elke U., Ann-Renée Blais, and Nancy E Betz. (2002). A Domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, 15, 263–290.

Werner, Bönte, "Gender differences in competitive preferences: new cross-country empirical evidence," *Applied Economics Letters*, forthcoming.
Wik, M., Kebede, T.A., Bergland, O. and S. Holden, 2004, "On the measurement of risk aversion from experimental data," *Applied Economics*, 36 (21) 2443-2451.
Woolley H.T., 1914, "The Psychology of Sex," Psychological Bulletin, 11, 353-379.
Wozniak, David, William T. Harbaugh, and Ulrich Mayr, "The Menstrual Cycle and Performance Feedback Alter Gender Differences in Competitive Choices," *Journal of Labor Economics*, Vol. 32, No. 1 (January 2014), pp. 161-198
Zhang, Y. Jane. 2012a. "Can Experimental Economics Explain Competitive Behavior Outside the Lab?" Unpublished manuscript.
Zhang, Y. Jane. 2012b. "The Communist Experiment in China: Narrowing the Gender Gap in Competitive Inclination." Unpublished manuscript.
Ziller, Robert C. "Vocational Choice and Utility for Risk", *Journal of Counseling Psychology*, Vol 4, No 1, 1957, 61-64.