WORK IN PROGRESS. Please do not circulate.

Does Relative Grading Help Male Students? Evidence from a Field Experiment in the Classroom

Eszter Czibor*, Sander Onderstal*, Randolph Sloof* and Mirjam van Praag**

*University of Amsterdam
**Copenhagen Business School

This version: April 2014

Abstract

We perform a direct comparison of the two most commonly used grading practices: the absolute (i.e. criterion-referenced) and the relative (norm-referenced) grading schemes in a large-scale field experiment at a university. We test whether relative grading, by creating a rank-order tournament in the classroom, provides stronger incentives for male students than absolute grading. In the full sample, we only find weak support for our hypothesis. Among the more motivated students we find evidence that men score significantly higher on the test when graded on the curve. Female students, irrespective of their motivation, do not increase their scores under relative grading. Since in our setting women slightly outperform men under absolute grading, grading on a curve actually narrows the gender gap in performance. The increased performance of male students can be attributed to higher effort provision during the exam. Our results are relevant for the design of policies addressing the gender gap in educational outcomes.

JEL codes: I21, I23, A22, D03, C93
Keywords: education, test performance, grade incentives, competition, gender, field experiment

1 Introduction

Over the last decades, girls have come to outperform boys at all levels of education in the developed world[1]. This phenomenon has led some governments worrying about the poor outcomes of boys to consider policies aimed at improving their performance (The Economist, 1999). The increasing gender gap has also triggered the attention of popular academic book writers sometimes referred to as belonging to the 'boys' movement', producing best-selling titles like "Boys Adrift" (Sax, 2007), "The War Against Boys" (Hoff Sommers, 2000) and "The End of Men" (Rosin, 2012). A variety of reasons has been put forward to explain the under-achievement of men, ranging from biological differences in learning styles between the sexes to allegedly "feminine" educational practices favoring girls. From an economist's point of view it is natural to consider another potential explanation: it is possible that grade incentives offered by the educational system do not provide sufficient motivation for boys to do well.

The majority of grading practices that are in use worldwide can be classified into two schemes: absolute and relative grading. Under absolute grading, grades solely depend on students' own individual test outcomes, independent of the performance of their classmates on the same test. This scheme is also known as criterion-referenced grading, because the score of the student is compared to an objective criterion. 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. In the United States, colleges typically implement relative grading[2], while in continental Europe the absolute scheme prevails (Karran, 2004).

A key difference between the two grading schemes is that relative grading induces direct competition between peers. In organizational economics lingo, absolute grading is analogous to a piece rate compensation scheme, while relative grading creates a rank-order tournament in the classroom. The advantageous and disadvantageous incentive effects of competitive reward schemes have been studied more broadly within this literature. 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)). Effort incentives can thus be provided with lower risk exposure (i.e. in a cheaper way). 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 (cf. Bull et al. (1987); van Dijk et al. (2001)).

Another common empirical finding is that tournament incentives increase male, but not female performance in a mixed-sex environment. This gender gap in response to tournament

[1] A recent OECD report states: "Compared to girls, boys have become more likely to underperform in secondary education and less likely to complete higher education degrees" (Salvi del Pero and Bytchkova, 2013, p. 7). Moreover, Guiso et al. (2008) find that in more gender-equal cultures, there is no gender gap in math performance among 15-year-olds, while girls largely outperform boys in reading.

[2] 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

incentives has been first documented by Gneezy et al. (2003) who found 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. Their result has been replicated using both laboratory (e.g. Gunther et al. (2010)) and field experiments (e.g. Gneezy and Rustichini (2004)) as well as naturally occurring data (e.g. Price (2008)). [3].

Based on the empirical stylized fact of gender differences in response to tournaments we hypothesize that grading on a curve, by offering competitive incentives, provides better motivation and leads to higher performance for male students than absolute grading. We expect no such difference for female students. This paper provides a test for the above hypothesis by means of a field experiment. We empirically compare absolute versus relative grading in the naturalistic setting of a Bachelor course with a large number of participants at the University of Amsterdam. Students in our experiment are randomly divided into two treatment groups. All students have to take the same midterm exam, but the grading of this exam differs by treatment: in one group the midterm test is graded on the curve, in the other group on an absolute scale. (The grading schemes are reversed for the end-term exam.) Based on various measures capturing effort provision in preparation for and performance in the midterm exam, we provide a direct empirical comparison of the two grading schemes.

For the full sample of participating students we only obtain weak support for the hypothesis that relative grading helps male students without being disadvantageous for females. As expected, women slightly outperform men under the absolute scheme while the gender gap shrinks under the relative scheme. However, these differences are small in size and statistically insignificant. Only when correcting for demographics and ability and preference variables, we find that *ceteris paribus*, men respond more favorably to relative grading than women. We further observe that the two grading schemes do not differ considerably in terms of the study effort and exam preparation time they induce. We therefore tentatively attribute the increase in test scores of male students to higher effort provision during the exam itself.

Obviously, in order for students to be responsive to 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 but beyond that do not attach importance to their grade per se, the incentive effect of grading on the curve is likely to be limited. In policy reports and the popular press, the Dutch educational system is often criticized for its prevailing culture of 'just pass' (*'zesjescultuur'*)[4]. Indeed, international comparisons of primary school pupils reveal that although almost all Dutch pupils attain the minimum requirements, only a few excel and meet the truly advanced benchmarks (Jaspers, 2012). The phenomenon is not constrained to primary schools: Dutch university students are found to provide insufficient study

[3]Please refer to Niederle and Vesterlund (2011) and Croson and Gneezy (2009) for detailed reviews of studies on gender and competition.

[4]A recent article (in Dutch) covering the 2012/13 edition of 'The state of education' (*'De staat van het onderwijs'*) report claims already in its title that Dutch pupils are undermotivated. The lead researcher of the report is quoted to say that pupils are satisfied with getting a 6, the lowest passing grade: http://www.nrc.nl/nieuws/2014/04/16/rapport-nederlandse-leerlingen-zijn-niet-gemotiveerd/.

effort (Leuven et al., 2010), leading to high fail rates and long study durations[5]. The response of students in our sample to the grade incentives may thus very well vary in the importance they attach to grades. To take account of the potential heterogeneity in this dimension, we also test our hypothesis among the subsample of "responsive" students that are conjectured to care about the level of their grade: students following the international program. In this subgroup, male students score significantly higher when graded on the curve. Female students, whether in the Dutch or international program, do not respond to relative grading.

Overall we conclude that the incentive effect of relative grading is limited when measured using the sample that includes the entire group of students. However, test scores among a subsample of students that are arguably and verifiably motivated by grade incentives show a picture in line with expectations: male students increase their performance when graded on the curve, whereas the performance of female students is unaffected by the grade scheme used. These findings suggest that absolute grading does not provide sufficiently strong motivation for male students. Especially the more motivated men could be given stronger incentives by grading them competitively on the curve. This would not harm the absolute performance of female students, as they do not respond to the induced competition. Unfortunately, however, relative grading does not help to incentivize the less motivated males, the arguably most problematic group.[6]

A number of recent empirical studies also focus on the effect of competition in education and obtain gender differences in line with our results. Jurajda and Munich (2011) find that male and female students of the same ability were equally likely to be admitted to universities in the case of little competition, but men were significantly more likely to be accepted when competition was high. Ors et al. (2013) observe that females outperformed men in a "noncompetitive" national exam while in the same cohort, men outperformed women at the very competitive HEC admission exam later on. Morin (forthcoming) studies the effect of intensified competition resulting from a legislative change that created a "double cohort" in Ontario and observes that men's relative performance increases. Bigoni et al. (2011) find that competition induces higher effort among male but not among female students in the case of low-stakes homework assignments. Jalava et al. (2013) examine various non-financial incentive schemes for primary school children in low-stakes tests and conclude that boys only increase performance when faced with relative grading. In contrast, De Paola et al. (2013) does not find gender differences in terms of entry into a tournament or performance under competition in a setting where university students self-selected into a competitive scheme in a midterm exam to obtain bonus points.

To the best of our knowledge, this paper is the first to empirically compare absolute and

[5]From the cohort of students who entered higher education in 2007, less than a third received their Bachelor degree after four years, the nominal duration of the program. Five years after the start, the share of those who successfully graduated was still less than half (Inspectie van het Onderwijs, 2014).

[6]Other incentive instruments may potentially prove helpful here. Monetary incentives to improve student achievement have been studied by Leuven et al. (2010), Fryer (2011) and Braun et al. (2011), while Levitt et al. (2012) and Jalava et al. (2013) look at non-pecuniary incentives such as certificates and trophies. Grove and Wasserman (2006) study whether making assignments count towards the final grade improves student learning and exam performance.

relative grading in a naturalistic, high-stakes experimental setting[7]. Our field experiment includes a 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 students study for the exam.

The remainder of this paper is organized as follows. In Section 2, we describe the details of our experimental design. In Section 3, we provide an overview of our data and some summary statistics. Section 4 presents our results. Section 5 contains a further discussion of the results. We conclude in Section 6.

2 Context and design

2.1 Context

We conducted a framed field experiment (Harrison and List, 2004) among students of the University of Amsterdam (UvA). The experiment took place in the 2$^{\text{nd}}$ year BSc course *Economics of Markets and Organizations* (EMO) during the first block of the 2013/2014 academic year.[8] The course covered topics from Organizational Economics and Industrial Organizations in a simple game-theoretic framework. Over 500 students enrolled in the course. The large sample size is desirable not only because it allows us to detect potentially small 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[9]. The class was compulsory for the majority of the enrolled students, ensuring relatively low attrition rates. The course grade was calculated as the unweighted average of grades from a mid- and an end-term exam. The two exams covered roughly the same amount of study material and were designed to be of comparable difficulty. Both tests consisted of multiple-choice questions and were corrected by machines, thus grading was by construction unbiased. In addition, students could earn a bonus point on top of their exam grade by handing in homework assignments in teams of three or four people. During study weeks, students could participate in a lecture (focusing mostly on theory) and a tutorial (discussing exercises, homework solutions and mock exam questions). Class attendance was voluntary.

The course was offered with identical content in both Dutch and in English, the latter for students following the English-language Bachelor study (referred to as the "international program"). The majority of students in the international program are foreigners (typically from Central-Eastern Europe, China and Germany), but the program is also open for aspiring Dutch students. While there are no entry requirements for attending the Dutch program (all applicants who complete the pre-university track in secondary education and pass the standardized national

[7]The above mentioned papers, instead of actual grading on a curve, focus on competitive grading in the form of comparison to a randomly chosen opponent or rewarding only the top performers.

[8]At the UvA, the academic year is divided into 6 blocks. The first block runs over 8 weeks in September and October.

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

school-leaving exam are admitted to the study), students have to qualify for the international program. Participants in the two programs thus differ in their ability and arguably also in their level of motivation. For foreign students, 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. Students in the Dutch program, on the other hand, are on average less motivated and provide insufficient study effort (Leuven et al., 2010). Their attitude can be described by the *zesjescultuur*, discussed in the Introduction: instead of striving for excellence, the majority of them merely aim to pass their courses with minimal effort provision.

2.2 Design of the experiment

Participants were randomly assigned to one of the two treatment conditions (communicated to students as the "yellow group" and the "blue group" in order to maintain a neutral framing). All students, regardless of their treatment group assignment, sat the same midterm and end-term exams at the same time and venue. As mentioned earlier, both exams counted with equal weights towards the final course grade and thus represented high stakes for students. The difference between the groups lay in the *grading schemes used in translating exam scores into exam grades*: students in one group were graded under a relative scheme in the midterm and an absolute scheme in the end-term exam while the schemes were reversed in the other group, as shown in Table 1. This design allows for a clean comparison of the effect of the two grading schemes on midterm exam performance while maintaining an *ex ante* fair and equal treatment of students in the two groups.

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

Table 1: Overview of treatment groups and grading schemes

We had access to a rich set of demographic variables including gender, age and nationality of students already before the start of the course. It was therefore possible to perform a stratified randomization along the dimensions we suspected would influence the response to the grading schemes. The strata we chose were: gender, study program and Mathematics ability. The course has been taught over several years with only small changes in the content, thus previous years' grade distributions could be taken into account when constructing the curve for relative grading.

The timeline of the experiment is shown in Table 2. Students were informed about 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 1: Sections from the Course Manual) and were also announced during the lectures and tutorials. Students were required to form homework teams within their treatment group (in order to reduce potential spill-overs), which also increased students' awareness of their treatment assignment. Homework results were not published until week 5, so students received no feedback on their relative performance until the midterm exam. Right before the midterm exam, students were requested to fill out a short questionnaire testing their understanding of the grading schemes and collecting information on time spent on the course.

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**

Table 2: Timeline of the experiment

Our main variable of interest is the score (i.e. the number of correct answers) on the midterm exam. Using a between-subject design, we compare the midterm scores of students in the absolute and relative grading groups. Since treatment assignment is random, we can attribute any difference we observe between the outcomes of the two groups to the difference in the grading scheme they experience. Besides the midterm exam scores, in our analysis we also consider the effect of relative grading on a number of proxies for effort provision *in preparation for* the exam. These proxies are: lecture and tutorial attendance during the study weeks (collected by an assistant and by the tutors), handing in homework assignments, grades of homework assignments and self-reported study time (collected in the aforementioned questionnaire).

2.3 Incentivized survey

We conducted an online survey to collect preference, confidence and ability measures from students. We included the survey among the compulsory course requirements which ensured a very high response rate (92%). The survey was incentivized: 5 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 were €215.67, with a minimum of €100 and a maximum of €457). Respondents spent on average 21 minutes completing the survey which was designed and pre-tested to take about 15-20 minutes, suggesting the majority of students took the task seriously and were not answering at random. The survey was programmed using the software Qualtrics.

The survey was framed as assessing familiarity with prerequisites for the course, and contained a timed multiple-choice quiz with ten questions related to first-year Mathematics and

Microeconomics courses (e.g. simple derivations, perfect competition, Nash-equilibria, etc.)[10]. Performance on the quiz serves as an important ability measure in our analysis. Before solving the quiz, students were required to choose the reward scheme applied to their quiz performance by reporting their switching point between a constant piece rate and a tournament scheme with an increasing prize. This serves as our proxy for competitive preferences. Moreover, we collected four different measures of overconfidence (ex ante and ex post; absolute and relative): students were asked to report, both before and after solving the quiz, their expected absolute score and relative rank on the quiz. In addition, risk and ambiguity preferences of participants were measured by eliciting switching points in Holt and Laury (2002)-style choice menus (see Figure 6 in Appendix) 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 towards norm- and criterion-referenced grading practices.

3 Data

This section contains an overview of our data. Panel A of Table 3 presents basic demographic information. In total, 529 students registered for the course, a quarter of them following the international program. The share of female students in the sample is relatively low, just over a third, reflecting the 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 for the past academic achievement of the students in our sample. Here we present descriptive statistics for the two measures that we found most important in our analysis: 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 ever took at the UvA (including failed tests), is a fairly good predictor for the final grade in the EMO course: the correlation between the two is 0.499. This measure indicates very low average performance: the mean of the variable, 5.88 is barely above the minimum requirement for passing[11]. The second indicator is calculated as the number of retake exams over all the courses the student ever registered for. We find that on average, students fail and have to repeat approximately one out of five exams[12].

Panel B of Table 3 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. (A homework grade of 5.5 or above ensured the bonus point.)

[10] For an example of a quiz question, please refer to Figure 5 in the Appendix

[11] In the Netherlands, grades range from 1 (lowest) to 10 (highest), and the minimum passing grade is either 5.5 (when half grades are awarded) or 6.

[12] Note that neither of these ability measures based on past performance are available for the 32 exchange students who visited the UvA only for the semester when our study was conducted.

Students reported to have spent 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-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-term exam[13]. 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 on top of the unweighted average of the two exam grades.

Results from the incentivized online survey are presented in Panel C of Table 3. As mentioned earlier, the response rate for the survey was very high. We observe a relatively low average performance on the quiz measuring knowledge in prerequisites (4.67 correct answers out of 10 questions), which is likely explained by the intense time pressure students were subject to during the quiz (25 seconds per question). 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 student's 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. 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 risk-averse (the risk-neutral switching point is at decision 5) but seeking competition (a risk-neutral agent who believes he/she will win half of the times is indifferent between piece rate and tournament at decision 7). The overconfidence of students is also reflected in their grade expectations exceeding their realized final grades (an average of 7.04 vs. 6.65) and their relative performance in terms of grades (students on average guess 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: students rarely face explicit relative grading in the Dutch educational system. Still, students are not opposed to relative grading: on average they rated their attitude towards grading on the curve as slightly positive (5.33 where 5 corresponds to neutral).

Section I of Table 4 proves that the randomization has been successful. The two treatment groups are balanced not only along the dimensions we have stratified on (study program, gender and Mathematics grades), but also with respect to other demographic, ability and preference variables. Section II compares the male and female students in our sample. We observe that women are more likely than men to follow the international program and are thus less likely to be born in the Netherlands. There is also a gender difference in past academic performance: on average, women obtained significantly higher Math grades and had to retake fewer exams than their male peers[14]. We find no such difference in the number of correct quiz questions, possibly

[13]We argue that this decrease does not reflect a difference in the difficulty of the exams but is rather a result of the way the final grade for the course was determined: a high midterm grade and a homework bonus point would ensure that a student passes the course simply by showing up at end-term exam, providing weaker incentives for effort provision in the end-term. We further discuss this finding in the Results section.

[14]The difference is not driven merely by the higher share of international students among women. Even after controlling for study program, women obtain significantly higher grades than men in our sample.

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 towards risk, with women being significantly more risk averse. This finding is in line with results from other studies (Croson and Gneezy, 2009)[15]. Contrary to our expectations, we find no significant gender differences in the willingness to enter tournaments. This may be explained in part by women in our sample being as confident as male students. Moreover, the continuous measure we designed (eliciting a switching point between a piece rate and a competitive payment scheme with varying prizes) has not been verified by other studies to reliably capture competitive preferences and could have potentially been confusing for our subjects. Section III of Table 4 shows the differences between students following the Dutch and the international program. We find that students in the English-language program outperform their Dutch peers: they have significantly higher Mathematics grades, have fewer retakes and solve more quiz questions correctly. It is therefore understandable that they also have higher grade expectations than students in the Dutch study program.

4 Results

4.1 Selection

Before comparing outcomes of students in the two treatment group, we need to alleviate concerns related to non-random attrition. Students 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 are likely to shy away from competition. We would thus expect to see lower midterm show-up in the relative grading group, driven by fewer female students participating in the exam. We find no support for this hypothesis in our data: show-up is actually slightly *higher* under relative grading (a raw difference of 4.9 percentage points, significant at the 5% level), and there is no gender difference in the propensity to participate under the two schemes. 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. We thus argue that exam participation is likely unrelated to our treatment, so our results are not biased by non-random selection.

4.2 Exam performance

4.2.1 Full sample

We start our analysis with a simple comparison of midterm scores under the different grading schemes. The mean number of correct answers was 19.196 under absolute and 19.366 under relative grading (with standard deviations of 3.790 and 3.810, respectively). According to a two-sample t-test with unequal variances, the difference is insignificant (p-value: 0.622). As Figure

[15]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.

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 0.988). We proceed to test whether the response to grade incentives differs by gender. Figure 2 compares the mean number of correct answers on the midterm exam by gender and treatment group. While there is an indication of women slightly outperforming men under the absolute scheme and the gender gap shrinking under the relative scheme due to higher scores of male students, these differences are negligible in size. An OLS regression (with standard errors clustered on the homework team level) confirms that there is no significant difference in midterm scores between the treatment groups, and the gender interaction is also insignificant (see columns 1 and 2 of Table 5).

A raw comparison of scores may, however, be misleading: as we have shown in Table 4, men and women in our sample are different along many dimensions that could possibly influence their response to the grading schemes. We therefore repeat the analysis controlling for demographic, ability and preference variables[16] and discover that *all else equal*, men respond more favorably to relative grading than women. Column 3 in Table 5 shows that the coefficient of the interaction term between relative grading and the male dummy is positive and significant at the 5% level. The result is remarkably robust to the inclusion of further interaction terms (see Table 10 in the Appendix). Analyzing the subsamples of men and women separately (columns 4 and 5 in Table 5) we find that the point estimate for the effect of relative grading on exam scores is negative for female and positive for male students, *ceteris paribus*, but these estimates are not statistically significant (p-values: 0.17 and 0.16 for men and women, respectively).

4.2.2 Responsiveness to grade incentives

In our analysis so far we have compared whether one *type* of grade incentive works better than the other. In doing so, 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 to differences in grading schemes. As we have discussed in the Introduction, students in the Netherlands are typically only interested in passing their courses and are not striving to achieve high grades. It is therefore possible that the small difference we observe in the outcomes of the two treatment groups is explained by the majority of students in our sample not being sufficiently responsive to grade incentives.

Besides the examples discussed in the Introduction, our data also contains evidence for the *'zesjescultuur'*. We show that even higher ability students tend to be satisfied with the lowest passing grade by focusing on the group of students who achieved a grade 7 or higher in the midterm exam and received a homework bonus point (in the following referred to as "safe" students). Due to the specific design of the grading schemes[17], these students knew they could pass the course simply by showing up at the end-term exam, without actually solving

[16] The control variables included in the regressions are the following: international program, age, Dutch born, average Math grade, quiz performance, overconfidence, risk aversion, ambiguity aversion, competition aversion and attitude towards relative grading.

[17] The lowest grade awarded under both schemes was not a 1 but a 2, see Appendix 1.

any questions correctly. We find that many of these students, instead of striving for a high final grade, provided low effort in the final exam: while the mean score of "safe" students was 23.358 in the midterm exam, it decreased in the end-term test to 19.182. Tellingly, 108 out of the 148 "safe" students solved fewer questions correctly in the end-term than in the midterm exam. This drop in performance can not be explained by differences in exam difficulty, since "non-safe" students (who were not yet sure to have passed the course before the end-term) only scored 1 point lower on average on the end-term compared to the midterm, and more than half of them actually performed better in the second test[18].

The above evidence illustrates that many students are unwilling to work hard for a higher grade. Since they are not worried about failing the course (less than 2% of them report an expected grade below 6), it is unlikely that they will be responsive grade incentives in general. We will therefore continue our analysis to test whether the response to absolute vs. relative grading is heterogeneous with respect to sensitivity to grade incentives. The proxy we use to measure such "responsiveness" is the study program students follow: we argue that those enrolled in the international program place more weight on the grades in the course.

4.2.3 International program

In this subsection we test whether we find a differential effect of relative grading among the 126 students following the international program. As we have seen in Table 4, these students performed better in their previous courses and have higher grade expectations in the EMO course than their Dutch peers[19]. Furthermore, they are significantly more likely to hand in homework assignments, they receive higher homework grades and report spending more time preparing for the course than students in the Dutch-language program, supporting our claim that they care more about the grade they receive.

Figure 3 shows that the response to relative grading indeed differs by study program: while the performance of students in the Dutch program is unaffected by the grading schemes, male students in the international program achieve significantly higher scores when graded on the curve. Table 6, presenting estimates from OLS regressions confirm that our results are robust to the inclusion of control variables: men in the international program respond significantly more positively to relative grading than female students, while no such difference is observed in the Dutch program. The effect of relative grading on the exam performance of male international students is estimated to be approximately 1.5 to 2.2 points (out of 30) depending on the

[18]We find more indication for the 'just pass' attitude by analyzing results from first-year courses. If all students worked hard to get the highest grade possible and thus the only explanation for low grades would be a lack of ability, than we would expect lower and more variable grades in courses that are Mathematics-intensive than in those that are not (assuming that in the latter effort provision can compensate for skill differences more). We, however, do not find lower average grades in Math-heavy courses such as Micro- and Macroeconomics than in less mathematical classes such as Marketing or Organization and Management, and the difference between the standard deviations is also small.

[19]Even after controlling for past Mathematics grades or performance on the quiz questions, students in the international program have significantly higher grade expectations than those in the Dutch program. We attribute this to differences in ambitious rather than in overconfidence, especially because students in the two programs did not differ in their overconfidence measured in the online incentivized survey.

specification (Table 11 in the Appendix shows the effect of relative grading in the international program, splitting the sample by gender). This corresponds to roughly a third to a half of a standard deviation of the midterm score in the international program.

Since the mean score of women following the international program does not differ between the two schemes, we find that male *relative* performance is higher under grading on the curve than under absolute grading. While women in this subsample outperform men under the absolute scheme (a difference of 1.8 points, p-value = 0.07 from a two-sample t-test with unequal variances), there is no significant gender difference under relative grading.

We continue by analyzing how students perform under the two schemes *compared to their predicted outcomes*. We use correlations from the previous cohort (students who followed the course in the academic year 2012/2013) between students' characteristics and their EMO grades to calculate a grade prediction for students in our sample based on their observables. We then compare the effect of grading schemes on the midterm scores, controlling for the predicted grade. Table 7 shows that among students who care about grade incentives (as proxied by international program), men perform significantly better than predicted when graded on a curve. The corresponding point estimates are negative but insignificant for women. This suggests that men in the international program profited from relative grading while women were at best unaffected by it.

4.3 Robustness checks

In this section we test whether the robustness of our findings. We start by repeating our analysis on a subsample where we exclude students who do not have a clear understanding of the treatments. As mentioned in the Design subsection, directly before writing the midterm exam students were required to answer multiple-choice questions about the color of the group they were assigned to, the grading scheme applied to their midterm test and the interpretation of the schemes. The propensity to report misunderstanding the grading schemes did not differ by treatment assignment. People in the absolute grading group were, however, less likely to complete the questionnaire. As our first robustness check, we exclude the 65 students who either reported to be unsure about the grading schemes or who gave the wrong answers. The first two columns of Table 8 show that our results do not change qualitatively when restricting the sample in this way. The estimates are no longer significant in the international program subsample, probably due to the reduction in sample size.

We also consider whether our findings are sensitive to the regression specifications we use, in particular to the ability measures we include as controls. We have rerun our estimations replacing our ability measures (average Mathematics grades and performance in the quiz) with the average number of retake exams. Using this specification, we no longer find a significant gender effect in the full sample (column 3 of Table 8), while the result becomes even stronger in the international program (column 4, Table 8).

4.4 Preparation

Since assignment to the treatment groups was announced at the beginning of the block, students could take into account their midterm grading scheme when choosing the time and effort to invest in studying. In the following, we compare preparation behavior between the two grading groups. Figure 4 displays students' self-reported time devoted to studying for the EMO course. It indicates no effect of relative grading on study times for women: the time spent on reading lecture notes, solving exercises, homework assignments or practice exams is the same in the two treatment groups. Men, however, report having spent more time on homework when assigned to relative grading (a t-test shows the difference to be significant at the 5% level). Conversely, the higher effort of men in the relative group did not translate into higher homework grades (calculated as the unweighted average of the first two assignments), while female students did perform slightly but significantly better.

Table 9 shows that the propensity to hand in homework assignments was uninfluenced by the treatment group assignment (results from a probit estimation). Even after the inclusion of a rich set of control variables, we find that attendance behavior and (an aggregate measure of) study time does not significantly differ between students in the absolute and the relative groups. We find evidence for a marginally significant difference in homework grades: all else equal, students facing relative grading received higher homework grades. The effect size is relatively small, 0.4 points out of 10, corresponding to approx. one third of a standard deviation. We can conclude that if anything, relative grading had a very small positive impact on the preparation behavior of students in our sample: men report to have spent slightly more time on homework and women prepared higher-quality assignments in anticipation of being graded on the curve. We find the same tendencies when we restrict the analysis to the international program (see Table 12 in the Appendix).

5 Discussion

This section is devoted to the discussion of two issues. First, we explore what mechanism could lead to higher exam scores among responsive men under relative than under absolute grading. Second, we consider the effect of the two grading schemes on the gender gap in performance.

In the previous subsection we found no consistent effect of relative grading on behavior *before* the exam: according to most measures, students in the two treatment groups differed little in how hard they prepared for the course. This result remains true if we limit the analysis to the subsample of international students. This leads us to argue that the increase we observe in the test scores of male international students graded on the curve is not necessarily explained by a better knowledge of the course material. Instead, we speculate that relative grading boosts the effort provision on these students *during the exam itself.* Even though the treatment was announced weeks ahead, it was made very salient for the exam: directly before the test, students were asked to complete a questionnaire related to the treatments, focusing their attention on the particular grading scheme that applied to them. We also included details of the relevant grading schemes on the front page of the exam booklets. Students were thus aware of the grade incentives when working on the test questions.

This explanation is in line with the results of Jalava et al. (2013) and Levitt et al. (2012) who find that incentives introduced unexpectedly before an exam (that could not possibly affect preparation) can have a significant impact on test scores. Similarly, in laboratory experiments we also observe that subjects respond to different incentive schemes by changing their effort provision, influencing their task performance even though their level of skills or knowledge is unaffected (see for instance van Dijk et al. (2001)). The fact that we observe a relatively small effect size in our data (motivated men solve 1.5-2.2 exam questions more correct out of 30) is compatible with the above explanation: on a test designed to capture course-specific knowledge, there is limited room to increase scores by working harder during the exam itself.

The second issue we discuss is how the gender gap in performance is affected by relative grading. We have consistently found a gender difference in response to relative grading: men react more positively to grading on the curve than women. Female students, irrespective of being classified motivated by our proxies, do not significantly differ in their preparation behavior or exam performance under the two grading schemes. Since motivated men do score higher when graded on the curve, relative grading increases male relative performance in these subsamples. This change is most evident in the international program, where women significantly outperform men under the absolute, but not under the relative scheme, as shown in Figure 3. Therefore, higher male relative performance under grading on a curve actually *narrows* the gender gap in exam performance.

We also examine the effect of relative grading on the gender balance among the top and bottom performers. We find that the share of women in the top quartile of midterm scores is higher under absolute than under relative grading (39.2% vs. 33.3%), but this difference is not significant. Similarly, the share of women among in bottom quartile is slightly and insignificantly lower under absolute than under relative grading (26.7% vs. 34.4%). Since the share of female students is 34.3% in the full sample, grading on a curve ensures a more balanced representation of the genders among the best and worse performers. A similar picture emerges if we focus on the subsample of students in the international program.

6 Conclusions

This paper reviews evidence from a field experiment comparing the two most commonly used grading practices: the absolute and the relative scheme. We find that grade incentives have little impact on the preparation behavior of students, suggesting the necessity to explore alternative methods in order to increase their study effort. Focusing on exam scores, we see evidence for gender differences in response to competitive incentives: the performance of female students is the same under absolute and relative grading, while men who are motivated by grade incentives perform better under rank-based grading. These findings suggest that absolute grading does not provide sufficient incentives for men to exert effort in the exam. Relative grading improves the exam outcomes of responsive male students (both in absolute terms and in comparison to their predicted performance) while leaving the exam scores of women unchanged. Our findings also show the limitations of competitive grade incentives: they do not increase the test scores of unmotivated male students, the group policy makers would primarily like to target.

We find it an important direction for future research to test whether our findings replicate in different environments. First, students in our sample were accustomed to an absolute grading scheme and for the most part never experienced relative grading during their university studies. It is possible that incentives work differently among students for whom the "default" scheme is competitive grading. Second, women in our study, although underrepresented in the Bachelor program, slightly outperformed men during their previous courses and under absolute grading. A decrease in female relative performance would be much more alarming in environments where the opposite was true, as discussed by Niederle and Vesterlund (2010).

7 Tables

Variable	Mean	Std. Dev.	Min.	Max.	N
Panel A					
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					
lecture attendance *(out of 3)*	1.21	0.94	0	3	517
tutorial attendance *(out of 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-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-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 complete *(0/1)*	0.92	0.28	0	1	529
quiz 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)*	6.10	1.91	1	11	487
ambiguity aversion *(scale 0-10)*	7.44	3.14	1	11	487
competition aversion *(scale 0-10)*	6.42	2.83	1	11	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)*	8.88	1.82	1	11	485
attitude relative grading *(scale 0-10)*	5.33	2.75	1	11	485

Table 3: Summary statistics for demographic variables, preparation and exam performance and survey outcomes

	Section I.			Section II.			Section III.		
	“Blue”	**“Yellow”**	**Diff.**	**Men**	**Women**	**Diff.**	**Dutch**	**Int.**	**Diff.**
Demographics									
int. program	0.259	0.247		0.199	0.348	***	-	-	
	(0.027)	(0.027)		(0.022)	(0.036)				
female	0.341	0.346		-	-		0.299	0.477	***
	(0.029)	(0.030)					(0.023)	(0.044)	
age	20.890	20.782		20.874	20.766		20.880	20.677	
	(0.147)	(0.117)		(0.113)	(0.170)		(0.110)	(0.176)	
Dutch born	0.738	0.748		0.795	0.652	***	0.924	0.224	***
	(0.027)	(0.027)		(0.022)	(0.036)		(0.014)	(0.036)	
Ability									
Math grade	5.890	5.861		5.740	6.141	***	5.717	6.504	***
	(0.096)	(0.101)		(0.084)	0.122		(0.077)	(0.148)	
num. retakes	0.226	0.220		0.237	0.196	*	0.240	0.153	***
	(0.014)	(0.015)		(0.013)	(0.017)		(0.012)	(0.022)	
quiz questions	4.730	4.612		4.738	4.517		4.579	4.936	**
	(0.102)	(0.112)		(0.096)	(0.120)		(0.087)	(0.151)	
Preferences									
overconfidence	16.857	19.609		17.299	20.093		17.743	19.640	
	(1.847)	(1.952)		(1.700)	(2.194)		(1.554)	(2.680)	
risk aversion	6.111	6.086		5.892	6.465	***	6.019	6.328	
	(0.126)	(0.119)		(0.108)	(0.141)		(0.099)	(0.175)	
ambig. aversion	7.660	7.222		7.519	7.308		7.409	7.536	
	(0.202)	(0.201)		(0.179)	(0.238)		(0.166)	(0.278)	
comp. aversion	6.369	6.471		6.364	6.517		6.460	6.304	
	(0.184)	(0.180)		(0.153)	(0.232)		(0.148)	(0.261)	
expected grade	6.988	7.087		7.074	6.959		6.850	7.576	***
	(0.057)	(0.056)		(0.051)	(0.065)		(0.043)	(0.079)	
N	270	259		346	181		395	134	

Standard errors in parentheses. Significance of difference from two-sample t-test with equal variances

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Comparison of means between groups

midterm score	**Full sample**			**Subsamples**	
	No controls	Gender interact.	With controls	Men	Women
relative	0.170	-0.064	-0.646	0.563	-0.721
	(0.376)	(0.607)	(0.482)	(0.400)	(0.518)
relative*male		0.300	1.223**		
		(0.729)	(0.594)		
male		-0.431	-0.864*		
		(0.504)	(0.448)		
Demographic controls	-	-	✓	✓	✓
Ability controls	-	-	✓	✓	✓
Preference controls	-	-	✓	✓	✓
N	483	482	419	272	147
R^2	0.000	0.002	0.275	0.280	0.342

Clustered standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: OLS regressions explaining midterm exam performance

midterm score	**International**		**Dutch**	
	No controls	With controls	No controls	With controls
relative	-0.333	-1.101	0.015	-0.389
	(1.072)	(0.770)	(0.706)	(0.583)
relative*male	2.576*	2.578**	-0.259	0.657
	(1.480)	(1.101)	(0.776)	(0.667)
male	-1.830*	-2.101**	0.233	-0.377
	(0.932)	(0.780)	(0.583)	(0.504)
Demographic controls	-	✓	-	✓
Ability controls	-	✓	-	✓
Preference controls	-	✓	-	✓
N	126	90	356	329
R^2	0.053	0.500	0.001	0.206

Clustered standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: International vs. Dutch program

midterm	**Male students**		**Female students**	
	Full sample	International	Full sample	International
relative	0.356	1.840*	-0.346	-0.541
	(0.424)	(0.917)	(0.475)	(0.787)
predicted grade	1.676***	2.170***	1.733***	2.335***
	(0.192)	(0.452)	(0.225)	(0.541)
N	281	46	150	47
R^2	0.182	0.373	0.253	0.369

Clustered standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Discussion: predicted and actual performance

midterm score	**Restricted sample** Full sample	International	**Different ability control** Full sample	International
relative	-0.606	-0.806	-0.511	-1.178
	(0.477)	(0.830)	(0.577)	(1.088)
relative*male	1.230**	1.714	0.965	3.453**
	(0.619)	(1.201)	(0.720)	(1.296)
male	-0.887**	-1.209	-0.783*	-2.127*
	(0.446)	(0.938)	(0.469)	(1.063)
Demographic controls	✓	✓	✓	✓
Ability controls †	✓	✓	-	-
Number of retakes	-	-	✓	✓
Preference controls	✓	✓	✓	✓
N	365	79	426	90
R^2	0.275	0.406	0.160	0.296

Clustered standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

†: average Mathematics grades and performance in the quiz

Table 8: Robustness checks

	handing in HW	***avg. HW***	***attendance***	***study time***
relative	0.345	0.665**	0.011	-0.054
	(0.257)	(0.307)	(0.248)	(0.129)
relative*male	-0.096	-0.252	0.033	0.072
	(0.300)	(0.326)	(0.275)	(0.166)
male	-0.301	0.124	-0.014	-0.239*
	(0.195)	(0.275)	(0.212)	(0.126)
Demographic controls	limited	✓	✓	✓
Ability controls	-	✓	✓	✓
Preference controls	-	✓	✓	✓
Tutor controls	-	✓	-	-
N	527	363	429	334
R^2		0.206	0.026	0.109

Clustered standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Regressions explaining preparation

8 Figures

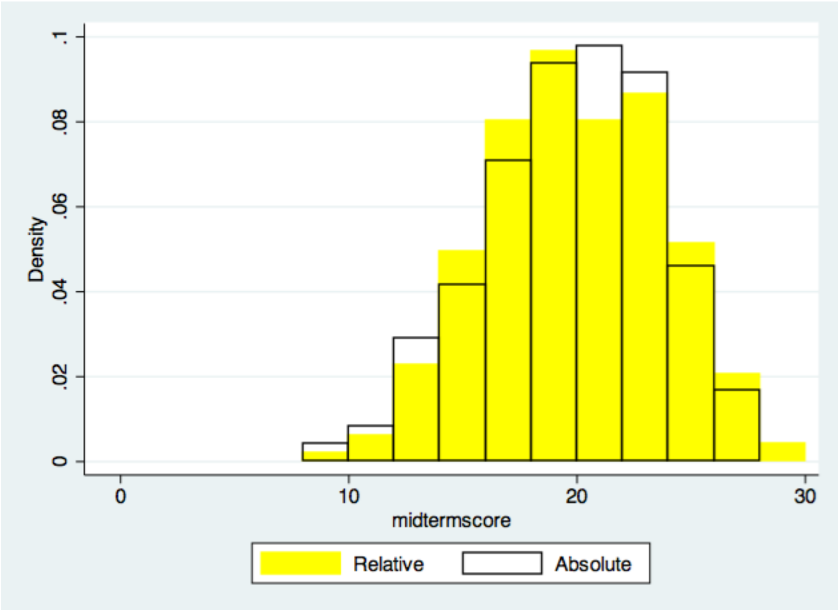


Figure 1: Distribution of midterm exam scores by treatment group

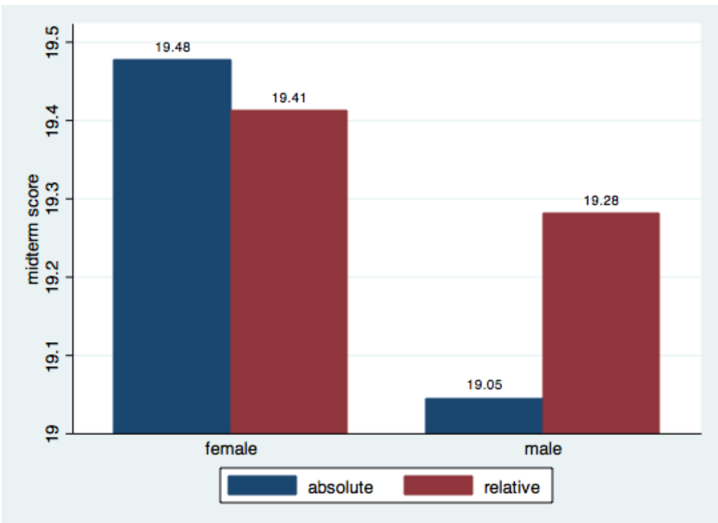


Figure 2: Comparison of midterm exam scores by treatment group and gender

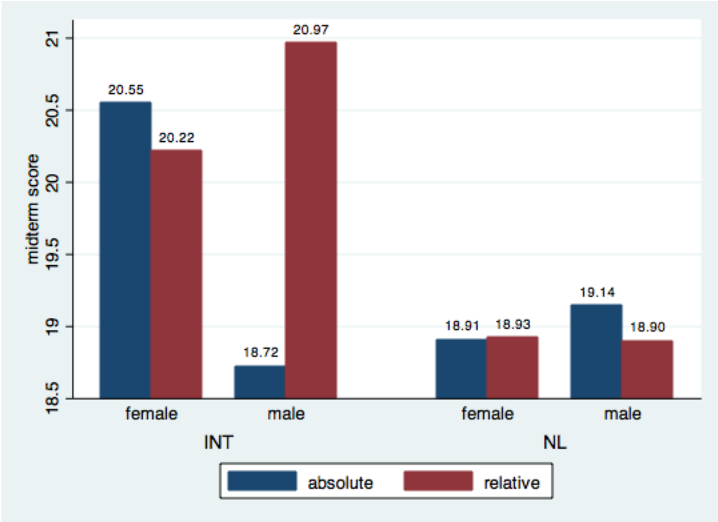


Figure 3: Midterm exam scores by treatment group and gender, separately by study program

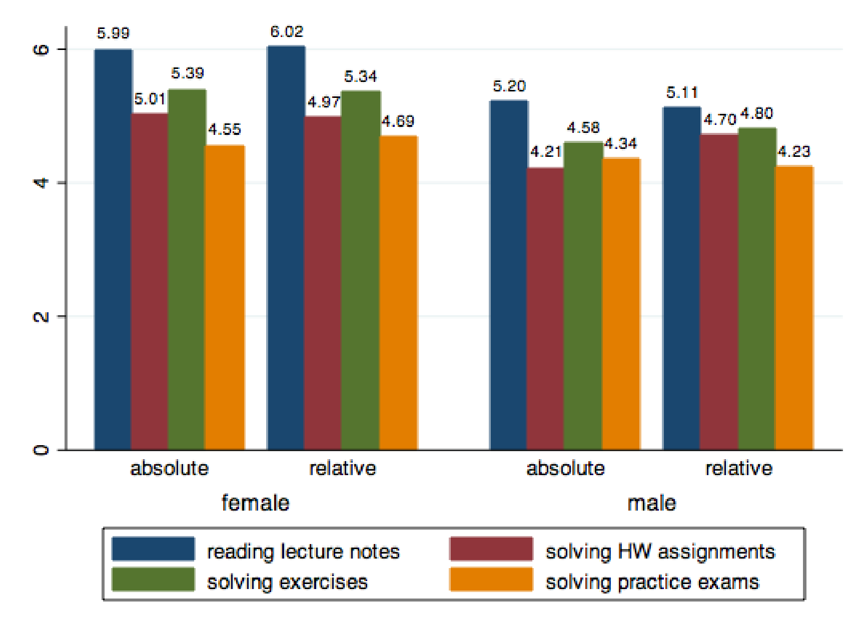


Figure 4: Self-reported study times on different activities by treatment group and gender

References

Bigoni, Maria, Margherita Fort, Mattia Nardotto, and Tommaso Reggiani, "Teams or Tournaments? A Field Experiment on Cooperation and Competition among University Students," IZA Discussion Papers 5844, Institute for the Study of Labor (IZA) July 2011.

Braun, Henry, Irwin Kirsh, and Kentaro Yamamato, "An Experimental Study of the Effects of Monetary Incentives on Performance on the 12th-Grade NAEP Reading Assessment," *Teachers College Record*, 2011, *113* (11), 2309–2344.

Budryk, Zack, "Dangerous Curves," *Inside Higher Ed*, 2013, *12 February.*

Bull, Clive, Andrew Schotter, and Keith Weigelt, "Tournaments and Piece Rates: An Experimental Study," *Journal of Political Economy*, February 1987, *95* (1), 1–33.

Croson, Rachel and Uri Gneezy, "Gender Differences in Preferences," *Journal of Economic Literature*, 2009, *47* (2), 448–74.

del Pero, Angelica Salvi and Alexandra Bytchkova, "A Bird's Eye View of Gender Differences in Education in OECD Countries," *OECD Social, Employment and Migration Working Papers*, 2013, (149).

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jurgen Schupp, and Gert G. Wagner, "Individual risk attitudes: measurement, determinants and behavioral consequences," *Journal of the European Economic Association*, 2011, *9* (3), 522–550.

Filippin, Antonio and Paolo Crosetto, "A Reconsideration of Gender Differences in Risk Attitudes," Departmental Working Papers 2014-01, Department of Economics, Management and Quantitative Methods at Università degli Studi di Milano January 2014.

Fryer, Roland G., "Financial Incentives and Student Achievement: Evidence from Randomized Trials," *The Quarterly Journal of Economics*, 2011, *126* (4), 1755–1798.

Gneezy, Uri and Aldo Rustichini, "Gender and Competition at a Young Age," *American Economic Review*, May 2004, *94* (2), 377–381.

_, Muriel Niederle, and Aldo Rustichini, "Performance In Competitive Environments: Gender Differences," *The Quarterly Journal of Economics*, August 2003, *118* (3), 1049–1074.

Green, Jerry R and Nancy L. Stokey, "A Comparison of Tournaments and Contracts," *Journal of Political Economy*, 1983, *91* (3), 349–64.

Grove, Wayne A. and Tim Wasserman, "Incentives and Student Learning: A Natural Experiment with Economics Problem Sets," *The American Economic Review*, 2006, *96* (2), pp. 447–452.

Guiso, Luigi, Ferdinando Monte, Paola Sapienza, and Luigi Zingales, "Culture, Gender, and Math," *Science*, 2008, *320* (5880), 1164–1165.

Gunther, Christina, Neslihan Arslan Ekinci, Christiane Schwieren, and Martin Strobel, "Women can't jump?–An experiment on competitive attitudes and stereotype threat," *Journal of Economic Behavior & Organization*, September 2010, *75* (3), 395–401.

Harrison, Glenn W. and John A. List, "Field Experiments," *Journal of Economic Literature*, 2004, *42* (4), pp. 1009–1055.

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

Holt, Charles A. and Susan K. Laury, "Risk Aversion and Incentive Effects," *American Economic Review*, 2002, *92* (5), 1644–1655.

Inspectie van het Onderwijs, "De staat van het onderwijs. Onderwijsverslag 2012/2013," *Inspectie van het Onderwijs*, 2014, *April.*

Jalava, Nina, Juanna Schrøter Joensen, and Elin Maria Pellas, "Grades and Rank: Impacts of Non-Financial Incentives on Test Performance," Technical Report, Available at SSRN December 2013.

Jaspers, Arnout, "Een voldoende, zonder uitschieters," *Wetenschap 24*, 2012, *12 December.*

Jurajda, Stepan and Daniel Munich, "Gender Gap in Performance under Competitive Pressure: Admissions to Czech Universities," *American Economic Review*, 2011, *101* (3), 514–18.

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

Lazear, Edward P. and Sherwin Rosen, "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, 1981, *89* (5), pp. 841–864.

Leuven, Edwin, Hessel Oosterbeek, and Bas van der Klaauw, "The Effect of Financial Rewards on Students' Achievement: Evidence from a Randomized Experiment," *Journal of the European Economic Association*, 2010, *8* (6), 1243–1265.

Levitt, Steven D., John A. List, Susanne Neckermann, and Sally Sadoff, "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance," Working Paper 18165, National Bureau of Economic Research June 2012.

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

Niederle, Muriel and Lise Vesterlund, "Do Women Shy Away from Competition? Do Men Compete Too Much?," *The Quarterly Journal of Economics*, 08 2007, *122* (3), 1067–1101.

_ **and** _ , "Explaining the Gender Gap in Math Test Scores: The Role of Competition," *Journal of Economic Perspectives*, 2010, *24* (2), 129–44.

_ **and** _ , "Gender and competition," *Annual Review of Economics*, September 2011, *3*, 601–630.

Ors, Evren, Frédéric Palomino, and Eloïc Peyrache, "Performance Gender Gap: Does Competition Matter?," *Journal of Labor Economics*, 2013, *31* (3), pp. 443–499.

Paola, Maria De, Francesca Gioia, and Vincenzo Scoppa, "Are Females Scared of Competing with Males? Results from a Field Experiment," IZA Discussion Papers 7799, Institute for the Study of Labor (IZA) December 2013.

Price, Joseph, "Gender Differences in the Response to Competition," *Industrial and Labor Relations Review*, April 2008, *61* (3), 320–333.

Rosin, Hanna, *The End of Men: And the Rise of Women*, Riverhead Books. Penguin Group, 2012.

Sax, Leonard, *Boys Adrift: The Five Factors Driving the Growing Epidemic of Unmotivated Boys and Underachieving Young Men*, Basic Books, 2007.

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

Sommers, Christina Hoff, *The War against Boys*, Simon & Schuster, 2000.

The Economist, "The trouble with boys," *The Economist*, 1999, *May 27.*

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

Appendix 1: Sections from the Course Manual

Grading the exams

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:

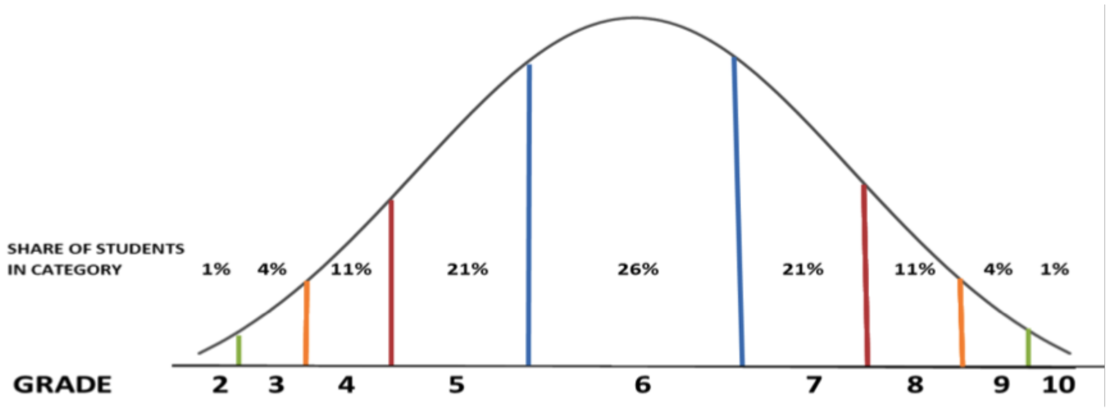
$$\text{Grade exam} = 10 - 0.4^{*}(\text{number of errors})$$

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.

Appendix 2: Screenshots from the survey

Figure 5: Example of a multiple-choice quiz 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 6: 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%
[illegible]	[illegible]	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

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

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

Appendix 3: Additional tables

midterm score	(1)	(2)	(3)	(4)	(5)
relative	-0.010	0.614	-0.046	-1.701**	-0.500
	(0.722)	(1.441)	(1.079)	(0.816)	(0.498)
relative*male	1.344**	1.158*	1.167*	1.269**	1.199**
	(0.590)	(0.606)	(0.601)	(0.609)	(0.593)
male	-0.908**	-0.833*	-0.826*	-0.869*	-0.858*
	(0.455)	(0.450)	(0.450)	(0.454)	(0.450)
relative*Dutch	-0.894				
	(0.674)				
relative*Math		-0.203			
		(0.210)			
relative*risk			-0.092		
			(0.150)		
relative*comp.				0.158	
				(0.100)	
relative*overconf.					-0.007
					(0.008)
Demographic controls	✓	✓	✓	✓	✓
Ability controls	✓	✓	✓	✓	✓
Preference controls	✓	✓	✓	✓	✓
N	419	419	419	419	419
R^2	0.277	0.276	0.275	0.278	0.275

Clustered standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 10: OLS regressions including interactions

	Men		**Women**	
midterm score	No controls	With controls	No controls	With controls
relative	2.243**	1.497*	-0.333	-0.823
	(1.005)	(0.769)	(1.075)	(0.725)
Demographic controls	-	✓	-	✓
Ability controls	-	✓	-	✓
Preference controls	-	✓	-	✓
N	65	45	61	45
R^2	0.089	0.524	0.002	0.573

Clustered standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 11: Sample split by gender in the international program

	handing in HW	***avg. HW***	***attendance***	***study time***
relative	4.039	0.756**	0.144	0.136
	(267.403)	(0.278)	(0.358)	(0.205)
relative*male	-3.631	-0.460	0.055	0.217
	(267.404)	(0.336)	(0.634)	(0.292)
male	-0.092	0.635**	-0.038	-0.402**
	(0.458)	(0.284)	(0.378)	(0.180)
Demographic controls	limited	✓	✓	✓
Ability controls	-	✓	✓	✓
Preference controls	-	✓	✓	✓
N	132	90	90	74
R^2		0.442	0.119	0.283

Clustered standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 12: Regressions explaining preparation in the international program