Heterogeneity in response to incentives: Evidence from field data

Czibor, E.

Publication date
2015

Document Version
Final published version

Link to publication

Citation for published version (APA):
Czibor, E. (2015). *Heterogeneity in response to incentives: Evidence from field data*. [Thesis, fully internal, Universiteit van Amsterdam].

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	RELATIVE GRADING
	Exam score	Relative rank
	(=points earned)	*(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).

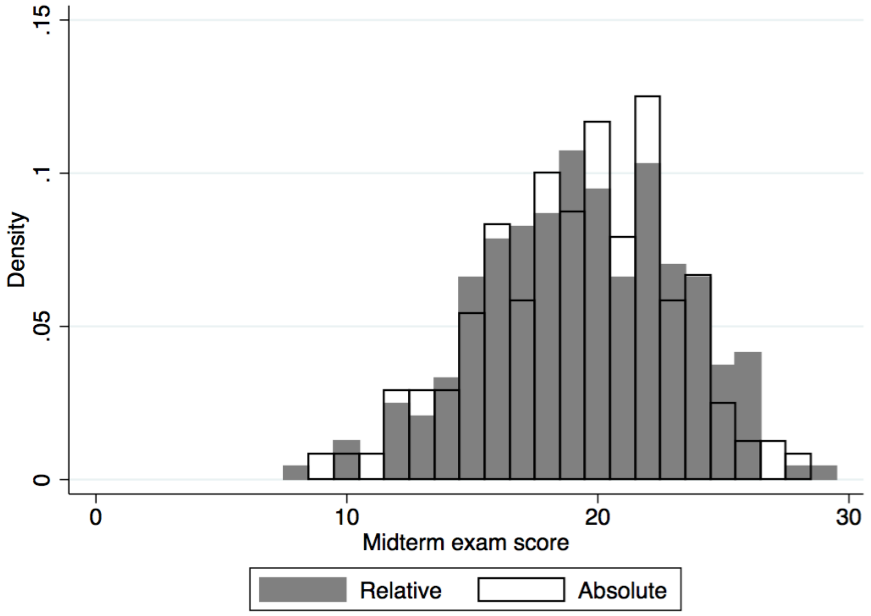


Figure 2.1: Distribution of midterm exam scores by grading scheme

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.

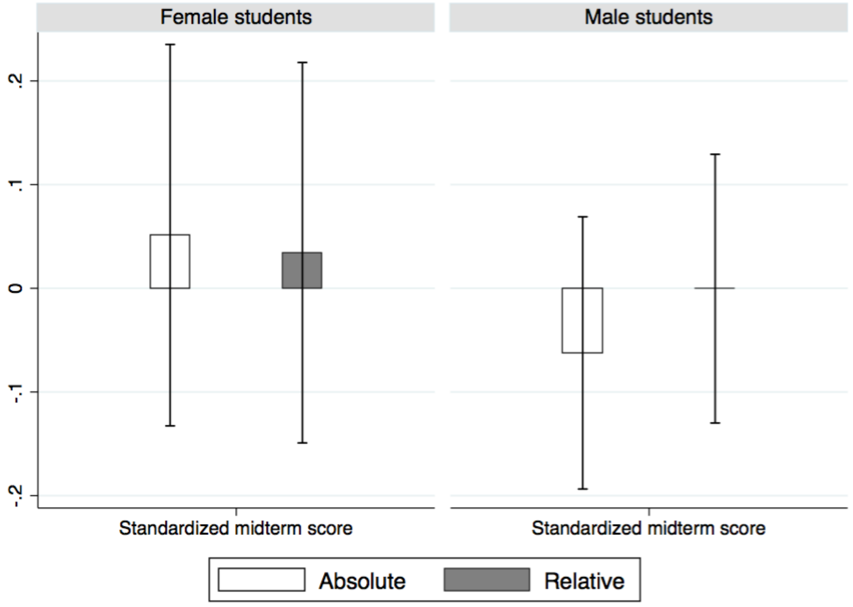


Figure 2.2: Comparison of midterm exam scores by grading scheme and gender (the height of the bars represent the mean and the error bars show the 90% confidence intervals)

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

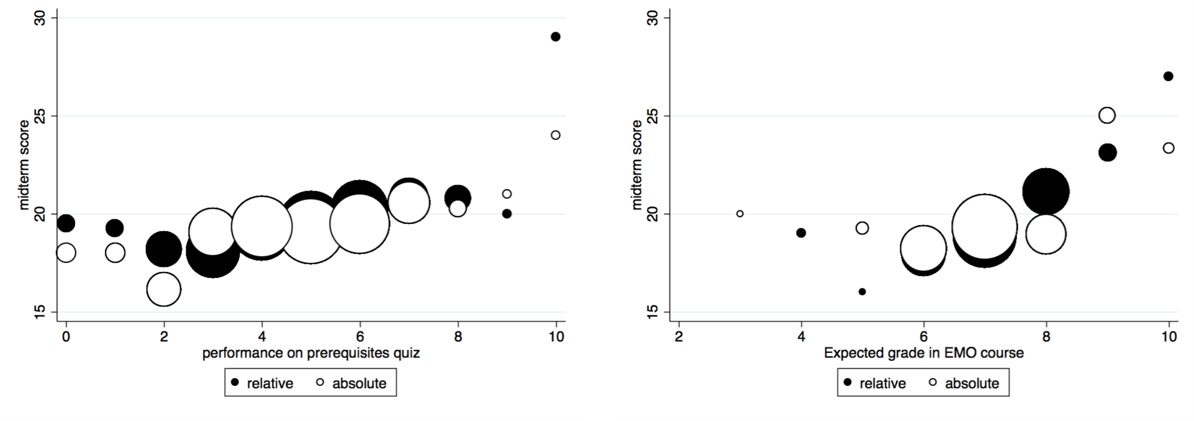
In our analysis so far we only included ability variables as covariates to increase the precision of our estimates. Inspired by the findings of Paredes (2012) and Müller and Schotter (2010) we now consider whether response to grade incentives is heterogeneous with respect to students' skills as captured 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 (1)	Risk aversion (2)	Competitiveness (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

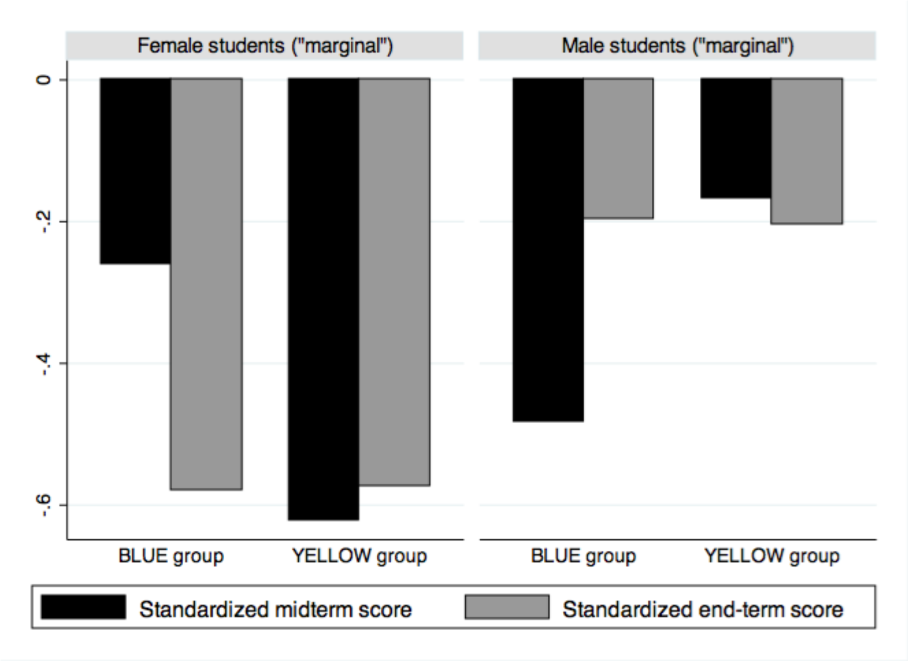


Figure 2.4: Subsample of marginal students: mean mid- and end-term (standardized) exam scores by treatment group and gender

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

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

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

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

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 ($\varphi = 1$), $H(\beta) = F\left(\frac{\beta}{1+\bar{\rho}}\right)$. As a consequence, $F^{-1}(H(\beta)) = \frac{\beta}{1+\bar{\rho}}$. The equilibrium bidding curve is given by

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

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

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

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:

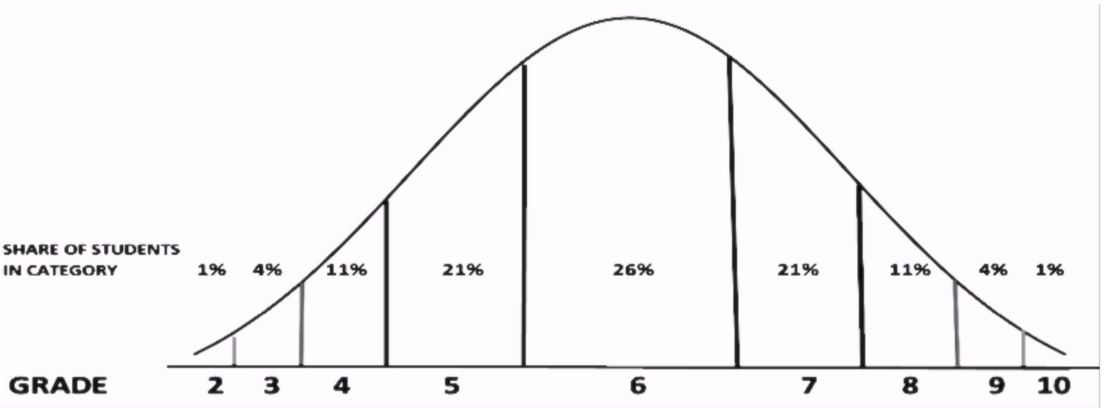
Grade exam = 10 0.4*(number of errors)

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

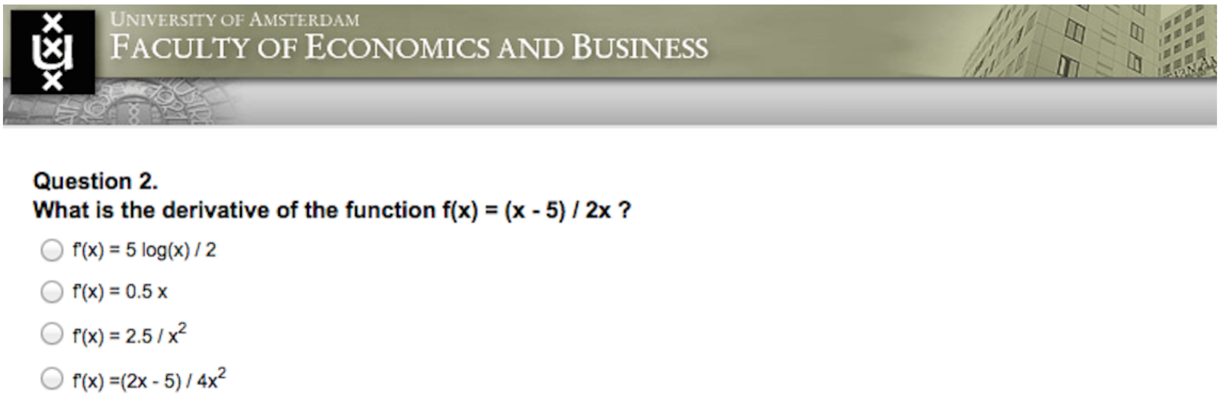
YELLOW group: midterm exam graded under relative, final exam graded under absolute scheme

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



UNIVERSITY OF AMSTERDAM
FACULTY OF ECONOMICS AND BUSINESS

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

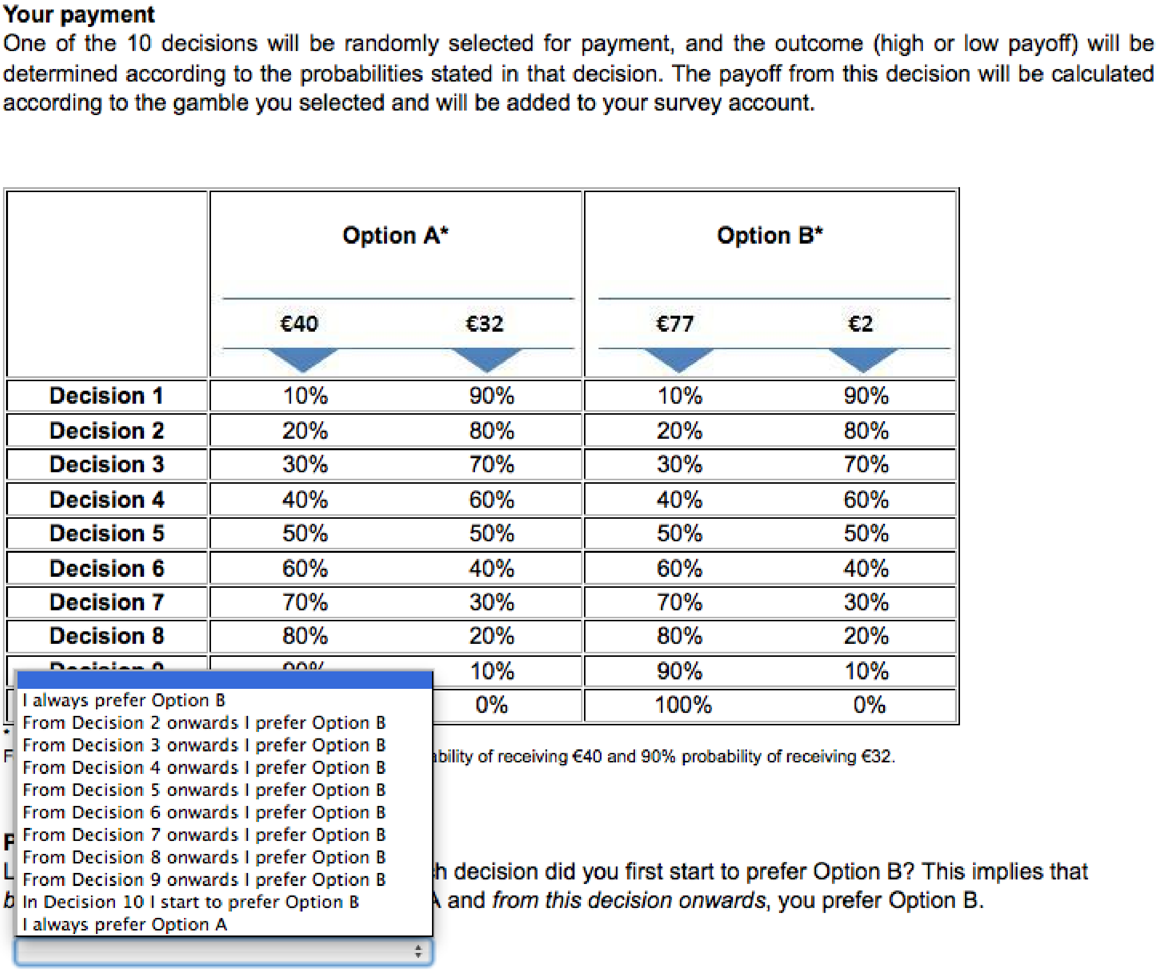
○ f'(x) = 5 log(x) / 2

○ 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%
[illegible]	[illegible]	0%	100%	0%

*

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

P

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.

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

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

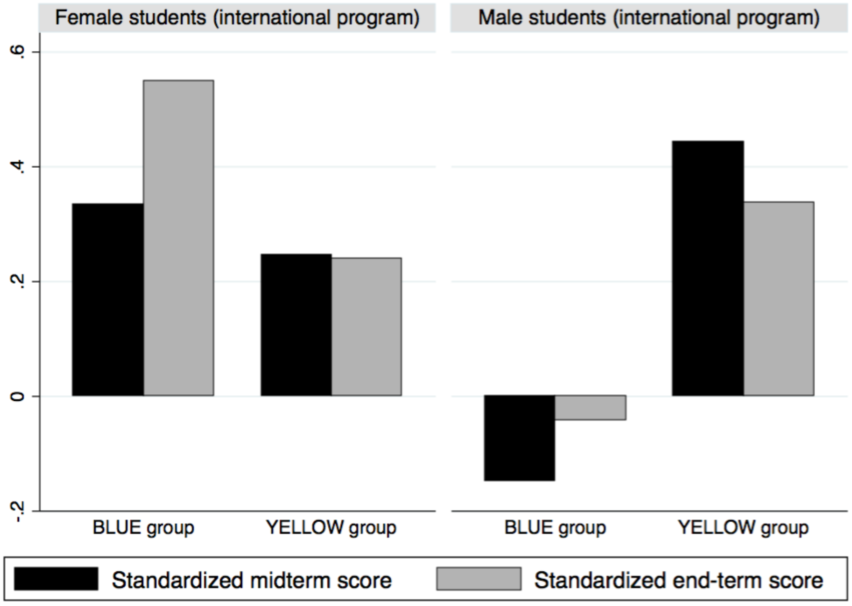


Figure D1: International program: mean mid- and end-term (standardized) exam scores by treatment group and gender