Heterogeneity in response to incentives: Evidence from field data

Czibor, E.

Publication date
2015
Document Version
Final published version

Heterogeneity in Response to Incentives: Evidence from Field Data

ISBN 978 90 5170 699 4

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. **632** of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

Heterogeneity in Response to Incentives: Evidence from Field Data

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het College voor Promoties
ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op vrijdag 20 november 2015, te 10:00 uur

door

Eszter Czibor

geboren te Zalaegerszeg, Hongarije

Promotiecommissie:

Promotores:	Prof. dr. C.M. van Praag	Copenhagen Business School
	Prof. dr. R. Sloof	Universiteit van Amsterdam
Overige leden:	Dr. T. Buser	Universiteit van Amsterdam
	Prof. dr. A.J. Dur	Erasmus Universiteit Rotterdam
	Prof. dr. U.H. Gneezy	Universiteit van Amsterdam
	Prof. dr. J.H. Sonnemans	Universiteit van Amsterdam
	Prof. dr. L. Vesterlund	University of Pittsburgh

Faculteit Economie en Bedrijfskunde

Apának, Anyának, Sancinak és Gergőnek

Acknowledgements

With my defense approaching and my time as a doctoral student in Amsterdam coming to an end, I would like to express my gratitude to the people who made these five years unforgettable. First and foremost, I would like to thank my supervisors, Mirjam and Randolph. You encouraged me to pursue my own ideas but you were also always available for discussion, and you constantly surprised me with your great insights and ideas. You were also a perfect pair of supervisors, Randolph's sharp logic, scientific rigor and attention to the minutest detail complementing Mirjam's focus on the bigger picture and her enthusiasm for exciting research questions. Randolph, I will miss walking down the corridor to your office any time I needed guidance - or just felt like chatting. Mir, what a source of inspiration you have been, both on the professional and the personal level!

I would also like to thank Uri Gneezy, my scientific role model and mentor. I am grateful for the time he spent reading my work, for his famously honest, brief but spot-on comments, for his support and motivation, for the fantastic people he introduced me to, and for the fun times we shared in Amsterdam, San Diego and Modica. I am thankful to John List for his inspiring PhD course in Bergen that turned my attention towards field experiments, and for endorsing my grant application and making it possible for me to continue my work at the University of Chicago. I am grateful to Jörg Claussen for making me feel welcome in Copenhagen and for teaching me so much, from Stata programming techniques to the rules of the Sauspiel card game.

I owe a lot to my colleagues at the University of Amsterdam, especially to Erik, Jeroen, Joep, Joeri, Philipp, Theo, Thomas, and most of all Silvia. I really enjoyed our lunches and coffee

breaks, and I am indebted for the comments and suggestions you gave for my work. Laura and Martin, you were the best office roommates and co-authors I could have asked for. Sander, it was a great experience to work with you, both on our joint research project but also as your teaching assistant.

I feel privileged to be part of the research community of the Tinbergen Institute. TI introduced me to a crowd of smart, inspiring young people. Anghel, Anita, Dave, Dávid, Gosia, Janna, Lydia, Mark, Max, Sabina, Sait, Sanne, Steffen, Swapnil, Tomasz, Violeta, it was so much fun having you around! My wonderful paranymphs, Inez and Karolina, deserve special thanks for being the best of friends. Lukas, our conversations meant so much to me. My friends in Amsterdam outside academia: Carl, Francesca, Linas, Loredana, Mariana, Michael, Monika, Nathalie, Peter, Sef, Tomek, and my CRS sisters all over the world, especially Dana, Els, Karen, Olga, Sofia, Stefa and Uyum, you made my life complete these past years.

I am grateful for my family for raising me with the belief that anything was possible. Our home has been the "best platform from which to jump beyond myself". My little brother actually beat me to the doctor title - Sanci, I am immensely proud of you. I am thankful to my friends from Hungary who stayed with me despite the distance: Böbi, Eszti, Fazi, Gábor, Győző, Kata, Matyi, Petra, Sára, I hope never to lose touch.

Save the best for last: Gergő, none of this would have been possible without you. Thank you for your endless support, for the interest you have shown in my research, for your patience and understanding, and most of all, for all the amazing moments we shared. I am looking forward to many more to come!

Eszter Czibor

Amsterdam, October 2015

Contents

Chapter 1

Introduction

"Underneath all, individuals,

I swear nothing is good to me now that ignores individuals..."

Walt Whitman

This dissertation explores whether *observable* individual characteristics such as gender, level of education and occupation are good predictors of people's response to competitive and cooperative incentive schemes. This inquiry is motivated by the belief that such characteristics can serve as proxies for the *unobservable* personality traits and attitudes that actually influence the reaction to incentives. The idea that there are systematic differences in behavior between individuals has been extensively researched. Several dimensions of heterogeneity have been identified, for instance gender (Croson and Gneezy, 2009), age (Harbaugh et al., 2002), field of study (Frank et al., 1993), racial identity (Burns, 2012), religion (Noussair et al., 2013), etc. These results, however, have primarily been obtained from laboratory experiments, and their relevance for field outcomes has been scantly researched (Bertrand, 2011). This dissertation adds to the existing literature by using field data to assess the relationship between certain observable personal characteristics and the response to incentives in natural or realistic settings.

To be more specific, the three core chapters of this dissertation aim to answer the following research questions: (1) Do male and female students respond differently to competitive grade

incentives? (2) Does the gender gap in risky and competitive choices hinder the success of women? (3) How do individual characteristics influence the decision to sort into team incentive schemes?

The first research question is addressed in Chapter 2. This chapter is based on the paper "Does relative grading help male students? Evidence from a field experiment in the classroom", co-authored with Sander Onderstal, Randolph Sloof and Mirjam van Praag (IZA Discussion Paper, No. 8429). It is inspired by the growing literature on gender differences in response to competition (Croson and Gneezy, 2009; Niederle and Vesterlund, 2011). Following the seminal paper of Gneezy et al. (2003), several studies have replicated the finding that men react more positively to tournament-style incentives than women. However, much less attention has been devoted to the implications of this result for incentive provision in education. While there is some evidence that grade incentives can increase student learning (Grove and Wasserman, 2006), it is unclear what type of grading practices induce the highest performance, and whether the response is heterogeneous by student characteristics such as gender. The two most commonly used grading schemes, absolute (i.e. criterion-referenced) and relative (i.e. norm-referenced) grading differ in the level of competition they generate. Relative grading by design creates a rank-order tournament in the classroom, while absolute grading is analogous to a piece-rate scheme (Becker and Rosen, 1992; Landeras, 2009). There is no consensus about the grading scheme that should be preferred: in continental Europe, absolute grading prevails, while in the US grading on a curve is more popular, especially in higher education.

Chapter 2 of this dissertation reports results from a framed field experiment that provides a direct empirical comparison of the impact of absolute and relative grading on male and female students' preparation effort and exam performance. Students following a Bachelor course at a Dutch university are randomly assigned to one of the two treatment conditions that differ only in the grading scheme used to translate exam scores into exam grades. The data collected in the experiment include exam scores, different proxies for preparation effort as well as a rich set of individual characteristics such as demographic information, preferences, confidence and ability

measures. The results show no clear difference either in effort provision or exam performance under the two grading schemes. There is no evidence for heterogeneity in response by gender, ability or competitive preferences, either. These findings are likely explained by an overall lack of ambition and a general disinterest in obtaining high grades, a manifestation of the often criticized "just-pass" attitude of Dutch students. This explanation is supported by the analysis of those students who are conjectured to care most about the way their exam scores are mapped into grades, i.e. students close to the pass-fail threshold. While this subgroup is relatively small, so inferences should be made with caution, there is a strong indication for a gender gap in response to relative grading among such "marginal" students. The results reported in Chapter 2 thus relate to the discussion on tackling the problem of unmotivated students, a challenge that educators and policymakers are confronted with worldwide (OECD, 2015).

Chapter 3 of this dissertation is based on the paper "Women do not play their aces - The consequences of shying away", co-authored with Jörg Claussen and Mirjam van Praag. It studies the second research question, namely the consequences of gender differences in entering tournaments. It is provoked by the massive and persistent underrepresentation of women at the top of organizational hierarchies worldwide (Catalyst Org., 2015). Anecdotal and survey evidence from competitive firms suggests that women's more cautious approach to promotion contests contributes to their slower career advancement (Barsh and Yee, 2012; Institute for Leadership and Management, 2011; Sandberg, 2013). There is indeed robust scientific evidence showing that women are more averse towards risk and uncertainty than men (Charness and Gneezy, 2012; Eckel and Grossman, 2008) and that women tend not to enter tournaments they could win (Niederle and Vesterlund, 2007). However, these findings originate predominantly from laboratory experiments. Notable exceptions are the studies by Buser et al. (2014) and Zhang (2013) who show that laboratory measures of competitiveness can predict real life outcomes. Flory et al. (2014) demonstrate in a field experiment that competitive remuneration schemes deter female applicants.

Chapter 3 of this dissertation complements the above-listed studies by analyzing in a natural setting whether female players continue to take less risk and initiate less contests than men even after self-selecting into an uncertain and competitive environment. It is based on a large set of naturally occurring data from an online card-game community that contains information from over 4 million games, generated by more than fifteen thousand individual players. The data include players' choices regarding the level of risk and competition they prefer to bear in each round, and also their subsequent playing performance in the resulting tournaments. The data show that even though players sort into the community, this self-selection does not close the gender gap in choices related to risk taking and competition. Female players are still less likely than males to increase the stakes of the given round or to actively initiate games. We argue that female players' differential choices are to a large extent attributable to gender differences in risk and competitive preferences. As a result of their selection choices, women end up more often in the difficult opponent position, and even when they initiate and win games, their earnings are lower due to the smaller stakes. Consequently, women accumulate lower scores in the game than men do. This gender gap in scores is not a reflection of differences in card playing ability: controlling for the type of the game and their role, women are as good as men in winning games. In sum, this study demonstrates the negative consequence of "shying away": despite no gender differences in on-task performance, women end up lagging behind men as a result of their lower propensity to take risk and to initiate tournaments.

Chapter 4 of this dissertation is based on the paper "Risks, gains and autonomy: An experimental analysis of sorting into teams", co-authored with Martin Koudstaal and Laura Rosendahl Huber. It explores the third research question, namely the role of individual characteristics in the decision to participate in team incentive schemes. This chapter is motivated by the pervasiveness of team-based remuneration offered by firms (Lazear and Shaw, 2007). While the effectiveness of team incentive schemes has been extensively researched (e.g. Bandiera et al. (2013); Hamilton et al. (2003); van Dijk et al. (2001)), less attention has been devoted to the sorting decision into such schemes. The success of team incentives, however, crucially depends

on the characteristics of individuals who decide to select into companies offering team pay. While a handful of papers have looked into the participation decision of individuals, certain important aspects have not yet been addressed. First, a key feature of real-life team production is the possibility of synergies arising from complementarities between the teammates. Most empirical papers that analyze the selection into teams have thus far ignored this aspect and modeled teamwork as a simple revenue sharing contract, e.g. an equal split of the pooled total output of the members (Bäker and Mertins, 2013; Dohmen and Falk, 2011; Herbst et al., 2015), or added an arbitrary, pre-defined mark-up on top of the joint output (Cooper and Saral, 2013; Kuhn and Villeval, 2014). Furthermore, existing literature on sorting into teams has analyzed the issues of team production and team decision making separately. The study we discuss in Chapter 4 combines these two aspects of teamwork in a single framework and thus allows us to assess the impact of shared decision rights on the willingness to join teams. It therefore informs the discussion on the value of authority and control (Bartling et al., 2014; Owens et al., 2014).

Chapter 4 is based on a large-scale, incentivized online survey conducted among Dutch entrepreneurs, mangers and employees. The survey elicits respondents' preferences for joining a team in two treatment conditions: in the Baseline treatment, the team option only entails joint production, while in the Joint Decision treatment it includes *both* joint production and a joint investment choice. In both treatments, team production is modeled to allow for synergies arising from complementarities between teammates. Results from the survey replicate several findings from related lab studies on sorting into team incentive schemes, such as the relevance of confidence and risk preferences. A novel finding presented in this chapter is the importance of education for individuals' participation decision in teams. Controlling for differences in task performance, confidence and risk preferences, higher education is associated with a greater willingness to pay for the team option. There is suggestive evidence that this heterogeneity is explained by differences in evaluating the team option: while participants with higher levels of education tend to primarily consider the potential gains from team pay, lower-educated respondents focus more on the risks associated with teams. As a consequence, lower educated

people miss out on the sizable efficiency gains that the team option entails. Participants are also found to be heterogeneous in their response to a potential compromise in decision making. In particular, entrepreneurs are shown to be averse to joint decision making when they predict that it moves them away from their individual optimal choice, while no such effect is observed for managers or employees, surprisingly.

Importantly, this dissertation does not try to suggest that the behavioral differences we demonstrate are necessarily *caused* by the observable individual characteristics we have analyzed. In case of self-employment, it is likely that the same underlying traits that make people self-select into entrepreneurship also explain their different choices in our survey. The finding that education affects team choices is consistent with the reasoning that schooling changes the way individuals evaluate situations involving strategic risk, but our setting does not rule out the alternative explanation that those who self-select into higher education are inherently different from others in this aspect. While differential sorting is not of concern in case of gender, it is still debated whether women's lower tolerance of risk and competition is due to nature or nurture (Buser, 2012; Gneezy et al., 2009). To sum up, we find it useful to study differences in response to incentives by observable characteristics not because we believe these factors drive people's choices but because they are correlated with the unobservable traits that do so. Consequently, by taking observable dimensions of heterogeneity between individuals into account we can improve our predictions of people's responses to incentives and thus design more efficient incentive schemes.

Chapter 2

Gender and competition in education

2.1 Introduction

Educators and policy makers worldwide are struggling to address the challenge posed by unmotivated students (OECD, 2015). Low motivation can cause insufficient study effort, increased drop-out rates and longer study durations, imposing a heavy toll on society in the form of extra expenditures on education and forgone productivity (Garibaldi et al., 2012; Leuven et al., 2010). Several recent policy reports (e.g. OECD (2013)) show that this problem is particularly severe among young men: boys are now more likely to underperform in secondary education than girls, and they are also less likely to complete higher education degrees than their female peers (Salvi del Pero and Bytchkova, 2013). *"New gender gaps in education are opening,"* warns the OECD (2015, p.13): at the lower end of the achievement spectrum it is boys who suffer more from poor motivation and a lack of ambition, and who consequently lag behind girls.

The study presented in this chapter aims to test whether the problem of insufficient student motivation could be tackled with the help of competitive grade incentives. In a large-scale field

This chapter is based on Czibor et al. (2014). Financial support of the Research Priority Area Behavioral Economics of the University of Amsterdam is gratefully acknowledged. We thank the Examination Board of the Faculty of Economics and Business at the University of Amsterdam for authorizing the field experiment.

experiment conducted among university students we compare effort provision and exam performance under the two most commonly used grading schemes: absolute and relative grading. Under absolute grading, grades depend solely on students' own individual test outcomes, independent of the performance of their classmates. Under relative grading, students' grades depend on their positions in the score distribution of the class. The scheme is also known as "grading on a curve," referring to the bell-shaped curve of the normal distribution.[1] A key difference between the two grading schemes is that relative grading induces direct competition between peers. We hypothesize that grading on a curve, by introducing a rank-order tournament in the classroom, provides more motivation for students to exert effort than absolute grading which is analogous to a piece rate incentive scheme. Based on the empirical stylized fact of gender differences in response to tournaments we expect the effect to be heterogeneous by gender. In sum, our goal is to test empirically whether introducing competitive grade incentives in a setting where absolute grading is the default can help motivate low-achieving boys without harming girls. To this end, we conduct a field experiment among Bachelor students of the University of Amsterdam, a sample that has been found to provide insufficient study effort under absolute grade incentives.[2]

To our knowledge, the study discussed in this chapter is the first to provide an experimental comparison of absolute and relative grading in a naturalistic, high-stakes environment.[3] In our field experiment we randomly assign students to grading schemes while keeping everything else (exam time, location and content) the same for both treatment groups. Our study includes a

[1]The practice of absolute grading is also known as "criterion-referenced grading" because the student's score is compared to an objective criterion. Relative grading is often referred to as "norm-referenced" grading. In the United States, colleges typically implement relative grading (as an example, consider the 2005 overview of law school grading curves by Andy Mroch for the Association of American Law Schools: http://www.aals.org/deansmemos/Attachment05-14.pdf), while in continental Europe, the absolute scheme prevails (Karran, 2004).

[2]Scholars analyzing the behavior of Economics and Business Bachelor students at the University of Amsterdam explain students' underperformance with a lack of effort provision: *"[...] the consensus is that the low pass-rate in the first year (and the long actual study durations) should be attributed to insufficient student effort and not to the program being too demanding"* (Leuven et al., 2010, p.1247).

[3]A few recent papers, discussed in more detail in the Literature section, present results from field experiments with competitive grading in the form of comparison to a randomly chosen opponent or a reward for the top performers only. We believe using an actual grading curve makes our setting more realistic and reproduces better the incentives of rank-order tournaments observed in classrooms in practice.

large sample of university students for whom we collect a rich set of control variables (including preferences as well as course-specific and general ability). We also observe different measures for the preparation behavior of students, so we can test whether grade incentives affect how much they study for the course.

Our results show no clear difference in effort provision or exam performance under the two grading schemes. We do not find a significant response to competitive grade incentives among male students, either. We argue that this result is not driven by students' lack of understanding of the treatment or by the particular design we used, nor is it likely that students were already on their effort frontier under absolute grading. Instead, we believe that an overall lack of ambition drives our findings: students in our sample are mainly interested in passing the course with minimal effort provision and do not attach much importance to obtaining high grades. We claim that students in our sample place little importance on grades *per se* (beyond passing) and this makes them unresponsive to changes in the *type* of grading scheme they face. We find support for this explanation when analyzing the subsample of students who are close to the pass-fail threshold and are thus conjectured to care most about grade incentives. In this group of 'marginal' students we indeed observe the expected gender difference in response to relative grading, with boys performing relatively better than girls when graded on the curve.

The behavior of students in our sample is in line with the "just pass" attitude (the so-called *zesjescultuur*) of Dutch pupils and students that has been widely criticized in policy reports and the media in the Netherlands.[4] Over 20% of university students in the Netherlands are insufficiently committed to their studies (where commitment includes, amongst other factors, the willingness to work hard for higher grades), and the share of very motivated students is low, particularly in the field of Economics where it is below 15% (van den Broek et al., 2009). Brennan et al. (2009) find that among thirteen European countries surveyed, Dutch students are

[4]The term *zesjescultuur* literally means 'culture of the six' (referring to the lowest grade typically required for passing), but online dictionaries including Google Translate suggest 'culture of mediocrity' as a translation. The term is widely used also on social media channels: #zesjescultuur is a popular hashtag on Twitter. A great illustration of the phenomenon is the smartphone application 'Zesjescultuur' that calculates what test mark students are required to get in order to achieve an average final grade of six.

the least likely to strive for the highest possible marks and the third least likely to work more than what is required for passing.

The finding that grade incentives have only limited effect is not specific to our sample. Grove and Wasserman (2006) analyze whether students at the Syracuse University, a large, private university in the state of New York work harder when their problem sets are graded. They find that only one particular type of students are affected: freshmen, while other students do not respond to grade incentives. (Note that our sample consists of second year students.)

The remainder of this chapter is organized as follows. Section 2.2 shortly reviews the related literature and states our contributions. In Section 2.3, we describe our setting, formulate our hypotheses and discuss the details of our experimental design. Section 2.4 provides an overview of our data and some summary statistics. In Section 2.5, we present our results. Section 2.6 contains a further discussion of the findings. We conclude in Section 2.7.

2.2 Literature review

This chapter contributes to the broader literature on piece rate vs. tournament-style compensation schemes. The advantageous and disadvantageous incentive effects of competitive reward schemes have been studied extensively. Early theoretical contributions by Lazear and Rosen (1981) and Green and Stokey (1983) develop the argument that tournament-style incentives may outperform piece rates because under relative performance evaluation "common shocks" are filtered out (see also Holmstrom (1982)). Depending on assumptions about the utility function and the structure of the risk term, relative performance-based schemes may thus provide effort incentives with lower risk exposure. Empirical studies on the incentive effect of competition typically find evidence in line with tournament theory, although the variance in effort levels is much higher than under piece rate incentives (Bull et al., 1987; Harbring and Irlenbusch, 2003; van Dijk et al., 2001).

A few theoretical studies focus specifically on the comparison between relative and absolute

grading. Becker and Rosen (1992) and Landeras (2009) bring the tournament model to the classroom and show that with an appropriate reward scheme, grading on a curve can induce higher performance than absolute grading in the presence of "systemic" noise or correlated individual error terms. Dubey and Geanakoplos (2010) also compare the two grading schemes and find that absolute grading provides better incentives for students to work, provided student outcomes are independent. Paredes (2012) predicts that the response to grading systems differs by ability: in her model where students only care about passing the exam, low-ability students exert less and high-ability students exert more effort under absolute than under relative grading.

Recent contributions from the behavioral economics literature emphasize the importance of competitive preferences: people derive payoff from obtaining a higher rank even in the absence of any tangible benefits (Charness and Rabin, 2002). Azmat and Iriberri (2010) and Tran and Zeckhauser (2012) provide convincing field evidence that feedback on relative performance can increase performance.[5] Recent experiments have consistently shown gender differences in competitive preferences. The gender gap in response to tournament incentives was first documented by Gneezy et al. (2003), who find that male participants solve significantly more mazes under a competitive reward scheme than under piece rate, while no such increase is observed for female subjects in a mixed-sex environment. Their result has been replicated using both laboratory (e.g. Günther et al. (2010)) and field experiments (e.g., Gneezy and Rustichini (2004)) as well as naturally occurring data (e.g., Price (2008)). Niederle and Vesterlund (2011) and Croson and Gneezy (2009) provide detailed reviews of studies on gender and competition.

Empirical studies focusing on the effect of competition in education are still scarce. Jurajda and Münich (2011) and Örs et al. (2013) compare the gender gap in performance at non-competitive and highly competitive tests and find that female students perform worse than men in the competitive situations but not otherwise. Similarly, Morin (2015) observes that men's relative performance increases in response to intensified competition. However, none of these studies are able to separate whether the observed gender gap results from an increase in male

[5]Barankay (2012), on the other hand, finds in a field experiment that removing rank feedback *increases* the performance of male employees.

and/or a decrease in female absolute performance. Bigoni et al. (forthcoming) analyze in a field experiment students' performance on relatively low-stakes homework assignments and find that competition induces higher effort than piece rate among male but not among female students. Jalava et al. (2015) examine various non-financial incentive schemes for primary school children in low-stakes tests and conclude that both girls and boys increase performance when faced with competitive reward schemes. De Paola et al. (2015) do not find gender differences in terms of entry into a tournament or performance under competition in a setting where university students self-select into a competitive scheme to obtain bonus points. Buser et al. (2014) show a strong link between competitiveness and study track choice among Dutch high school students.

The study included in this chapter contributes to the empirical literature on competitive grade incentives by experimentally comparing absolute and relative grading using a design with several advantages. Uniquely, relative grading in our setting involves an actual grading curve where a student's exam grade is determined by his or her place in the class score distribution, perfectly resembling real-life grading practices. The experiment is conducted in a naturalistic setting among students attending a university course ("in the classroom"). The number of participants is high, and students are randomly assigned to treatments (no self-selection) that only differ from each other in the schemes used to translate exam scores to grades. Exams represent high stakes and there is no subjectivity in their evaluation. Administrative data on student characteristics are available, as well as measures of preferences from an incentivized survey. Students' study effort is also observed, allowing us to test whether any change in exam performance is attributable to differences in preparation under the two schemes.

2.3 Context and design

2.3.1 Context

We conducted a framed field experiment (Harrison and List, 2004) among students of the University of Amsterdam (UvA), authorized by the Examination Board of the Faculty of Economics

and Business. The experiment took place in the 2nd year BSc course *Economics of Markets and Organizations* (EMO) during the first block of the 2013/2014 academic year.[6] The course covered topics from Organizational Economics and Industrial Organization in a simple game-theoretic framework, based on lectures notes now published as "Economics of Organizations and Markets" (Onderstal, 2014).

Over 500 students enrolled in the course and thus participated in our experiment. The large sample size was desirable not only because it allowed us to detect potentially small or heterogeneous effect sizes but also because it made it nearly impossible for students in the relative grading group to collude against the experimenters by collectively providing low effort.[7] The attrition rate was low (only 9%) since the class was compulsory for the majority of the enrolled students. The course was offered with identical set-up and content in both Dutch and English, the latter for students following the English-language Bachelor program (in the following referred to as the "international program").

During each study week of the EMO course, students could participate in a three-hour plenary lecture (focusing mostly on theory) in either Dutch or in English, and a three-hour tutorial (discussing exercises, homework solutions and mock exam questions). For the tutorials, students were separated into smaller groups of 15-35 people. Lecture and tutorial attendance was voluntary. Even though students were required to officially register for one of the tutorials before the start of the course, they could in practice attend any of the classes they preferred, so the composition of the tutorial groups varied week by week.

The final grade students obtained for the course depended on their performance on the midterm and end-of-term exams, administered in weeks 4 and 8, respectively. The two exams covered roughly the same amount of material (the midterm exam that took place in week 4 included the topics of the first three weeks while the end-of-term exam focused on the material studied in weeks 5-7) and were designed to be of comparable difficulty. In both exams, students

[6] At the UvA, the academic year is divided into six blocks. The first block runs over eight weeks in September and October.

[7] Budryk (2013) reports a case where students successfully boycotted curved grading, using various social media tools to arrange the collusion.

had 90 minutes to answer 30 multiple-choice questions (calculations, theory, and literature-related, with four possible answers per question). Both exams were corrected by machines, thus grading was by construction unbiased. In addition to the exam grades, students could earn a bonus point (worth one grade point) by handing in four sets of homework assignments in teams of three or four people in weeks 3, 4, 6 and 7. Assignments were graded under an absolute scheme. Students obtained the bonus point if the average grade of their four homework assignments was 5.5 or above (in the Dutch system, the grading scale runs from 1 to 10). The final course grade was calculated as the unweighted average of the midterm and end-of-term exam grades, augmented by the bonus point when obtained. In order to pass the course, students had to have a final grade higher or equal to 5.5.[8]

2.3.2 Design of the experiment

Our experimental design involved randomly assigning course participants to one of the two treatment conditions (communicated to students as the "Yellow" and the "Blue" group in order to maintain a neutral framing). All students, regardless of this assignment, sat the same midterm and end-of-term exams at the same time and in the same venue. As mentioned earlier, both exams counted with equal weight towards the final course grade. The difference between the treatment groups lay in the *grading schemes used for translating exam scores into exam grades*. As shown in Table 2.1, students in the "Blue" group were graded under an absolute scheme in the midterm and under a relative scheme in the end-of-term exam, while the schemes were reversed in the "Yellow" group.[9] We performed a stratified randomization along the dimensions we suspected would influence the response to the grading schemes, i.e., gender, study program, and mathematics ability (this information, together with other demographic variables, was available to us prior to the start of the classes).

[8]Students who did not pass the course after the first attempt could take a resit exam in January that covered the complete course material. Homework bonus points were not carried over to the retake, so a resit exam grade of at least 5.5 was required to pass the course. Those who also failed the resit exam had to retake the course the following academic year.

[9]The reversal of grading schemes is required to perform the experiment while ensuring *ex ante* fair treatment of our subjects, a necessary requirement for approval by the Examination Board.

Table 2.1: DESIGN OVERVIEW: Treatment groups and grading schemes

	"BLUE" group	"YELLOW" group
Midterm exam	absolute	relative
End-term exam	relative	absolute

This design allows for a clean comparison of the effect of the two grading schemes on exam performance and study effort while maintaining an *ex ante* fair and equal treatment of students in the two groups. Using a between-subject design, we can compare the midterm exam outcomes between students in the absolute and the relative grading groups.[10] Moreover, we can take advantage of the within-subject nature of our design by analyzing changes in performance between the mid- and end-of-term exams within each treatment group.

Our main variable of interest is the score (i.e., the number of correct answers) on the midterm exam. We also consider several proxies for effort provision in preparation for the exam: lecture and tutorial attendance during the study weeks (collected by an assistant and by the tutors), handing in homework assignments, grades for homework assignments, and self-reported study time.

The timeline of the experiment is shown in Table 2.2. Students were informed of their treatment group assignment by e-mail and also by posts on the course Intranet page containing all study materials and course-related information. Detailed instructions regarding the grading schemes were included in the Course Manual (see Appendix 2.B) and were also announced during the lectures and tutorials. During the first week, preference and ability information

[10]Our identification relies on the assumption that students, when preparing for and taking the midterm exam, only focus on the midterm grading scheme they are assigned to and do not consider the end-of-term scheme that awaits them. If the fact that schemes are reversed for the second exam simply dilutes the incentives experienced by students, our results might be biased towards zero. It is more problematic if the reversal induces effort substitution between the two exams. In particular, for our identification strategy to produce clean results, we need the tendency of students to substitute effort between the two exams to be uncorrelated with their competitive preferences. We revisit this assumption in Section 2.5.2.

was collected from students in an online survey (discussed in more detail in Section 2.3.5). Students were required to form homework teams with others from the same treatment group (in order to reduce potential spillovers). This also increased their awareness of the treatment assignment. Homework results were not published before week 5, so students did not receive any feedback on their relative performance before the midterm exam. Right before the midterm exam, students were required to fill out a short questionnaire testing their understanding of the grading schemes and collecting information on the time they spent studying for the course.

Table 2.2: TIMELINE OF THE EXPERIMENT

Week 1	Study week	**Announce treatment group assignment**
Week 2	Study week	Deadline for survey; forming homework teams
Week 3	Study week	Deadline homework 1
Week 4	Exam week	Deadline homework 2; Questionnaire & **Midterm exam**
Week 5	Study week	Results homework 1-2 published
Week 6	Study week	Deadline homework 3
Week 7	Study week	Deadline homework 4
Week 8	Exam week	Results homework 3-4 published, **Final exam**

2.3.3 Hypothesis

In order to derive our hypothesis, we briefly review the theoretical considerations underlying our design of the grading schemes. For a meaningful comparison between a criterion- and a norm-referenced grading system, Landeras (2009) emphasizes that both schemes should be implemented *efficiently*, allowing us to compare the highest optimal effort under each scheme. In practice, however, it is hardly feasible to derive the optimal grading standard and curve with multiple different grade categories while taking into account the heterogeneity in student ability (see Moldovanu and Sela (2001)).[11] We therefore follow a different approach in our study and set the grading curve such that it imposes the same distribution of exam grades as we expect under absolute grading.

[11] Consider also the discussion in van Dijk et al. (2001) on choosing the payoffs in the tournament condition.

We analyze by means of a simple theoretical model (presented in Appendix 2.A) the utility maximization problem of students under absolute and relative grading when the curve is set such that the grade distribution is 'forced' to be the same under the two schemes. The model accounts for heterogeneity in student ability and assumes an effort-dependent noise term when translating effort into exam scores. We first show in a general version of the model that the two grading schemes should lead to the same optimal effort level. We then consider a special case of the model where students are assumed to have *competitive preferences*, modeled as an extra term in the utility function that is increasing in one's relative rank. In case a subsample of students have competitive preferences, the model predicts that these students will exert more effort under relative than under absolute grading, while their peers without competitive preferences are expected to exert less effort when graded on the curve. We combine these results with the standard empirical finding of gender differences in competitive preferences to obtain our hypothesis.

Hypothesis: Grading on a curve induces higher effort provision and better exam performance among male students than absolute grading. Female students, on the other hand, provide less effort and do worse under relative than under absolute grading.

2.3.4 Details of the grading schemes

We continue by discussing how the two grading schemes were implemented in practice. The course *Economics of Markets and Organizations* has been taught at the University of Amsterdam for several years with only small changes in the content. The observable characteristics of the student pool participating in the course have also been relatively stable over the recent years. Previous years' grade distributions could thus be taken into account when designing the specific details of the grading schemes in our experiment. Just like most other courses at the university, the EMO course had been graded under an absolute scheme in the years before our intervention.

Under absolute grading, students' exam score must pass a pre-specified standard in order for them to obtain a given grade. In our experiment we chose to use the standard that had been in place also in the previous years in the EMO course. Students' exam scores were translated to exam grades using the following formula:

$$\text{Grade exam} = 10 - 0.4*(\text{number of incorrectly answered questions})$$

With 30 exam questions in total, this formula leads to the standards described in the first column of Table 2.3.

Table 2.3: THE GRADING SCHEMES

GRADE	ABSOLUTE GRADING Exam score *(=points earned)*	RELATIVE GRADING Relative rank *(calculated from the top)*
10	29 - 30	1%
9	27 - 28	2 - 5%
8	24 - 26	6 - 16%
7	22 - 23	17 - 37%
6	19 - 21	38 - 63%
5	17 - 18	64 - 84%
4	14 - 16	85 - 95%
3	12 - 13	95 - 99%
2	0 - 11	99 - 100%

Under relative grading, students' exam grade is determined by their position in the score distribution. A pre-specified norm (the "curve") is used to assign an exam grade to any given rank. We decided to set the curve to mimic the overall realized grade distribution of the previous two years: with a mean of 6 and a standard deviation of 1.5.[12] Assuming that student ability and exam difficulty is unchanged over time, a curve based on previous years' grade distribution should lead to equivalent effort provision under the two grading schemes in our experiment, in the absence of competitive preferences. Our design choice, besides being prompted by theoretical considerations, was also motivated by fairness concerns: we did not want to *ex ante*

[12]To be precise, the means (standard deviations) of EMO exam grades in the academic years 2011-12 and 2012-13 were 6.33 (1.55) and 5.60 (1.7), respectively. An alternative option was to use a curve with a normal distribution, with its parameters determined by the actual mean and standard deviation that occur in the absolute grading group. However, we wanted to avoid the complexity and uncertainty that this design would have entailed.

impose a stricter or more lenient standard on either treatment group. For practical purposes we designed the curve to be symmetric around the mean.[13] The resulting grading norm is presented in the second column of Table 2.3.

As mentioned in Section 2.3.2, we communicated all the details of the grading schemes to the students at the very beginning of the course. They were not informed, however, that we set the curve to closely resemble the grade distribution of the years before, so as not to unintentionally bias students' beliefs and perceptions about the schemes.

2.3.5 Incentivized survey

We conducted an online survey to collect preference, confidence, and ability measures from students that might influence their response to the two grading schemes (see e.g., Niederle and Vesterlund (2007) or Gneezy et al. (2003)). We included the survey among the compulsory course requirements, ensuring a very high response rate (92%). The survey was incentivized with monetary rewards: five respondents were randomly chosen at the end of the course and were paid according to their performance and their choices in the survey (average earnings of the prize winners were €215.67, with a minimum of €100 and a maximum of €457). Respondents spent 21 minutes on average to complete the survey (designed and pre-tested to take about 15-20 minutes), suggesting the majority of students took the task seriously and did not answer at random. The survey was programmed using the online survey software Qualtrics.

The survey was framed as assessing familiarity with the prerequisites for the course, and contained a timed multiple-choice test with 10 questions related to first-year mathematics and microeconomics courses (e.g., simple derivations, perfect competition, Nash-equilibria, etc.).[14] Performance on the test serves as an ability measure in our analysis. Before solving this test, students were required to choose the reward scheme to be applied to their test performance by

[13] As a result, the lowest grade awarded under this curve is not a 1, but a 2. To keep the two schemes comparable, we also adjusted the absolute grading standard such that students "automatically" receive a grade 2 even if they do not answer any questions correctly. A side effect of not awarding grades below 2 is that students who obtain a grade of 7 or higher on the midterm exam and receive the homework bonus point can pass the course simply by showing up at the end-of-term exam ($(7+2)/2+1=5.5$, the lowest passing grade).

[14] For an example of a test question, please refer to Figure C1 in Appendix 2.C.

reporting their switching point between a constant piece rate and a tournament scheme with an increasing prize (similar to the design of Petrie and Segal (2014)). This measure serves as a proxy for competitive preferences. Besides, we also elicited an unincentivized, self-reported rating of "competitiveness in general". Moreover, we collected four different measures of overconfidence[15] (*ex ante* and *ex post*; absolute and relative): students were asked to report their expected absolute score and relative rank both before and after taking the test. In addition, risk and ambiguity preferences of participants were measured by eliciting switching points in Holt and Laury (2002)-style choice menus (see Figure C2 in Appendix 2.C), and also by asking students to rate their willingness to take risk in general (Dohmen et al., 2011). Finally, students reported their expectations regarding their absolute and relative performance in the course and also their attitudes toward norm- and criterion-referenced grading practices.

2.4 Data

This section contains an overview of our data. Panel A of Table 2.4 presents basic demographic information based on administrative data provided by the University of Amsterdam. In total, 529 students registered for the course, with a quarter following the international program. The share of female students in the sample is relatively low, just over a third, reflecting the general gender composition of the Economics and Business Bachelor program. The average age is 20.8 with relatively low variance. The majority of the participants were born in the Netherlands and are Dutch citizens. Our dataset contains several indicators of the past academic achievement of the students in our sample, most notably the average mathematics grade and the number of retake exams. The first, constructed as the unweighted average of any mathematics- or statistics-related exam a student had ever taken at the UvA (including failed tests), is a fairly good predictor of the final grade in the EMO course: the correlation between the two is 0.50 and is highly significant. This math-grade based measure indicates very low average performance:

[15] We define an agent as overconfident when her perceived ability exceeds her true ability. For a discussion on different definitions of overconfidence from an economics perspective, please refer to Hvide (2002).

the mean of the variable, 5.88, is barely above the minimum requirement for passing (i.e. 5.5). The second indicator is calculated as the number of retake exams over all the courses the student ever registered for. On average, students repeat approximately one out of five exams.[16]

Panel B of Table 2.4 provides an overview of the preparation behavior and performance of students in the EMO course. Attendance rates were relatively low during the study weeks preceding the midterm exam: out of the three lectures and tutorials, students participated on average 1.21 and 1.45 times, respectively. The majority of students handed in homework assignments and obtained fairly good homework grades (a mean of 6.95 out of 10), varying in the range between 3.45 and 9.45. (An average homework grade of 5.5 or above ensured the bonus point.) Students reported spending on average 10 hours per week on studying and practicing for the course. The show-up rate at both of the exams was very high, 91% at the midterm and 87% at the end-of-term exam. The average number of correct answers on the midterm exam was 19.28 out of 30, which decreased to 17.41 in the end-of-term exam. Analyzing the final grades, note that it was theoretically possible to get a grade 11 in this course (two students indeed received a calculated grade of 10.5) because the homework bonus point was added to the unweighted average of the two exam grades.

Descriptive results from the incentivized online survey are presented in Panel C of Table 2.4. The relatively low average performance on the test measuring knowledge in prerequisites (4.67 correct answers out of 10 questions) is likely explained by the intense time pressure students were subjected to during the test (25 seconds per question). Performance on the test is significantly correlated with the final grade of the course (corr $= 0.23^{***}$). Students are on average overconfident according to all confidence measures we have elicited. In the table we present the *ex ante* relative overconfidence variable, based on a comparison between the students' guessed and actual relative performance. A correct guessed rank would correspond to a score of zero on our overconfidence scale, and any positive number indicates overconfidence.

[16]Note that values for demographic and ability variables are missing for a number of students in our sample. We deal with the issue of missing covariates in regressions by replacing missing values with zeros and including indicator variables in all regressions indicating whether the given observation has a missing value for the given covariate. (We do not impute missing values for gender. The two observations for whom the gender information is missing are dropped from our analysis.) Our results are not sensitive to the method of imputation we use.

Table 2.4: SUMMARY STATISTICS: Demographic variables, course and survey outcomes

	MEAN	STD. DEV.	MIN.	MAX.	N
PANEL A: DEMOGRAPHICS					
international program	0.25	0.44	0	1	529
female	0.34	0.48	0	1	527
age	20.84	2.08	18	35	485
Dutch-born	0.74	0.44	0	1	517
Dutch nationality	0.79	0.41	0	1	517
avg. math grade	5.88	1.49	1.13	10	463
avg. number of retakes	0.22	0.23	0	1.43	475
PANEL B: COURSE OUTCOMES					
lecture attendance*(scale 0-3)*	1.21	0.94	0	3	517
tutorial attendance *(scale 0-3)*	1.45	1.00	0	3	529
handing in HW *(0/1)*	0.81	0.39	0	1	529
average HW grade *(scale 0 - 10)*	6.95	1.13	3.45	9.45	427
self-reported study time *(scale 1-5)*	2.42	0.77	1	5	385
midterm show-up *(0/1)*	0.91	0.28	0	1	529
end-of-term show-up *(0/1)*	0.87	0.34	0	1	529
midterm score *(scale 0-30)*	19.28	3.8	8	29	483
end-of-term score *(scale 0-30)*	17.41	4.27	4	27	461
final grade *(scale 1 - 11)*	6.65	1.33	2.5	10.5	461
PANEL C: SURVEY OUTCOMES					
survey complete *(0/1)*	0.92	0.28	0	1	529
test questions *(scale 0-10)*	4.67	1.67	0	10	486
overconfidence *(scale -100 to 100)*	18.23	29.65	-78	100	487
risk aversion *(scale 0-10)*	5.10	1.91	1	11	487
ambiguity aversion *(scale 0-10)*	6.44	3.14	1	11	487
competitiveness *(incentivized, scale -10-10)*	-0.05	3.76	-8	10	485
competitiveness *(self-reported, scale 0-10)*	6.79	1.92	0	10	486
expected grade *(scale 0-10)*	7.04	0.89	3	10	485
expected rank *(scale 0-100)*	37.37	17.81	0	100	485
attitude absolute grading *(scale 0-10)*	7.88	1.82	1	11	485
attitude relative grading *(scale 0-10)*	4.33	2.75	1	11	485

As mentioned in the previous section, students' risk, ambiguity, and competitive preferences were measured in Holt and Laury (2002)-style choice lists. We find respondents to be on average risk-neutral: the mean switching point is 5.10 and the risk-neutral subject should switch at decision 5. We have calculated a proxy for students' competitiveness by comparing their "optimal" switching point (based on their relative performance guess, assuming risk neutrality) with their actual switching point between piece rate and tournament incentives. This measure shows large variability in competitiveness with an average of -0.05 and standard deviation of 3.763 on a scale that runs between -10 (maximal aversion to competition) and 10 (maximal preference for competition). According to the self-reported measure, students on average consider themselves competitive (a mean rating of 6.79). Students' overconfidence is also reflected in their grade expectations exceeding their realized final grades (an average of 7.04 vs. 6.65) and in their guessed relative performance in terms of grades (students guess on average that out of 100, only 37.37 of their peers will do better than them). Students report a more positive attitude towards absolute than towards relative grading, which is likely due to their inexperience with the latter scheme. Still, students are not strongly opposed to relative grading: the mean rating for grading on a curve was 4.33 on a scale of 0 to 10 (where 5 corresponds to neutral).

Table 2.5 proves that the randomization was successful by comparing observable characteristics of students between the two treatment groups (separately by gender). The groups are balanced not only along the dimensions used for stratification (study program and mathematics grades), but also with respect to other demographic, ability, and preference variables. Male students in the "Blue" group do not differ from men in the "Yellow" group, except (marginally) in terms of their ambiguity aversion. Women in the two treatment groups are not significantly different in their demographic and ability characteristics, although female students in the "Yellow" group report higher expected grades. Once we apply the Bonferroni correction for multiple comparisons, neither of these differences remains significant.

Table 2.5: COMPARISON OF MEANS BETWEEN TREATMENT GROUPS, BY GENDER.

	MEN			WOMEN			GENDER
	BLUE	YELLOW	Diff.	BLUE	YELLOW	Diff.	Diff.
DEMOGRAPHICS							
int. program	0.219	0.179		0.337	0.360		***
	(0.031)	(0.030)		(0.050)	(0.051)		
age	20.951	20.795		20.774	20.759		
	(0.175)	(0.142)		(0.270)	(0.208)		
Dutch born	0.779	0.812		0.659	0.644		***
	(0.032)	(0.031)		(0.050)	(0.052)		
ABILITY							
Math grade	5.782	5.697		6.108	6.174		***
	(0.115)	(0.122)		(0.172)	(0.173)		
num. retakes	0.237	0.237		0.203	0.189		*
	(0.018)	(0.019)		(0.024)	(0.024)		
test questions	4.830	4.643		4.541	4.494		
	(0.132)	(0.138)		(0.157)	(0.183)		
PREFERENCES							
overconfidence	16.201	18.426		18.082	22.057		
	(2.337)	(2.477)		(3.015)	(3.187)		
risk aversion	4.881	4.903		5.541	5.391		***
	(0.160)	(0.145)		(0.195)	(0.204)		
ambig. aversion	6.830	6.200	*	6.341	6.276		
	(0.250)	(0.253)		(0.340)	(0.335)		
competitiveness *(incentivized)*	-0.132	0.033		0.365	-0.391		
	(0.273)	(0.311)		(0.441)	(0.416)		
competitiveness *(self-report)*	7.264	6.994		6.153	6.230		***
	(0.138)	(0.146)		(0.220)	(0.217)		
overconfidence	16.201	18.426		18.082	22.057		
	(2.337)	(2.476)		(3.015)	(3.187)		
expected grade	7.090	7.058		6.800	7.115	**	
	(0.075)	(0.069)		(0.083)	(0.096)		
expected rank	36.260	37.00		40.671	37.057		
	(1.413)	(1.489)		(2.037)	(1.654)		
N	178	168		92	89		

Notes: Significance of differences calculated from two-sample t-test with unequal variances. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.5 also allows for gender comparisons. We observe that women are more likely than men to follow the international program and are thus less likely to have been born in the Netherlands. There is also a gender difference in past academic performance: on average, women obtained higher math grades and had to retake fewer exams than their male peers (a two-sided t-test with unequal variances confirms that these differences are significant at the 1% and the 10% level, respectively).[17] We find no such difference in the number of correct test questions, possibly due to the intense time pressure in the survey (Shurchkov, 2012). In terms of preferences, we find that men and women differ in their attitudes toward risk, with women being significantly more risk averse. This finding is in line with results from other studies (Croson and Gneezy, 2009).[18] Contrary to our expectations, we find no significant gender differences in competitiveness, as measured by the incentivized choice between piece rate and tournament. This finding may be explained by women in our sample being as confident as men, and no more ambiguity averse than male students, either. However, men rate themselves significantly higher on the self-reported competitiveness scale than women.

2.5 Results

2.5.1 Selection

Having shown that there are no concerning pre-intervention differences between the treatment groups, we need to alleviate concerns related to non-random attrition. Students assigned to relative grading who are particularly averse to competition may decide to skip the midterm exam or to drop out of the course entirely, biasing our estimation results. The findings of Niederle and Vesterlund (2007) and several replications suggest that even high-ability women

[17] The difference is not driven merely by the higher share of international students among women. Even after controlling for the study program, women obtain significantly higher grades than the men in our sample. Using the Bonferroni correction to account for multiple testing, the differences between men and women in the number of retakes is no longer significant.

[18] The review and meta-analysis by Filippin and Crosetto (2014) suggests, however, that the gender differences in risk-taking observed in the literature are sensitive to the methods of elicitation and are often economically insignificant.

are likely to shy away from competition. We would thus expect to see lower midterm show-up among females in the relative grading group. We find no support for this hypothesis in our data: there is no gender difference in the propensity to participate in the exam (a t-test yields a p-value of 0.23). Selection does not ruin the balancedness of the two treatment groups, and the actual number of non-participants is very low: 16 vs. 30 in the relative and absolute group, respectively.[19] We thus argue that non-random exam participation is unlikely to bias our results.

2.5.2 Effect of relative grading in the full sample

Preparation behavior

We start our analysis by comparing preparation behavior between the treatment groups in the weeks leading up to the midterm exam. We test whether treatment assignment influenced students' propensity to hand in homework assignments, their homework grades, their lecture and tutorial attendance and their self-reported study times. Panel A of Table 2.6 summarizes our results. Competitive grade incentives had little impact on preparation behavior prior to the midterm exam: students in both groups were equally likely to hand in assignments (column (1)), to attend classes (column (3)) and to spend time studying for the course (column (4)) during the first three weeks. Relative grading had a marginally significant positive impact on the quality of homework assignments (column (2)): the grade average of the first two assignments was 0.57 higher for students in the "Yellow" group.

An analysis of preparation behavior could also shed some light on the extent to which treatment assignment induced differential effort substitution between the midterm and end-of-term exams. Panel B of Table 2.6 shows that preparation efforts in the weeks preceding the end-of-term exam did not differ between the treatment groups.[20]

[19] Show-up is thus slightly *higher* under relative grading (a raw difference of 4.9 percentage points, significant at the 5% level).

[20] Note that our baseline category in the models in Panel A contains students assigned to absolute grading on the midterm exam, i.e. students in the Blue treatment group, and the coefficient associated with *relative* grading measures the impact of being assigned to the "Yellow" group instead. When presenting results in Panel B, in order to treat the same students as "baseline" as in Panel A and to be able to compare the impact of belonging to the "Yellow" group between the panels, we decided to report the impact of *absolute* grading on students' effort. Self-reported measures of study time were only collected before the midterm exam and therefore cannot be included in Panel B.

Table 2.6: THE EFFECT OF RELATIVE GRADING ON PREPARATION BEHAVIOR.

	hand in HW (1)	*avg. HW grade* (2)	*attendance* (3)	*prep. time* (4)
PANEL A: WEEKS 1-3 (prior to midterm exam)				
relative	0.474	0.565*	0.026	-0.055
	(0.525)	(0.295)	(0.217)	(0.129)
male	-0.506	0.186	0.063	-0.289**
	(0.388)	(0.268)	(0.189)	(0.113)
relative*male	-0.023	-0.483	0.020	0.135
	(0.604)	(0.329)	(0.268)	(0.160)
Demographic controls	✓	✓	✓	✓
Ability controls	✓	✓	✓	✓
Constant	3.559*	6.342***	3.062***	2.413***
	(1.862)	(0.856)	(0.900)	(0.587)
N	527	426	516	384
(Pseudo-)R^2	0.250	0.082	0.026	0.119
PANEL B: WEEKS 5-7 (prior to end-of-term exam)				
absolute	-0.281	0.033	-0.190	
	(0.384)	(0.336)	(0.245)	
male	-0.661**	0.006	-0.019	
	(0.328)	(0.318)	(0.215)	
absolute*male	0.620	-0.257	0.065	
	(0.460)	(0.401)	(0.303)	
Demographic controls	✓	✓	✓	
Ability controls	✓	✓	✓	
Constant	6.913***	6.536***	1.904*	
	(1.723)	(1.113)	(1.020)	
N	527	408	516	
(Pseudo-)R^2	0.262	0.272	0.043	

Notes: The table displays estimated coefficients from (1): logistic and (2)-(4): OLS regressions. Dependent variables Panel A: (1): hand in HW 1&2, (2): avg. grade HW 1&2, (3): attendance weeks 1-3, (4): self-reported study time. Dependent variables Panel B: (1): hand in HW 3&4, (2): avg. grade HW 3&4, (3): attendance weeks 5-7. Covariates (1)-(4): int. program, age, Dutch born, Math grades, num. retakes, test questions. In all specifications, indicator variables for missing covariates included. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Students in the two groups were equally likely to hand in the third and fourth homework assignments (column (1)) and to attend classes in the study weeks prior to the second exam (column (3)). Moreover, we see no evidence of students in the "Yellow" group (now preparing for an

exam that is graded under the absolute scheme) receiving lower homework grades (column (2)). Overall, these findings support our identifying assumption as they suggest that the different grading schemes did not cause students to substitute effort away from one exam to the other.

Between-subject analysis

We continue our analysis with a comparison of midterm exam performance between the two treatment groups. Since students were randomly assigned to the grading schemes, a simple comparison of the groups' scores shows us whether students performed differently under relative than under absolute grading. The mean number of correct answers was 19.20 under absolute and 19.37 under relative grading (with standard deviations of 3.79 and 3.81, respectively) out of 30 questions. According to a two-sample t-test with unequal variances, the difference is insignificant (p-value: 0.62). As Figure 2.1 shows, the distributions of outcomes in the two treatment groups also look very similar. A Kolmogorov-Smirnov test does not reject the equality of the two distributions (exact p-value: 0.99).

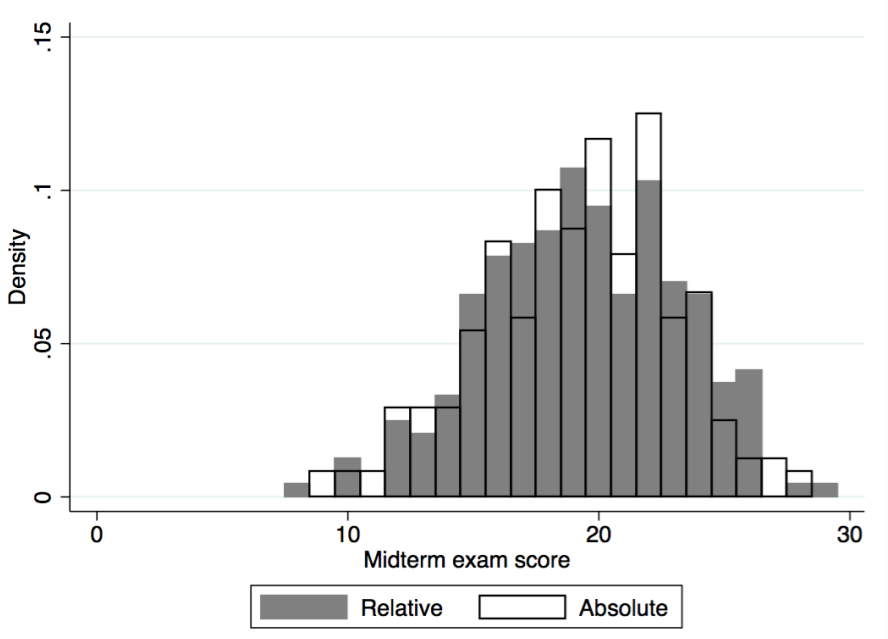


Figure 2.1: Distribution of midterm exam scores by grading scheme

We proceed to test whether the response to grade incentives differs by gender. Figure 2.2 compares the mean number of correct answers on the midterm exam by gender and treatment group (the figure depicts standardized scores). While there is a slight indication of men performing better under the relative than under the absolute scheme, the difference is small in size and not statistically significant. No treatment difference is observed for female students, either.

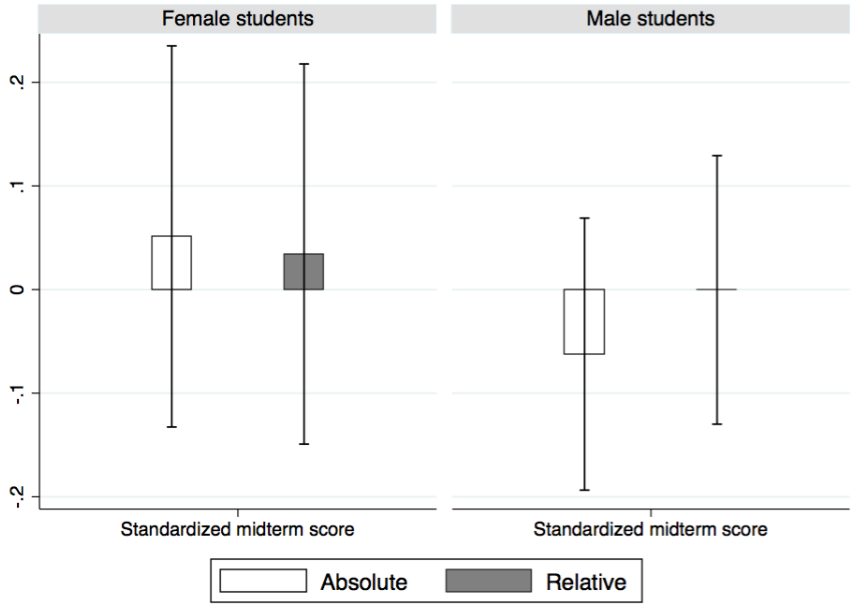


Figure 2.2: Comparison of midterm exam scores by grading scheme and gender (the height of the bars represent the mean and the error bars show the 90% confidence intervals)

These findings are also supported by OLS regressions. Results are presented in Table 2.7: column (1) confirms the finding that there is no overall difference between the scores by grading schemes, while column (2) shows that the interaction term between relative grading and male is also insignificant. In column (3) we see that adding covariates largely improves the explanatory power of our model (the R^2 increases to 0.279 when we include controls for demographic and ability variables). The point estimate for the effect of relative grading is negative and the coefficient associated with *relative*male* is positive; however, both are small and not significantly different from zero.

Table 2.7: THE EFFECT OF RELATIVE GRADING ON MIDTERM SCORES.

midterm score	No covariates	Gender interaction	With covariates
	(1)	(2)	(3)
relative	0.170	-0.064	-0.292
	(0.346)	(0.582)	(0.503)
relative*male		0.300	0.746
		(0.722)	(0.625)
male		-0.431	-0.281
		(0.512)	(0.446)
Demographic controls			✓
Ability controls			✓
Constant	19.196***	19.476***	14.426***
	(0.245)	(0.413)	(2.352)
N	483	482	482
R^2	0.001	0.002	0.279

Notes: The table displays estimated coefficients from OLS regressions. Covariates in column (3): int. program, age, Dutch born, Math grades, num. retakes, test questions. In column (3), indicator variables for missing covariates included. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Within-subject analysis

In this section we consider how students' performance changed between the mid- and the end-of-term exams. In total, 461 students showed up at the end-of-term exam, rather evenly divided between the two treatment groups (226 from the "Blue" and 235 from the "Yellow" group). Comparing the end-of-term exam scores shows no treatment effect: students in the "Blue" group (graded on the curve) received on average 17.16 points, while students in the "Yellow" group (graded on the absolute scale) scored on average 17.65 points.[21] Such a simple comparison, however, is not very informative, since students' effort provision and motivation in the second exam is probably affected by their experience in the midterm exam. We therefore continue by analyzing how the gender gap changes *within* each treatment group when we move from the midterm to the end-of-term exam.

[21] We should note that even though students' scores did not differ significantly between the two treatment groups at either the first or the second exam, the grading schemes did affect students' exam *grades*. Since students assigned to absolute grading performed worse than expected on both exams (their mean grade was 5.69 from the midterm and 5.10 from the end-of-term exam, as opposed to the mean grade of 6 pre-set under relative grading), grading on a curve resulted in higher grades at both occasions.

Table 2.8: CHANGE IN PERFORMANCE BETWEEN THE MID- AND END-TERM EXAMS.

	BLUE group		YELLOW group	
end-term score	No covariates	With covariates	No covariates	With covariates
	(1)	(2)	(3)	(4)
male	-0.425	-0.499	-0.212	0.080
	(0.572)	(0.541)	(0.512)	(0.522)
midterm score	0.446***	0.199**	0.479***	0.339***
	(0.076)	(0.083)	(0.068)	(0.077)
Demographic controls		✓		✓
Ability controls		✓		✓
Constant	8.719***	3.926	8.402***	3.378
	(1.578)	(4.427)	(1.395)	(4.232)
N	225	225	234	234
R^2	0.137	0.308	0.176	0.255

Notes: The table displays estimated coefficients from OLS regressions. Covariates in columns (2) & (4): int. program, age, Dutch born, Math grades, num. retakes, test questions. In columns (2) and (4), indicator variables for missing covariates included. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Since students in the "Blue" group experienced absolute grading in the first and relative grading in the second exam, we expect that males in this subsample respond to the competitive grade incentives by improving their performance in the end-of-term exam. Phrasing it differently, controlling for midterm exam scores, boys in the "Blue" group are predicted to do better than girls at the end-of-term exam. In the "Yellow" group, we expect the opposite: moving from tournament-style to non-competitive incentives, we expect boys to perform relatively worse in the second exam. Table 2.8 presents results from an OLS regression explaining end-of-term exam scores by gender, controlling for midterm scores (columns (1) and (2) show estimates for the "Blue" group and columns (3) and (4) for the "Yellow" group). We find no support for our predictions in the data: there is no significant gender difference in how students' performance changed between the two exams in either of the treatment groups. If anything, the point estimates go in the opposite direction than we expected: controlling for midterm outcomes, boys in the "Blue" group actually do (insignificantly) worse than girls on the competitively graded second exam. Adding controls for demographic and ability characteristics does not change these findings.

2.5.3 Heterogeneity in response

Ability and preferences

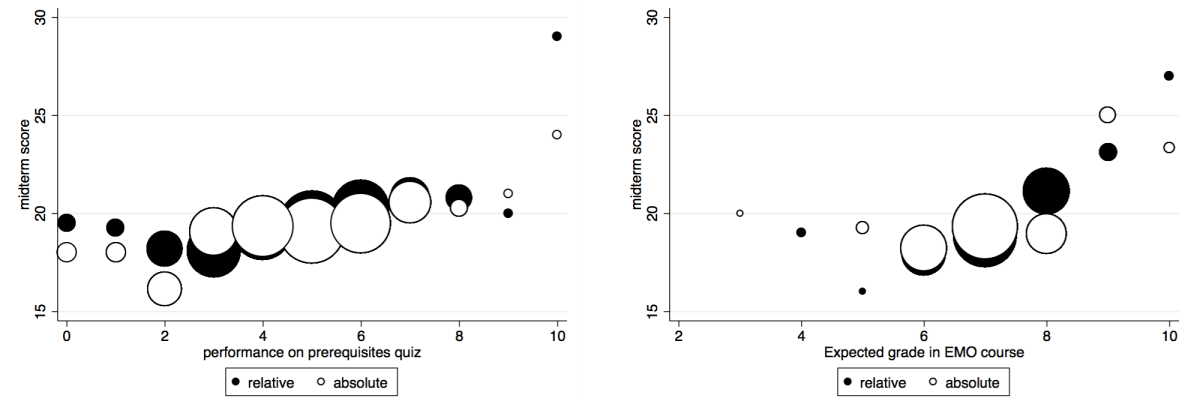
In our analysis so far we only included ability variables as covariates to increase the precision of our estimates. Inspired by the findings of Paredes (2012) and Müller and Schotter (2010) we now consider whether response to grade incentives is heterogeneous with respect to students' skills as captured by their previous mathematics grades. Column (1) of Table 2.9 shows that this is not the case in our sample: the interaction term between relative grading and math grades is insignificant in explaining midterm scores.[22] To account for the fact that previous grades not only reflect skills but also motivation, we test alternative proxies for ability, such as the number of correct answers on the test questions or students' expected grade in the EMO course. As panels (a) and (b) of Figure 2.3 illustrate, these measures of ability do not seem to influence the reaction to grade incentives, either.

Table 2.9: THE IMPACT OF ABILITY AND PREFERENCES ON THE RESPONSE TO RELATIVE GRADING.

midterm score	Ability	Risk aversion	Competitiveness
	(1)	(2)	(3)
relative	0.723	1.065	0.176
	(0.793)	(0.803)	(0.295)
relative * Math	-0.098		
	(0.136)		
relative * risk aversion		-0.178	
		(0.150)	
relative * competitiveness			-0.029
			(0.080)
Demographic controls	✓	✓	✓
Ability controls	✓	✓	✓
Constant	13.905***	15.182***	14.544***
	(2.404)	(2.405)	(2.337)
N	482	482	482
R^2	0.277	0.288	0.267

Notes: The table displays estimated coefficients from OLS regressions. Covariates column (1): male, int. program, age, Dutch born, Math grades, num. retakes, test questions; column (2): (1) + risk aversion; column (3): (1) + competitiveness. In all specifications, indicator variables for missing covariates included. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

[22] We also find no significant effect when testing for a non-linear relationship by including either a squared term for math grades or dummies for the four math quartiles, and their interaction with relative grading.



(a) Ability proxy: test performance (b) Ability proxy: expected grade

Notes: The size of the circles is proportionate to the number of observations in the given category.

Figure 2.3: Mean midterm score in the two grading groups, by different ability levels

We can also make use of the preference measures we elicited in the incentivized survey to test whether less risk averse or more competitive students react more positively to relative grading. In Columns (2) and (3) of Table 2.9 we see no such difference: *relative*risk aversion* and *relative*competitiveness* have no significant effect on midterm scores. In the regressions presented in the paper we used the incentivized measures both for risk and competitive preferences. Results are unchanged when we include the unincentivized, self-reported measures for both traits. Overconfidence and ambiguity preferences do not seem to affect the response to relative grading, either.

Up till now we have compared whether one *type* of grade incentive works better than the other. In so doing, we have implicitly assumed that all students are motivated by grade incentives in the first place. Those students, however, who place little or no weight on the actual level of their grades are unlikely to respond strongly to differences in grading schemes even if they have competitive preferences. We therefore continue our analysis by identifying and examining a subsample of students within our experimental population who are conjectured to be particularly responsive to grade incentives: students who are on the margin of passing or failing the course.[23]

[23]Another subsample of interest is the group of students participating in the international program. Students following the international program surpass their peers in the Dutch-language program both in terms of ability

Marginal students

As we have discussed in the Introduction, Dutch pupils and students are often accused of having a 'just pass' attitude (*"zesjescultuur"*). If students in our sample are also mainly interested in passing the course and not so much in obtaining high grades, then it is plausible to expect the strongest reaction to a change in grade incentives from those students who are close to the pass-fail margin. We identify marginal students by focusing on *predicted grades*. Analyzing data from students who attended the EMO course in the academic year preceding our experiment (i.e. in 2012/2013), we estimate the effect of observable student characteristics on the final course grade. We find that study program, age and math grades can fairly accurately predict course outcomes: they explain 27% of the variation in grades. Using these correlations between student characteristics and course grades from the year before, we create grade predictions for students in our experiment. We find our forecasts to work well: the correlation between predicted and actual course grades is 0.498 and highly significant. We can therefore identify marginal students by focusing on course participants whose predicted grade is near the passing threshold of 5.5.[24]

Table 2.10 analyzes the impact of relative grading on exam performance among marginal students. While there is no overall difference in midterm exam scores between the two treatment groups (Table 2.10, column (1)), the effect seems to be heterogeneous with respect to gender (column (2)): in this sample men respond significantly more positively to competitive grade incentives than women. This result is remarkably robust to the inclusion of control variables (column (3)): the gender gap in response to relative grading is approximately 2.5 on a scale of 0 to 30 (almost two thirds of a standard deviation).

and motivation and are therefore conjectured to be more responsive to grade incentives. Midterm exam results show that in this subsample boys indeed do better under relative than under absolute grading, while there is no difference among girls. A within-subject analysis, however, raises doubts whether the observed difference among male students is the result of our treatment or is due to imperfect randomization. For a detailed analysis, please refer to Appendix 2.D.

[24] In the regressions presented in this chapter, we use the cutoff $4.75 < predgrade < 6.25$. This leads to a sample of 150 students (45 female) of whom 141 showed up for the midterm exam. Note that this subsample is relatively small, so inferences should be made with caution. Our results are robust to different specifications of the marginal category. A different possibility for identifying "marginal" students would have been to rely on self-reported grade expectations (collected in the online survey in the first week of the course). However, this measure was collected after students learned their group assignment, so it could potentially be influenced by the treatment.

Table 2.10: ANALYZING THE SUBSAMPLE OF MARGINAL STUDENTS.

	No covariates	Gender interact.	Demog.& ability	Preferences	Preparation
	(1)	(2)	(3)	(4)	(5)
relative	0.393	-1.371	-1.557	-1.649	-1.627
	(0.626)	(1.145)	(1.134)	(1.097)	(1.127)
male		-0.847	-1.148	-1.434	-1.467
		(1.039)	(1.041)	(1.010)	(1.035)
relative*male		2.571*	2.553*	2.868**	2.856**
		(1.365)	(1.356)	(1.317)	(1.338)
Demographic controls			✓	✓	✓
Ability controls			✓	✓	✓
Preference controls				✓	✓
Preparation controls					✓
Constant	17.672***	18.294***	9.373**	8.515*	8.818
	(0.463)	(0.890)	(4.384)	(5.108)	(5.470)
N	141	141	141	141	141
R^2	0.003	0.034	0.146	0.237	0.239

Notes: The table displays estimated coefficients from OLS regressions. Covariates in column (3): int. program, age, Dutch born, Math grades, num. retakes, test questions. Covariates in column (4): (3) + risk aversion, ambiguity aversion, residual competitiveness, overconfidence. Covariates in column (5): (4) + dummy: hand in HW and self-reported preparation time. In all specifications, indicator variables for missing covariates included. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The sample is too small to get precise estimates when splitting it by gender, but there is suggestive evidence that the observed gender difference is the result of boys doing better and girls doing worse when graded on the curve, as shown in Figure 2.4. The left and right panels depict standardized exam scores of female and male students, respectively. The black bars, corresponding to midterm exam results suggest that average female performance is higher under absolute than under relative grading, while the opposite holds for men.

To explore whether preference differences can explain the gender gap in response to competitive grade incentives, we included measures of risk and ambiguity aversion, overconfidence and competitiveness in our model (column (4)). Contrary to our expectations, the gender gap did not close nor even decrease: the point estimate of the coefficient associated with *relative*male* actually increased slightly and became more significant. This suggests that preferences, as proxied by the incentivized measures we have collected, are not the drivers of the differential response of male and female "marginal" students to relative grading.[25]

[25]Note that our incentivized measure of competitiveness is related to the *propensity to enter* a competitive

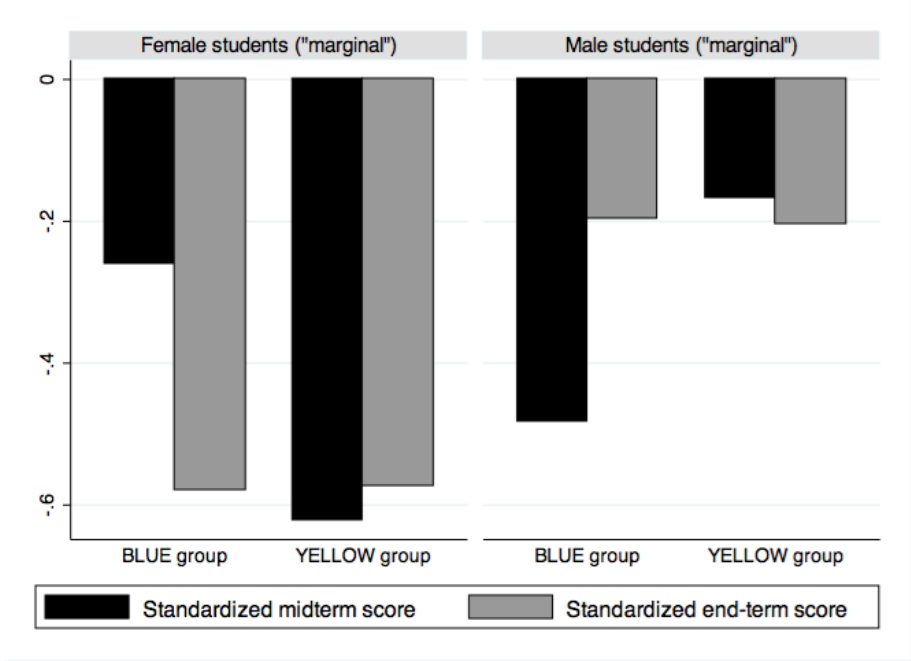


Figure 2.4: Subsample of marginal students: mean mid- and end-term (standardized) exam scores by treatment group and gender

Differences in effort provision cannot explain our results, either. While relative grading increases the propensity to hand in homework assignments in this subgroup, and male marginal students also report spending significantly more time studying for the course when graded on the curve, column (5) of Table 2.10 shows that controlling for these factors in a regression framework does not close the gender gap in reaction to relative grading.

Observing the scores on the second exam (gray bars in Figure 2.4, depicting standardized end-of-term exam scores by treatment assignment and gender), the difference between the grading groups disappears. This is not surprising: students who were predicted to be close to the pass-fail threshold before the midterm are not necessarily the same who are considered "marginal" at the end-term.[26]

environment, which is not necessarily the same as the *ability to perform* in tournaments. Competitive environments might induce choking under pressure such that higher motivation actually leads to lower performance (Ariely et al., 2009). Azmat et al. (2014), studying gender differences in response to high stakes, find result consistent with females responding worse to such pressure.

[26]We have tried different approaches to identifying "marginal" students at the second exam, but we found no clear effect of the different grading schemes on their performance at the end-of-term exam.

2.6 Discussion

Our results indicate that students in our sample do not respond to the treatment: there is no significant difference in preparation effort or exam performance between students experiencing absolute or relative grading schemes. Among those conjectured to be the most sensitive to grade incentives (marginal students), we find some evidence for boys responding better and girls worse to competitive grading. In this section we discuss potential explanations for these findings.

An obvious reason for students not reacting to the different grading schemes could be confusion: participants in the experiment may not have been aware of what the treatments entailed. The questionnaire we conducted before the midterm exam rules out this explanation: out of the 483 students who showed up for the exam, 403 answered these questions and 84% of them understood the treatment. When we exclude those who did not fill out the questionnaire or gave the wrong answers, our results are qualitatively unchanged.

We may not find an effect of relative grading due to the specific design of our experiment: students knew in advance that they will experience both grading schemes. However, we found no indication in our data that this knowledge caused students to shift effort between the two exams in response to their treatment group assignment. It is also unlikely to have diluted our incentives. For students with competitive preferences, the return on effort is higher under relative than under absolute grading because on top of the utility resulting from obtaining a good grade, relative grading also provides them with "rank-utility". This is true regardless of the reversal of the grading schemes at the end-of-term exam.

Another potential explanation could be that students are already at their effort frontier under absolute grading so there is no scope for them to improve in response to competitive grade incentives. This is unlikely to be the case, given the overall low attendance and preparation effort among our participants. Moreover, Leuven et al. (2010) show that students are able to improve their performance in response to financial incentives. Their study pertains to the same population (Economics & Business Bachelor students at the University of Amsterdam) some years prior to our experiment.

A plausible explanation for the lack of response to our treatments is the low competitiveness of students in our sample. Buser et al. (2014) studies the link between competitiveness and track choices among Dutch high school pupils (high school track choices are strongly correlated with the choice of major in tertiary education). They found that those with relatively low levels of competitiveness tend to select into the Economics and Society academic track. Our incentivized measure of tournament choice confirms their finding: on average, both male and female students in our sample are averse to competition. While this might explain why we find no overall effect of relative grading on exam scores in the full sample, it is not clear why those with higher levels of competitiveness do not respond to the treatment either.[27]

Finally, as we have discussed in the Introduction and in Section 2.5.3, in order for students to be responsive to the (differences in) grade incentives, they should be interested in the level of their grades in the first place. If students are mainly interested in passing the course with minimal effort provision and do not attach importance to their grade *per se*, the incentive effect of grading on a curve is likely to be limited. In line with this explanation, we find an effect of competitive grade incentives among the subsample of "marginal" students. These students are predicted to be close to the pass-fail margin and are thus conjectured to care more about grade incentives. In this group, male students seem to react positively to relative grading while females perform better under absolute grading. (Note that the subsample of marginal students is relatively small so inferences must be made with caution.)

2.7 Conclusion

In this chapter we have set out to test a potential remedy for the low performance of male students: competitive grade incentives. In a large-scale field experiment we compared student effort provision and exam performance under absolute and relative grading schemes. Our re-

[27] A possible reason could be the difference between the two aspects of competitiveness: while in our survey we measured the *propensity to enter* a competitive situation, response to relative grade incentives rather depends on the *ability to perform* in a competitive environment. The two do not necessarily coincide: e.g. Niederle and Vesterlund (2007) find a substantial gender gap in the willingness to enter competition even though they record no gender difference in the increase in performance resulting from tournament incentives.

sults show that competitive grade incentives are unable to induce students to work harder in an environment where students care mainly about passing and not so much about excelling: competition in the classroom is ineffective when the prizes (i.e. high grades) are not considered valuable. Policy makers interested in raising student motivation should thus focus on making high grades more attractive. This could be achieved for instance through linking academic performance to financial incentives, by e.g. tying tuition fees to grade point averages. Leuven et al. (2010), also studying the sample of Bachelor students at the University of Amsterdam, have shown that financial rewards can successfully induce higher performance even in an otherwise unambitious sample, and without crowding out intrinsic motivation.

We have shown that tournament-style grade incentives are not the cure for the "new gender gap" that occurs among low-achieving students. Given our findings, we believe it is important to replicate our study in a setting where students are more ambitious and care more about the level of their grades. It is an interesting avenue for future research to focus on the top of the distribution and test experimentally whether competitive grading hinders the academic performance of females, especially in more mathematics-related subjects as suggested by Niederle and Vesterlund (2010).

Appendix

2.A Theoretical model

This section presents a theoretical model that considers the utility maximization problem of students and derives their optimal effort provision under absolute and relative grading. We discuss a setting where the relative grading curve is set in a way that the distribution of grades is 'forced' to be the same under the two schemes. We first consider a general version of the model, then discuss a special case with competitive preferences.

Effort under absolute and relative grading

A continuum of students decide how much effort to exert on an exam. A student is characterized by $\alpha \in \mathbb{R}^n$, $n \in \mathbb{N}$, denoting a vector of student characteristics (e.g. ability). Students' utility is a function $U(e, g, \alpha)$ of their effort $e \geq 0$, their exam grade g, and α. For the moment, we do not make any assumptions on the shape of U. In the literature, U is typically assumed to be additively separable in grade and effort, where U is strictly decreasing in e (as effort is assumed to be costly for the students), increasing in a one-dimensional ability parameter α (the higher a student's ability the lower her effort costs or, equivalently, the higher her grade at any fixed level of effort), and weakly increasing in g (students care about passing the course or the level of the grade). The model also captures a setting where students are bound by an effort frontier (which could be modelled by letting $U(e, g, \alpha) = -\infty$ for effort exceeding a student's effort frontier).

Under both absolute grading and relative grading, a student's exam grade is determined by her exam score s, which is determined by her effort and effort dependent noise $\epsilon(e) \geq 0$, in the following way: $s = e + \epsilon(e)$. When choosing her effort, the student does not observe the realization of $\epsilon(e)$. All students are expected utility maximizers.

Absolute grading

Under absolute grading, a student's grade is a strictly increasing function g of her score s. Each student maximizes utility by picking

$$\begin{aligned} e^A(\alpha) \in \arg\max_e E\left\{U(e, g, \alpha)\right\} = \arg\max_e E\left\{U(e, g(s), \alpha)\right\} = \\ = \arg\max_e E\left\{U(e, g(e + \epsilon(e)), \alpha)\right\}. \end{aligned} \tag{2.1}$$

Before discussing relative grading, we make several simplifying assumptions.

A1 The effort maximization problem under absolute grading has a solution for all students, which is unique.

Let $\sigma = e^A(\alpha) + \epsilon\left(e^A(\alpha)\right)$ denote a student's exam score and let F denote the distribution of σ over the student population and f the corresponding density function. L

A2 For each type α, the fraction of students for whom $\epsilon\left(e^A(\alpha)\right)$ is less than any $\widehat{\epsilon} \in \mathbb{R}$ equals the ex ante probability that $\epsilon\left(e^A(\alpha)\right)$ is less than $\widehat{\epsilon}$.

Assumption A2 implies that the distribution F of grades is fully deterministic. Let $\overline{\sigma}$ denote the highest possible score under absolute grading.

A3 $F(0) = 0$; $F(\overline{\sigma}) = 1$; $f(\sigma) > 0$ for all $\sigma \in (0, \overline{\sigma})$.

Assumption A3 implies that F is invertible on the interval $[0, \overline{\sigma}]$. Let $F^{-1} : [0, \overline{\sigma}] \to [0, 1]$ represent the inverse function of F.

Relative grading

Under relative grading, a student's grade is determined by her rank r in the score distribution where r equals the fraction of students in the entire student population whose score is below hers. Now, consider a scheme of relative grading where a student's grade $G(r)$ as a function of her rank is determined by the score distribution under absolute grading in the following way:

A4 $G(r) = g(F^{-1}(r))$.

Assumption A4 'forces' the grade distribution under relative grading to be the same as under absolute grading. The next proposition shows that as a consequence of this, students will exert the same effort under relative grading as under absolute grading.

Proposition 1 *Under assumption A4,* $e^R(\alpha) = e^A(\alpha)$ *constitutes a Bayesian Nash equilibrium for relative grading.*

Proof. Consider a student characterized by type α. Suppose all other students choose effort $e = e^A(\hat{\alpha})$ if their type is $\hat{\alpha}$. If the student chooses effort e, her score $s = e + \epsilon(e)$ will result in rank $r = F(s)$. The student best responds by choosing

$$\begin{aligned} e \in \arg\max_e E\left\{U(e, G(r), \alpha)\right\} &= \arg\max_e E\left\{U(e, g\left(F^{-1}(r)\right), \alpha)\right\} = \\ &= \arg\max_e E\left\{U(e, g\left(s\right), \alpha)\right\} = \\ &= \arg\max_e E\left\{U(e, g\left(e + \epsilon(e)\right), \alpha)\right\} \end{aligned} \tag{2.2}$$

Observe that maximization problems (2.1) and (2.2) coincide so that $e = e^A(\alpha)$ is indeed a best response. ■

Competitive preferences

In this section, we assume that some students not only care about their grades and their effect costs but also about their rank in the grade distribution. We make the additional assumption that

students only care about their rank if they are perfectly informed about this (e.g., because they can credibly inform fellow students about their rank). By definition, relative grading provides students with information about their rank. We assume that under absolute grading, students do not care about their rank as they only obtain imperfect information about it.

Consider a population of risk-neutral students whose utility, in the case of absolute grading, is given by

$$U(e, g, \alpha) = g - \frac{e^2}{2\alpha}.$$

Suppose that the students' α's are one-dimensional and distributed according to cumulative distribution function F over the interval $(0, \overline{\sigma}]$ that satisfies assumptions A1-A3. Suppose that under absolute grading, effort translates into a grade in the following way:

$$g(s) = s = e.$$

It is readily verified that a student's optimal effort equals

$$e^A(\alpha) = \alpha.$$

If we construct the relative grading scheme using assumption (A4), it follows that the grade distribution follows the students' effort distribution under absolute grading, which is F. As a consequence, $G(r) = g(F^{-1}(r)) = F^{-1}(r)$. (Note that assumptions A1-A3 guarantee that F^{-1} is well-defined for all $r \in [0, 1]$.) Now, suppose that under relative grading, a student's utility is modified to

$$\hat{U}(e, g, r, \alpha, \rho) = g + \rho F^{-1}(r) - \frac{e^2}{2\alpha}$$

where $\rho \geq 0$ is a parameter measuring how much the student cares about her relative rank. This is in line with the preference structure imposed by Moldovanu et al. (2007) in their paper on status classes; we assume a continuum of status classes. The particular functional form $F^{-1}(r)$ for the impact of relative rank is imposed so that we can translate a two-dimensional

problem into a one-dimensional one which, in turn, allows us to find a closed-form solution for the equilibrium effort curve. In addition, by making this assumption, we capture the essential feature that a student's utility is increasing in her rank. A fraction $\varphi \in [0,1]$ of students has competitive preferences in that sense that their rank parameter is strictly positive. We denote their rank parameter by $\overline{\rho} > 0$ and assume that it is constant for the entire subpopulation. The remaining fraction $1-\varphi$ of students does not have competitive preferences, i.e., we assume that their rank parameter equals zero. The probability that a student's rank parameter equals $\overline{\rho}$ is independent of her ability.

Proposition 2 *Let $\beta \equiv \alpha(1+\rho)$. Let H denote the cumulative distribution function of β. Under relative grading, the following effort function constitutes a Bayesian Nash equilibrium:*

$$e^R(\alpha,\rho) = \sqrt{2\int_0^{\beta} x dF^{-1}(H(x))}$$

Proof. The proof follows standard techniques to derive Bayesian Nash equilibria. Assume, for the moment, that the equilibrium effort can we written as $e^R(\alpha,\rho) = e(\beta)$, where e is a strictly increasing function with $e(0) = 0$. As a consequence, a student type β's equilibrium rank is $r = H(\beta)$, where the distribution H of β is given by

$$H(x) = P\{\beta \leq x\} = \begin{cases} \varphi P\left\{\alpha \leq \frac{x}{1+\overline{\rho}}\right\} + (1-\varphi) P\{\alpha \leq x\} = \varphi F\left(\frac{x}{1+\overline{\rho}}\right) + (1-\varphi) F(x) & \text{if } x \leq \overline{\sigma} \\ \varphi P\left\{\alpha \leq \frac{x}{1+\overline{\rho}}\right\} + (1-\varphi) = \varphi F\left(\frac{x}{1+\overline{\rho}}\right) + 1 - \varphi & \text{otherwise} \end{cases}. \tag{2.3}$$

Observe that

$$\hat{U}(e, g = G(r), r, \alpha, \rho) = G(r) + \rho F^{-1}(r) - \frac{e^2}{2\alpha} = (1+\rho) F^{-1}(r) - \frac{e^2}{2\alpha} = (1+\rho) F^{-1}(H(\beta)) - \frac{e^2}{2\alpha}.$$

Suppose that a student misrepresents her type β as $\hat{\beta}$. If all other students bid according to

equilibrium, her expected utility is given by

$$u(\beta, \hat{\beta}) = (1+\rho)\, F^{-1}(H(\hat{\beta})) - \frac{e(\hat{\beta})^2}{2\alpha}.$$

The equilibrium FOC is given by

$$\left.\frac{\partial u(\beta, \hat{\beta})}{\partial \hat{\beta}}\right|_{\hat{\beta}=\beta} = (1+\rho)\, \frac{dF^{-1}(H(\beta))}{d\beta} - \frac{e(\beta)e'(\beta)}{\alpha} = 0$$

at all points where H is differentiable, which is equivalent to

$$e(\beta)e'(\beta) = \beta \frac{dF^{-1}(H(\beta))}{d\beta}.$$

Imposing the boundary condition $e(0) = 0$, this differential equation is uniquely solved by

$$e(\beta) = \sqrt{2\textstyle\int_0^\beta x dF^{-1}(H(x))}. \tag{2.4}$$

■

If none of the students has competitive preferences, i.e., if $\varphi = 0$, it immediately follows that $H = F$, so that

$$e^R(\alpha, \rho) = \sqrt{2\textstyle\int_0^\beta x dF^{-1}(H(x))} = \sqrt{2\textstyle\int_0^\alpha x dx} = \alpha = e^A(\alpha).$$

In line with proposition 1, all students will exert the same effort under relative grading as under absolute grading. For the other extreme case in which the entire student population has competitive preferences ($\varphi = 1$), $H(\beta) = F\left(\frac{\beta}{1+\overline{\rho}}\right)$. As a consequence, $F^{-1}(H(\beta)) = \frac{\beta}{1+\overline{\rho}}$. The equilibrium bidding curve is given by

$$e^R(\alpha, \rho) = \sqrt{2\textstyle\int_0^{\alpha(1+\overline{\rho})} x dF^{-1}(H(x))} = \alpha\sqrt{1+\overline{\rho}} > \alpha = e^A(\alpha).$$

In this case, relative grading induces all students to exert more effort than absolute grading. We obtain the following results for the intermediate case where $0 < \varphi < 1$.

Proposition 3 *If* $0 < \varphi < 1$, $e^R(\alpha, \overline{\rho}) > e^R(\alpha, 0)$.

Proof. The result follows immediately from $e^R(\alpha, \rho)$ being strictly increasing in $\beta = \alpha(1 + \rho)$. ■

An interpretation of this proposition is that a student from the subpopulation having competitive preferences will exert more effort than the students from the subpopulation without such preferences.

The following proposition shows that under some smoothness condition on F, the subpopulation without competitive preferences will exert less effort under relative grading than under absolute grading.

Proposition 4 *If* $0 < \varphi < 1$ *and* $\frac{dF^{-1}(H(\beta))}{d\beta} < 1$ *for all* $\beta \in (0, \overline{\sigma}(1 + \rho)]$, $e^R(\alpha, 0) < e^A(\alpha)$.

Proof. The result straightforwardly follows from

$$e^R(\alpha, 0) = \sqrt{2\int_0^{\beta} x dF^{-1}(H(x))} = \sqrt{2\int_0^{\alpha} x dF^{-1}(H(x))} < \sqrt{2\int_0^{\alpha} x dx} = \alpha = e^A(\alpha),$$

where the condition $\frac{dF^{-1}(H(\beta))}{d\beta} < 1$ implies the inequality in the above chain. ■

The intuition behind this proposition is obtained by considering the distribution of modified types $\beta = \alpha(1 + \rho)$. By introducing competitive preferences, the type distribution of β's is obtained by 'stretching' the type distribution of α's. The condition $\frac{dF^{-1}(H(\beta))}{d\beta} < 1$ guarantees that this is done 'smoothly' in the sense that a type $\rho = 0$ faces fewer β-types in their marginal neighborhood compared to a setting where competitive preferences were absent. In the latter case, the student would expend the same effort as under absolute grading according to proposition 1. Because all types can 'relax' in the case of competitive preferences relative to their neighboring types, in equilibrium, the entire subpopulation of $\rho = 0$ types will exert less effort than under absolute grading.

The opposite result pertain to the subpopulation with competitive preferences as the next proposition shows.

Proposition 5 *If* $0 < \varphi < 1$ *and* $\frac{dF^{-1}(H(\beta))}{d\beta} > \frac{1}{(1+\overline{\rho})^2}$ *for all* $\beta \in (0, \overline{\sigma}(1+\rho)]$, $e^R(\alpha, \overline{\rho}) > e^A(\alpha)$.

Proof. The result straightforwardly follows from

$$e^R(\alpha, \overline{\rho}) = \sqrt{2\int_0^{\beta} x dF^{-1}(H(x))} = \sqrt{2\int_0^{\alpha(1+\overline{\rho})} x dF^{-1}(H(x))} > \sqrt{2\int_0^{\alpha(1+\overline{\rho})} x dx / (1+\overline{\rho})^2} = \alpha = e^A(\alpha),$$

where the condition $\frac{dF^{-1}(H(\beta))}{d\beta} > \frac{1}{(1+\overline{\rho})^2}$ implies the inequality in the above chain. ■

Intuitively, type $\rho = \overline{\rho}$ faces two opposing forces: On the one hand, she has an incentive to exert more effort than under relative grading because her modified type β is greater than her original type α. On the other hand, she has a reason to put in less effort as all fellow students exert less effort under relative grading than under absolute grading if their original type α were equal to their modified type β (see Proposition 4). The smoothness condition $\frac{dF^{-1}(H(\beta))}{d\beta} > \frac{1}{(1+\overline{\rho})^2}$ guarantees that the first force is stronger than the second for all types.

2.B Excerpt from the Course Manual on Grading Schemes

The lecturers of the University of Amsterdam are constantly striving to improve their teaching and evaluation practices. As part of this initiative, during the EMO course we will test two different grading schemes that are recognized by the university: all students will experience both an absolute and a relative grading scheme. These grading schemes determine how exam scores are translated into grades.

Absolute grading

Under an absolute scheme, students' grades depend solely on their individual absolute performance in the exams. Specifically, the exam grade is calculated as follows:

Grade exam = 10 0.4*(number of errors)

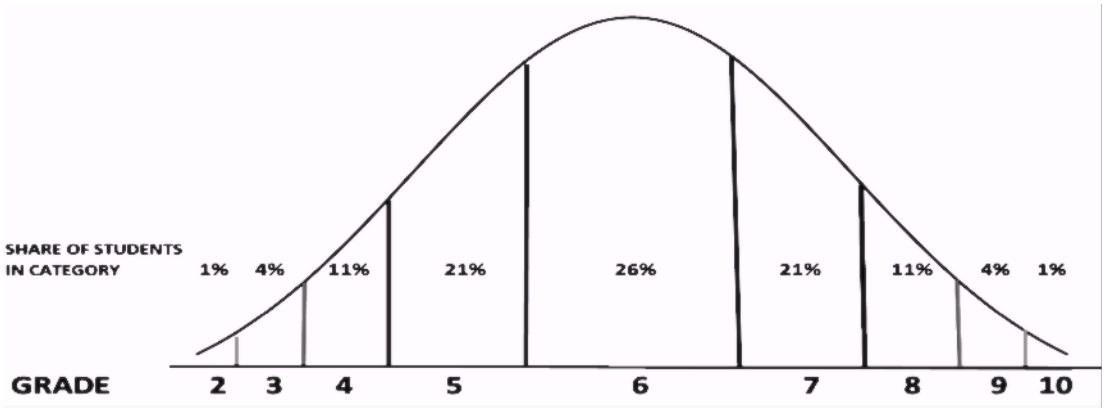
We round the grade to the nearest integer and we do not assign a grade below 2. This implies that exam scores translate into exam grades according to the table below:

Exam score *(=points earned)*	**Grade**
29 - 30	10
27 - 28	9
24 - 26	8
22 - 23	7
19 - 21	6
17 - 18	5
14 - 16	4
12 - 13	3
0 - 11	2

Relative grading

Under a relative grading scheme, or grading on a curve, students' grades depend on how well they perform in the exams compared to other students taking this course. It is not the individual score, but the students' position in the class score distribution (i.e., the students' rank among all students taking the exam) that determines the exam grade. For this course the curve is fixed so that the average score translates into an exam grade of 6, and the highest performing 1% of

students receive a grade 10 while the lowest performing 1% get a grade 2. We illustrate this scheme by the figure and the table below:



Relative rank *(calculated from the top)*	**Grade**
1%	10
2 - 5%	9
6 - 16%	8
17 - 37%	7
38 - 63%	6
64 - 84%	5
85 - 95%	4
95 - 99%	3
99 - 100%	2

Comparison of the schemes

In order to compare the two grading schemes, we will randomly divide all students into two grading groups: the blue group and the yellow group. Students in the two groups will take exams of the same difficulty level but will face different grading schemes:

BLUE group: midterm exam graded under absolute, final exam graded under relative scheme
YELLOW group: midterm exam graded under relative, final exam graded under absolute scheme

This way fairness is ensured: all students will experience both grading schemes, only the timing is different (remember: the midterm and final exams have equal weights and cover the same

amount of study material). The grades of students under the relative schemes are always determined compared to other exam takers in their grading group, not the whole class.

Before the start of the course, we will notify you of your grading group via e-mail and a Blackboard message. Please make sure you know which grading group you belong to, as it is important not only for your exam but also for the composition of homework groups.

2.C Screenshots from the survey

Figure C1: Example of a multiple-choice test question

Question 2.
What is the derivative of the function f(x) = (x - 5) / 2x ?

- f'(x) = 5 log(x) / 2
- f'(x) = 0.5 x
- f'(x) = 2.5 / x^2
- f'(x) =(2x - 5) / $4x^2$

Figure C2: Eliciting risk preferences

Your payment
One of the 10 decisions will be randomly selected for payment, and the outcome (high or low payoff) will be determined according to the probabilities stated in that decision. The payoff from this decision will be calculated according to the gamble you selected and will be added to your survey account.

	Option A*		Option B*	
	€40	€32	€77	€2
Decision 1	10%	90%	10%	90%
Decision 2	20%	80%	20%	80%
Decision 3	30%	70%	30%	70%
Decision 4	40%	60%	40%	60%
Decision 5	50%	50%	50%	50%
Decision 6	60%	40%	60%	40%
Decision 7	70%	30%	70%	30%
Decision 8	80%	20%	80%	20%
Decision 9	90%	10%	90%	10%
Decision 10	100%	0%	100%	0%

I always prefer Option B
From Decision 2 onwards I prefer Option B
From Decision 3 onwards I prefer Option B
From Decision 4 onwards I prefer Option B
From Decision 5 onwards I prefer Option B
From Decision 6 onwards I prefer Option B
From Decision 7 onwards I prefer Option B
From Decision 8 onwards I prefer Option B
From Decision 9 onwards I prefer Option B
In Decision 10 I start to prefer Option B
I always prefer Option A

ability of receiving €40 and 90% probability of receiving €32.

h decision did you first start to prefer Option B? This implies that
A and *from this decision onwards*, you prefer Option B.

2.D Analyzing the subsample of international students

About one third of the participants in our experiment were studying in the international program. These students have on average higher ability than their peers in the Dutch-language program. While there are no entry requirements for the Dutch program (all applicants who complete the pre-university track in secondary education and pass the standardized national school-leaving exam are automatically admitted to the study), students have to qualify for the international program by taking an English proficiency test and a mathematics entrance test. Only one in four applicants is admitted to the English-language Bachelor program. Students in the international program also tend to be more motivated. The English-language program is composed predominantly of foreign students (typically from Central-Eastern Europe, China, and Germany), but the program is also open to aspiring Dutch students. For foreign students in the international program, tuition fees and living expenses in Amsterdam often represent a comparatively much larger investment in education than for their Dutch peers, likely increasing the importance they attach to performing well in their studies. Dutch students choosing to comply with the selective entry criteria for the international program and to follow courses in English instead of their mother tongue also signal dedication and higher levels of aspiration. International students in our sample have significantly higher math grades, solve more test questions correctly and have fewer retake exams than those in the Dutch language program. They are also significantly more likely to hand in homework assignments, they receive higher homework grades and report spending more time preparing for the course. Moreover, even after controlling for past mathematics grades or performance on the test questions, students in the international program have significantly higher grade expectations than those in the Dutch program. We attribute this difference to international students being more ambitious rather than more overconfident, especially because students in the two programs did not differ in their overconfidence measured in the incentivized survey.

We test whether male students in this skilled and motivated group are responsive to competitive grade incentives. Results from the midterm exam, depicted with black bars in Figure D1

suggest that male students in the international program indeed perform better when facing competitive incentives, while female performance is unaffected. Columns (1) and (2) of Table D1 confirm that male students in the international program respond significantly more positively to relative grading than females on the midterm exam, and the effect is fairly large (a difference of approx. $2 - 2.5$ points). The picture becomes less clear when we also consider the end-of-term scores (gray bars in Figure D1): male students in the "Blue" group do not catch up with girls on the competitively graded second exam while boys in the "Yellow" group continue to do as well as girls when graded under the absolute scheme. Regression analysis confirms these findings: columns (3) and (4) in Table D1 show that controlling for their midterm scores, boys in the "Blue" group score significantly lower than girls on the end-of-term exam even though it is graded on the curve, while there is no significant gender gap among students in the "Yellow" group in terms of the change in performance between the two exams.

The above finding is consistent with male students in the "Blue" group becoming demotivated by their relatively low midterm performance and, as a consequence, providing less effort for the end-of-term exam. Patterns in preparation behavior provide only weak support for this explanation.[28] On the other hand, we can not rule out that the difference we observed in the midterm exam performance is not a result of the treatment but is rather driven by pre-existing differences between the groups due to imperfect randomization. Even though the two groups seem balanced with respect to observables, we should note that the most important ability proxy, the average math grade is missing for 41 students in the international program.

[28] Among international students, the drop in lecture and tutorial attendance after the midterm was significantly larger for boys than for girls in the "Blue" group, while the decrease was the same for both genders in the "Yellow" group. Focusing on homework grades, we find no evidence for lower effort provision after the midterm by male international students in the "Blue" group. However, these assignments were prepared in mixed-gender teams, so it is not clear to what extent demotivation of male students could lead to lower team performance. We can not analyze the change in self-reported study time as these measures were only collected once, after the midterm exam.

Table D1: ANALYZING THE SUBSAMPLE OF INTERNATIONAL STUDENTS.

	FULL SAMPLE		BLUE group	YELLOW group
	midterm score		*end-term score*	
	(1)	(2)	(3)	(4)
relative	-0.333	-0.611		
	(0.965)	(0.814)		
relative*male	2.576*	2.069*		
	(1.346)	(1.148)		
male	-1.830*	-0.964	-2.101**	0.721
	(0.939)	(0.823)	(1.035)	(1.014)
midterm score			0.208	0.377**
			(0.153)	(0.180)
Demographic controls		✓	✓	✓
Ability controls		✓	✓	✓
Constant	20.552***	13.065***	13.527	3.262
	(0.699)	(4.820)	(8.223)	(9.425)
N	126	126	63	60
R^2	0.053	0.395	0.425	0.333

Notes: The table displays estimated coefficients from OLS regressions. Covariates in columns (2)-(4): age, Dutch born, Math grades, num. retakes, test questions. In all specifications, indicator variables for missing covariates included. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

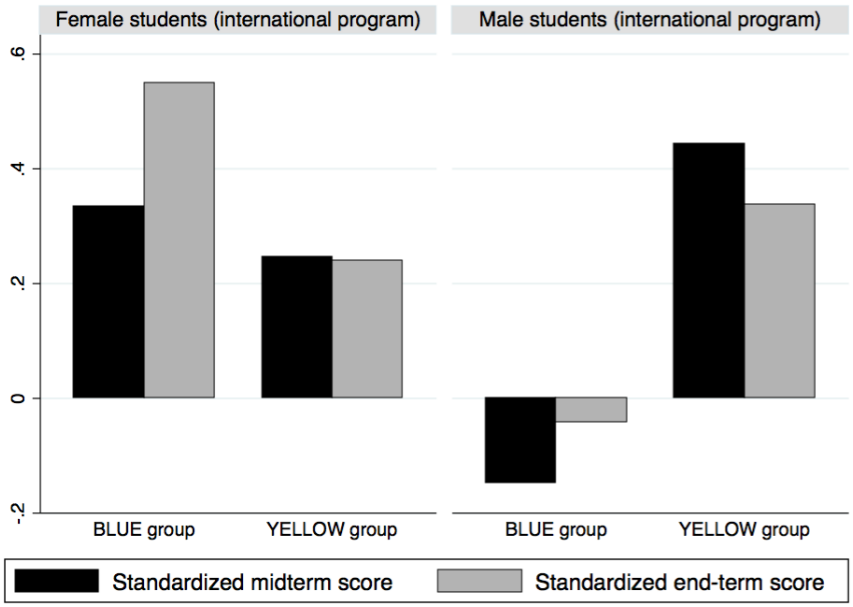


Figure D1: International program: mean mid- and end-term (standardized) exam scores by treatment group and gender

Chapter 3

The consequences of shying away

3.1 Introduction

Women are missing from the top of organizational hierarchies around the world. As of January 2015, a mere 4.6% of CEO positions at S&P 500 companies are held by women (Catalyst Org., 2015) and 21.9% of national parliamentarians are female (UN Women, 2015). Interestingly, even women who self-select into prestigious and competitive education follow different career trajectories than men: observing high-potential graduates from elite MBA programs, Carter and Silva (2010) find that women lag behind men in advancement and compensation starting from their first jobs. According to the European Commission (2012), the pattern is no different in academia: while there was a complete gender balance at the PhD level, only one fifth of full professors were women in 2010, and the progress towards gender equality has been slow.

A popular explanation for the absence of women in leadership positions concerns gender differences in attitudes towards risk and competition. Survey evidence from companies suggests that female managers' careers are stalled because they enter promotion contests less often

This chapter is based on a joint work with Jörg Claussen and Mirjam van Praag. We are grateful for *sauspiel.de* for making the dataset analyzed in this study available for the purpose of our research.

than men do (Barsh and Yee, 2012; Institute for Leadership and Management, 2011; Sandberg, 2013). Lloyds TSB Bank found, for instance, that although female employees are more likely than men to meet or exceed performance expectations, they tend not to apply for promotion (Desvaux et al., 2008). Such anecdotal evidence is often quoted to support the claim that women's less positive attitudes towards risk and competition not only keep them from entering uncertain and competitive environments but continue to hinder their chances of advancement even after they have sorted into such settings. Observational data from (competitive) companies rarely allow assessing the causal impact of "shying away" on the career of women who have self-selected into such firms.[1] This is partly due to a lack of objective performance measures but also due to confounding factors such as potential discrimination against women (Altonji and Blank, 1999) or gender differences in actual experience and labor force interruptions (Kim and Polachek, 1994).

We therefore use a very large set of naturally occurring data from a simple and abstract setting. We analyze the behavior of registered users in an online card game community. The game has some appealing features resembling the environment of interest in various aspects: there is real, strategic and repeated interaction (with feedback) between male and female players who can encounter each other as opponents or as teammates over several rounds. Uniquely, the data allow us to make a distinction between what we call 'selection' and 'playing' behavior: we observe players' choices regarding the level of risk and competition they prefer to bear in each round, and also their subsequent playing performance (i.e. the tricks they win and the points they collect) in the resulting tournaments. Players' performance is measured exactly and depends on skill, strategic behavior and luck. Since there are no real monetary incentives involved, we conjecture that members of the card game community are highly intrinsically motivated. While we acknowledge possible concerns related to the external validity of the setting, our study allows an internally valid assessment of the impact of gender differences in

[1] A notable exception is the study of Card et al. (2015) who decompose the impact of firm-specific pay differentials on the gender wage gap into sorting and bargaining effects. Using data from Portugal, they show that women only receive about 90% of the firm-specific premiums men collect.

risky and competitive choices on achievement.

We demonstrate that even women who self-select into a competitive environment subsequently refrain from initiating contests they could win. Women do not only initiate fewer contests than men, they are also less likely to play "Solo" (instead of in a team) or increase the stakes of the game. Multiple rounds of repetition and feedback about past performance do not change this behavior. As a consequence, female players accumulate lower scores and appear less competent than their male competitors, despite no actual skill difference in the playing stage (that we also demonstrate). We find that the performance of female players is at least as high as males' once their different 'selection' choices have been taken into account. Finally, we argue that women's different choices in the selection stage likely result from gender differences in preferences.

The topic of gender differences in preferences and attitudes and their potential impact on labor market outcomes has received considerable attention among economic scholars over the past decade. Several studies have found that women indeed differ from men along multiple dimensions, such as attitudes towards risk and competition (see for overviews Croson and Gneezy (2009), Niederle and Vesterlund (2011), Bertrand (2011) and Azmat and Petrongolo (2014)). In particular, females are shown to be more risk averse (Charness and Gneezy, 2012; Eckel and Grossman, 2008), less likely to enter a tournament (Niederle and Vesterlund, 2007) and less responsive to competitive incentives (Gneezy et al., 2003) than males. While the existence of a gender gap in preferences has been robustly established, its measured magnitude and importance is context-sensitive. For instance, the gender gap in competitiveness is much larger when the tournament is between individuals rather than between teams (Dargnies, 2012; Healy and Pate, 2011).[2]

Evidence for a gender gap in risk attitudes and competitiveness comes predominantly from

[2]Further relevant factors affecting the observed gender gap in competitiveness include the gender composition of the group of opponents (Gneezy et al., 2003; Gneezy and Rustichini, 2004; Ivanova-Stenzel and Kübler, 2011) or of the competing teams (Delfgaauw et al., 2013) and the nature of the task at hand (Grosse et al., 2014; Wieland and Sarin, 2012). Considering risk aversion, Filippin and Crosetto (2014) show that the gender gap is sensitive to the method of elicitation. Adams and Funk (2012) find in a large, unincentivized survey that female directors are actually more risk loving than their male counterparts.

laboratory experiments, and the actual relevance of this gap for field outcomes has been scantly researched (Bertrand, 2011). A few recent studies have used laboratory measures of behavioral preferences to explain real life choices. Competitiveness is indeed shown to be a strong predictor for outcomes such as secondary school students' choices of academic tracks (Buser et al., 2014) or the decision to take a selective high school entrance exam (Zhang, 2013). Flory et al. (2014) demonstrate in a field experiment that the share of female job applicants decreases with the competitiveness of the offered compensation scheme. These studies directly link gender differences in competitive preferences to occupational segregation in real life.

The study discussed in this chapter complements the above-listed papers by analyzing in a natural environment the consequences of less risky/competitive choices for the outcomes of females, given their participation in a competitive game. Our results support previous findings that females tolerate less competition and risk than men. Our main contribution is that we show using naturally occurring data that even skilled women who self-select into a merit-based, discrimination-free but competitive environment will end up lagging behind men because they are less inclined to take risk and to initiate competition. While prior results have often been obtained in the lab or in smaller sized field experiments, ours has been obtained in more natural circumstances using a large and diverse sample of people (from various age groups and backgrounds). The realistic (longer term) game environment allowed us, for instance, to reveal the substantial consequences over time of seemingly innocent and small gender differences in behavior. This would not have been possible in a lab experiment. Our study highlights the importance of research on gender differences in preferences by presenting evidence that females' differential selection choices that impede their success are to a large extent driven by their attitudes towards competition and risk.

We argue that our results carry relevant implications for organizations that strive for a gender balance in their population of employees, including the top of the hierarchy. For instance, firms often use systems that rely, more or less explicitly, on employees initiating competitions, such as self-nomination for promotions or employee-initiated negotiation for pay raises. These firms

might have to mentor female employees to take a more pro-active or competitive approach. Alternatively, they might design mechanisms where self-selection and self-nomination are less important in the promotion process.

The rest of the chapter is structured as follows. In Section 3.2, we provide an overview of the context and our dataset. Section 3.3 contains our results, followed by a discussion in Section 3.4. Section 3.5 concludes.

3.2 Context and data

3.2.1 The game

We use data from an online community for playing *Schafkopf*: a popular, traditional Bavarian card game. The game is known in other regions and countries as well (though with minor variations), e.g. *Doppelkopf, Skat* or *Sheepshead*. *Schafkopf* is a zero-sum game, played by four participants at a (virtual) table, using the unique German/Bavarian deck of cards.[3] Each game consists of a selection stage in which players announce their willingness to initiate a game and have the option to raise stakes, and the actual playing stage during which all distributed cards are played out trick by trick. Points collected in the playing stage determine the winner(s) of the game. The way points map into earnings ('cents') depends on the chosen stakes. The aim of the game is to gain as many cents as possible since the sum of the cents collected constitutes a player's score.

The game begins with the dealer (rotating one position clockwise each round) distributing four cards to each player. The players then evaluate the strength of their first four cards and decide whether to double the stakes of the game by knocking on the table. Afterwards, the remaining four cards are distributed to each player. Players can then take the offensive role by actively initiating a game, or the passive role by playing as a partner/opponent in a game initiated

[3]The Bavarian deck has eight different values (in increasing rank: 7, 8, 9, 10, Jack (Unter), Queen (Ober), King (König) and Ace) in four different suits (in increasing rank: acorn, grass, heart, bell) each, with the Queens, Jacks and heart cards being trumps.

by someone else at the table. Players initiating a game have a choice between three game types: the standard two-against-two-players *Sauspiel* game and the more risky and competitive one-against-three *Wenz* and *Solo* game types.[4] Figure A1 in the Appendix shows a screenshot from the selection stage in the online game: each player is asked if they want to initiate a *Sauspiel*, a *Wenz* or a *Solo* game, or if they prefer not to initiate a game at all (i.e. to "pass"). Passing does not result in dropping out of the round: as long as at least one person initiates a game, all four players at the table will join as partners or opponents. The highest announced game type will be played (*Solo* being the highest game type, followed by *Wenz* and *Sauspiel*) and if multiple players want to initiate the same game type, the player closest to the dealer will be given priority. If nobody initiates a game, the cards are reshuffled and a new round begins.

The game is played in eight tricks and every player has to contribute one card to each trick. Before playing the first trick, players in the opponent role are allowed to give a "Contra", which doubles the stakes in the game. The player sitting to the left of the dealer starts the first trick and subsequent tricks are started by the winner of the last trick. The other players then all have to lay down a card in clockwise order.[5] The player who contributes the card with the highest rank wins the trick.[6]

To determine the winner of the game, each card is allocated a point value, with a total of 120 points for the whole card deck.[7] To win the game, the initiator (together with his/her partner in case of a *Sauspiel* game) needs to win tricks worth at least 61 points in total. Therefore it is possible to win five out of eight tricks but still lose the game if these tricks do not contain enough points. Each player of the losing party has to pay 10 cents to the winner(s) when losing

[4] *Sauspiel* thus involves a competition between teams, while *Solo* and *Wenz* require the individual to compete alone against all other players at the table. Consequently, the stakes are also higher in the latter two types. The difference between Solo and Wenz is that only Jacks count as trumps for a Wenz game while the Solo is more similar to the Sauspiel in that Queens, Jacks and one designated color count as trumps. The partner of a Sauspiel's initiator cannot be freely chosen but is randomly determined by the initiator calling a specific suit of ace. Team composition in Sauspiel games is thus only revealed when the specific ace is played out. Team composition for the more competitive Wenz or Solo game is immediately revealed as the initiator plays against the three other players.

[5] Players have to contribute a card of the same color as the first played card of the trick, but can play any other card if they don't possess a card of the same color. If the first played card of the trick is a trump card, the other players also have to play a trump card if they still possess one.

[6] If a trick contains multiple Queens or Jacks, their rank order is determined by suit.

[7] 11 for the Ace, 10 for the 10, 4 for the King, 3 for the Queen, 2 for the Jack, and 0 for all other cards.

a *Sauspiel* game and 50 cents when losing a *Wenz* or *Solo*. The amounts are higher when players win with a large margin.[8] Finally, as mentioned earlier, each knock on the table after the distribution of the first four cards and each Contra results in a doubling of payouts. The cents that players accumulate over the rounds constitute their score.

After each round, each player decides if they want to stay in for another game or leave the table. Distribution of new cards only starts when four players sit at the table.

3.2.2 The online platform

Our data was provided by *sauspiel.de*, the largest online *Schafkopf* gaming community. The platform was founded in 2007 by a group of four students with the goal of bringing this traditional card game to the online world. The founders implemented the online version of the game with exactly the same rules as for regular Schafkopf, lowering entry costs for experienced players of the game. The platform became quickly popular and has already hosted more than 500 million games as of 2015.

First-time players have to register a profile on the platform, which includes choosing a user name and customizing a male or female avatar.[9] Registration and the use of the platform is free. Virtual 'cents' collected in the online games have no value outside the platform. As *Schafkopf* is always played by a group of four players, each player joins a virtual table of four. Figure A2 in the Appendix shows the table selection stage: players can either set up a new table by clicking on the plus symbol and wait until three other players join, or join a table with less than four players. As the tables usually fill up within a few seconds and because at this stage no information on other players' past performance is displayed, there is little room for strategically selecting the set of players to share the table with.

Once four players have joined a table, the regular game begins: as described in the previous

[8]The above amounts are increased by 10 cents if the losing party obtained less than 30 points and by 20 cents if no tricks were won by the losing party. Furthermore, if the winning party had a sequence of at least the three highest trumps, the sum is increased by 10 cents per trump.

[9]Players can voluntarily also register their gender, date of birth and their ZIP code. In our sample, 22% of all players provided this information.

section, the players decide if they want to raise the stakes by knocking after the first four cards are distributed, then announce if they want to initiate a game after all eight cards have been dealt, and if at least one player announces a game, all eight tricks are played out consecutively. Players make their choice of initiating a game being informed of their own card quality, the decision of the other players at the table to raise stakes (which can serve as an indicator of their card quality), the past performance of other players (indicated by their cumulative scores displayed below their user names, as shown in Figure A1), and the gender composition at the table (shown by the avatars). The median game duration including the selection stage is only 81 seconds, so players usually stay at the same table for multiple game rounds. If one player leaves the table, the remaining players can stay at the same table and wait until a new participant joins them.

We observe all games played between September 5th 2007 and January 9th 2008. For each round during these five months, we observe the type of game played (*Sauspiel*, *Wenz* or *Solo* or no game if everyone "passed"), who initiated the game and who has been in a partner (for *Sauspiel*) or opponent role. We can also see who raised the stakes of the game by "knocking" and in which order players announced their game choice (i.e. players' position at the table). For each initiator we observe whether they were given "Contra" by their opponent(s), but the identity of the exact opponent(s) who gave the "Contra" is not recorded. Therefore the individual decision to give "Contra" can not be analyzed in our setting.

If a game is played, our dataset also contains the exact cards played out by each player, allowing us to assess whether good or bad cards were dealt to the players. We calculate a joint measure of card quality by regressing multiple indicators of card value[10] on the probability of winning a *Sauspiel* and then using the coefficients to derive the joint measure of card quality.[11] If all players decide to "pass" and cards are therefore not played out, the distribution of cards is not recorded.

[10]The indicators are the cumulative point value of all owned cards as well as dummy variables for the number of trump cards, the number of suits a player does not possess (as this allows to go in with a trump if another player plays this suit), the number of aces, and the consecutive number of highest trumps in a row.

[11]The results are stable to including the different measures of card quality separately instead of the joint card quality measures in the regressions performed later.

For each game, we observe the winner(s) and the number of points achieved (between 0 and 120) as well as the final score in cents for each participant. As we have seen in Figure A1, each player's cumulative score is publicly displayed on the screen during the games and thus influences the status in the community. The displayed score provides an imperfect indication of a player's skills and performance because these scores can be reset to zero whenever they are falling below zero. Our dataset does not contain these displayed scores. However, we do observe the actual scores and can use them to infer the displayed scores of all players.

3.2.3 Data

This section introduces our dataset and the most important variables in more detail. Table 3.1 provides an overview of our data. The dataset contains over 4 million games recorded from the perspective of each player at the table, resulting in 16,655,344 observations in total. Our data is generated by more than 15 thousand individual players. The share of female players[12] in the community is low, around 8.5%, reflecting Schafkopf being a traditionally "masculine" activity. Female players are, however, very active on the platform: since the number of games per player is much higher for women (mean: 1,745) than for men (mean: 1,033), they generate 13.52% of all observations (see also Figure A3 for the distribution of the number of games per player).

As mentioned earlier, our dataset does not contain much information about the characteristics or background of the players. For a subsample of 3,323 registered users we observe their age and the ZIP code of their residence. Based on this subsample we do not observe a substantial difference between men and women in terms of their age: the mean age is 30.19 for men and 29.80 for women. Figure A4 in the Appendix, depicting the age distribution of our players, shows that our sample is very diverse, with some players younger than 15 and others older than 70 years. Only a minority of the registered users come from Munich, and women are slightly more likely than men to live in the state capital.

[12] In our analysis we talk about female and male players when discussing users who registered a female or male avatar. We discuss in more detail to what extent this proxy provides an accurate representation of players' true gender in Section 3.2.4.

Table 3.1: OVERVIEW OF THE DATASET

Panel A					
	MALE	%	FEMALE	%	TOTAL
Number of observations	14,402,768	86.48	2,252,576	13.52	16,655,344
Number of players	13,941	91.52	1,291	8.48	15,232
Panel B					
		MALE		FEMALE	
		Mean	Median	Mean	Median
Number of games per player		1,033	238	1,745	447
Age		30.19	28	29.80	27
Location Munich		0.113	-	0.178	-

Table 3.2 tabulates the gender composition of tables for all rounds. In more than half of the cases there were only male players at the table, approx. 36% of games involved one female player, and less than 0.03% of the rounds were played by four women at the table (this latter share, however, still corresponds to 1,110 games).

Table 3.2: GENDER COMPOSITION OF GAMES

Women at table	N	%
0	2,308,689	55.45
1	1,491,212	35.81
2	333,553	7.96
3	31,273	0.75
4	1,110	0.03
Total	4,163,837	100

Even though we do not observe the employment or marital status of the players in our sample, we can still draw some cautious inferences regarding their lifestyles based on the hour of the day when they play games on the online platform. Figure A5 in the Appendix suggests that there is no systematic difference between men and women in the game community: both are most likely to play in the evening hour. Moreover, we find no indication for women playing more often during the typical working hours, suggesting that they are as likely as men to be employed.

3.2.4 Evaluation of our setting

Both the large dataset at hand and the particular features of the *Schafkopf* game are important ingredients for our contribution to the literature. A major advantage of our context is that we study naturally occurring data. The experimenter demand effect is absent and participants' choices are based on a good understanding of the rules of the game they find very natural. The large dataset, both in terms of the number of individual players and the number of observations per player, allows us to detect small effect sizes and heterogeneous effects. It also enables the convincing measurement of potential null-effects. This kind of precise measurement would be difficult to obtain in a typical laboratory experiment with a limited number of participants. Moreover, unlike most lab experiments that are based on student samples, our dataset covers a wider age range and most likely also different levels of education. People self-select in the sample and, due to the character of the game, this spontaneously results in a male-dominated group (which is a good approximation for the environment at the top in corporate hierarchies in most labor markets). The features that participants earn only "virtual" cents and that players can be identified by their user names (which enables reputation building) ensure that players are very highly intrinsically motivated and care about their ranks.

The *Schafkopf* game has certain appealing features for studying gender differences in competition and risk taking and their likely consequences for performance and success. In our setting, players face real interaction: the actions taken by a player's opponent and partner actually affect the player's chances of winning. Success in this set-up thus depends on strategic uncertainty (the behavior of other players at the table), non-strategic uncertainty (the distribution of card quality between players) and individual ability (game-specific skills). This feature is missing from related laboratory studies where tournament winners are typically selected based on their performance in an individually executed task.[13]

[13] In our setting, players do not make a choice between tournament and piece rate as common in lab experiments. In Schafkopf, the alternative of not initiating a game is either participating as partner or opponent in a game initiated by someone else, or playing no game at all. We find that this feature brings our context closer to real-life situations where a "safe option" or a non-competitive track is often unavailable. It also allows us to assess the performance of the less competitive players under tournament incentives.

Moreover, in our data the interaction between players is repeated hundreds, often thousands of times, and players receive feedback after each round. At the end of each game they learn whether or not they have won and by how much their scores increased/decreased as a consequence. As mentioned before, players' ranks in terms of displayed cumulative scores are common knowledge. Therefore, players do not only receive updates on their own absolute scores, but also on their position compared to other players. Feedback on one's performance both in absolute and relative terms is a characteristic feature of labor and other markets. It is also a relevant feature of the data when studying gender differences in competitiveness because there is evidence that feedback induces more competitive choices among high-ability women (Wozniak et al., 2014).

Despite all these appealing features of the collected data, we also have to deal with a number of limitations. First, little is known about the characteristics of our subjects (as discussed before, self-reported information on age and residence is available, but only for a sub-sample of the players). Our measures would be biased if female players had systematically different unobserved characteristics than males and these characteristics were correlated with the preferences we study.

Second, the gender measure we use (the gender of the avatar chosen by the players) is self-reported and could potentially be misreported. For a subsample of 3,323 individuals we also observe the gender they administer at registration and can compare this with the gender of their avatar. In 97 percent of the cases the two are identical.[14] Based on this we assume that most peoples' avatars are a truthful representation of their sex. Nevertheless it might be theoretically possible that people misrepresent their sex both at registration and when choosing their avatar. This would increase the overall percentage of misrepresented avatars somewhat. It might also be that people push the wrong button when registering, while their avatar (that can be updated all the time) is the correct representation of their gender. That would decrease the share of misrepresented avatars in the total sample. In any case, misrepresentation seems quite

[14] Approximately two-thirds of the "mismatches" result from players registered as female choosing a male avatar.

low. If it occurred at random, its only consequence would be less precise estimates. However, our estimates would be biased upwards if the most competitive female players systematically played with a male avatar and the least competitive male players played with a female avatar. Anecdotal evidence from *Schafkopf* players suggests that such gender misrepresentation is rare, but this claim is admittedly untestable given our data.

A third, more minor, disadvantage of our setting is that the game is too complex for us to derive the optimal strategy and behavior in each situation. Therefore we can not conclude whether women initiate games "too little" and men "too much". Instead, we can infer the *relative profitability* of their playing strategies by comparing the scores collected by men and women (controlling for their playing ability).

3.3 Results

In this section we test whether female players in our dataset behave differently than male players when it comes to risk taking and competition. In particular, we focus on two decision players make in the selection stage of the game: whether to increase the stakes of the game by "knocking", and whether to actively initiate a game (and which type). We then consider whether there are skill differences in the playing stage between men and women in our sample, where 'skill' refers to players' ability to win games and collect points *given their role and the game type*. We also compare the success of male and female playing strategies by analyzing the scores players accumulate over the games. Finally, we speculate whether women's differential selection choices may be attributed to gender differences in risk aversion and competitiveness.

Throughout the analysis, we use the following terminology: "focal player" refers to the individual whose perspective we consider in the analysis (remember that all games are recorded from the viewpoint of all four participants), and "opponents" denote the three other players sitting at the table. We decided to use the general term "opponent" because even though in the *Sauspiel* game one of the fellow players will actually be the initiator's partner, in the selection stage it is not yet known who this partner may be.

3.3.1 Descriptive evidence

Table 3.3 provides simple descriptive evidence for gender differences in playing behavior. It shows a large and significant gender gap in the propensity to raise the stakes: while female players only "knock" in 20.52% of the rounds they play, men do so in 28.11% of the cases. Women also play significantly less often in the initiator role than men: the difference is 3.34 percentage points (this amounts to a gender gap of approximately 22%, given that women play as initiators 15% of the rounds). Consequently, women are more likely than men to play as opponents or as partners or not to play any game in the given round.

Table 3.3 also shows winning probabilities for players in each role. Female players are somewhat more likely than males to win games in the initiator and partner role (the difference is 2.02 and 0.37 percentage points, respectively), and underperform slightly (by 0.4 percentage points) compared to men in the opponent role.

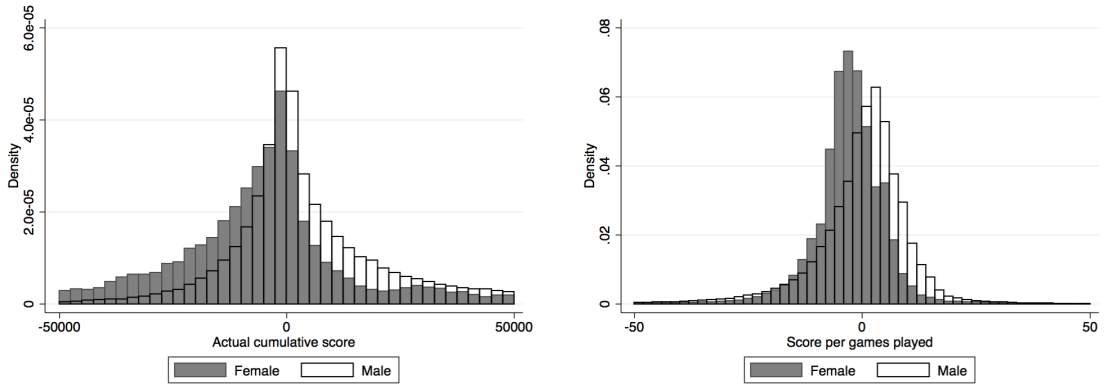
Table 3.3: DESCRIPTIVE EVIDENCE: Gender differences

	MALE	FEMALE	Difference
Raise stakes ("knock"), %	28.11	20.52	-7.59***
Role in game, %			
Initiator	18.37	15.03	-3.34***
Partner	12.14	12.61	0.47***
Opponent	41.44	42.22	0.78***
None	28.05	30.14	2.09***
Games won in role, %			
Initiator	76.39	78.41	2.02***
Partner	79.82	80.19	0.37***
Opponent	24.39	23.99	-0.40***
Score (mean)			
Cumulative	12,692	-7,504	-20,196***
Displayed	20,021	12,142	-7,879***

Notes: Significance of differences from t-tests with unequal variances; *** p<0.001, ** p<0.01, * p<0.05

Finally, Table 3.3 shows that despite playing more games and having similar winning probabilities, women on average accumulate substantially lower scores (both displayed and actual) than men in our sample: the mean score of male players is 12,692 while it is -7,504 for women. Figure 3.1a displays the distribution of actual cumulative scores by gender, while 3.1b depicts

the scores collected per games played. Both figures confirm that female players are on average less successful in the community than males.



(a) Cumulative scores

(b) Scores per games played

Figure 3.1: Distribution of cumulative scores, by gender

3.3.2 Behavior in the selection stage

We continue by analyzing in more detail the finding that men and women play differently in the selection stage. Table 3.4 reports marginal effects[15] calculated from probit models analyzing the propensity to initiate a game (columns (1)-(3)) and to raise the stakes by "knocking" (column (4)).[16] The table also contains the "baseline" initiation/knocking rates, i.e. the predicted probabilities for male players, evaluating all coefficients (other than gender) at their mean values. Column (1) displays results from a specification that only includes the focal player's position at the tables as a control variable. We find that there is indeed a significant and sizable gender gap in the willingness to initiate games: while male players on average have an 18.3% likelihood of starting a game, being female is associated with a 3.3 percentage point drop in initiation rates. Column (2) and (3) show that this gender gap is remarkably robust to the inclusion of covariates. Controlling for the presence of females among the opponents, the number of opponents who "knocked" in the selection stage, the focal player's relative rank in terms of displayed

[15] Estimated coefficients are displayed in Table A1 in the Appendix.

[16] To account for the fact that observations in our sample may be correlated in two non-nested dimensions (we observe multiple observations for the same individual, and four people play the game and interact at a single table), we cluster standard errors at the individual level, and add table level controls (see Cameron and Miller (2015)).

cumulative scores or the focal player's experience measured by the (natural logarithm of) the number of games he/she has played before, the estimated gender difference in the likelihood to initiate a game remains in the range of $2.9 - 3.4$ percentage points.

Table 3.4: LIKELIHOOD OF INITIATING A GAME AND RAISING STAKES

VARIABLES	Initiate a game			Increase stakes
	(1)	(2)	(3)	(4)
Female player	-0.033***	-0.034***	-0.029***	-0.068***
	(0.003)	(0.003)	(0.003)	(0.008)
Position at table = 2	-0.022***	-0.021***	-0.021***	-0.010***
	(0.000)	(0.000)	(0.000)	(0.001)
Position at table = 3	-0.031***	-0.030***	-0.030***	-0.010***
	(0.000)	(0.000)	(0.000)	(0.001)
Position at table = 4	-0.035***	-0.035***	-0.035***	-0.008***
	(0.001)	(0.001)	(0.001)	(0.000)
Female opponent		-0.003***	-0.004***	-0.006***
		(0.000)	(0.000)	(0.001)
Num. opponents knocked = 1		-0.110***	-0.109***	
		(0.000)	(0.000)	
Num. opponents knocked = 2		-0.170***	-0.169***	
		(0.001)	(0.001)	
Num. opponents knocked = 3		-0.202***	-0.201***	
		(0.001)	(0.001)	
Num. games played (log)			-0.005***	0.003
			(0.001)	(0.001)
Rank in score = 2			-0.013***	-0.029***
			(0.001)	(0.003)
Rank in score = 3			-0.020***	-0.052***
			(0.002)	(0.004)
Rank in score = 4			-0.024***	-0.081***
			(0.002)	(0.005)
Predicted prob. male players	0.183***	0.176***	0.175***	0.279***
	(0.001)	(0.001)	(0.001)	(0.002)
Observations	16,655,344	16,655,344	16,655,344	16,655,344
Number of players	15232	15232	15232	15232
Pseudo R^2	0.00230	0.0312	0.0317	0.00748

Notes: The table displays marginal effects at the means from probit models (for factor variables, calculated as the discrete change from the base level). Dependent variable in columns (1) - (3): initiate any game type; column (4): increase the stakes of the game by "knocking". As a baseline, we present predicted probabilities for male players (evaluating all other variables at their means). Omitted categories for covariates: position at table = 1; female opponent = 0; rank in score = 1; num. opponents knocked = 0. Standard errors are clustered on player ID and are reported in parentheses. *** p<0.001, ** p<0.01, * p<0.05

As we see in column (4), the gender gap in knocking behavior is even larger: all else equal, female players are 6.8 percentage points less likely to increase the stakes, compared to a "baseline" predicted probability of 27.9% for men.

As noted before, cards dealt to players in rounds where no-one initiated a game are not recorded in our dataset. It is therefore not possible to include card quality as a control variable in the above regression explaining initiation decisions. We can, however, exploit the very large number of observations in our dataset and compare the theoretical card quality distribution that occurs when cards are dealt at random with the actual card quality distribution observed for initiated games. This allows us to make inferences about the way card quality influences the entry decision on average in our sample. Figure 3.2 shows that the gender gap in game initiation rates persists throughout the card quality distribution, including the best cards as well.[17]

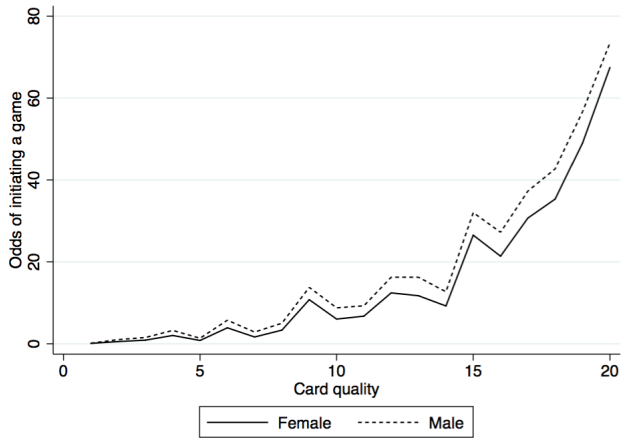


Figure 3.2: Odds of initiating a game by card quality, separately by gender

We consider how the gender gap in initiation rates differs by game type. As discussed in Section 3.2.1, players can choose between three game types: *Sauspiel* involves a team competition, whereas in *Wenz* and *Solo* games the initiator competes alone against all three other players at the table. In the latter two game types the stakes are also higher. We conjecture that the gender gap in initiation rates should be higher in the more competitive and risky *Wenz* and

[17]The apparent non-monotonicity of initiation rates by card quality results from the way we created the proxy for card quality. The card quality measure is calculated from several proxies for card quality, but is likely still missing some information that is related to the choice of initiating a game. So while the odds of initiating a game generally increase with card quality, uncaptured differences in card quality that are correlated with the captured measures can lead to the observed non-monotonic relationship.

Solo games than in *Sauspiel*. To test our hypothesis, we estimate a multinomial logit model explaining the choice to initiate one of the three game types (using the category *initiate no game* as our baseline). Table 3.5 displays marginal effects estimated from this model. We find that in case of *Sauspiel*, male players' predicted probability of initiating this game type is 12.4%, and the gender gap is 1.4 percentage points. In case of *Solo* games, the gender gap is 0.9 percentage points - a large difference given that men on average initiate this game type only 3.6% of the rounds. Similarly, a gender gap of 0.6 percentage points in *Wenz* initiation rates amounts to a sizable disparity compared to men's "baseline" initiation rate of 2.3%.

Table 3.5: GENDER GAP IN INITIATION RATES, BY GAME TYPES

	Sauspiel	Wenz	Solo
VARIABLES	(1)	(2)	(3)
Female player	-0.014***	-0.006***	-0.009***
	(0.002)	(0.001)	(0.001)
Predicted prob. male players	0.124***	0.023***	0.036***
	(0.001)	(0.000)	(0.000)
Covariates included		✓	
Observations		16,655,344	
Number of players		15232	
Pseudo R^2		0.00220	

Notes: The table displays marginal effects from a multinomial logit model. The dependent variable is *game type initiated* with four categories: Sauspiel; Wenz; Solo; none (baseline). Covariates included (unreported): position at table, female opponent, experience (log. number of games played), rank in score, num. opponents knocked. Omitted categories for covariates: position at table = 1; female opponent = 0; rank in score = 1; num. opponents knocked = 0. Standard errors are clustered on player ID and are reported in parentheses. *** p<0.001, ** p<0.01, * p<0.05

We also test whether the overall gender gap in initiation rates results from female players responding differently to the characteristics or behavior of their opponents. Table A2 in the Appendix, showing estimated coefficients[18] from probit models with gender interaction terms, suggests that this is not the case. We find that neither men nor women respond strongly to the presence of women at the table (having at least one female among the three opponents

[18]We report estimated coefficients - instead of marginal effects - because obtaining and interpreting marginal effects is less straightforward in case of models including interaction terms of factor variables. The coefficients in this table may be compared to the estimates reported in column (3) of Table A1.

increases the initiation rates very slightly, and the gender interaction is insignificant).[19] Players are influenced by their relative rank at the table: both men and women are less likely to initiate games when they are ranked worse than their opponents, and women are especially deterred from initiating when they have the lowest accumulated score at the table. On the other hand, there is no gender difference in the response to opponents' knocking behavior: male and female players are equally discouraged from initiating a game when their opponents raise the stakes.

Finally, Figure 3.3 demonstrates that experience affects the behavior of men and women differently. Initiation rates decrease for women with the number of games they have played, while the relationship is U-shaped for men. As a result, the gender gap is largest among the most experienced players: practice and feedback, instead of closing the gap, actually widens it. This divergent pattern by experience could tentatively be explained by gender differences in response to losing a competition (Buser, forthcoming): if losses induce men but not women to subsequently take more challenges, the gender gap in game initiation rates should widen over time.[20]

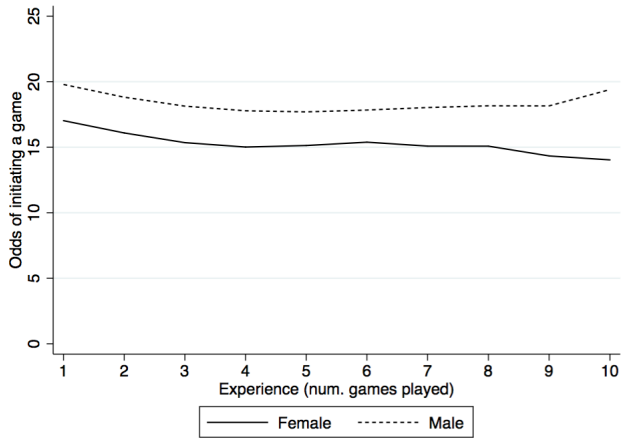


Figure 3.3: Odds of initiating a game by experience, separately by gender

[19] We also test whether we can replicate the finding that women are more competitive in a single-sex environment (Gneezy et al., 2003) by studying the initiation rates of women at female-only tables. We find no evidence in our data for female players being more likely to initiate a game in the absence of men.

[20] To correct for the fact that a large share of our observations comes from a minority of players who play thousands of games in the period we observe (thus the most active and experienced players have a disproportionately large influence on our results), we estimate a model explaining initiation behavior using weights that are inversely proportional to the total number of games played by each individual. The gender gap remains significant and similar in size even after this correction.

3.3.3 Performance in the playing stage

Having considered players' behavior in the selection stage, we continue by analyzing their performance in the playing stage. We study whether the large score difference between men and women in our sample can be attributed to differences in playing skills.

Table 3.6: LIKELIHOOD OF WINNING SELF-INITIATED GAMES

VARIABLES	No controls (1)	With controls (2)	Card quality (3)
Female player	0.020***	0.014***	0.001
	(0.003)	(0.003)	(0.003)
Position at table = 2		-0.008***	-0.021***
		(0.001)	(0.001)
Position at table = 3		0.014***	-0.003***
		(0.001)	(0.001)
Position at table = 4		0.034***	0.015***
		(0.001)	(0.001)
Num. games played (log)		0.015***	0.015***
		(0.001)	(0.000)
Female opponent		0.002***	-0.003***
		(0.001)	(0.001)
Card quality			0.151***
			(0.003)
Opponents' card quality			-1.073***
			(0.005)
Partner's card quality			0.250***
			(0.003)
Contra			-0.239***
			(0.001)
Predicted prob. male players	0.764***	0.766***	0.799***
	(0.001)	(0.001)	(0.001)
Observations	2,984,123	2,984,123	2,984,123
Number of players	14302	14302	14302
Pseudo R^2	0.000214	0.00473	0.136

Notes: The table displays marginal effects at the means from probit models (for factor variables, calculated as the discrete change from the base level). Dependent variable: odds of winning self-initiated games. As a baseline, we present predicted probabilities for male players (evaluating all other variables at their means). Omitted categories for covariates: position at table = 1. For the variable partner's card quality, missing values (in case of Solo and Wenz games) are imputed with zeros, and an additional indicator variable for missing values is included. Standard errors are clustered on player ID and are reported in parentheses. *** p<0.001, ** p<0.01, * p<0.05

We start by focusing on winning probabilities of players in games they initiated. The first column of Table 3.6 repeats the result from Section 3.3.1 in a regression framework: women are slightly but significantly more likely to win games they started themselves. While the predicted probability of winning for male players is 76.4%, being female is associated with a 2.0 percentage points increase in winning odds. When we include covariates such as position at the table, the presence of a female opponent or the number of rounds played on the platform before, the gender difference is reduced to 1.4 percentage points (column (2)). Controlling for the initiator's and opponents' quality of cards we find that male and female players are equally likely to win: the estimated coefficient for female player is reduced to zero and loses significance. The combined results of this table suggest that women have on average better cards than men when they initiate games. More importantly, column (3) shows that given card quality, there is no gender difference in the ability to win self-initiated games.[21] This finding is illustrated in Figure 3.4.

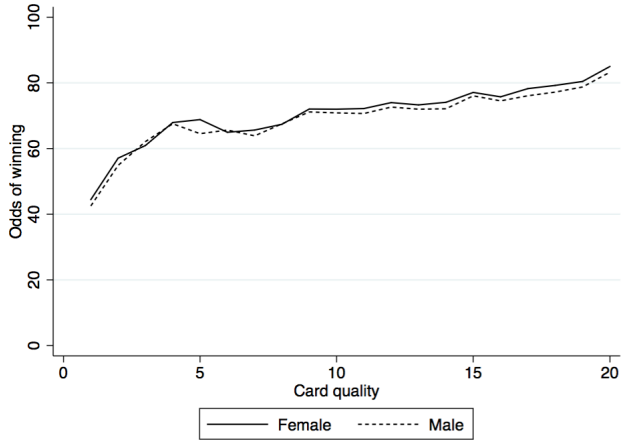


Figure 3.4: Odds of winning a game (conditional on initiating it) by card quality, separately by gender

Besides winning a game, we can also analyze points collected in a given round. Column (1) of Table 3.7 confirms that women are as successful players as men according to this continuous measure of performance: female players collect as many points per initiated game as men.

[21]This result is unchanged when we correct for the fact that players in our dataset differ in the total number of rounds they played. Estimating a probit model with weights that are inversely proportional to the total number of games played, the difference between the winning odds of male and female players is small and only marginally significant.

Table 3.7: DIFFERENT PERFORMANCE MEASURES

VARIABLES	*Full sample* Points in initiator role (1)	*Players in 4th position* Winning Solo in opponent role (2)
Female player	0.176	0.001
	(0.158)	(0.002)
Position at table = 2	-1.516***	
	(0.035)	
Position at table = 3	-0.440***	
	(0.035)	
Position at table = 4	0.587***	
	(0.035)	
Num. games played (log)	0.927***	0.001
	(0.033)	(0.001)
Female opponent = 1	-0.297***	-0.004**
	(0.032)	(0.002)
Card quality	19.901***	0.074***
	(0.168)	(0.004)
Opponents' card qual.	-58.542***	-0.636***
	(0.262)	(0.005)
Partner's card quality	27.250***	
	(0.148)	
Contra	-17.080***	0.350***
	(0.081)	(0.003)
Constant	117.183***	
	(0.296)	
Predicted prob. male players		0.290***
		(0.001)
Observations	2,984,123	427,093
Number of players	14302	12086
Adjusted/Pseudo R^2	0.258	0.0863

Notes: The table displays estimated coefficients from an OLS (column (1)) and marginal effects at the means from a probit model (column (2), for factor variables, calculated as the discrete change from the base level). Dependent variable: column (1): points collected in a self-initiated game; column (2): odds of winning a Solo game as opponent. Column (2) reports results from a model estimated on the subsample of players in the fourth position at the table. As a baseline, we present predicted probabilities for male players (evaluating all other variables at their means). For the variable partner's card quality, missing values (in case of Solo and Wenz games) are imputed with zeros, and an additional indicator variable for missing values is included. Omitted categories for covariates: position at table = 1. Standard errors are clustered on player ID and are reported in parentheses. *** p<0.001, ** p<0.01, * p<0.05

Studying self-initiated games we have found that female players perform as well as men. This result, however, suffers from potential endogeneity problems. Players self-select into the initiator role and this selection is unlikely to be random: the unobservable characteristics that influence initiating behavior could also affect winning probabilities, and could be correlated with gender. We therefore also consider performance in a situation where the scope for self-selection is virtually absent: players in the fourth position at the table playing as opponents in *Solo* games. To see why the opponent role is exogenously assigned in these roles, consider the following. Players in the fourth position are the last to announce their intent to initiate games. Therefore the only way for them to prevent being an opponent in a game started by someone before them at the table is to call a higher game type than the one already initiated. Since *Solo* is the highest type, there is no way for the last player at the table to avoid the opponent role in this game type. Column (2) of Table 3.7 analyzes the winning odds of opponents in the fourth position of *Solo* games and confirms that there is no significant gender difference in performance in these exogenously assigned roles, either.[22] We can thus conclude that given their cards and their role in the game, male and female players in our sample perform equally well: there is no gender gap in "on-task" performance.

3.3.4 Explaining the gender difference in scores

In the previous two subsections we have established that female players indeed "shy away": they make less competitive and risky choices in the selection stage. Furthermore, we have shown that there is no gender difference in winning probabilities in the playing stage once we control for card quality and roles. This latter result is puzzling given the difference in cumulative scores between male and female players (see Figure 3.1). In the following we explain two

[22] Another way to address the endogeneity of winning probabilities in self-initiated games would be to estimate a Heckman sample selection model. However, we did not find a variable in our dataset that would influence game initiation but not performance in the game. In the absence of such an exclusion restriction, identification of the estimated parameters depends on untestable functional form assumptions. Moreover, the most important predictor of winning, card quality, is only observed in cases where a game is actually initiated. Adding it as a control variable in the main but not in the selection equation would potentially make it an endogenous covariate. Based on these considerations we decided not to include a probit model with sample selection in our analysis.

ways in which gender differences in choices lead to lower scores for women despite no gender difference in actual winning ability.

First, as a result of their lower propensity to initiate a game, female players end up more often in the difficult opponent role and less often in the easier initiator role. Consequently, even though the winning probabilities are the same for male and female players *in each role*, the *overall* winning rates of women are lower.

Moreover, scores collected in the game are calculated as the sum of the cents earned, and women are less effective in turning the points they win into cents. Note that the conversion between points and cents depends on the game type (*Wenz* and *Solo* pay more) and the chosen stakes of the game (determined by the number of 'knocks' and 'Contras'). Since women disproportionately shy away from the more competitive (and more lucrative) game types and fail to increase the stakes even when their cards are good, they profit less from winning than men do.

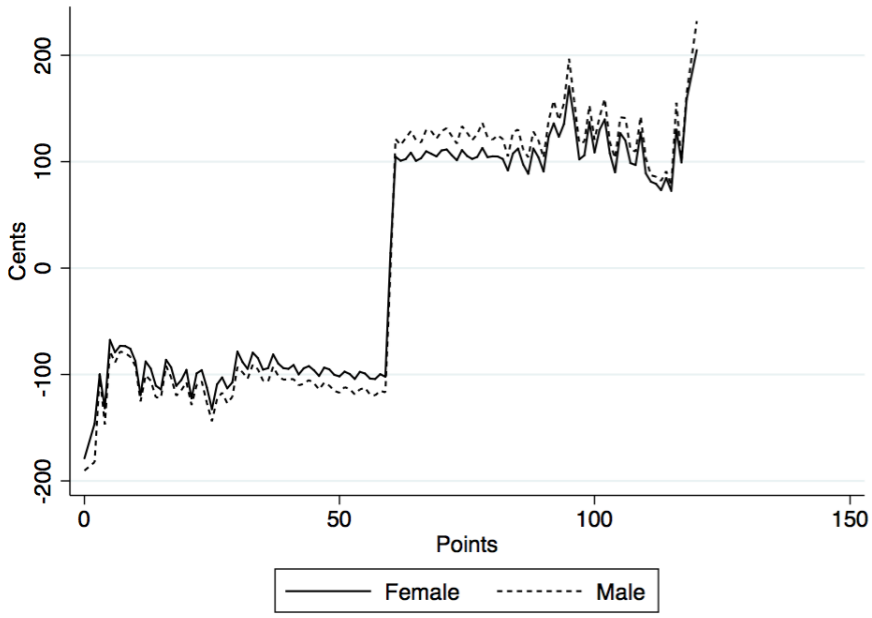


Figure 3.5: Conversion between points and cents, separately by gender

Figure 3.5, plotting the mean cents earned against the points collected, illustrates the above reasoning. It shows that female players gain less from winning and lose less from losing a game

than males (note that collecting at least 61 points out of 120 ensures winning). The difference is smaller in absolute terms in case of winning than in case of losing: as winners, men earn on average 19.5 cents more than women for the same number of points, while as losers women lose 14.3 cents less than men, given equal points. Over the course of several games, the difference accumulates (exacerbated by the above-discussed result that women have lower overall winning rates).

3.3.5 Drivers of the gender gap in choices

In our analysis we have shown that female players have a different playing style than men: they have a lower propensity to initiate a game and to raise stakes. We now consider whether these differences can be attributed to gender differences in preferences towards risk and competition, or they are rather driven by alternative forces.

We first discuss whether "shying away" is a result of discrimination against female players. Table 3.8 suggests that this is unlikely to be the case: men do not play more aggressively against female opponents. To the contrary, male players are actually slightly *less* likely to initiate the most competitive game type (*Solo*) (column (1)) and are *less* inclined to the raise the stakes (column (2)) when there is a woman at the table (both differences are negligible in size). In column (3) we consider players who have an all-male group of opponents and show that female players are less likely to receive "Contra". We thus find no sign of women being 'penalized' for participating in this traditionally masculine game. Additionally, we find (in unreported analysis) the gender gap in initiation rates to be the largest in the first position where it is certainly not a response to the actions of others at the table. Taken together, these results imply that women's different choices are not a response to negative discrimination on the side of male players.

Table 3.8: REACTION TO FEMALE PLAYERS

VARIABLES	*Male players* Initiate Solo (1)	Raise stakes (2)	*Players with all-male opponents* Receive Contra (3)
Female player			-0.007***
			(0.001)
Female opponent	-0.002***	-0.013***	
	(0.000)	(0.001)	
Position at table = 2	-0.004***	-0.009***	-0.000
	(0.000)	(0.001)	(0.000)
Position at table = 3	-0.004***	-0.010***	-0.000
	(0.000)	(0.001)	(0.000)
Position at table = 4	-0.001***	-0.008***	-0.000
	(0.000)	(0.001)	(0.000)
Num. games played (log)	0.000*	0.013***	-0.002***
	(0.000)	(0.001)	(0.000)
Num. opponents knocked = 1	-0.025***	-0.118***	0.012***
	(0.000)	(0.001)	(0.000)
Num. opponents knocked = 2	-0.045***	-0.219***	0.024***
	(0.000)	(0.001)	(0.000)
Num. opponents knocked = 3	-0.060***	-0.281***	0.033***
	(0.001)	(0.003)	(0.001)
Predicted prob. male players	0.033***	0.276***	0.040***
	(0.000)	(0.002)	(0.000)
Observations	14,402,768	14,402,768	10,725,967
Number of players	13941	13941	14618
Pseudo R^2	0.0258	0.0258	0.00720

Notes: The table displays marginal effects at the means from probit models (for factor variables, calculated as the discrete change from the base level). Dependent variables: column (1): initiate Solo; column (2): increase the stakes of the game by "knocking", column (3): receive "Contra". Columns (1) and (2): subsample of male players; column (3): subsample of players with all-male opponents. As a baseline, we present predicted probabilities for male players (evaluating all other variables at their means). Omitted categories for covariates: position at table = 1; num. opponents knocked = 0. Standard errors are clustered on player ID and are reported in parentheses. *** p<0.001, ** p<0.01, * p<0.05

A gender gap in confidence is another candidate explanation for the differences observed between male and female playing styles (Kamas and Preston, 2012). As shown in Table A2, women indeed react more to negative feedback on their relative performance than men do (they are particularly deterred from initiating when they have the worst rank at the table). Controlling for the interaction term between female and rank (column (3)), the gender gap in game initiation

rates reduces slightly but remains significant, suggesting that confidence alone cannot account for the differences between male and female behavior in the selection stage.

It could also be argued that female players are simply less skilled than males in identifying rounds when it is beneficial to initiate a game or to increase the stakes. While we can not rule out this explanation, observing the initiation rates by card quality in Figure 3.2 suggests that this is not the main driver. The skill of identifying opportunities should play the most important role in game initiation decisions in case of intermediate quality cards: even the least able player is unlikely to start a game with obviously bad cards and will probably not fail to initiate when he/she is dealt the best possible cards. The gender gap in initiation rates, however, is not the largest in case of medium quality cards (in absolute terms, it is largest for the highest, in relative terms, for the lowest quality cards).

Finally, women's less offensive playing style could be driven by their relative disinterest in winning. Women may initiate and raise the stakes less often because they get less utility from high scores than men do. This reasoning holds especially since the game is played for "virtual cents", not for real money. However, an analysis of hazard rates suggests that women are even more likely to leave a table after a loss than men. This particularly strong negative response to losing a game suggests that women care at least as much about their performance in the game as male players.

To conclude, we believe none of the above alternative explanations can account for the observed gender gap in playing styles. This leads us to conjecture that women's lower game initiation and 'knocking' tendencies are to a large extent driven by gender differences in risk appetite and competitiveness. It is important to note, however, that women who have self-selected into the card game community are not averse to being *exposed to* risk and competition, but they prefer not to *initiate* such situations. Female players who do not start games themselves still enter the tournaments, only in the role of partner or opponent. Moreover, even though women are much less likely than men to increase the stakes, there is no difference in their response upon observing that their opponents 'knock' (see column (4) in Table A2). We find

this phenomenon to be in line with women's shame aversion (Ludwig and Thoma, 2012) or adherence to social norms (Gneezy et al., 2009).

3.4 Conclusions

Our results are consistent with the idea that even those women who have deliberately entered an uncertain tournament environment will subsequently "shy away" from actively seeking competition and risk. In the context we study, females are less inclined than males to initiate competitive games and to increase the stakes of the game. These findings replicate earlier results related to gender differences in preferences, albeit in a context and setup that is different from previous studies. Most notably, instead of a controlled laboratory experiment we obtain our results by studying a large set of naturally occurring data. The novelty of our study comes from showing that females' lower perceived performance (i.e. lower cumulative displayed scores) are not due to differences in playing skills but to women's less competitive and risky choices. Their more cautious playing behavior causes them to gain less and appear less competent in the game, despite the fact that given the games they play, female players do not underperform at all. We might - cautiously - extrapolate our findings to gender differences in behavior in real (labor) markets and thereby link gender differences in preferences to the slower advancement of women in organizational hierarchies.

Why is it problematic if women "shy away" and (therefore) have lower incomes and less steep career tracks in competitive labor markets? Perhaps women are willing to pay that price in order to avoid risk and competition. While this might explain the gender gap in choices, it is also possible that women are unaware of the *consequences* of their behavior in terms of forgone earnings and opportunities. For instance, Sandberg (2013) reports an example from Google where female engineers' self-nomination rates rose to the same level as men's after the management shared evidence with them about prior gender differences in nomination rates.

The consequences of gender differences in preferences are also important from a more general perspective. If qualified women refrain from using their full potential in terms of produc-

tivity by "shying away", this could lead to efficiency losses. For instance, if the best candidate for a CEO position was a woman who was deterred by the cut-throat competition to reach the top, and therefore a lower quality male CEO was appointed instead, not only would the female candidate miss out on a valuable career opportunity but also the company could suffer from lower performance.

All in all, it is worthwhile to study and develop instruments that encourage women to take more risk and competition. As mentioned before, giving advice or sharing evidence of gender differences in behavior and their consequences might already be helpful. Encouragingly, a recent laboratory experiment shows that advice increases entry into tournaments among high-performing women (Brandts et al., 2015). It is important to note, however, that a lower appetite for risk and competition might actually be beneficial in certain contexts. Eckel and Füllbrunn (2015) find, for instance, that increasing the number of women as investors in an experimental financial market reduces overpricing. They support their results with a large meta-analysis of previous experiments that also shows a substantial negative correlation between the share of female traders in the market and the magnitude of price bubbles that occur. Such results suggest that besides (or instead of) coaching women to tolerate uncertainty and competition better, we could consider promotion mechanisms that rely less on self-initiated competition and risk taking.

Appendix

3.A Additional tables and figures

Figure A1: Screenshot from the online game

Figure A2: Table selection in the online card game community

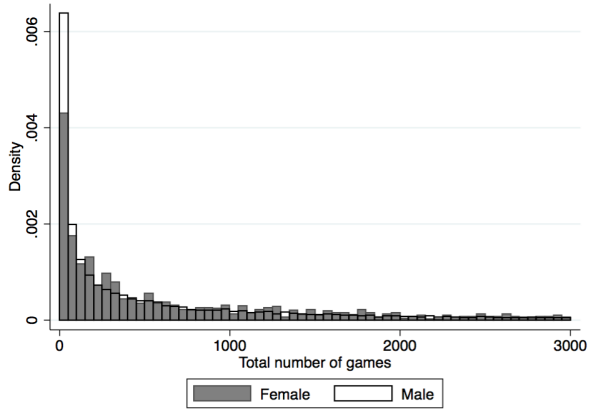


Figure A3: Distribution of total number of games played, by gender

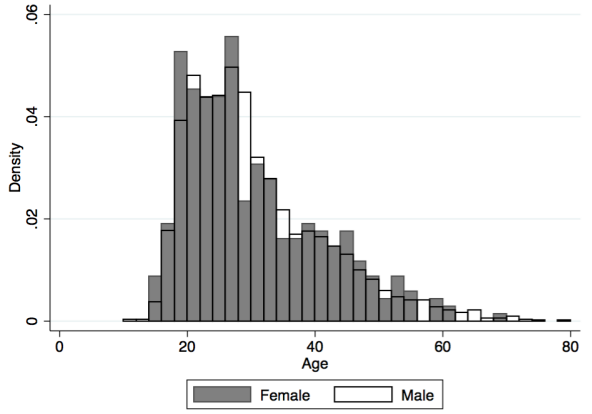


Figure A4: Distribution of age, by gender

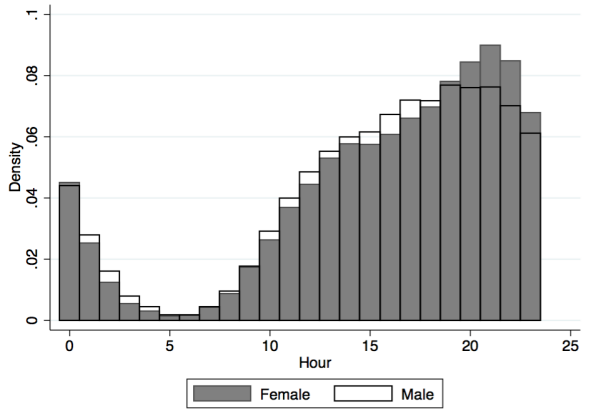


Figure A5: Distribution of the hour of the day when games are played, by gender

Table A1: LIKELIHOOD OF INITIATING A GAME AND RAISING STAKES - Estimated coefficients

VARIABLES	Initiate a game			Increase stakes
	(1)	(2)	(3)	(4)
Female player	-0.134***	-0.139***	-0.117***	-0.218***
	(0.012)	(0.013)	(0.012)	(0.028)
Position at table = 2	-0.080***	-0.080***	-0.080***	-0.029***
	(0.001)	(0.001)	(0.001)	(0.001)
Position at table = 3	-0.116***	-0.117***	-0.117***	-0.031***
	(0.001)	(0.001)	(0.001)	(0.002)
Position at table = 4	-0.133***	-0.135***	-0.135***	-0.023***
	(0.002)	(0.002)	(0.002)	(0.001)
Female opponent		-0.013***	-0.014***	-0.018***
		(0.002)	(0.002)	(0.003)
Num. opponents knocked = 1		-0.392***	-0.390***	
		(0.001)	(0.001)	
Num. opponents knocked = 2		-0.698***	-0.694***	
		(0.002)	(0.002)	
Num. opponents knocked = 3		-0.931***	-0.926***	
		(0.007)	(0.007)	
Num. games played (log)			-0.019***	0.008
			(0.002)	(0.004)
Rank in score = 2			-0.048***	-0.084***
			(0.005)	(0.009)
Rank in score = 3			-0.077***	-0.155***
			(0.006)	(0.012)
Rank in score = 4			-0.094***	-0.246***
			(0.007)	(0.015)
Constant	-0.820***	-0.538***	-0.348***	-0.495***
	(0.004)	(0.004)	(0.014)	(0.033)
Observations	16,655,344	16,655,344	16,655,344	16,655,344
Number of players	15232	15232	15232	15232
Pseudo R^2	0.00230	0.0312	0.0317	0.00748

Notes: The table displays estimated coefficients from probit models. Dependent variable in columns (1) - (3): initiate any game type; column (4): increase the stakes of the game by "knocking". Omitted categories for covariates: position at table = 1; female opponent = 0; rank in score = 1; num. opponents knocked = 0. Standard errors are clustered on player ID and are reported in parentheses. *** p<0.001, ** p<0.01, * p<0.05

Table A2: GENDER INTERACTIONS

VARIABLES	Female opponent (1)	Opponents knocked (2)	Rank (3)	Full model (4)
Female player	-0.115***	-0.115***	-0.083***	-0.079***
	(0.011)	(0.012)	(0.020)	(0.020)
Position at table = 2	-0.080***	-0.080***	-0.080***	-0.080***
	(0.001)	(0.001)	(0.001)	(0.001)
Position at table = 3	-0.117***	-0.117***	-0.117***	-0.117***
	(0.001)	(0.001)	(0.001)	(0.001)
Position at table = 4	-0.135***	-0.135***	-0.135***	-0.135***
	(0.002)	(0.002)	(0.002)	(0.002)
Female opponent	-0.014***	-0.014***	-0.014***	-0.013***
	(0.002)	(0.002)	(0.002)	(0.002)
Num. opponents knocked = 1	-0.390***	-0.390***	-0.390***	-0.390***
	(0.001)	(0.001)	(0.001)	(0.001)
Num. opponents knocked = 2	-0.694***	-0.692***	-0.694***	-0.692***
	(0.002)	(0.002)	(0.002)	(0.002)
Num. opponents knocked = 3	-0.926***	-0.922***	-0.926***	-0.922***
	(0.007)	(0.008)	(0.007)	(0.008)
Num. games played (log)	-0.019***	-0.019***	-0.018***	-0.018***
	(0.002)	(0.002)	(0.002)	(0.002)
Rank in score = 2	-0.048***	-0.048***	-0.048***	-0.048***
	(0.005)	(0.005)	(0.005)	(0.005)
Rank in score = 3	-0.077***	-0.077***	-0.073***	-0.073***
	(0.006)	(0.006)	(0.006)	(0.006)
Rank in score = 4	-0.094***	-0.094***	-0.084***	-0.084***
	(0.007)	(0.007)	(0.007)	(0.007)
Female player * female opponent	-0.006			-0.009
	(0.006)			(0.006)
Female player * 1 opp. knocked		-0.001		0.000
		(0.003)		(0.003)
Female player * 2 opp. knocked		-0.013*		-0.011
		(0.006)		(0.006)
Female player * 3 opp. knocked		-0.038		-0.035
		(0.023)		(0.023)
Female player * 2nd rank			-0.009	-0.009
			(0.015)	(0.015)
Female player * 3rd rank			-0.036	-0.036
			(0.020)	(0.020)
Female player * 4th rank			-0.069**	-0.069**
			(0.023)	(0.023)
Constant	-0.348***	-0.348***	-0.355***	-0.356***
	(0.014)	(0.014)	(0.014)	(0.014)
Observations	16,655,344	16,655,344	16,655,344	16,655,344
Number of players	15232	15232	15232	15232
Pseudo R^2	0.0317	0.0317	0.0318	0.0318

Notes: The table displays estimated coefficients from probit models. Dependent variable: initiate any game type. Omitted categories for covariates: position at table = 1; rank in score = 1; num. opponents knocked = 0. Standard errors are clustered on player ID and are reported in parentheses. *** p<0.001, ** p<0.01, * p<0.05

Chapter 4

The drivers of selection into teams

4.1 Introduction

Different incentive schemes lead to different output levels not only through the incentive effect but also through sorting. A famous example is the study by Lazear (2000) showing that half of the productivity gains at Safelite Glass may be attributed to the self-selection of more productive agents into the company after the introduction of a piece rate incentive scheme (see also Leuven et al. (2011) who disentangle the incentive and sorting effect of tournaments). Nowadays, many firms rely on teamwork, or use some sort of team-based reward structure (Bandiera et al., 2013; Hamilton et al., 2003; Lazear and Shaw, 2007). The productivity of the teams and the success of the use of team incentives strongly depend on the sorting that will occur, that is, on the characteristics of the individuals that self-select into the organization offering these types of incentives (Hamilton et al., 2003). Although the effectiveness of team remuneration schemes has been extensively researched (see e.g. van Dijk et al. (2001), Hamilton et al. (2003), Danilov et al. (2013) and Bandiera et al. (2013)), most papers on team decision making or team

This chapter is based on a joint work with Martin Koudstaal and Laura Rosendahl Huber. We are grateful to the Research Priority Area Behavioral Economics of the University of Amsterdam for their generous financial support and to the Amsterdam Center for Entrepreneurship for the help they provided in conducting our experiment.

production take the composition of the team as a given, and do not take the drivers and the effect of sorting into team incentives into account. Our aim is to contribute to the literature on team incentives by studying the actual participation decision of individuals.

Team-based incentive schemes affect different aspects of the individual's decision problem: they influence both the expected monetary gains and the uncertainty associated with the payoffs. The participation decision is therefore governed by the individual's ability, confidence, risk preferences and beliefs about the potential partners.[1] The extent to which these factors matter for sorting depends largely on the design of the team remuneration scheme. Most papers focusing on self-selection into teams model the team option as a simple revenue sharing contract, e.g. an equal split of the pooled total output of the members (Bäker and Mertins, 2013; Dohmen and Falk, 2011; Herbst et al., 2015). A few studies acknowledge potential team efficiency gains by adding some pre-defined, automatic mark-up on top of the joint output (Cooper and Saral, 2013; Kuhn and Villeval, 2014). A distinguishing feature of real life team production, however, is the possibility for synergies: gains from the team option are *ex ante* unknown and depend on the degree of complementarities between the team members. The study discussed in this chapter addresses this issue by using a team production function that allows potential synergy gains.

Moreover, in real life team situations the potential gains from joint production often come at a price: team members have to partially give up their authority and make joint decisions. Even when the instrumental value of the decision rights is not too high (in the sense that the joint choice is not far from the individual's optimum), the loss of power resulting from joint decision making could still make the team choice unattractive for some (Bartling et al., 2014; Fehr et al., 2013; Sloof and von Siemens, 2014). The existing evidence suggests that people differ in the intrinsic value they attach to decision rights. Self-employed, for instance, are claimed to value control more than non-entrepreneurs (Benz and Frey, 2008; Reynolds and Curtin, 2008). It is unclear, however, whether they are motivated by a desire for payoff autonomy (i.e. they prefer

[1] Kuhn and Villeval (2014) and Teyssier (2008) show that inequity aversion could also play a role in the choice to select team remuneration schemes. This factor is not in the focus of our study.

to retain an independent control over their own outcomes, see Charness and Gneezy (2010); Owens et al. (2014)) or authority (i.e. the right to make decisions, see (Bartling et al., 2014)). Our study intends to separate the impact of these two drivers on the choice to enter teams.

To summarize, this chapter aims to contribute to the existing economics literature on team preferences in two ways. The first objective is to shed light on the factors that influence the choice between individual and team incentives in a setting where the teamwork includes potential synergy gains. We test for heterogeneous choices by individual characteristics such as age, gender, education, income and occupational categories, and see if these heterogeneities are attributable to differences in performance, beliefs or preferences. Secondly, we want to determine the importance of preferences for control for the self-selection into teams that involve both joint production and joint decision making. To this end, we specifically designed two team treatments reflecting these different team characteristics. We conduct a large scale lab-in-the-field experiment in which individuals are randomly assigned to one of the two team treatments, combining a controlled experimental environment with the flexibility and reach allowed by an online survey. Our sample contains 1,164 individuals (entrepreneurs, manager and employees) and is very diverse in terms of age, education, experience and income.

The experiment consists of four incentivized parts. In the first part, we measure the productivity level of each individual in a real-effort task (i.e. solving Raven matrices (Raven et al., 2003)). In the second part, we elicit participants' risk preferences following Gneezy and Potters (1997). In Part 3, we use the BDM mechanism (Becker et al., 1964) to elicit participants' willingness to pay for the team incentive scheme. In the fourth part of the experiment we measure the beliefs of each participant regarding his/her own performance, and the performance and risk preferences of a random survey participant. These four incentivized parts are followed by a short questionnaire to collect information about the rationale behind the team choice and some background characteristics. At the beginning of the survey we randomly assign each individual to one of two team treatments conditions. In the *Baseline* treatment the team option only involves joint production (as in e.g. Cooper and Saral, 2013; Kuhn and Villeval, 2014). In

the second (*Joint Decision*) treatment, the team option includes both joint production and joint decision making. The comparison between the willingness to pay for the team option in the two treatments allows us to examine if and to what extent the loss of autonomy in decision making affects self-selection into teams.

We analyze the determinants of team choice by first focusing on the subsample of participants in the Baseline treatment (joint production without joint investment). Contrary to the findings from previous laboratory experiments (Dargnies, 2012; Kocher et al., 2006; Teyssier, 2008), we find that the willingness to be in a team is unrelated to task performance, i.e. in our study low performers do not prefer team pay more than high performers. To understand why we observe no adverse selection in this setting, we evaluate several factors that potentially affect the choice for team incentives. In line with previous studies (Dohmen and Falk, 2011; Kuhn and Villeval, 2014) we find that relative and absolute confidence and risk aversion all influence the selection into team incentives, in the expected directions. All else equal, confidence in one's own absolute performance decreases the willingness to join a team. Higher relative performance expectations also lead to lower bids for the team option. Moreover, we find that risk aversion has a negative impact on the willingness to be in a team. Thus, similar to the results from a recent study by Bäker and Mertins (2013), we find that the uncertainty about (the performance and the effort of) the teammate is more important than the reduction in the exposure to individual idiosyncratic shocks.

Interestingly, we find that education is strongly positively correlated with preferences for team incentives (even when controlling for actual and guessed performance). That is, all else equal, higher educated individuals in our sample bid significantly more for the team option. This result is surprising given that low educated people on average perform worse on the task and could thus gain even more from the team option than higher educated respondents. Looking further into this positive education effect, we find that the negative impact of risk aversion on the preference for teams decreases with the education level: the higher a participant's level of education, the weaker the link between her risk preferences and her team sorting decision. This

suggests that the level of education is correlated with the way people *evaluate* team incentives. While lower-educated individuals tend to concentrate on the risky aspect of team pay, those with a higher level of education focus more on the potential gains and synergies from teamwork. This interpretation of our findings stems from self-reported explanations of participants for their team bids and is also confirmed by a regression analysis of the impact of risk preferences on the bids for the team option. Given that the experimental task (i.e. solving Raven puzzles) is a proxy for cognitive ability, our results are consistent with the idea that education (above and beyond its impact on task performance) affects preferences for teams by changing the way people weigh the associated risks and gains. This brings to mind the finding of Heckman et al. (2013) that an education intervention may improve life outcomes through its impact on non-cognitive skills and motivation even when IQ is unchanged. An alternative interpretation of the "education effect" we observe is that people who self-select into obtaining high education are inherently different from others in the way they assess situations with (strategic) risks and potential synergy gains.

In order to measure the effect of desire for authority on the willingness to be in a team, we compare the team choices between the Baseline and the Joint Decision treatments. The results show that the response to shared decision rights is heterogeneous with occupational categories. We find that managers' and employees' preferences for team pay are not significantly affected by the inclusion of joint decision making in the team option. Entrepreneurs, on the other hand, are averse to team decision making when they expect their potential teammate to make different investment choices than themselves. This result is in line with the findings of Masclet et al. (2009) who demonstrate that self-employed people have a stronger preference than employees for making decisions individually instead of in a team.[2] While previous studies demonstrate the non-pecuniary value attached to decision rights (e.g. Fehr et al. (2013)), our study provides no evidence for a positive willingness to pay for authority *per se*. This difference in findings might be explained by the fact that in our setting the power to make a decision was *shared with*,

[2]Relatedly, Reynolds and Curtin (2008) find, using survey data from the Panel Study of Entrepreneurial Dynamics II, that entrepreneurs are to a large extent motivated by a preference for autonomy.

not delegated to the other party: even in case of joint decision making, each individual member retained a large influence over the choice that was selected.

The remainder of this chapter is structured as follows. In Section 4.2 we provide an overview of the related literature. Section 4.3 describes the experimental context and the design, while Section 4.4 shows the descriptive statistics. Results are presented in Section 4.5. In Section 4.6 we summarize and conclude.

4.2 Related literature

As we have briefly discussed in the Introduction, limited research has been conducted to understand the factors underlying the preferences for individual and team incentives. Table 4.1 provides an overview of the existing studies on sorting into teams. A few insights emerge from the summary. First, none of the reviewed papers analyze team settings that involve joint production as well as joint decision making. These two aspects have only been studied separately so far. Moreover, none of the experimental studies on selection into team remuneration schemes model team production with a scope for synergies: the majority of studies impose an equal revenue sharing rule where individuals can only benefit from joint work if they are teamed up with a more productive partner than themselves. Such settings necessarily result in adverse selection into teams. Even studies that acknowledge the possibility of team efficiency gains do so in a rather artificial way, by adding an automatic mark-up over the sum of individual earnings. The production function suggested in our study, by making synergies possible but not certain, is a novel addition to the existing literature.[3] Finally, most of the above-mentioned papers study a 'traditional' subject pool of university students, and even those that include non-student participants have a rather limited number of observations.[4] Our large and diverse sample allows a more extensive analysis of the impact of experience, education and occupational categories on team preferences.

[3] A drawback of this design choice is that the size of the potential gains from teamwork are *a priori* unknown, and we do not observe participants' beliefs about the size of these gains. While this feature complicates the analysis we decided for it because in our opinion it models more accurately real life team situations.

[4] In the paper by Masclet et al. (2009) the total number of participants is 144, and less than half of them are non-students. Cooper and Saral (2013) study a sample of 184 individuals, 44 of whom are (full- or part-time) self-employed.

Table 4.1: Overview of the literature on sorting into teams

	Data source	Sample	Joint production	Real effort	Tournament[a]	Joint decision	Interaction[b]	Automatic EA[c]	Free-riding[d]	Mutual consent[e]
Hamilton et al. (2003)	Field	Factory workers	✓	✓			✓	✓	✓	✓
Kocher et al. (2006)	Laboratory	Students				✓	✓			✓
Teyssier (2008)	Laboratory	Students	✓						✓	✓
Masclet et al. (2009)	Laboratory	Students, salaried workers, self-employed				✓	✓			
Dohmen et al. (2011)	Laboratory	Students	✓	✓					✓	✓
Dargnies (2012)	Laboratory	Students	✓	✓	✓				*both*[f]	✓
Cooper and Saral (2013)	Lab-in-field (online survey)	Students, salaried workers, self-employed	✓	✓				✓	✓	
Bäker and Mertins (2013)	Laboratory	Students	✓	✓					✓	✓
Kuhn and Villeval (2014)	Laboratory	Students	✓	✓				*both*	*both*	*both*
Herbst et al. (2015)	Laboratory	Students	✓		✓				✓	*both*

[a] Are teams competing against each other?
[b] Is there actual interaction between the team members?
[c] Is there an automatic efficiency advantage incorporated into the team option? (The alternative is a simple split of the pooled joint output)
[d] Is there a possibility for free-riding in the team option?
[e] Does team formation require mutual consent from all team members?
[f] The notation *'both'* indicates studies that include treatments with and without the given features

In the following, we review in more detail the papers that are closest related to our study. In particular, we focus on papers that analyze the choice between an individual or a team piece rate scheme, and those that address preferences for individual vs. team decision making.

The paper by Hamilton et al. (2003) is among the first to report about self-selection into team incentives. Although the main focus of their field experiment is the effect of team composition on team performance, the authors also provide evidence on the characteristics of those who are the first to join the teams once the new incentive system is introduced. The task studied in this paper (i.e., sewing together pieces of garment in a factory) is primarily based on joint production and does not contain any specific elements of joint decision making. The team setting in this experiment provides a scope for synergies. Contrary to the theoretical predictions on adverse selection, Hamilton et al. (2003) find that the high-productivity workers are among the first to join the team.[5]

One of the first studies to look at sorting into teams in a controlled laboratory experiment is the paper by Kocher et al. (2006) who study the preferences for individual or group decision making using a beauty-contest game. In their setting there are monetary gains to be expected from teaming up. They find that 60% of individuals in their sample prefer to work in a team. There is also some indication for adverse selection. In order to measure the drivers of team preferences, the authors focused on the costs and benefits from the joint decision making process.[6] Higher profits turned out to be the most important reason for choosing the team option. They found that the endogenously formed teams performed significantly better than the individuals that decided to play alone. However, they also found that the individuals in both settings were very satisfied with their choice, despite the lower earnings for the individual players. Hence, the

[5]There are two factors that could be confounding the self-selection into the teams in this field experiment. First, as pointed out by Kocher et al. (2006), the team production framework is introduced over a period of three years, with the purpose that at the end of this period everybody is working in teams. Thus, the self-selection is not entirely voluntary. It could be that the high ability workers are the first to realize that they do not have a choice in the end and thus decide to opt for the team option at an early stage in order to secure their job. Secondly, the team option in the field experiment involved slightly higher piece rates. Hamilton et al. (2003) discuss that this increased intensity of incentives could also act as a confound.

[6]As the main drivers underlying this decision, the participants were offered a choice between two explanations: *"I want to act as an individual (in a team), because I want (do not want) to decide alone"* and *"I want to act as an individual/in a team because in this way I can earn more"* (Kocher et al., 2006, p.263).

authors argue that individuals are on average willing to pay a price for autonomy in decision-making.[7]

Kuhn and Villeval (2014) conduct a lab experiment to study gender differences in the choice between individual and team-based incentives. The participants have to perform a real-effort task for a piece rate. The experiment contains two different team treatments, both with equal revenue sharing but neither with joint decision making. In their baseline team treatment the piece rate is the same as in the individual setting, while in the efficiency advantage treatment (EA) there is an automatic 10% markup on the team pay.[8] As expected, due to the equal revenue sharing, the authors find evidence for adverse selection both in the baseline and in the EA treatment for both men and women. The increase in willingness to join the team between the baseline and the EA treatment is much larger for men than for women, thus eliminating the gender gap that was observed in the baseline treatment.[9]

Cooper and Saral (2013) conduct an online (lab-in-the-field) real-effort experiment on a diverse sample in terms of socio-economic characteristics. The aim of their paper is to test for differences in team preferences and free-riding tendencies between the different sub-groups. The set up of the experiment is such that self-regarding participants should weakly prefer the team option over the individual option. The results show that full-time entrepreneurs have a significantly stronger preference to be alone compared to all other employment categories. The results provide some evidence for adverse selection, i.e., individuals that performed better in the first part of the experiment bid (slightly) more to be alone. However, they find no differences in free-riding among the different occupational groups. Moreover, even though the experimental set-up did not involve any joint decision making, participants cite a fear for the loss of control

[7]Masclet et al. (2009), studying decision making in groups in a lottery-choice experiment, also address the question of selection. Their main result is that risk averse individuals then to bid more to avoid joint decision making. They also find that self-employed participants tend to have a higher willingness to decide alone.

[8]Each individual first performs the task twice: once alone and once the under the team payment scheme. Then each individual can choose between a team or individual incentives three times. First, when matched with the performance of the partner from the individual round. Second, when matched with the performance of the partner from the team incentive round, and finally, matched with the performance from the partner in that round, but only if both choose to be in a team (i.e., mutual consent).

[9]Dohmen and Falk (2011) also consider the impact of gender on sorting into team incentives schemes where the alternative is a flat wage option.

or a preference for self-reliance as a reason for choosing to play alone.

Finally, Bäker and Mertins (2013) investigate how risk influences the sorting into two different variable payment schemes, i.e., individual and team piece rates. The authors look at two types of risks that are associated with team pay. First, there is the risk of being matched with a low productivity co-worker, which would then decrease pay off. On the other hand, being matched with another worker reduces the risk from individual idiosyncratic shocks, such as luck, motivation or distraction, and thus may positively influence the desire to be in a team. The results from this experiment show that both types of risk indeed influence the selection into team piece rates: higher idiosyncratic risk is shown to increase the probability of choosing the team pay option, whereas the risk of being teamed up with a low ability worker reduces the probability of sorting into team pay. Because the marginal effect size and the significance of the strategic risk component is larger than the idiosyncratic risk component, the former seems to be more important for the sorting decision.

Other topics that have been studied more extensively in relation to self-selection into teams are social preferences such as cooperation (Dur and Sol, 2010; Kosfeld and Siemens, 2009), inequality aversion (Teyssier, 2008) and in-group favoritism (Herbst et al., 2015). Given the design of the experiment and the focus of our paper, we do not discuss them further here.

4.3 Context and Design

4.3.1 Context

Our study was conducted as the fourth wave of an extensive scientific project studying behavioral traits of entrepreneurs, managers and employees in The Netherlands.[10] Each wave of this

[10]We use the same definitions as Koudstaal et al. (2015). Entrepreneurs are those people who have founded, inherited or taken over a company that they are currently (co-)managing. Managers are people who are employed by an organization that they did not start up themselves and have at least two subordinates for whom they are directly responsible (i.e. "direct reports"). We also regard project managers as 'managers' whenever they have overall responsibility of their project and at least two direct reporting lines. Finally, employees are those who are employed by an organization they did not set up themselves, and have less than two two direct reporting lines. Participants who belonged to multiple occupational categories were instructed to select the one generating most of

project encompasses a large-scale lab-in-the-field (or, in the terminology of Harrison and List (2004), *artefactual field*) experiment in the form of an online incentivized survey. The waves all have a different focus: risk and uncertainty (Wave 1, October-November 2013), confidence and optimism (Wave 2, May 2014), intuitive vs. rational decision making (Wave 3, December 2014) and preference for teams (Wave 4, discussed in this chapter). A detailed description of the general project and the results of the first wave may be found in Koudstaal et al. (forthcoming).

To recruit participants for our study, we used the same channels as in the previous three waves.[11] We contacted entrepreneurs with the help of "Synpact", a large organizer of entrepreneurship events. Managers were contacted via "De Baak", a highly reputed training institute for managers. For the recruitment of employees, we collaborated with a Dutch market research agency. The invitations for the survey were sent out by e-mail on March 24, 2015, followed by a reminder after 7 days. The survey was open to respondents for 14 days. In total, close to 25,000 potential participants received an invitation and 1,164 individuals (400 entrepreneurs, 155 managers and 609 employees) completed our survey.[12]

Many respondents in the sample have very high income. Therefore, the relatively low earnings used in traditional laboratory experiments with student subjects were unlikely to provide proper incentives in our case. Instead, we decided to use very high payoffs and only pay out a subsample of our respondents. We therefore randomly selected 20 prizewinners from among all participants who completed the survey.[13] These prizewinners received the total amount they had earned in the survey. The payment structure was communicated very clearly to the participants

their income.

[11] As a result, there is considerable overlap in the subject pool of our study and the previous waves, especially among entrepreneurs and managers. Approximately half of our respondents (542 people) participated in one or more of the earlier waves, but only 67 individuals completed all four surveys of the general project. There is little difference in terms of demographic characteristics between respondents who participated before and those who were new to the project (the gender composition and the level of education is similar in the two groups, but "new" participants are on average 2.5 years younger), and we found no significant difference in their performance or choices in our survey.

[12] We conducted a pilot study among employees between March 11 and March 18, 2015 and received 192 complete responses. The aim of this pilot was to test the length of the survey and to compare different calibrations. Answers from the pilot study are not included in our main analysis.

[13] Such an approach is common in the literature and should produce similar results as when paying out all participants (see e.g. Gneezy and Rustichini (2000) and Laury (2006)).

at the beginning of our survey. Furthermore, to foster trust, the drawing of prizewinners was performed by a civil-law notary.

Prizewinners earned on average €330,58 with a minimum of €148 and a maximum of €785. The *ex post* chance of being paid out was approximately 1/58, but this was unknown to the participants (and ourselves) beforehand. However, to alleviate the concern that participants might hold different beliefs about the likelihood of being a prizewinner, we informed the participants at the beginning of the survey that the chance of being paid out had been approximately 1/100 in earlier research waves (as in Koudstaal et al. (forthcoming)).

Participants took on average 13 minutes to complete the survey that was designed and pretested to take approximately 15 minutes, suggesting that they took the tasks and choices seriously and read the instructions carefully. The original survey was conducted in Dutch.

4.3.2 Design

Our experiment contained four incentivized parts: a production phase, an investment phase, a choice between an individual or a team incentive scheme, and an evaluation phase; followed by an unincentivized background questionnaire. To analyze the impact of joint decision making on preferences for team incentives, respondents were randomly assigned to one of two treatment conditions: the *Baseline* and the *Joint decision* treatment that differed from each other in whether the team option entailed a potential compromise in a decision situation.[14] The details of each part of the experimental design and the two treatment conditions are described below.

[14]We performed a stratified randomization by gender and occupational category to ensure that we can analyze these subsamples separately. Our design contained also a within-subject element. Participants were asked to make a choice between individual and team remuneration schemes in two subsequent scenarios. The scenarios differed from each other in the characteristics of the potential teammate. In this paper we only discuss the first scenario where the teammate was randomly drawn from among all participants of the survey. Since the second scenario was only introduced after participants made their choice in the first one, their answers in the first scenario are unaffected.

Production phase (individual)

The first part of the experiment entailed a real-effort task. Participants were presented with 10 puzzles from the Raven Advanced Progressive Matrices (see Raven et al. (2003)) and were asked to solve as many of them as possible within a time frame of 10 minutes. This task required participants to complete puzzles consisting of three rows of three figures where the bottom-right figure was missing (see Figure A1 in Appendix 4.A). Raven test questions of varying difficulty are not uncommon as a production task in experimental economics (see e.g. Herz et al. (2014)). One of the main benefits of these puzzles for our study is that it is hard to find the correct solution on the Internet, which is a potential hazard when using an online survey. Moreover, performance on this task provides a proxy for cognitive ability, as Raven matrices are developed to serve as a "culture-free IQ test" (Herz et al., 2014, p.5).

To get participants acquainted with the set-up, we first provided them with an example question (without any time limit) and the general instructions as in Raven et al. (2003). Once participants indicated that they were ready, they were directed to the next page containing all 10 Raven puzzles one below another, as well as a timer showing the time remaining from the total 10 minutes. Participants were free to decide the order in which they solved the puzzles and they could go back and forth between the puzzles within the set time limit. After the allotted time was over, respondents were automatically directed to the next page (it was also possible to move on to the next page before the time was up). Participants faced individual piece rate incentives of €40 per correctly solved puzzle. There was no money deducted for incorrect answers. Participants were informed that they would receive feedback on the number of puzzles they solved correctly at the very end of the survey.

Investment phase (individual)

Part 2 measured participants' risk preferences. More specifically, following Gneezy and Potters (1997) we asked participants what share of their Part 1 earnings (0-100%) they were willing to invest in the following risky gamble:

- *a 2/3 (67%) chance that you lose the money you invested*
- *a 1/3 (33%) chance that you win two and a half times the amount you invested (on top of your investment).*

Subjects made their investment choice using a slider, as shown in Figure A2 in Appendix 4.A. Gneezy and Potters (1997) let their subjects make this investment choice several times in order to measure (myopic) loss aversion. In our design participants only answer this question once, and we use this investment choice as a proxy for their risk preferences. While this measure is not able to differentiate between risk loving and risk neutral subjects (the expected return on the gamble is positive, so already a risk neutral subject should invest everything), it is a simple, quick and easy-to-understand method for measuring different degrees of risk aversion.[15] In this paper, we calculate risk aversion as the share that the participant did *not* invest in the risky bet (i.e., $100\%-$share invested).

Team option

In Part 3, the key element of our experiment, participants were offered a choice between an individual and a team renumeration scheme. In particular, respondents had to decide whether they wanted to keep their individual piece rate earnings from Part 1 or whether they wanted to form a team with another survey participant instead. Note that by asking participants to choose incentive schemes *ex post* for their Part 1 performance, we have eliminated the possibility of free-riding. In our opinion, the elimination of this potential confound makes the design cleaner and the results easier to interpret. Moreover, while many theoretical papers on teamwork emphasize the incentives to free ride, empirical studies often find little evidence for such practice in real teams (e.g., Bäker and Mertins, 2013; Hamilton et al., 2003; Herbst et al., 2015).

As we have mentioned before, respondents were randomly assigned to one of the two treatment conditions that determined the content of the team option. Table 4.2 provides an overview of the design.

[15]This method has been used by e.g. Dreber et al. (2011) and Charness and Gneezy (2012) to measure risk aversion. Also, since risk aversion is in general considered to be a stable personality trait, we assume our measurement of risk preferences to be unaffected by participants' performance in Part 1. We revisit this assumption in Section 4.4.

Table 4.2: SUMMARY OF THE TREATMENT CONDITIONS

	Team option	
	Joint production	Joint investment
Baseline Treatment	✓	
Joint Decision Treatment	✓	✓

In the Baseline treatment, the team option only influenced the earnings from the real effort task but not the investment decision from Part 2. The Joint Decision treatment, on the other hand, entailed both joint production and a joint investment decision. In this treatment the share invested in the risky bet was determined as the unweighted average of the two teammates' individual choices in Part 2. Hence, the team option in this treatment entailed a potential compromise, with shared decision rights in the investment choice and the possibility of the being moved away from one's individual utility-maximizing risk exposure.[16]

A distinguishing feature of our design is that we model team incentives by introducing *conditional* efficiency gains. More specifically, payoffs in the team option were determined by the following production function with complementarities (the same in both treatment conditions):

$$P_i^{team} = P_j^{team} = \sum_{n=1}^{10}[max\{I_i^n, I_j^n\}],$$

where *i* and *j* denote the two team members, $n = 1, .., k, .., 10$ represents the question numbers from Part 1 and I_i^k and I_j^k are indicator variables showing whether participant *i* and *j* solved question *k* correctly. This production function allows participants to benefit from teaming up even with a less able partner provided that there are complementarities between their outputs, i.e. that their correct answers do not completely overlap.

Instead of showing participants the production function, we gave them the following explanation for team earnings: "*You get €40 for each puzzle that either you or your teammate solved*

[16]In our design, there is no scope for the 'wisdom of the crowds': since the decision is related to individual preferences, there is no 'correct' answer. Team decision making thus does not help the members to achieve a more efficient outcome: individual choice in our setting is always weakly preferred to the group choice. We do not model the bargaining process either: the compromise that results from the joint decision making is always the unweighted average of the two members' individual choices.

correctly in Part 1. Therefore your earnings in the team option are always at least as high as in the individual option, and higher in case your teammate solved more/different puzzles correctly than you did". To ensure that the overall set-up was clear to all participants in both treatments, we provided them with numerical examples on how the team option could affect their earnings and investment decisions (see Figure A3 in Appendix 4.A). After the example, we elicited our main measure of interest, i.e. the willingness to pay for the team option.

Instead of a binary choice between the individual and the team option, we elicited the willingness to pay for the team option by means of the BDM mechanism (Becker et al., 1964) which allows us to obtain a continuous measure of team preferences in an incentive-compatible manner. Specifically, we gave each respondent an endowment of €50 that they could either keep or use to bid for the possibility to be in a team. Participants were informed that the actual price of the team option would be randomly drawn from the interval [€1,€50] at the notary after the survey was closed. Teams were formed when both potential team members submitted a bid that was at least as high as the actual price. Team formation thus required mutual consent. We instructed participants that their teammate would be randomly drawn from the total sample of survey respondents. To fix beliefs, we explicitly mentioned that the teammate is equally likely to belong to either of the three occupational categories (entrepreneur, manager or employee). Participants received no feedback at any point in the survey about the identity, performance or bid of their potential teammate. Bids only had to be paid in case subjects actually formed a team. Participants were reminded that it was in their best interest to report their preferences truthfully.[17]

The team option in our setting did not involve an actual interaction between the teammates. This design choice was mostly due to practical constraints imposed by our data collection method: respondents of our online survey did not necessarily work on the questionnaire at the same time, so real-time interaction or communication would not have been possible. As

[17]In this setting the team option weakly dominates the individual option in terms of expected payoffs, and bidding zero always ensured that a respondent is payed on the individual basis, so we did not allow negative bids. In our analysis we account for the potential left- and right-censoring imposed by our elicitation technique that restricts the willingness to pay between the boundaries of €0 and €50 by estimating tobit models.

the literature overview in Table 4.1 shows, it is common to study sorting into team incentive schemes without allowing respondents to interact with each other. Moreover, there are also examples from real life that resemble the way we modeled teamwork. In the world of open-source software, developers often work individually and remotely on issues, submit their solutions, and the best suggestion gets accepted.[18] Similarly, in case of international scientific cooperations, the parties often already have ideas or preliminary results at the time when they decide to cooperate and pool their resources for a better final outcome.

Evaluation phase

In Part 4 of the survey we measured several factors that may influence the choice for the team option. Based on the existing literature, we identified three main candidates: (1) beliefs about own performance, (2) beliefs about the potential teammate's performance, and (3) value attached to decision rights (both instrumental and non-pecuniary). We therefore asked participants to submit their guesses for the following three questions:

1. the number of puzzles they solved correctly in Part 1;
2. the number of puzzles a random other survey respondent solved correctly in Part 1;
3. the share a random other survey respondent chose to invest in the risky gamble in Part 2.

All three evaluation questions were incentivized. The participants received €20 when their answer to the first question was correct, and €20 when their answer to the second question was correct. Finally, they could earn another €20 when their estimate in the third question was less than five percentage points away from the true value.

Background questionnaire

After the four incentivized parts, the final part of the survey included a questionnaire to gain insight into respondents' choices and to collect some background characteristics. First, we asked

[18] An example is the Linux kernel development process, see e.g. http://techblog.aasisvinayak.com/linux-kernel-development-process-how-it-works/.

participants to explain their choice in Part 3 by selecting from a list of possible explanations the option(s) they found most applicable:

- *I believed the team option could increase my earnings.*
- *I did not want to take too much risk.*
- *I thought I solved more puzzles correctly than other participants.*
- *I have calculated the expected gains.*
- *I wanted to be responsible for my earnings and not depend on others.*
- *I don't trust someone I don't know.*
- *It was just a guess.*
- *I followed my intuition.*

The list was based on the most common answers from the pilot survey (see footnote (12)) where respondents answered an open-ended question explaining their bid for the team option.

Respondents then answered background questions specific to their occupational categories.[19] All respondents were asked to report the years of work experience they had, and to select the income category they belonged to (with the option to keep this information private).

4.4 Descriptive statistics

This section provides an overview of the data collected in our survey and the most important variables used in the analysis. Table 4.3 shows the descriptive statistics for the total sample (columns (1) and (2)) and for the two treatment groups separately (columns (3)-(4) and (5)-(6)). Panel A describes the sample in terms of background characteristics while Panel B introduces the survey outcomes.

[19]Entrepreneurs reported the legal structure of their companies, whether they were founders, the number of their employees and the share they owned in their companies. Managers reported whether they were general or project managers, whether they were the CEOs of their organization and the number of their direct reports. Using this information we could specify groups of entrepreneurs and managers according to various (stricter) definitions used in the literature (see also Koudstaal et al. (2015)).

Table 4.3: Descriptive statistics

	Total sample (N = 1,168)		Baseline (N = 588)		Joint Decision (N = 576)	
	mean	std. dev.	mean	std. dev.	mean	std. dev.
Panel A: Background characteristics						
Age	46.34	11.14	46.69	11.20	45.98	11.07
Female (dummy)	0.40	0.49	0.40	0.49	0.40	0.49
Education (highest degree):						
- High school	11%		12%		11%	
- Lower vocational degree	24%		22%		26%	
- College education	41%		41%		40%	
- University	24%		25%		23%	
Work experience (years)	18.74	11.56	18.86	11.75	18.62	11.40
Income[1]						
- $\leq$ €25,000	25%		26%		24%	
- €25,001 - €50,000	40%		41%		40%	
- €50,001 - €75,000	16%		16%		17%	
- €75,001 - €125,000	12%		10%		13%	
- €125,001 - €200,000	4%		4%		4%	
- €200,001 - €300,000	1%		1%		1%	
- €300,001 - €400,000	1%		1%		0%	
- $>$ €400,000	1%		1%		1%	
Occupational category:						
- Entrepreneur	35%		34%		34%	
- Manager	13%		14%		13%	
- Employee	52%		52%		53%	
Panel B: Main variables						
Willingness to pay for the team (0-50)	€27.46	€16.08	€28.00	€16.32	€26.91	€15.83
Puzzles correct (actual) (0-10)	5.01	2.25	5.06	2.27	4.97	2.22
Puzzles correct (guess) (0-10)	5.73	1.98	5.67	2.03	5.78	1.92
Partner's correct (guess) (0-10)	5.54	1.48	5.54	1.49	5.53	1.47
Risk aversion (0-100)	53.72	27.06	54.26	27.79	53.17	26.31
Partner's risk aversion (guess) (0-100)	52.89	19.29	52.75	19.23	53.04	19.36

[1] For income, the number of observations drops to N = 779 (total sample), N = 381(Baseline), and N = 398 (Joint Decision).

Notes: Significance of differences between treatments from t-tests with unequal variances; *** p<0.001, ** p<0.01, * p<0.05

Table 4.3 suggests that the randomization worked well. Comparing the two treatment groups, we find no significant difference in terms of background characteristics (confirmed by two-sample t-tests and Kolmogorov-Smirnov tests). Panel A further shows that participants in our sample are on average 46.34 years old (the standard deviation is 11.14) and that 40% of our respondents are female. Respondents are most likely to have a college degree, but there is substantial variation in education levels. Participants have on average 18.74 years of work

experience. The modal (gross) income category is €25,001 - 50,000 per annum (as a comparison, the gross modal income was €33,500 in 2014 in The Netherlands (Netherlands Bureau for Economic Policy Analysis, 2014)).

Panel B of Table 4.3 shows the descriptive statistics for the survey outcomes. The mean bid for the team option is 27.46 with a standard deviation of 16.08 (a more detailed description of this outcome measure will be provided in the Results section). Participants solved on average 5.01 out of 10 puzzles. Figure B1a in Appendix 4.B depicts the distribution of the number of correct answers per participant and shows a large variance in puzzle performance. Reassuringly, only a very small fraction of respondents have zero correct answers, suggesting that participants took the task seriously and exerted effort. Figure B1b presents for each puzzle the share of respondents who solved the given question correctly. We see that there was a substantial difference in difficulty between the puzzles: while some questions were solved by close to 80% of the participants, this rate drops below 25% for other questions. Column (1) of Table B1 in Appendix 4.B demonstrates the importance of demographic variables for the performance on the Raven puzzles. All else equal, older respondents were less successful in our task, while more educated individuals solved significantly and substantially more puzzles. Even after controlling for age, gender and education, entrepreneurs and managers performed significantly better on the task (compared to employees).

Panel B of Table 4.3 further shows participants' guesses for their own performance as well as that of their potential teammate. The mean guess for the number of own correct answers was 5.73, while respondents on average estimated that their potential teammates solved 5.54 puzzles correctly. Figure B2 in Appendix 4.B sheds more light on the different aspects of individuals' confidence. In Figure B2a, we see the distribution of overestimation (i.e. the difference between own actual and guessed absolute performance (Moore and Healy, 2008)) among participants: the modal answer is 1 and the distribution is shifted to the right, suggesting that the majority of participants overestimate their performance. Figure B2b compares participants' guesses for their own and their partner's number of correct answers. Respondents are most likely to antici-

pate no performance difference. Among those who do predict a gap the majority expects to be better than a randomly chosen other survey participant.

Panel B of Table 4.3 also discusses outcomes related to risk preferences. We find that the average percentage invested in the risky gamble is 46.28% (not reported), which corresponds to a 'risk aversion' of 53.72 (with a standard deviation of 27.06) on scale of 0 to 100. On average, participants' guesses for the potential teammate's investment behavior is close to the respondents' own choice, but its variance is lower (the mean is 52.89 with a standard deviation of 19.29). Figure B3a in Appendix 4.B shows the distribution of the investment choices and confirms that the majority of respondents are risk averse. Figure B3b shows the estimated difference between own and partner's guessed investment choice. The majority of respondents believe there is no difference, and hardly anyone expects a gap larger than 50 percentage points. Finally, column (2) of Table B1 analyzes the relationship between demographics and risk aversion. It confirms the findings of Eckel and Grossman (2008) and Charness and Gneezy (2012) that women are more averse to risk than men. It further shows that both entrepreneurs and managers take more risk than employees (consistent with the results of Koudstaal et al. (forthcoming)). Age and education do not seem to play a role. Reassuringly, we find that risk aversion is unaffected by actual or perceived performance in the puzzle task, supporting the claim that our risk elicitation technique was not confounded by the outcomes in the preceding task.

Before analyzing in more detail participants' bids for the team option, we review the statements participants selected as explanations for their team bids in the background questionnaire at the end of the survey. The most popular answer was emphasizing the monetary benefits of the team option ("I believed the team option could increase my earnings"), selected by 39.3% of the respondents. Participants also frequently cited following their intuition (38.9%) and trying to avoid taking too much risk (23.6%). About a fifth of the respondents expressed a preference for control ("I wanted to be responsible for my earnings and not depend on others"), while 14.1% based their decision on the belief that they performed better than others. Reassuringly, only about 10 percent of respondents indicated that they made their team bid at random ("It was just

a guess"). The statements about calculating the expected gains and about not trusting strangers were the least frequently chosen, by about 5 percent of the respondents each.[20]

4.5 Results

This section presents our results on team preferences. For the analysis of the different factors that influence participants' willingness to pay for the team option we only consider the Baseline treatment (involving joint production but no joint decision making). This enables a more straightforward comparison of our findings with results of other studies that focus exclusively on team production. To estimate the effect of joint decision making on preferences for team pay, we then compare the willingness to pay for the team option between the Baseline and the Joint Decision treatments.

4.5.1 Determinants of team choice in the Baseline treatment

Figure 4.1 gives an overview of participants' willingness to pay for the team option in the Baseline treatment. The mean bid in this treatment group is 28.00 with a standard deviation of 16.32. Little more than 10% of the participants chose to bid zero and 18% was willing to pay the maximum possible amount, €50 for the team option. Even though bids were reported using a slider, respondents were still inclined to choose "round" numbers, i.e. multiples of five and especially ten. We see no indication for participants being biased by the slider's default setting: €25, the default option is only the fifth most common answer.

[20] The above shares do not add up to 100% since participants were allowed to select more than one explanation. The shares reported in the text are calculated based on the responses of participants in the Baseline treatment. Answers are very similar in the Joint Decision condition as well.

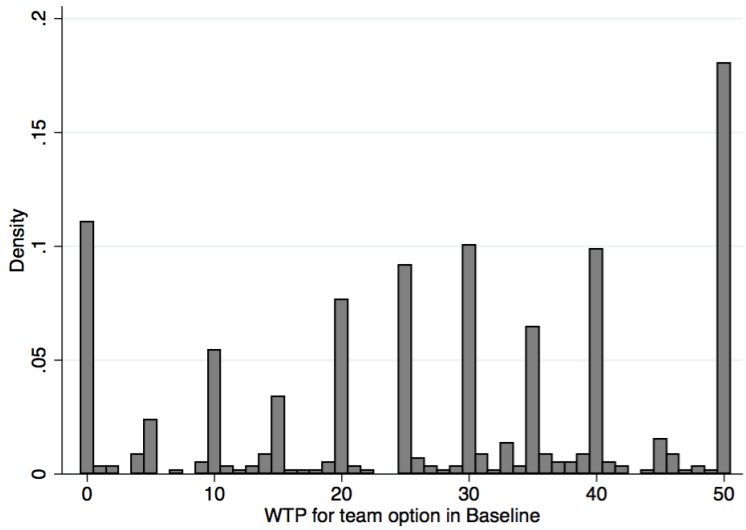


Figure 4.1: Distribution of bids for the team option in the Baseline

Table 4.4 displays pairwise correlation coefficients between participants' willingness to pay for the team option and variables we suspect might influence the bids: actual and guessed number of correct answers, the partner's estimated performance, risk aversion, gender, age, education and occupation categories.[21] We find that only a few of these factors are significantly correlated with preferences for team pay: beliefs about the potential teammate's performance are positively associated with team bids, while higher risk aversion corresponds to a lower willingness to pay. Among the demographic characteristics, age and education show up significant. At the same time, many of these covariates are highly significantly correlated with each other. We therefore continue to analyze each of these factors to estimate their impact on team preferences in isolation.

We first focus on the effect of task performance on sorting into the team pay option. Table 4.4 suggests that neither actual nor guessed performance is significantly related to the team bids. This result is puzzling given that low-performing participants have much more to gain from the team option. Figure 4.2 illustrates this point by plotting participants' actual and *optimal* bids against their true (Panel A) and guessed (Panel B) number of correct answers. The optimal bids

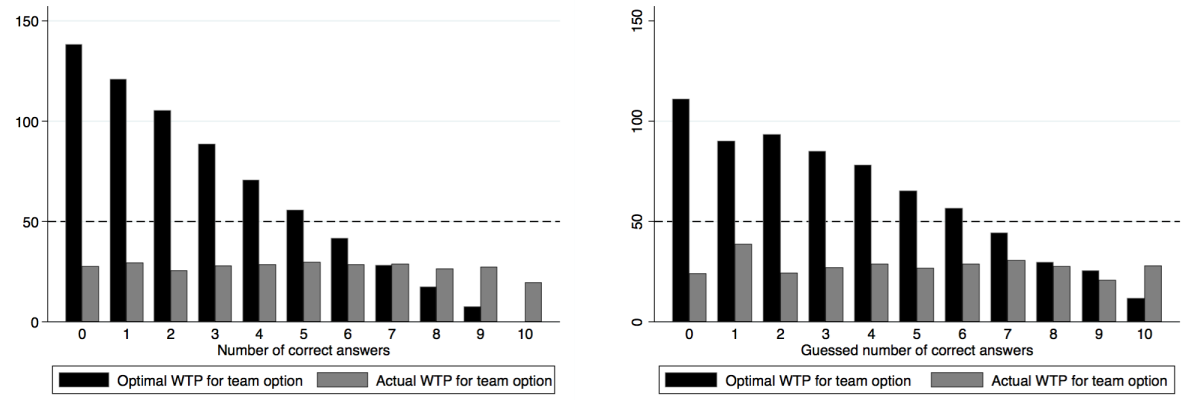
[21] In our survey we have also measured participants' work experience (in years). This measure is highly correlated with age, and is missing for 34 respondents, so we decided not to include it in our analysis and focus only on age. All the results presented in this section are robust to including experience instead of age in the analyses.

Table 4.4: Cross-correlation table

	WTP for team	Puzzles correct (actual)	Puzzles correct (guess)	Partner's correct (guess)	Risk aversion	Female	Age	Education	Entrepreneur	Manager
WTP for team	1.000									
Puzzles correct (actual)	-0.014	1.000								
Puzzles correct (guess)	-0.022	0.614***	1.000							
Partner's correct (guess)	0.162***	0.204***	0.444***	1.000						
Risk aversion	-0.227***	-0.040	-0.102**	-0.056	1.000					
Female	-0.054	-0.010	-0.111***	0.021	0.156***	1.000				
Age	0.079*	-0.156***	-0.050	-0.002	-0.042	-0.101**	1.000			
Education	0.132**	0.332***	0.200***	0.018	-0.082**	-0.021	0.032	1.000		
Entrepreneur	0.065	0.083**	0.153***	0.032	-0.134***	-0.217***	0.276***	0.173***	1.000	
Manager	0.023	0.205***	0.122***	0.050	-0.067	-0.016	0.035	0.255***	-0.285***	1.000

Notes: The table displays pairwise correlation coefficients between all variables. Significance levels indicated by stars: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

represent the *ex ante* expected gains from the team option for a perfectly informed participant.[22] Unsurprisingly, the optimal bids are highest for low-performing respondents and decrease with the number of correct puzzles. Actual bids, however, do not follow the same pattern: participants with a low (guessed) score do not pay more for the chance to be teamed up. As a result, we find no evidence for adverse selection in our setting: the sorting decision does not seem related to performance.



(a) Actual (b) Guessed

Figure 4.2: Optimal vs. actual team bids by actual and guessed number of correct puzzles

To separate the impact of true performance and confidence, we turn to a regression framework. Table 4.5 reports results from tobit models explaining the willingness to pay for the team option in the Baseline treatment. Column (1) confirms the finding that the actual number of correct answers does not affect the decision to join a team. It also shows that given true performance, confidence in ability matters: participants who guess they solved more puzzles correctly bid significantly less for the team option, *ceteris paribus*. Relative performance beliefs are also important determinants of the team choice: those respondents who expect their potential partner to answer more questions correctly bid more. A unit increase in the partner's estimated score is associated with an approx. €3 increase in the predicted willingness to pay, corresponding to about one fifth of a standard deviation.

[22]Optimal bids account for synergy possibilities: expected gains are calculated by assessing for each puzzle and for each potential teammate whether the given question was solved correctly by the teammate but not by the participant herself. Optimal bids are constructed assuming everyone who bids €25 or higher is a potential teammate. Weights are used to ensure that the likelihood of being teamed up with a manager, entrepreneur or employee is the same.

Table 4.5: WILLINGNESS TO PAY FOR TEAM OPTION IN THE BASELINE TREATMENT

	Performance	Risk	Demographics	Occupation
WTP for team	(1)	(2)	(3)	(4)
Puzzles correct (actual)	0.335	0.412	0.055	0.091
	(0.514)	(0.499)	(0.520)	(0.525)
Puzzles correct (guess)	-1.427**	-1.737***	-1.835***	-1.889***
	(0.631)	(0.615)	(0.615)	(0.619)
Partner's correct (guess)	3.014***	3.023***	3.190***	3.097***
	(0.702)	(0.683)	(0.680)	(0.682)
Risk aversion		-0.186***	-0.173***	-0.176***
		(0.033)	(0.033)	(0.033)
Female			-0.867	-0.847
			(1.851)	(1.916)
Age			0.082	0.092
			(0.080)	(0.086)
Education			3.175***	3.319***
			(0.961)	(1.012)
Entrepreneur				1.349
				(2.392)
Manager				-0.090
				(3.148)
Income categories				✓
Constant	18.623***	30.054***	18.501***	18.005***
	(3.848)	(4.247)	(6.189)	(6.759)
N	588	588	588	588
Pseudo-R^2	0.004	0.012	0.015	0.016

Notes: The table displays estimated coefficients from tobit models (lower limit 0, upper limit 50). Dependent variable: WTP for team option in Baseline treatment. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Previous studies identified risk aversion as an important factor in the sorting decision into teams. According to Table 4.4, risk preferences are strongly correlated with the team bids also in our setting. Column (2) of Table 4.5 shows that this result is robust to controlling for performance and confidence: more risk averse participants have a significantly lower willingness to pay for the team option. The effect is sizable: all else equal, participants who invest zero in the risky bet are predicted to bid on average €18.6 less for the team than those who choose to invest 100% of their earnings in the risky lottery in Part 2.[23] Our results suggest that players perceive

[23]Could the result that risk preferences are important determinants of team choice be merely a "side effect" of the particular elicitation method that we used to measure team preferences? As discussed in Section 4.3.2,

the team option as more risky than the individual piece rate. This is in line with the conclusions of Bäker and Mertins (2013) who find that the additional strategic risk associated with team pay matters more for the sorting decision than the reduction in the idiosyncratic components of risk.

One of the goals of our study is to explore how personal characteristics such as age, gender and level of education affect preferences to join teams. We therefore add these variables to our model explaining team choices (controlling for performance, beliefs and risk preferences). The resulting estimates are presented in column (3) of Table 4.5. Age does not seem to affect team bids: the marginally significant negative correlation observed in Table 4.4 disappears when we include other covariates in the regression.

Ex ante predictions for the impact of gender are ambiguous. On the one hand, women are less confident: even though there is no gender difference in actual task performance, female participants' guesses for the number of puzzles they solved correctly is lower (see Table 4.4). This lower confidence should increase their bids for the team option. On the other hand, their higher risk aversion (see Table B1) is predicted to decrease their willingness to participate in teams. Indeed, the raw correlation we observe between gender and team bids in Table 4.4 is insignificant. This is in line with the findings of Kuhn and Villeval (2014) who show no gender gap in sorting into teams in the presence of efficiency advantages associated with the team option. When we control for relative performance beliefs and risk aversion, we also find no gender difference in team bids (column (3) of Table 4.5), suggesting that in our sample female respondents do not differ from men in their inherent 'taste' for team remuneration schemes.

Existing research on sorting into teams has not analyzed the influence of education, a factor we find has a substantial influence on sorting into teams. Column (3) of Table 4.5 shows that above and beyond its impact on task performance, education has a significant, large and

participants received €50 which they could use to bid for the team option. The design thus involved a choice between keeping the "safe endowment" and investing (some share of) it in the risky team option. In a pilot study we compared this calibration with a different elicitation technique where respondents received no endowment but had to use their survey earnings to bid for the team option. There was no significant difference in bids between the two calibrations in the pilot survey. Moreover, risk aversion was found to be an important predictor of team choice in the "no-endowment" version as well. We thus believe the result that risk aversion matters substantially for sorting into teams is a general finding and is not driven by the specifics of our design.

positive effect on the willingness to pay for the team option.[24] Column (4) confirms that this effect is not driven by the correlation between education and occupational categories: including indicator variables for entrepreneurs and managers, the estimated coefficient for education remains virtually unchanged. It is interesting to note that entrepreneurs and managers do not seem to differ from employees in their team preferences once we control for differences in ability, beliefs and demographic characteristics.[25] Self-reported explanations do not indicate a larger desire for control among the self-employed: entrepreneurs in our sample were not any more likely than non-entrepreneurs to select the statement "I wanted to be responsible for my earnings and not depend on others" in the background questionnaire. This is in contrast with the findings of Cooper and Saral (2013) whose results indicate that self-employed have a stronger preference to work alone than others. Column (4) further shows that the positive association between education and team bids is robust to controlling for income differences between the high- and low-educated.[26]

Lastly we explore in more detail the role that education plays in shaping preferences for teams. To gain some insight into the respondents' decision making process, we analyze in more detail the explanations they gave for their team bids. Figure 4.3 compares the frequencies with which respondents with different levels of education chose the explanations related to risks ("I did not want to take too much risk") or potential gains ("I believed the team option could increase my earnings"). The figure indicates a pattern: the higher a respondent's level of education, the less likely she is to mention the former and the more likely she is to select the latter explanation. (The differences in frequencies are (marginally) significant when we compare respondents with either a college or university degree to those with primary/secondary or vocational education.) This finding suggests that higher educated participants are more likely to

[24]Results are very similar when instead of estimating a linear effect of education we include indicator variables for the different levels of education.

[25]We find no difference in the willingness to pay for the team option when, instead of the broad occupational categories, we make a distinction between founders and entrepreneurs who have taken over/inherited their companies, or 'people managers' and project managers.

[26]Note that only 381 participants in the Baseline treatment reported their income category. In the regression presented in Table 4.5, we treated non-respondents as the omitted category. Our results are similar when we focus only on the subsample who disclosed their income.

base their choice on the possible gains from teamwork, whereas the lower educated individuals tend to focus more on risk considerations. Results in Table B2 in Appendix 4.B support this interpretation by showing that the impact of risk aversion on team bids is heterogeneous with education. While risk preferences have a large impact on the willingness to pay for teams for respondents with primary/secondary education, they seem not to affect the choices of participants with a university degree at all.[27] As a result, participants with a lower level of education - who on average perform worse on the task than those with more education - miss out on the sizable potential benefits that the team option entails for them.

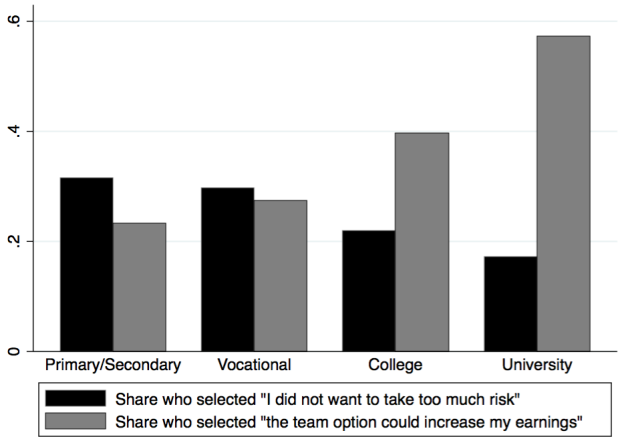


Figure 4.3: Self-reported explanations for team bids by education levels

4.5.2 Comparison of team choice between the treatments

In this section we analyze the effect of joint decision making on respondents' willingness to pay for the team option by contrasting the bids between the two treatment conditions: the Baseline where the team option only entails joint production and the Joint Decision treatment where the team option affects both the earnings from the real-effort task and the investment decision.

[27]To exclude the possibility that our findings are driven by lower educated people being confused about the specifics of the team option or the bidding procedure, we re-estimate the model presented in column (1) of Table B2 on the subsample of participants in the Baseline treatment who are most likely to have understood the scheme. In particular, we omit those participants who selected "It was just a guess" as an explanation for their bids at the end of the survey (we have 79 such participants in the Baseline treatment). In the resulting subsample we replicate the finding that the impact of risk is heterogeneous with respect to education. We therefore argue that misunderstanding is unlikely to drive the different choices of low- and high-educated participants.

Figure 4.4 depicts the distribution of bids in both treatments. In the Joint Decision treatment, respondents bid on average 26.91 (st. dev. 15.83) which is only slightly lower than the mean bid of 28.00 in the Baseline. A simple comparison by means of a t-test shows no difference between the two treatments (p-value$= 0.248$). A Kolmogorov-Smirnov test does not reject the equality of the two distributions, either (p-value$= 0.187$).

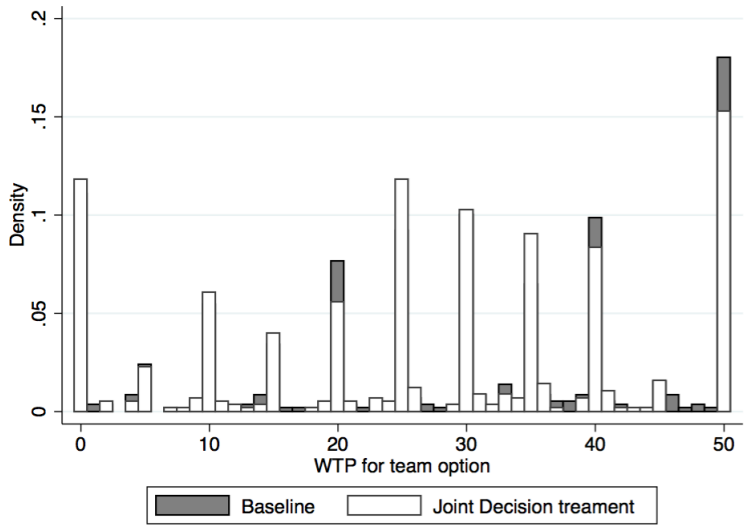


Figure 4.4: Comparison of WTP betwen the two treatments

Table 4.6 analyzes the impact of joint choice on team bids in a regression framework. Column (1) shows that even after controlling for potential differences in performance, beliefs, preferences and personal characteristics between the two groups, the bids for teams are not significantly different in the Joint Decision treatment. The fact that the inclusion of covariates does not change our results is unsurprising given that participants were randomly assigned to the treatment conditions and the two groups were balanced in terms of observables, as shown in Table 4.3.

Having found no *overall* effect of the treatment conditions, we check whether the response to joint decision making depends on respondents' guesses about the instrumental value of the right to decide. In particular, we test whether those respondents who predict a large gap between their own and their teammate's investment choice (and consequently expect the joint choice to

be far away from their own) bid less in the Joint Decision treatment. Column (2) of Table 4.6 shows that this is indeed the case: the greater the predicted (absolute) difference in risk taking between the teammates, the less appealing the team option in the Joint Decision treatment is compared to the Baseline. Respondents in our sample thus have a clear preference for keeping the decision rights in cases where they believe the team's choice would be different than their individual optimum.

Table 4.6: Comparison of the treatments

	Full sample		Entrepreneurs	Managers	Employees
WTP for team	(1)	(2)	(3)	(4)	(5)
Puzzles correct (actual)	0.547	0.515	0.656	1.855	0.327
	(0.354)	(0.353)	(0.640)	(1.290)	(0.435)
Puzzles correct (guess)	-2.244***	-2.152***	-0.371	-4.223***	-2.649***
	(0.445)	(0.445)	(0.865)	(1.539)	(0.539)
Partner's correct (guess)	2.635***	2.590***	0.987	2.960	3.268***
	(0.489)	(0.489)	(0.971)	(1.809)	(0.577)
Risk aversion	-0.189***	-0.191***	-0.233***	-0.004	-0.222***
	(0.024)	(0.024)	(0.044)	(0.066)	(0.032)
Female	-1.571	-1.650	-0.687	0.673	-2.594
	(1.308)	(1.305)	(2.654)	(3.981)	(1.579)
Age	0.089	0.086	0.263**	0.187	0.019
	(0.059)	(0.059)	(0.120)	(0.229)	(0.067)
Education	1.240*	1.298*	1.271	0.975	1.002
	(0.695)	(0.693)	(1.262)	(2.628)	(0.852)
Entrepreneur	0.461	0.519			
	(1.528)	(1.523)			
Manager	-0.753	-0.711			
	(2.063)	(2.057)			
Joint Decision treatment	-1.289	1.932	3.944	-1.502	1.508
	(1.229)	(1.810)	(3.388)	(5.607)	(2.259)
Abs. diff. RA		0.028	0.086	-0.075	-0.008
		(0.047)	(0.085)	(0.137)	(0.061)
Joint Decision * Abs. diff. RA		-0.171**	-0.304**	-0.025	-0.110
		(0.069)	(0.125)	(0.197)	(0.090)
Constant	27.504***	26.901***	16.908*	17.870	32.975***
	(4.508)	(4.670)	(10.180)	(17.892)	(5.585)
N	1163	1163	400	155	608
Pseudo-R^2	0.014	0.015	0.015	0.009	0.022

Notes: The table displays estimated coefficients from tobit models (lower limit 0, upper limit 50). Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Inspired by studies claiming that entrepreneurs have a particularly strong need for power, authority and control, we analyze whether including joint decision making in the team option has a particularly strong negative impact for entrepreneurs. In columns (4)-(6) of Table 4.6 we assess the impact of the Joint Decision treatment condition and its interaction with the predicted difference between investment choices separately for the three occupational categories. We find that the result of column (2) is driven entirely by the entrepreneurs in our sample: employees and managers, irrespective of their beliefs about their partner's risk preferences, do not differ in their willingness to pay for the team option between the two treatments. Entrepreneurs, on the other hand, respond to a predicted gap between their own and their teammate's investment choice by placing significantly lower bids in the Joint Decision treatment than in the Baseline.

4.6 Summary and conclusion

Our study reports results from a large-scale lab-in-the-field experiment analyzing the sorting decision into teams. We replicate several findings from related lab studies, using a larger and more diverse subject pool and a different modeling of teamwork. We confirm that absolute and relative confidence are important determinants of the willingness to pay for the team option (Herbst et al., 2015; Kuhn and Villeval, 2014) also in a setting where team production is not implemented as an equal split of the pooled output but entails potential synergy gains. We further confirm that strategic risk resulting from uncertainties about the teammate's performance has a large impact on the selection into teams (Bäker and Mertins, 2013) even when the possibility of free riding is eliminated by design. Moreover, we show that the result of no gender gap in team choices in the presence of efficiency gains (Kuhn and Villeval, 2014) persists also in a setting where such gains are not automatic but are conditional on the complementarities between the teammates.

Our results highlight that selection into teams is related to participants' level of education, and that entrepreneurs respond differently to joint decision making than managers and employ-

ees. These findings would not have been possible to obtain with a sample of college students typically participating in laboratory experiments. Thus, our paper demonstrates the added value of studying a 'non-traditional' sample, i.e. participants with diverse socio-economic backgrounds. In this aspect our study is similar to the work of e.g. Harbaugh et al. (2002) who analyze risk preferences in different age groups or Sutter and Kocher (2007) who focus on the relationship between age and trust.

A novelty of our paper is the finding that education is an important predictor for individuals' participation decision in teams. Controlling for differences in task performance (which can be viewed as a proxy for IQ), confidence and risk preferences, higher education is associated with a greater willingness to pay for the team option. We find suggestive evidence that this heterogeneity is explained by differences in evaluating the team option: while participants with higher levels of education tend to primarily consider the potential gains from team pay, lower-educated respondents focus more on the risks associated with teams. As a consequence, lower educated people miss out on the sizable efficiency gains that the team option entails. It is important to note that education does not affect risk preferences: educational attainment is uncorrelated with the share of earnings participants invest in the risky bet. It is the extent to which risk preferences matter for the selection into teams that is affected by education: the same level of risk aversion leads to a greater reduction in bids for the team option for the low- than the high-educated respondents in our sample. These inferences are consistent with the conclusions from the Perry Preschool program where education (in the form of an early childhood intervention) did not have a lasting effect on IQ but it still lead to improved life outcomes through its impact on non-cognitive skills (Heckman et al., 2013). Our conjectures also bring into mind the results of Choi et al. (2014) who find higher educated people make 'higher quality', i.e. more rational decisions than those with lower levels of education. We wish to emphasize that our data does not allow us tell whether education changes individuals' decision making process or whether those who obtain higher education are innately different from others in the way they evaluate situations involving strategic risk and synergy gains. We

find this question to represent an interesting avenue for future research.

Our study also assesses the influence of shared decision rights on the self-selection into team incentive schemes. We show that respondents are heterogeneous in their response to a potential compromise in decision making. In particular, we find that entrepreneurs are averse to joint decision making when they predict that it moves them away from their individual optimal choice. This results supports claims that entrepreneurs have a greater desire for control than non-entrepreneurs (Benz and Frey, 2008; Masclet et al., 2009; Reynolds and Curtin, 2008). We do not reproduce the finding that people attach a non-pecuniary value to decision rights *per se* (Bartling et al., 2014; Fehr et al., 2013): even among entrepreneurs, shared decisions only decrease the willingness to join teams in cases when the partner's choice is expected to deviate substantially from one's own optimum. Those who predict no difference between their own and their partner's investment decision bid the same for the team option in both treatments. We speculate that this is due to decisions being *shared*, not delegated. In our setting, instead of an obvious loss of power, a compromise is implemented, and the resulting team choice is still strongly influenced by each member's individual choice.

Appendix

4.A Excerpts from the online survey

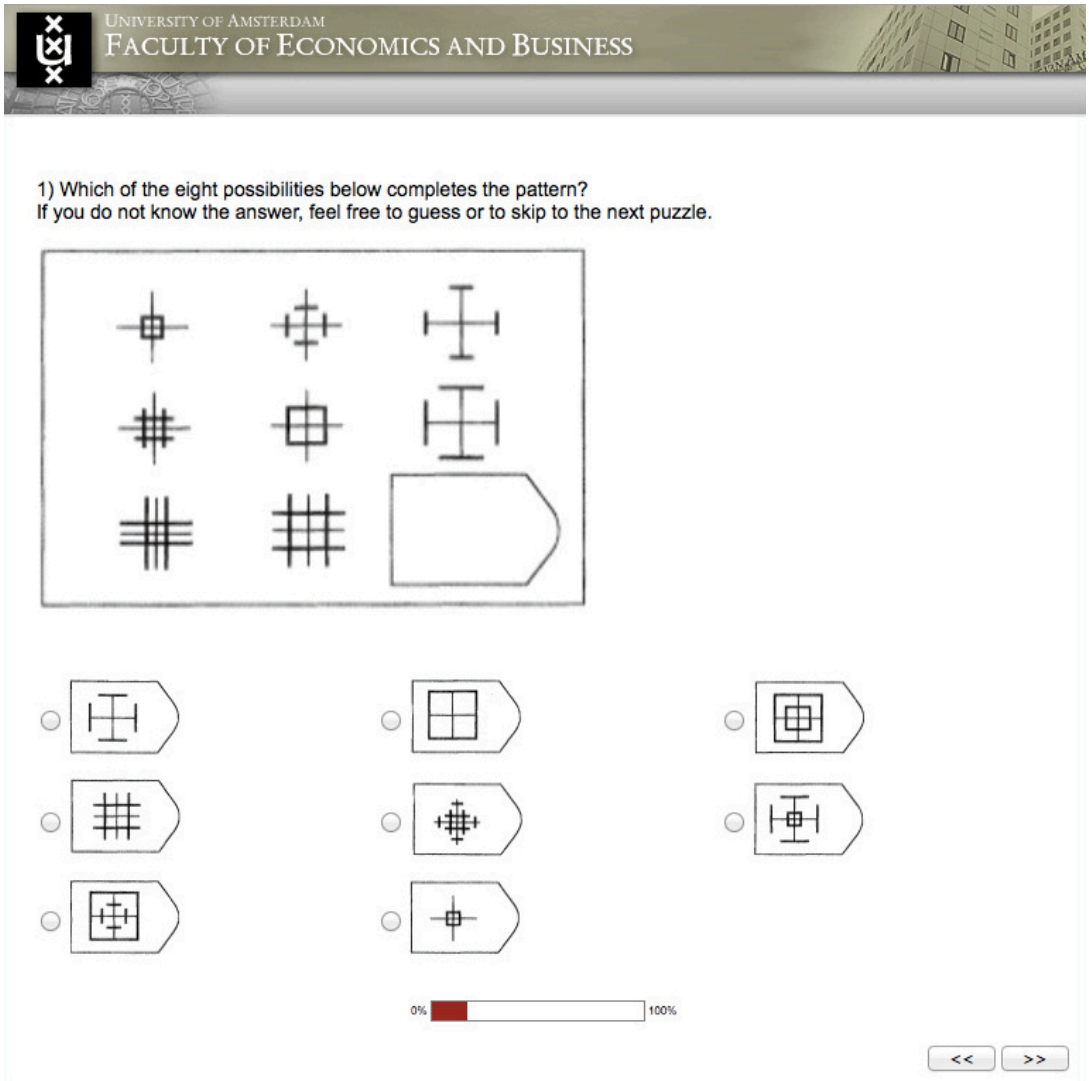


Figure A1: Example of a Raven puzzle

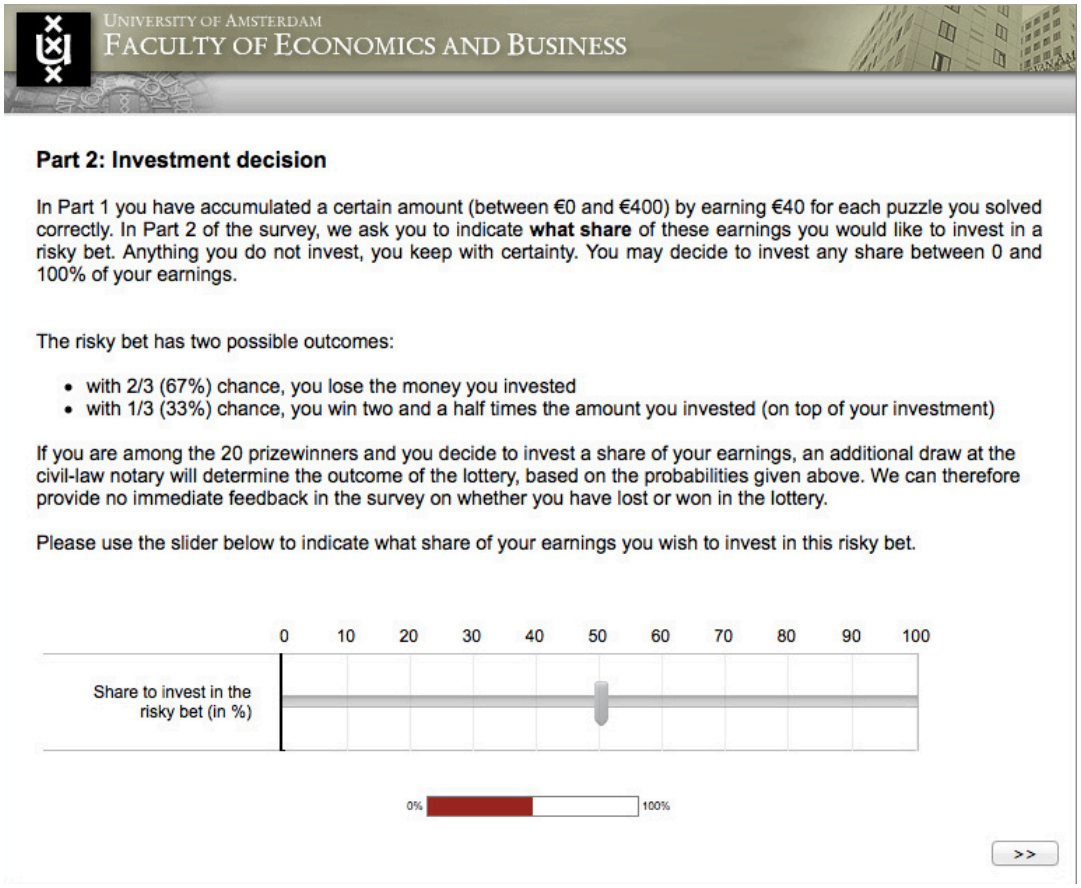


Figure A2: Measuring risk aversion

Example:

(Please note that the numbers in the example below are hypothetical and convey no information about the actual performance/choices of other respondents.)

- Imagine you solved puzzles 1 and 2 correctly (Part 1), and chose to invest 60% (Part 2).
- Participant B solved puzzles 2 and 3 correctly (Part 1), and chose to invest 20% (Part 2).
- Participant C solved puzzle 1 correctly (Part 1), and chose to invest 80% (Part 2).

The table below shows **your outcomes**, depending on your choice (team/individual option) and your randomly assigned partner (either B or C).

Your outcomes	**Individual**	**In a team**	
		with Participant B	with Participant C
# correct puzzles	2	3	2
Earnings	2*€40	3*€40	2*€40
Investment (share of earnings)	60%	60%	60%

(a) Baseline

Example:

(Please note that the numbers in the example below are hypothetical and convey no information about the actual performance/choices of other respondents.)

- Imagine you solved puzzles 1 and 2 correctly (Part 1), and chose to invest 60% (Part 2).
- Participant B solved puzzles 2 and 3 correctly (Part 1), and chose to invest 20% (Part 2).
- Participant C solved puzzle 1 correctly (Part 1), and chose to invest 80% (Part 2).

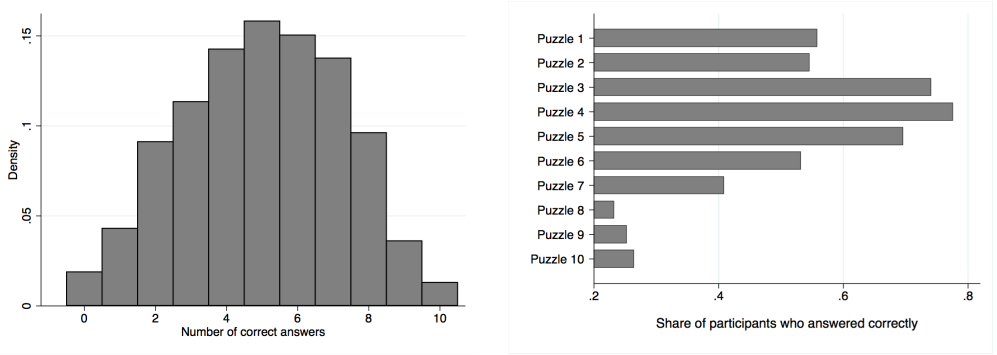
The table below shows **your outcomes**, depending on your choice (team/individual option) and your randomly assigned partner (either B or C).

Outcomes of Participant A	**Individual**	**In a team**	
		with Participant B	with Participant C
# correct puzzles	2	3	2
Earnings	2*€40	3*€40	2*€40
Investment (share of earnings)	60%	40%	70%

(b) Joint Decision treatment

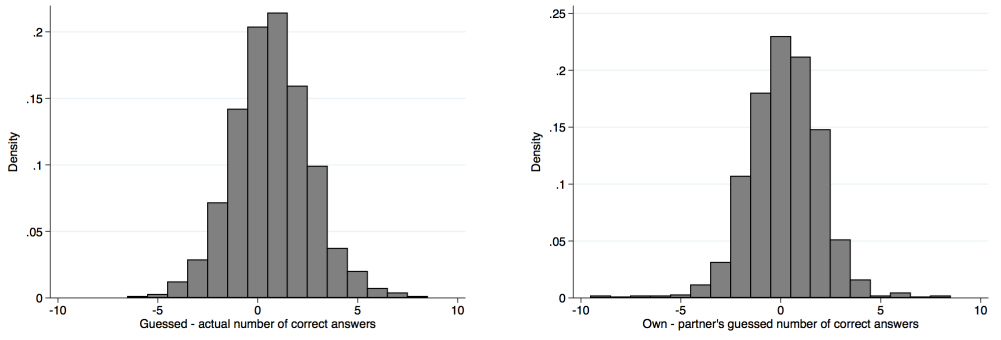
Figure A3: Explaining the team option in the survey

4.B Additional figures and tables



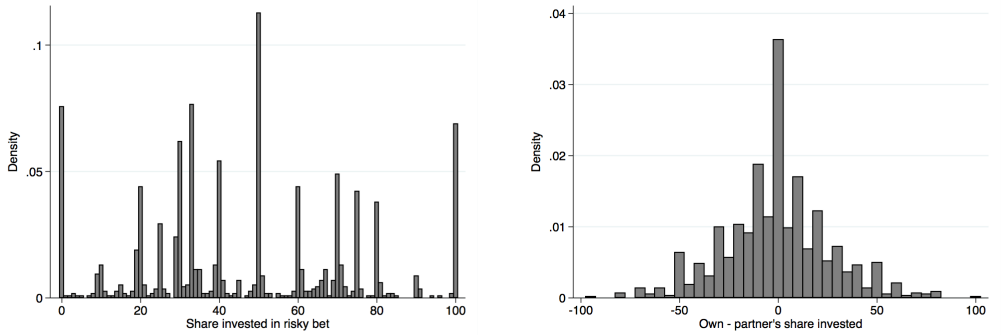
(a) Number correct per player (b) Share correct per puzzle

Figure B1: Distribution of correct answers on the Raven puzzles



(a) Absolute - actual performance (b) Own - partner's guessed performance

Figure B2: Absolute and relative performance guesses



(a) Investment choices (b) Estimated difference

Figure B3: Investment in the risky gamble

Table B1: Explaining task performance and risk aversion

	(1) Puzzles correct (actual)	(2) Risk aversion
Female	0.029	5.420***
	(0.130)	(1.631)
Age	-0.035***	-0.039
	(0.006)	(0.074)
Education	0.496***	-1.207
	(0.068)	(0.871)
Entrepreneur	0.704***	-9.558***
	(0.150)	(1.897)
Manager	1.209***	-9.594***
	(0.203)	(2.575)
Puzzles correct (actual)		0.441
		(0.444)
Puzzles correct (guess)		-0.772
		(0.551)
Partner's correct (guess)		-0.508
		(0.603)
Constant	4.834***	66.309***
	(0.333)	(5.187)
N	1163	1163
Adjusted R^2		

Notes: the table displays estimated coefficients from OLS models. Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table B2: The role of education

WTP for team	Full sample (1)	Primary/Secondary (2)	Vocational (3)	College (4)	University (5)
Puzzles correct (actual)	0.081	1.101	-0.261	-0.631	1.175
	(0.520)	(1.223)	(0.751)	(0.803)	(1.718)
Puzzles correct (guess)	-1.904***	-0.798	-0.459	-1.647*	-4.504***
	(0.614)	(1.645)	(0.996)	(0.952)	(1.673)
Partner's correct(guess)	3.269***	4.838***	2.104*	3.376***	2.838
	(0.677)	(1.472)	(1.119)	(1.073)	(1.864)
Risk aversion	-0.446***	-0.274***	-0.217***	-0.211***	-0.020
	(0.101)	(0.090)	(0.053)	(0.055)	(0.077)
Female	-0.553	-5.311	-2.467	-2.634	7.818*
	(1.873)	(4.243)	(3.112)	(3.140)	(4.584)
Age	0.075	-0.161	0.188	0.091	0.123
	(0.084)	(0.206)	(0.125)	(0.137)	(0.232)
Entrepreneur	0.426	1.699	-3.331	-0.130	3.287
	(2.180)	(5.908)	(3.620)	(3.297)	(6.001)
Manager	-1.307	-1.478	-7.278	0.096	-2.935
	(2.915)	(11.392)	(9.112)	(4.546)	(6.211)
Education	-2.069				
	(2.092)				
Education * RA	0.095***				
	(0.033)				
Constant	33.772***	22.690*	25.075***	30.498***	31.087*
	(8.227)	(13.146)	(9.227)	(9.602)	(16.066)
N	588	73	128	242	145
Pseudo-R^2	0.017	0.053	0.026	0.017	0.016

Notes: the table displays estimated coefficients from tobit models (lower limit 0, upper limit 50). Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Chapter 5

Summary

This dissertation consists of a series of studies that explore individual differences in response to incentives. Chapter 2 and 3 focus on gender differences in reaction to tournaments, while Chapter 4 concentrates on other dimensions of heterogeneity such as education and occupational categories in relation to sorting into teams.

The aim of Chapter 2 is to study in a natural yet controlled environment the impact of absolute vs. relative grading on student effort and performance, with a focus on potential gender differences in reaction to the grading schemes. To this end, a large-scale framed field experiment is conducted among students following a Bachelor course at a Dutch university. Students are randomly assigned to either absolute or relative grading at the midterm exam (grading schemes are reversed for the final exam). The random assignment of students to grading schemes allows causal inferences about the effect of grading schemes on student outcomes, and a clean estimation of the potential gender differences in response to competitive grade incentives. The data collected in the experiment include exam scores, different proxies for preparation effort as well as a rich set of individual characteristics such as demographic information, preferences, confidence and ability measures.

The results show no clear difference either in effort provision or exam performance under the two grading schemes. Even though the exams represent high stakes, and comprehension of the treatment conditions is adequate, students in this sample do not react differently to grading

on the curve than to absolute grading. There is no evidence for a heterogeneity in response by gender, ability or competitive preferences, either. These findings are likely explained by an overall lack of ambition and a general disinterest in obtaining high grades, a manifestation of the often criticized "just-pass" attitude of Dutch students. This explanation is supported by the analysis of those students who are conjectured to care most about the way their exam scores are mapped into grades, i.e. students close to the pass-fail threshold. While this subgroup is relatively small, so inferences should be made with caution, there is a strong indication for a gender gap in response to relative grading among such "marginal" students. Overall, the results of this chapter suggest that competitive grade incentives can not solve the issue of insufficient student motivation: a tournament is ineffective when the prizes (i.e. high grades) are not considered valuable by the majority of the contestants.

Chapter 3 exploits a large set of naturally occurring data from an online card-game community to study gender differences in risk taking and competitive choices. The data contain information from over 4 million games, generated by more than fifteen thousand individual players. While the setting is simplified, it nonetheless has several appealing features, such as self-selection into an uncertain and competitive environment as well as real, strategic and repeated interaction (with feedback) between the players. The data allow us to make a distinction between what we call 'selection' and 'playing' behavior: we observe players' choices regarding the level of risk and competition they prefer to bear in each round, and also their subsequent playing performance (i.e. their ability to win tricks) in the resulting tournaments. Players' performance is measured exactly and depends on skill, strategic behavior and luck.

The data show that even though players sort into the community to participate in a game that effectively consists of a series of tournaments, this self-selection does not close the gender gap in subsequent choices related to risk taking and competition. Female players are still less likely than males to increase the stakes of the given round or to actively initiate games. The gap in initiating is largest for the game types involving individual (as opposed to team) competition. As a result of their selection choices, women end up more often in the difficult opponent po-

sition, and even when they initiate and win games, their earnings are lower due to the smaller stakes. Consequently, women accumulate lower scores in the game than men do. This gender gap in scores is not a reflection of differences in ability in the playing stage. Controlling for the type of the game and their role, women are as good as men in winning games. This chapter thus demonstrates the negative consequence of "shying away": despite no gender differences in on-task performance, women end up lagging behind men because of their lower propensity to take risk and to launch tournaments. The data further suggests that female players' differential choices are to a large extent attributable to gender differences in risk and competitive preferences.

Chapter 4 analyzes the determinants of sorting into teams based on results from a large-scale lab-in-the-field (i.e. artefactual field) experiment: an incentivized online survey among Dutch entrepreneurs, managers and employees. The sample of respondents consists of more than one thousand individuals and is diverse in terms of gender, age, experience, level of education and income. The survey features a real-effort task (i.e. solving Raven matrices (Raven et al., 2003)) and measures participants' risk preferences, confidence and beliefs. In the key part of the survey, participants' willingness to form a team with another respondent is elicited by means of an incentive-compatible mechanism. Respondents are randomly assigned to one of the two treatments that affect the content of the team option. In the Baseline condition, the team option only entails joint production, while in the Joint Decision treatment it includes *both* joint production and a joint investment choice. In both treatments, earnings from the real effort task under the team option are determined by a production function that recognizes complementarities between the partners, and thus allows for but does not guarantee efficiency gains.

Results presented in Chapter 4 replicate several findings from related lab studies on sorting into team incentive schemes. Absolute and relative confidence are confirmed to be important determinants of the willingness to pay for the team option also in a setting where team production is not implemented as an equal split of the pooled output but entails potential synergy

gains. It is further shown that strategic risk resulting from uncertainties about the teammate's performance has a large impact on the selection into teams even when the possibility of free riding is eliminated by design. A novel finding presented in this chapter is the importance of education for individuals' participation decision in teams. Controlling for differences in task performance, confidence and risk preferences, higher education is associated with a greater willingness to pay for the team option. There is suggestive evidence that this heterogeneity is explained by differences in evaluating the team option: while participants with higher levels of education tend to primarily consider the potential gains from team pay, lower-educated respondents focus more on the risks associated with teams. As a consequence, lower educated people miss out on the sizable efficiency gains that the team option entails. Finally, Chapter 4 of this dissertation finds that participants are heterogeneous in their response to a potential compromise in decision making. In particular, entrepreneurs are shown to be averse to joint decision making when they predict that it moves them away from their individual optimal choice, while no such effect is observed for managers or employees.

Based on the three studies discussed in this dissertation we can conclude that there are systematic differences between individuals' reactions to incentives, and such heterogeneities have economically important consequences. The first two chapters discuss gender differences in response to tournament incentives. Chapter 2 presents suggestive evidence that among "marginal" students (who are conjectured to care about grade incentives) competitive grading increases the exam scores of male students but decreases the performance of female students. Chapter 3 demonstrates persistent differences between the choices of female and male card game players in relation to risk taking and competition, and also the negative consequences of these differences for women's relative success in terms of accumulated scores. Finally, Chapter 4 suggests that lower-educated individuals tend to focus more strongly on the risks associated with team incentive schemes than those with higher education, and as a result, miss out on the sizable potential gains from cooperation. Chapter 4 also confirms that entrepreneurs are different from the rest of the population in their approach to shared decision rights. Taken together, these

findings imply that observable individual characteristics can be helpful in predicting people's responses to certain incentives, and taking them into account when designing incentive schemes could increase efficiency. However, we should caution against interpreting our results as causal: observable individual characteristics do not necessarily induce choices, rather they have the potential to serve as proxies for the unobservable traits, attitudes and beliefs that are the real drivers of behavior.

It is interesting to note that some important results of the dissertation would not have been possible to obtain in a lab setting. In Chapter 3 the large dataset, both in terms of the number of individual players and the number of observations per player, allows us to detect small effect sizes and heterogeneous effects. It also enables us to evaluate the long-term consequences of seemingly small and innocent differences each round. This kind of precise measurement would be difficult to acquire in a typical laboratory experiment with a limited number of participants. Additionally, the findings of Chapter 4 that selection into teams is related to participants' level of education, and that entrepreneurs respond differently to joint decision making than managers and employees would not have been possible to observe with a sample of college students typically participating in laboratory experiments. This dissertation thus highlights the importance of complementing laboratory experiments with scientific projects based on field data.

Bibliography

Adams, R. B. and Funk, P. (2012). Beyond the glass ceiling: Does gender matter? *Management Science*, 58(2):219–235.

Altonji, J. G. and Blank, R. M. (1999). Race and gender in the labor market. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 3, chapter 48, pages 3143–3259. Elsevier.

Ariely, D., Gneezy, U., Loewenstein, G., and Mazar, N. (2009). Large stakes and big mistakes. *The Review of Economic Studies*, 76(2):451–469.

Azmat, G., Calsamiglia, C., and Iriberri, N. (2014). Gender differences in response to big stakes. LSE Research Online Documents on Economics 60607, London School of Economics and Political Science, LSE Library.

Azmat, G. and Iriberri, N. (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, 94(7-8):435–452.

Azmat, G. and Petrongolo, B. (2014). Gender and the labor market: What have we learned from field and lab experiments? *Labour Economics*, 30:32 – 40.

Bäker, A. and Mertins, V. (2013). Risk-sorting and preference for team piece rates. *Journal of Economic Psychology*, 34:285–300.

Bandiera, O., Barankay, I., and Rasul, I. (2013). Team incentives: evidence from a firm level experiment. *Journal of the European Economic Association*, 11(5):1079–1114.

Barankay, I. (2012). Rank incentives: Evidence from a randomized workplace experiment. Working paper, University of Pennsylvania.

Barsh, J. and Yee, L. (2012). *Unlocking the full potential of women at work*. McKinsey & Company.

Bartling, B., Fehr, E., and Herz, H. (2014). The Intrinsic Value of Decision Rights. *Econometrica*, 82:2005–2039.

Becker, G. M., DeGroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral science*, 9(3):226–232.

Becker, W. E. and Rosen, S. (1992). The learning effect of assessment and evaluation in high school. *Economics of Education Review*, 11(2):107–118.

Benz, M. and Frey, B. S. (2008). The value of doing what you like: Evidence from the self-employed in 23 countries. *Journal of Economic Behavior & Organization*, 68(3-4):445–455.

Bertrand, M. (2011). New Perspectives on Gender. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 4, chapter 17, pages 1543–1590. Elsevier.

Bigoni, M., Fort, M., Nardotto, M., and Reggiani, T. (forthcoming). Cooperation or competition? A field experiment on non-monetary learning incentives. *The B.E. Journal of Economic Analysis & Policy*.

Brandts, J., Groenert, V., and Rott, C. (2015). The impact of advice on women's and men's selection into competition. *Management Science*, 61(5):1018–1035.

Brennan, J., Patel, K., and Tang, W. (2009). Diversity in the student learning experience and time devoted to study: a comparative analysis of the UK and European evidence. Report to HEFCE by Centre for Higher Education Research and Information, The Open University.

Budryk, Z. (2013). Dangerous curves. *Inside Higher Ed*, 12 February.

Bull, C., Schotter, A., and Weigelt, K. (1987). Tournaments and Piece Rates: An Experimental Study. *Journal of Political Economy*, 95(1):1–33.

Burns, J. (2012). Race, diversity and pro-social behavior in a segmented society. *Journal of Economic Behavior & Organization*, 81(2):366–378.

Buser, T. (2012). The impact of the menstrual cycle and hormonal contraceptives on competitiveness. *Journal of Economic Behavior & Organization*, 83(1):1–10.

Buser, T. (forthcoming). The impact of losing in a competition on the willingness to seek further challenges. *Management Science*.

Buser, T., Niederle, M., and Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics*, 129(3):1409–1447.

Cameron, C. A. and Miller, D. L. (2015). A practitioners guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372.

Card, D., Cardoso, A., and Kline, P. (2015). Bargaining, Sorting, and the Gender Wage Gap: Quantifying the Impact of Firms on the Relative Pay of Women. NBER Working Paper 21403.

Carter, N. M. and Silva, C. (2010). Pipeline's broken promise. In *The Promise of Future Leadership: A Research Program on Highly Talented Employees in the Pipeline*. Catalyst Org.

Catalyst Org. (2015). Knowledge center: Women ceos of the sp500. `http://catalyst.org/knowledge/women-ceos-sp-500`. Accessed 30 January 2015.

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

Charness, G. and Gneezy, U. (2012). Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization*, 83(1):50–58.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):pp. 817–869.

Choi, S., Kariv, S., Müller, W., and Silverman, D. (2014). Who is (more) rational? *American Economic Review*, 104(6):1518–50.

Cooper, D. J. and Saral, K. J. (2013). Entrepreneurship and team participation: An experimental study. *European Economic Review*, 59:126–140.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–74.

Czibor, E., Onderstal, S., Sloof, R., and van Praag, M. C. (2014). Does relative grading help male students? Evidence from a field experiment in the classroom. IZA Discussion Paper 8429, Institute for the Study of Labor (IZA).

Danilov, A., Biemann, T., Kring, T., and Sliwka, D. (2013). The dark side of team incentives: Experimental evidence on advice quality from financial service professionals. *Journal of Economic Behavior & Organization*, 93(C):266–272.

Dargnies, M.-P. (2012). Men too sometimes shy away from competition: The case of team competition. *Management Science*, 58(11):1982–2000.

De Paola, M., Gioia, F., and Scoppa, V. (2015). Are females scared of competing with males? Results from a field experiment. *Economics of Education Review*, 48:117 – 128.

Delfgaauw, J., Dur, R., Sol, J., and Verbeke, W. (2013). Tournament incentives in the field: Gender differences in the workplace. *Journal of Labor Economics*, 31(2):305 – 326.

Desvaux, G., Devillard-Hoellinger, S., and Meaney, M. C. (2008). A business case for women. *The McKinsey Quarterly*.

Dohmen, T. and Falk, A. (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *The American Economic Review*, pages 556–590.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schüpp, J., and Wagner, G. G. (2011). Individual risk attitudes: measurement, determinants and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550.

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

Dubey, P. and Geanakoplos, J. (2010). Grading exams: 100,99,98,... or A,B,C? *Games and Economic Behavior*, 69(1):72–94.

Dur, R. and Sol, J. (2010). Social interaction, co-worker altruism, and incentives. *Games and Economic Behavior*, 69(2):293–301.

Eckel, C. C. and Füllbrunn, S. C. (2015). Thar SHE Blows? Gender, Competition, and Bubbles in Experimental Asset Markets. *American Economic Review*, 105(2):906–20.

Eckel, C. C. and Grossman, P. J. (2008). Differences in the Economic Decisions of Men and Women: Experimental Evidence. In Plott, C. R. and Smith, V. L., editors, *Handbook of Experimental Economics Results*, volume 1, chapter 57, pages 509–519. Elsevier.

European Commission (2012). *She Figures. Gender in Research and Innovation: Statistics and Indicators*. Luxembourg: Office for Official Publications of the European Communities.

Fehr, E., Herz, H., and Wilkening, T. (2013). The lure of authority: Motivation and incentive effects of power. *American Economic Review*, 103(4):1325–59.

Filippin, A. and Crosetto, P. (2014). A reconsideration of gender differences in risk attitudes. IZA Discussion Papers 8184, Institute for the Study of Labor (IZA).

Flory, J. A., Leibbrandt, A., and List, J. A. (2014). Do competitive workplaces deter female workers? A large-scale natural field experiment on job entry decisions. *The Review of Economic Studies*, 82(1):122–155.

Frank, R. H., Gilovich, T., and Regan, D. T. (1993). Does studying economics inhibit cooperation? *Journal of Economic Perspectives*, 7(2):159–171.

Garibaldi, P., Giavazzi, F., Ichino, A., and Rettore, E. (2012). College cost and time to complete a degree: Evidence from tuition discontinuities. *The Review of Economics and Statistics*, 94(3):699–711.

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

Gneezy, U., Niederle, M., and Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, 118(3):1049–1074.

Gneezy, U. and Potters, J. (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics*, pages pp. 631–645.

Gneezy, U. and Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, 115(3):791–810.

Gneezy, U. and Rustichini, A. (2004). Gender and competition at a young age. *American Economic Review*, 94(2):377–381.

Green, J. R. and Stokey, N. L. (1983). A comparison of tournaments and contracts. *Journal of Political Economy*, 91(3):349–64.

Grosse, N. D., Riener, G., and Dertwinkel-Kalt, M. (2014). Explaining gender differences in competitiveness: Testing a theory on gender-task stereotypes. Working paper, University of Mannheim.

Grove, W. A. and Wasserman, T. (2006). Incentives and student learning: A natural experiment with economics problem sets. *The American Economic Review*, 96(2):pp. 447–452.

Günther, C., Ekinci, N. A., Schwieren, C., and Strobel, M. (2010). Women can't jump? An experiment on competitive attitudes and stereotype threat. *Journal of Economic Behavior & Organization*, 75(3):395–401.

Hamilton, B. H., Nickerson, J. A., and Owan, H. (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *Journal of Political Economy*, 111(3):465–497.

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

Harbring, C. and Irlenbusch, B. (2003). An experimental study on tournament design. *Labour Economics*, 10(4):443–464.

Harrison, G. W. and List, J. A. (2004). Field experiments. *Journal of Economic Literature*, 42(4):1009–1055.

Healy, A. and Pate, J. (2011). Can teams help to close the gender competition gap? *The Economic Journal*, 121(555):1192–1204.

Heckman, J., Pinto, R., and Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, 103(6):2052–86.

Herbst, L., Konrad, K. A., and Morath, F. (2015). Endogenous group formation in experimental contests. *European Economic Review*, 74:163–189.

Herz, H., Schunk, D., and Zehnder, C. (2014). How do judgmental overconfidence and overoptimism shape innovative activity? *Games and Economic Behavior*, 83:pp. 1–23.

Holmstrom, B. (1982). Moral Hazard in Teams. *Bell Journal of Economics*, 13(2):324–340.

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

Hvide, H. K. (2002). Pragmatic beliefs and overconfidence. *Journal of Economic Behavior & Organization*, 48(1):15 – 28.

Institute for Leadership and Management (2011). Ambition and gender at work. ILMAGW/0211.

Ivanova-Stenzel, R. and Kübler, D. (2011). Gender differences in team work and team competition. *Journal of Economic Psychology*, 32(5):797–808.

Jalava, N., Joensen, J. S., and Pellas, E. (2015). Grades and rank: Impacts of non-financial incentives on test performance. *Journal of Economic Behavior & Organization*, 115(C):161–196.

Jurajda, S. and Münich, D. (2011). Gender gap in performance under competitive pressure: Admissions to Czech universities. *American Economic Review*, 101(3):514–18.

Kamas, L. and Preston, A. (2012). The importance of being confident: Gender, career choice, and willingness to compete. *Journal of Economic Behavior & Organization*, 83(1):82 – 97. Gender Differences in Risk Aversion and Competition.

Karran, T. (2004). Achieving Bologna convergence: Is ECTS failing to make the grade? *Higher Education in Europe*, 29(3):411–421.

Kim, M.-K. and Polachek, S. W. (1994). Panel estimates of male-female earnings functions. *The Journal of Human Resources*, 29(2):pp. 406–428.

Kocher, M., Strauß, S., and Sutter, M. (2006). Individual or team decision-making: causes and consequences of self-selection. *Games and Economic Behavior*, 56(2):259–270.

Kosfeld, M. and Siemens, F. A. (2009). Worker self-selection and the profits from cooperation. *Journal of the European Economic Association*, 7(2-3):573–582.

Koudstaal, M., Sloof, R., and van Praag, C. (2015). Are entrepreneurs more optimistic than managers? Evidence from a large lab-in-the-field experiment. Unpublished manuscript.

Koudstaal, M., Sloof, R., and van Praag, C. (forthcoming). Risk, uncertainty and entrepreneurship: Evidence from a large lab-in-the-field experiment. *Management Science*.

Kuhn, P. and Villeval, M. C. (2014). Are women more attracted to co-operation than men? *The Economic Journal*, 125(582):115–140.

Landeras, P. (2009). Student effort: standards vs. tournaments. *Applied Economics Letters*, 16(9):965–969.

Laury, S. K. (2006). Pay One or Pay All: Random Selection of One Choice for Payment. Experimental Economics Center Working Paper Series 2006-24, Experimental Economics Center, Andrew Young School of Policy Studies, Georgia State University.

Lazear, E. P. (2000). Performance pay and productivity. *The American Economic Review*, 90(5):1346–1361.

Lazear, E. P. and Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, 89(5):pp. 841–864.

Lazear, E. P. and Shaw, K. L. (2007). Personnel economics: The economist's view of human resources. *The Journal of Economic Perspectives*, pages 91–114.

Leuven, E., Oosterbeek, H., Sonnemans, J., and Klaauw, B. v. d. (2011). Incentives versus sorting in tournaments: Evidence from a field experiment. *Journal of Labor Economics*, 29(3):pp. 637–658.

Leuven, E., Oosterbeek, H., and van der Klaauw, B. (2010). The effect of financial rewards on students' achievement: Evidence from a randomized experiment. *Journal of the European Economic Association*, 8(6):1243–1265.

Ludwig, S. and Thoma, C. (2012). Do Women Have More Shame than Men? An Experiment on Self-Assessment and the Shame of Overestimating Oneself. Discussion Papers in Economics 12905, University of Munich, Department of Economics.

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

Moldovanu, B. and Sela, A. (2001). The Optimal Allocation of Prizes in Contests. *American Economic Review*, 91(3):542–558.

Moore, D. A. and Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2):502–517.

Morin, L.-P. (2015). Do Men and Women Respond Differently to Competition? Evidence from a Major Education Reform. *Journal of Labor Economics*, 33(2):443 – 491.

Müller, W. and Schotter, A. (2010). Workaholics and dropouts in organizations. *Journal of the European Economic Association*, 8(4):717–743.

Netherlands Bureau for Economic Policy Analysis (2014). Table main economic indicators 2012-2015. http://www.cpb.nl/cijfer/kerngegevenstabel-2012-2015-voor-het-centraal-economisch-plan-2014. Accessed: 2015-07-08.

Niederle, M. and Vesterlund, L. (2007). Do Women Shy Away from Competition? Do Men Compete Too Much? *The Quarterly Journal of Economics*, 122(3):1067–1101.

Niederle, M. and Vesterlund, L. (2010). Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives*, 24(2):129–44.

Niederle, M. and Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics*, 3:601–630.

Noussair, C., Trautmann, S., Kuilen, G., and Vellekoop, N. (2013). Risk aversion and religion. *Journal of Risk and Uncertainty*, 47(2):165–183.

OECD (2013). *Education at a Glance 2013,*. OECD Indicators. OECD Publishing.

OECD (2015). *The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence,*. PISA. OECD Publishing.

Onderstal, S. (2014). *Economics of Organizations and Markets*. Pearson, Amsterdam.

Örs, E., Palomino, F., and Peyrache, E. (2013). Performance gender gap: Does competition matter? *Journal of Labor Economics*, 31(3):pp. 443–499.

Owens, D., Grossman, Z., and Fackler, R. (2014). The control premium: A preference for payoff autonomy. *American Economic Journal: Microeconomics*, 6(4):138–61.

Paredes, V. (2012). Grading system and student effort. Unpublished manuscript.

Petrie, R. and Segal, C. (2014). Gender differences in competitiveness: The role of prizes. GMU Working Papers in Economics No. 14-47.

Price, J. (2008). Gender differences in the response to competition. *Industrial and Labor Relations Review*, 61(3):320–333.

Raven, J., Raven, J. C., and Court, J. H. (2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. San Antonio, TX: Harcourt Assessment.

Reynolds, P. D. and Curtin, R. T. (2008). Business Creation in the United States: Panel Study of Entrepreneurial Dynamics II. Initial Assessment . *Foundations and Trends in Entrepreneurship*, 4(3):155 – 307.

Salvi del Pero, A. and Bytchkova, A. (2013). *A Bird's Eye View of Gender Differences in Education in OECD Countries*. OECD Social, Employment and Migration Working Papers, No. 149. OECD Publishing.

Sandberg, S. (2013). *Lean In. Women, Work and the Will to Lead*. Alfred A. Knopf. New York.

Shurchkov, O. (2012). Under Pressure: Gender Differences In Output Quality And Quantity Under Competition And Time Constraints. *Journal of the European Economic Association*, 10(5):1189–1213.

Sloof, R. and von Siemens, F. (2014). Illusion of Control and the Pursuit of Authority. CESifo Working Paper Series 4764, CESifo Group Munich.

Sutter, M. and Kocher, M. G. (2007). Trust and trustworthiness across different age groups. *Games and Economic Behavior*, 59(2):364 – 382.

Teyssier, S. (2008). Experimental evidence on inequity aversion and self-selection between incentive contracts. GATE working paper 0710.

Tran, A. and Zeckhauser, R. (2012). Rank as an inherent incentive: Evidence from a field experiment. *Journal of Public Economics*, 96(910):645 – 650.

UN Women (2015). Facts and figures: Leadership and political participation. http://www.unwomen.org/en/what-we-do/leadership-and-political-participation/facts-and-figures. Accessed 30 January 2015.

van den Broek, A., Wartenbergh, F., Hogeling, L., Brukx, D., Warps, J., Kurver, B., and Muskens, M. (2009). Studentenmonitor hoger onderwijs 2007. ResearchNed Nijmegen.

van Dijk, F., Sonnemans, J., and van Winden, F. (2001). Incentive systems in a real effort experiment. *European Economic Review*, 45(2):187 – 214.

Wieland, A. and Sarin, R. (2012). Domain specificity of sex differences in competition. *Journal of Economic Behavior & Organization*, 83(1):151–157.

Wozniak, D., Harbaugh, W. T., and Mayr, U. (2014). The Menstrual Cycle and Performance Feedback Alter Gender Differences in Competitive Choices. *Journal of Labor Economics*, 32(1):161 – 198.

Zhang, J. (2013). Can experimental economics explain competitive behavior outside the lab? Unpublished manuscript.

Samenvatting (Summary in Dutch)

In dit proefschrift wordt een aantal studies beschreven die laten zien hoe bepaalde externe prikkels individueel gedrag kunnen beïnvloeden. In de hoofdstukken 2 en 3 worden de verschillen in gedrag tussen mannen en vrouwen binnen een toernooimodel onderzocht en in hoofdstuk 4 wordt gekeken naar het verband tussen de keuze om in een team te werken en individuele karakteristieken, zoals opleidingsniveau of beroepsgroep.

Het doel van hoofdstuk 2 is om te onderzoeken wat het effect is van absolute versus relatieve beoordelingsmethoden op de inzet en de prestatie van studenten in hun natuurlijke (maar tegelijkertijd gecontroleerde) studie-omgeving. De vraag is of mannen en vrouwen verschillend reageren op de twee verschillende beoordelingsmethoden. Om dit te onderzoeken is er een grootschalig veldexperiment, een zogenaamd framed field experiment, uitgevoerd onder bachelorstudenten aan een Nederlandse universiteit. Samengevat laten de uitkomsten zien dat de verschillende beoordelingsmethoden geen verschillen veroorzaken in de inzet of de tentamencijfers van studenten. Ondanks het belang van de studie-uitkomsten voor de studenten en ondanks hun begrip van de verschillen in de beoordelingsmethoden, zien we toch geen verschil in gedrag tussen de studenten die beoordeeld worden op basis van absolute cijfers en de groep die beoordeeld wordt op basis van relatieve cijfers. Bovendien vinden we geen verschillen in gedrag tussen mannen en vrouwen, tussen studenten met verschillende bekwaamheidsniveaus of studenten met verschillende voorkeuren voor competitie. Wellicht is een plausibele verklaring voor deze resultaten dat bij Nederlandse studenten veelal de ambitie om hoge cijfers te halen ontbreekt (de zogenaamde zesjes-cultuur). De uitkomsten van dit onderzoek laten zien

dat het invoeren van relatieve beoordeling geen oplossing lijkt te zijn voor het stimuleren van de motivatie van studenten: een toernooi is ineffectief als de prijs, in dit geval hoge cijfers, voor het grootste deel van de deelnemers niet als waardevol wordt gezien.

Het doel van hoofdstuk 3 is om sekseverschillen in risicohouding en competitief gedrag te onderzoeken op basis van een analyse van data van een online-kaartspel. De dataset bevat data van meer dan 4 miljoen verschillende spellen van meer dan 15.000 verschillende individuele spelers. In dit in de Duitse streek Beieren populaire online-kaartspel (Schafkopf) kunnen spelers verschillende keuzes maken die meer of minder competitief zijn en die hun uitkomsten beïnvloeden. De uitkomsten van de studie laten zien dat vrouwen minder competitief en meer risico-avers spelen dan mannen en dat de verschillen aanzienlijk zijn. Vrouwen zijn bijvoorbeeld minder geneigd om hun inzet te verhogen binnen een spel of om actief een nieuw spel te initiëren. Als gevolg hiervan behalen vrouwen bij dit spel lagere scores dan mannen, ook al zijn ze ten minste zo bekwaam in het spel als mannen. Met andere woorden, vrouwen winnen spellen niet minder vaak dan mannen gegeven het soort spel en de rol van de persoon binnen een spel. Dit hoofdstuk van mijn proefschrift laat dus zien wat de negatieve gevolgen voor vrouwen zijn van het terugschrikken voor competitie. Bij gelijke prestaties tijdens het spel, blijft de score van vrouwen achter op die van de mannen doordat ze minder risico nemen en competitie meer uit de weg gaan.

In hoofdstuk 4 wordt onderzocht wat de determinanten zijn van iemands keuze om in een team te werken. Twee scenarios worden bestudeerd: bij het ene deelt het team slechts de opbrengsten van het teamwerk, bij het andere worden ook de beslissingsbevoegdheden gedeeld. De resultaten in dit hoofdstuk zijn gebaseerd op een grootschalig online onderzoek (met financiële prikkels) onder ondernemers, managers en werknemers in Nederland: een zogenaamd lab-in-the-field experiment. De steekproef bestaat uit ruim duizend deelnemers en is zeer divers met betrekking tot geslacht, leeftijd, werkervaring, opleidingsniveau en inkomen. Met deze steekproef repliceren we bevindingen uit eerder laboratoriumonderzoek. De resultaten tonen het belang aan van zelfvertrouwen en risicohouding bij de keuze voor teambeloning in plaats

van individuele beloning. Een nieuwe bevinding uit het onderzoek in dit hoofdstuk is de rol die opleidingsniveau speelt bij deze keuze voor teambeloning. Rekening houdend met de individuele prestatie, zelfvertrouwen en risicohouding, is er een positieve relatie tussen het opleidingsniveau en de bereidheid om in een team te werken. De studie laat als verklaring zien dat mensen met verschillende opleidingsniveaus de toegevoegde waarde van samenwerking in teamverband verschillend beoordelen. Mensen met een hoog opleidingsniveau focussen voornamelijk op de mogelij-ke voordelen van teambeloning (zoals complementariteit tussen de teamgenoten), terwijl mensen met een lager opleidingsniveau vooral kijken naar de mogelijke risicos die verbonden zijn aan groepswerk (zoals onzekerheid over de kwaliteit en de inzet van de teamgenoot). Daarnaast tonen de resultaten van hoofdstuk 4 uit dit proefschrift aan dat mensen verschillen in hun reactie/keuze als het gaat om potentieel gedeelde besluitvorming. In het bijzonder blijkt dat ondernemers afwijzend staan tegenover gezamenlijke besluitvorming als ze denken dat de voorkeur van hun teamgenoot erg zal afwijken van hun eigen optimale keuze. Dit verschil in voorkeur bij gedeelde besluitvorming wordt niet gevonden voor managers en werknemers.

This dissertation is based on the following studies:

Chapter 2:

Does relative grading help male students? Evidence from a field experiment in the classroom

Co-authors: Sander Onderstal, Randolph Sloof and Mirjam van Praag

Published as: IZA Discussion Paper, No. 8429.

Contribution of the doctoral candidate: She is the main contributor to the study. She initiated the research project, was to a large extent responsible for the design of the experiment, conducted the econometric analysis, and wrote substantial parts of the working paper.

Chapter 3:

Women do not play their aces - The consequences of shying away

Co-authors: Jörg Claussen and Mirjam van Praag

Unpublished manuscript.

Contribution of the doctoral candidate: She is one of the main contributors to the study. She and her co-authors jointly formulated the research question. She was to a large extent responsible for the econometric analysis, and wrote substantial parts of the manuscript.

Chapter 4:

Risks, gains and autonomy: An experimental analysis of sorting into teams

Co-authors: Martin Koudstaal and Laura Rosendahl Huber

Unpublished manuscript.

Contribution of the doctoral candidate: She is one of the main contributors to the study. She and her co-authors jointly formulated the research question and designed the experiment. She conducted the econometric analysis, and wrote substantial parts of the manuscript.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

582. X. SHEN, Essays on Empirical Asset Pricing

583. L.T. GATAREK, Econometric Contributions to Financial Trading, Hedging and Risk Measurement

584. X. LI, Temporary Price Deviation, Limited Attention and Information Acquisition in the Stock Market

585. Y. DAI, Efficiency in Corporate Takeovers

586. S.L. VAN DER STER, Approximate feasibility in real-time scheduling: Speeding up in order to meet deadlines

587. A. SELIM, An Examination of Uncertainty from a Psychological and Economic Viewpoint

588. B.Z. YUESHEN, Frictions in Modern Financial Markets and the Implications for Market Quality

589. D. VAN DOLDER, Game Shows, Gambles, and Economic Behavior

590. S.P. CEYHAN, Essays on Bayesian Analysis of Time Varying Economic Patterns

591. S. RENES, Never the Single Measure

592. D.L. IN T VELD, Complex Systems in Financial Economics: Applications to Interbank and Stock Markets

593. Y. YANG, Laboratory Tests of Theories of Strategic Interaction

594. M.P. WOJTOWICZ, Pricing Credits Derivatives and Credit Securitization

595. R.S. SAYAG, Communication and Learning in Decision Making

596. S.L. BLAUW, Well-to-do or doing well? Empirical studies of wellbeing and development

597. T.A. MAKAREWICZ, Learning to Forecast: Genetic Algorithms and Experiments

598. P. ROBALO, Understanding Political Behavior: Essays in Experimental Political Economy

599. R. ZOUTENBIER, Work Motivation and Incentives in the Public Sector

600. M.B.W. KOBUS, Economic Studies on Public Facility use

601. R.J.D. POTTER VAN LOON, Modeling non-standard financial decision making

602. G. MESTERS, Essays on Nonlinear Panel Time Series Models

603. S. GUBINS, Information Technologies and Travel

604. D. KOPNYI, Bounded Rationality and Learning in Market Competition

605. N. MARTYNOVA, Incentives and Regulation in Banking

606. D. KARSTANJE, Unraveling Dimensions: Commodity Futures Curves and Equity Liquidity

607. T.C.A.P. GOSENS, The Value of Recreational Areas in Urban Regions

608. .M. MAR, The Impact of Aid on Total Government Expenditures

609. C. LI, Hitchhiking on the Road of Decision Making under Uncertainty

610. L. ROSENDAHL HUBER, Entrepreneurship, Teams and Sustainability: a Series of Field Experiments

611. X. YANG, Essays on High Frequency Financial Econometrics

612. A.H. VAN DER WEIJDE, The Industrial Organization of Transport Markets: Modeling pricing, Investment and Regulation in Rail and Road Networks

613. H.E. SILVA MONTALVA, Airport Pricing Policies: Airline Conduct, Price Discrimination, Dynamic Congestion and Network Effects.

614. C. DIETZ, Hierarchies, Communication and Restricted Cooperation in Cooperative Games

615. M.A. ZOICAN, Financial System Architecture and Intermediation Quality

616. G. ZHU, Three Essays in Empirical Corporate Finance

617. M. PLEUS, Implementations of Tests on the Exogeneity of Selected Variables and their

Performance in Practice

618. B. VAN LEEUWEN, Cooperation, Networks and Emotions: Three Essays in Behavioral Economics

619. A.G. KOPNYI-PEUKER, Endogeneity Matters: Essays on Cooperation and Coordination

620. X. WANG, Time Varying Risk Premium and Limited Participation in Financial Markets

621. L.A. GORNICKA, Regulating Financial Markets: Costs and Trade-offs

622. A. KAMM, Political Actors playing games: Theory and Experiments

623. S. VAN DEN HAUWE, Topics in Applied Macroeconometrics

624. F.U. BRUNING, Interbank Lending Relationships, Financial Crises and Monetary Policy

625. J.J. DE VRIES, Estimation of Alonsos Theory of Movements for Commuting

626. M. POPAWSKA, Essays on Insurance and Health Economics

627. X. CAI, Essays in Labor and Product Market Search

628. L. ZHAO, Making Real Options Credible: Incomplete Markets, Dynamics, and Model Ambiguity

629. K. BEL, Multivariate Extensions to Discrete Choice Modeling

630. Y. ZENG, Topics in Trans-boundary River sharing Problems and Economic Theory

631. M.G. WEBER, Behavioral Economics and the Public Sector