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

Eszter Czibor[1], Sander Onderstal[2], Randolph Sloof[2] and Mirjam van Praag[2,3]

[1]University of Chicago [2]University of Amsterdam [3]Copenhagen Business School

September 2016

Abstract

We conduct a framed field experiment in a Dutch university to compare student effort provision and exam performance under the two most prevalent evaluation practices: absolute (criterion-referenced) and relative (norm-referenced) grading. Based on the empirical stylized fact of gender differences in competitiveness we hypothesize that the rank-order tournament created by relative grading will increase male, but not female, performance. Contrary to our expectations, we find no impact of competitive grading on preparation behavior or exam scores among either gender. Our result may be attributed to the low value students in our sample attach to academic excellence.

JEL codes: I21, I23, A22, D03, C93

Keywords: grade incentives, competition, education, gender, field experiment

*Acknowledgments: We are grateful to the Examination Board of the Faculty of Economics and Business at the University of Amsterdam for authorizing the field experiment, and to the Research Priority Area Behavioral Economics of the University of Amsterdam for the generous financial support they provided. We wish to thank Alexander Cappelen, Uri Gneezy, John List, Hessel Oosterbeek, Erik Plug, Andrew Schotter, Bertil Tungodden and Lise Vesterlund for their invaluable advice and inspiration, Sanne Blauw, Thomas Buser, Adam Eric Greenberg, Laura Rosendahl Huber and Dana Sisak for their helpful comments and insights, and Inge Beks for excellent research assistance.

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). This study aims to test whether the problem of poor student motivation could be tackled with the help of competitive grade incentives, and whether the response to such incentives differs by gender. In a large-scale field experiment conducted among students at the Economics department of a Dutch university - a population shown to provide insufficient study effort (Leuven *et al.*, 2010) - 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.[1] A key difference between the two grading schemes is that relative grading induces direct competition between peers: grading on a curve introduces a rank-order tournament in the classroom while absolute grading is analogous to a piece rate incentive scheme.

The advantageous and disadvantageous incentive effects of competitive reward schemes have been studied extensively. Early theoretical contributions by Lazear & Rosen (1981) and Green & 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)). 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 & Irlenbusch, 2003; van Dijk *et al.*, 2001).

A few theoretical studies focus specifically on the comparison between absolute and relative grading. Becker & Rosen (1992) and Landeras (2009) show that grading on a curve can induce higher performance than absolute grading in the presence of "systemic" noise or correlated individual error terms. Dubey & Geanakoplos (2010) find that absolute grading provides better incentives in case student outcomes are independent. Paredes (forthcoming) predicts that low-ability students exert less and high-ability students exert

[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 or "grading on a curve," referring to the bell-shaped curve of the normal distribution. In the United States, colleges typically implement relative grading (as an example, consider the overview by Levit & Downs (1997) who report that 84% of the surveyed law schools have some form of grade normalization policy), while in continental Europe, the absolute scheme prevails (Karran, 2004).

more effort under absolute than under relative grading. While most studies derive optimal effort provision under the (in practice rather unrealistic) assumption that schemes are calibrated efficiently, we compare outcomes in case the grading curve is set such that it imposes the same distribution of exam grades as we expect under absolute grading. On the basis of a theoretical model we hypothesize that in such a setting students with competitive preferences (i.e. who care about their rank) will exert more effort under relative than under absolute grading.

Recent contributions from the behavioral economics literature emphasize the importance of competitive preferences: people derive utility from obtaining a higher rank even in the absence of any tangible benefits (Charness & Rabin, 2002).[2] Attitudes towards competition, however, have been shown to differ by gender. 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.[3] Combining the predictions of our theoretical model with the empirical stylized fact of gender differences in preferences, we expect male students to perform better and female students to perform worse when graded on the curve. Our study thus aims to test whether competitive grade incentives help male students while being disadvantageous for females, as argued by e.g. Niederle & Vesterlund (2010).[4]

Empirical studies on the effect of competition in education are still scarce. Jurajda & 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 and/or a decrease in female absolute performance. Bigoni *et al.* (2015) 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 con-

[2] Azmat & Iriberri (2010a); Blanes i Vidal & Nossol (2011); Kosfeld & Neckermann (2011); Tran & Zeckhauser (2012) provide convincing field evidence that feedback on relative performance can increase performance. Barankay (2012), on the other hand, finds in a field experiment that removing rank feedback increases the performance of male employees.

[3] Their result has been replicated using both laboratory (e.g. Günther *et al.* (2010)) and field experiments (e.g., Gneezy & Rustichini (2004)) as well as naturally occurring data (e.g., Price (2008)). Niederle & Vesterlund (2011) and Croson & Gneezy (2009) provide detailed reviews of studies on gender and competition.

[4] Gender differences in response to incentives in education have been documented in a variety of non-competitive settings, see e.g. Angrist *et al.* (2009); Lindo *et al.* (2010).

clude 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.

This study 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 is implemented as an actual grading curve where a student's exam grade is determined by his or her place in the class score distribution, closely resembling real-life grading practices.[5] The experiment is conducted in a natural setting among students attending a university course ("in the classroom"). The number of participants is high, and students are randomly assigned to treatments 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.

Our results show no clear difference in preparation effort, exam participation or test performance under the two grading schemes. Contrary to our expectations, we do not find a gender difference in response to being graded on the curve: neither male nor female students in our sample seem responsive to competitive grade incentives. There is no heterogeneity in reaction to our intervention by ability or preferences, either. This essentially null result is obtained across different model specifications and identification techniques.

We argue that our findings are 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 our findings may be explained a low willingness to compete for high grades (consistent with the results of Buser *et al.* (2014)). This apparent lack of interest in academic excellence is consistent 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.[6]

[5] The above-mentioned field experiments implemented competitive grading in the form of comparison to a randomly chosen opponent or a reward for the top performers only.

[6] 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'

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

We interpret our finding as proof that competitive grade incentives do not solve the motivation crisis in a sample that does not sufficiently value academic excellence. Our results highlight the need for policies that either increase the importance of grades for students or offer additional incentives to study.

The remainder of this paper is organized as follows. In Section 2 we derive our hypotheses and introduce the context and details of our experimental design. Section 3 provides an overview of our data. Results are presented in Section 4. We discuss our findings in Section 5 and conclude in Section 6.

2 Context and design

2.1 Hypotheses

In this section we briefly review the theoretical considerations underlying the design of our experiment. For a meaningful comparison between a criterion- and a norm-referenced grading system, Landeras (2009) emphasizes that both schemes should be implemented *efficiently*, allowing researchers 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.[7] 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.

We analyze by means of a simple theoretical model (presented in Appendix A) the utility maximization problem of students under absolute and relative grading. The model accounts for heterogeneity in student ability and assumes an effort-dependent, idiosyncratic noise term when translating effort into exam scores. We first consider a general version of the model where students' utility depends on the level of their grade, their

that calculates what test mark students are required to get in order to achieve an average final grade of six.

[7] Consider Moldovanu & Sela (2001), and the discussion in van Dijk *et al.* (2001) on choosing the payoffs in the tournament condition.

rank in terms of grades, their effort cost and their ability parameter. We show that in case the curve is set such that the grade distribution is 'forced' to be the same under the two schemes, the two grading schemes should induce the same equilibrium effort level, regardless ability (Proposition 1 in Appendix A). This leads us to formulate our null hypothesis:

Null hypothesis: Students exert the same effort and perform equally well on exams graded under the absolute and the relative schemes.

We then adapt the model by letting rank utility depend on whether or not students obtain explicit information about their rank: the more information students obtain about their relative performance, the more utility they gain from a high rank (e.g. because they can credibly inform fellow students or future employers about their relative performance). For ease of exposition, we discuss the extreme case where students only care about their rank if they are explicitly informed about it. It follows from this assumption that the extent to which students care about their rank in the grading distribution depends on the scheme used to determine their grades. In particular, students under absolute grading do not obtain rank utility as they only receive explicit information about their absolute performance: marks themselves carry no explicit information on rank, and no list of grades or cohort averages are posted publicly. It is thus up to the students to collect information on how their grades compare to that of their peers, which in case of a large class size can only result in noisy estimates even if everyone reports their grades truthfully. On the other hand, relative grading by definition provides students with explicit information about their rank as grades in this scheme correspond directly to students' relative performance.

Under this additional assumption, our model predicts that the subset of students with competitive preferences 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 (Propositions 4 and 5 in Appendix A). We combine these results with the standard empirical finding of gender differences in competitive preferences to obtain our alternative hypothesis:

Alternative 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.

2.2 Context

To test our hypotheses, we conducted a framed field experiment (Harrison & List, 2004) among students of the University of Amsterdam (UvA). The experiment was authorized by the Examination Board of the Faculty of Economics and Business. Our intervention took place in the 2$^{\text{nd}}$ year BSc course *Economics of Markets and Organizations* (EMO) during the first block of the 2013/2014 academic year.[8] The course covered topics from Organizational Economics and Industrial 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.[9] The attrition rate was low (only 9% of students missed the midterm exam) 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.

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 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 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. 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.[10]

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

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

[10] 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

2.3 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 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. This reversal of grading schemes is required to perform the experiment while ensuring *ex ante* fair and equal treatment of our subjects, a necessary requirement for approval by the Examination Board. 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).

Table 1: DESIGN OVERVIEW: Treatment groups and grading schemes

	"Blue" group	"Yellow" group
Midterm exam	absolute	relative
End-term exam	relative	absolute

Our design allows for two different techniques to identify the impact of relative vs. absolute grading on preparation effort and exam scores. First, in a between-subject comparison we can contrast the midterm exam outcomes of students in the two treatment groups. Since students were randomly assigned to grading schemes, any difference between the groups' midterm behavior is attributable to our treatment.[11] Second, we can take advantage of the within-subject nature of our design by pooling together results from the mid- and end-term exams and estimating regressions with individual and exam fixed effects. Both identification strategies rely on certain assumptions. The between-subject comparison requires that students optimize their effort provision for the midterm exam by focusing only on the midterm grading scheme they are assigned to, without considering the end-of-term scheme that awaits them. The within-subject identification assumes

had to retake the course the following academic year.

[11] A simple comparison of the treatment groups' outcomes on the end-term exam could lead to biased outcomes since students assigned to the two treatments may no longer be comparable: as a result of the different conversion of exam scores to grades, the two groups could potentially receive different midterm grades, leading to systematic differences between them already prior to 'being treated' with the end-term grading scheme.

that there are no order effects: the response to relative grading in the second exam is the same as it is in the first exam.[12]

Our main variable of interest is the score, i.e., the number of correct answers obtained on the midterm (and for the within-subject analysis, also the end-of-term) exam. We also consider several proxies for effort provision in preparation for the exams: 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.

Table 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, **End-of-term exam**

The timeline of the experiment is shown in Table 2. Students were informed of their treatment group assignment during the first week of the course 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 B) and were also announced during the lectures and tutorials. The next week, preference and ability information was collected from students in an online survey (discussed in more detail in Section 2.5). Homework assignments were due in weeks 3, 4, 6 and 7. 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. The midterm exam took place in the fourth week of the course. Right before the midterm exam, students were required to fill out a short questionnaire testing their understanding of the grading schemes and collecting

[12] Neither of these assumptions are directly testable given the data available to us: we lack a control group that was subject to absolute grading on both exams, and an 'order of treatment' indicator variable would be perfectly collinear with the combination of the exam fixed effect and the relative grading dummies. If the fact that schemes are reversed for the second exam simply dilutes the incentives experienced by students, our results might be biased toward 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.

information on the time they spent studying for the course. Homework results were not published before week 5, so students did not receive any feedback on their relative performance before the midterm exam. The end-of-term exam was conducted in the final week of the course.

2.4 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 (EMO)* 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 (where the grade is rounded to the nearest integer, with a minimum of 2):

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

With 30 exam questions in total, this formula leads to the standards described in the first column of Table 3.

Table 3: THE GRADING SCHEMES

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

Under relative grading, students' exam grades are 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.[13] Under the assumption that student ability and exam difficulty are unchanged over time, setting the curve this way "forces" the grade distribution to be the same under the two schemes. As discussed in Section 2.1, as a consequence of this design choice we expect students to exert the same effort in both treatment groups if rank utility is independent of whether grades convey explicit information on relative performance (our null hypothesis). In case rank utility depends on how explicit signals grades are of relative performance, our model predicts that competitive students increase and non-competitive students decrease their effort provision under relative grading (our alternative hypothesis).

Our design choice, besides being prompted by theoretical considerations, was also motivated by fairness concerns: we did not want to *ex ante* impose a stricter or more lenient standard on either treatment group. For practical purposes we designed the curve to be symmetric around the mean.[14] The resulting grading norm is presented in the second column of Table 3. Students were instructed that they would be competing against all other students in their treatment group.

As mentioned in Section 2.3, we communicated all the details of the grading schemes to the students at the very beginning of the course. However, we have not informed students explicitly about the fact that the curve was set to closely resemble the grade distribution of the years before, so as not to unintentionally bias their beliefs and perceptions about the schemes. The exact instructions are presented in Appendix B.

2.5 Incentivized survey

Response to tournament-like incentives may depend on several factors such as ability, confidence, as well as attitudes towards uncertainty and competition. In order to measure these characteristics, we conducted an incentivized online survey among course participants.[15] 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

[13] To be precise, in the academic years 2011-12 and 2012-13 combined the mean of the EMO grades was 5.989, with a standard deviation of 1.662. 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.

[14] 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).

[15] The survey was programmed using the online survey software Qualtrics.

prize winners were €215.67, with a minimum of €100 and a maximum of €457). Respondents spent 21 minutes on average completing the survey (designed and pre-tested to take about 15-20 minutes), suggesting that the majority of students took the task seriously and did not answer at random.

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 concepts from first-year mathematics and microeconomics courses, e.g. simple derivations, perfect competition, Nash-equilibria, etc.[16] Students had 25 seconds to answer each question. Performance on this test serves as an ability measure in our analysis. Prior to completing the test, we assessed students' willingness to compete by requiring them to choose the reward scheme to be applied to their multiple-choice test outcomes. In particular, students reported their switching point between a piece rate scheme and a winner-take-all tournament against one randomly chosen opponent. Similar to the design of Petrie & Segal (2014), each item on the choice menu offered the same constant piece rate while tournament prizes were increasing. As shown in Figure C3 in Appendix C, risk-neutral subjects should be indifferent between the piece rate and the tournament at Decision 7. Besides this incentivized measure of the willingness to enter tournaments, we also asked students to indicate their "competitiveness in general" on a Likert-type scale.

We collected incentivized measures of students' expected scores and their relative rank compared to other test-takers both before and after taking the multiple choice test, resulting in different measures of confidence (ex ante and ex post, absolute and relative).[17] In addition, we measured participants' risk and ambiguity preferences by eliciting switching points in Holt & Laury (2002)-style choice menus between gambles.[18] Students also rated 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 (in terms of the final course grade) and also their attitudes toward norm- and criterion-referenced grading practices.

3 Data

This section contains an overview of our data. Panel A of Table 4 presents basic demographic information based on administrative data provided by the University of Amster-

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

[17] 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 e.g. Hvide (2002).

[18] As shown in Figure C2 in Appendix C, risk-neutral subjects should switch from the less risky gamble to the more risky one at Decision 5. Similarly, our measure for ambiguity aversion involved a menu of choices between two gambles, one with known, the other with unknown probabilities.

dam. In total, 529 students registered for the course, a quarter following the international program.

Table 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 W1-4 *(scale 0-3)*	1.21	0.94	0	3	517
tutorial attendance W1-4 *(scale 0-3)*	1.45	1.00	0	3	529
handing in HW W1-4 *(0/1)*	0.81	0.39	0	1	529
average HW grade W1-4 *(scale 0 - 10)*	6.95	1.13	3.45	9.45	427
self-reported study time W1-4 *(scale 1-5)*	2.42	0.77	1	5	385
midterm show-up *(0/1)*	0.91	0.28	0	1	529
midterm score *(scale 0-30)*	19.28	3.8	8	29	483
lecture attendance W5-8 *(scale 0-3)*	0.57	0.93	0	3	517
tutorial attendance W5-8 *(scale 0-3)*	0.83	1.09	0	3	529
handing in HW W5-7 *(0/1)*	0.61	0.49	0	1	529
average HW grade W5-8 *(scale 0 - 10)*	5.83	1.57	1.60	9.85	409
end-of-term show-up *(0/1)*	0.87	0.34	0	1	529
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 performance *(scale 0-10)*	4.67	1.67	0	10	486
overconfidence *(ex ante relative, scale -100 to 100)*	18.23	29.65	-78	100	487
risk preferences *(incentivized, scale -10-10)*	-1.10	1.91	-6	4	487
risk attitude *(unincentivized, scale 0-10)*	5.41	1.98	0	10	487
ambiguity aversion *(incentivized, scale 0-10)*	6.44	3.14	0	10	487
competitive preferences*(incentivized, scale -10-10)*	0.02	3.78	-8	10	486
competitive attitude *(unincentivized, scale 0-10)*	6.79	1.92	0	10	486
expected course grade *(scale 0-10)*	7.04	0.89	3	10	485
expected rank *(in terms of grade, scale 0-100)*	37.37	17.81	0	100	485
attitude absolute grading *(scale 0-10)*	7.88	1.82	0	10	485
attitude relative grading *(scale 0-10)*	4.33	2.75	0	10	485

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 good predictor of the final grade in the EMO course: the correlation between the two is 0.50 and is highly significant. The second indicator, calculated as the number of retake exams over all the courses the student ever registered for, is also significantly related to one's final grade ($corr = -0.39$). On average, students repeat approximately one out of five exams.[19]

Panel B of Table 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. Students reported spending on average 10 hours per week on studying and practicing for the course. According to all our measures, students decreased their preparation efforts after the midterm exam: we observe a drop in lecture and tutorial attendance as well as in homework outcomes. The show-up rate at both the mid- and end-of-term exams was very high, 91% at the former and 87% at the latter. 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. Students on average performed quite well in the course: the mean final grade (calculated as the unweighted average of the two exam grades, augmented by the bonus point when applicable) was 6.65.[20] Close to 17% of students failed the course at the first try (not reported in the table).

Results from the incentivized online survey are presented in Panel C of Table 4. The survey was included among the compulsory course requirements, ensuring a very high response rate (92%). The average performance on the test measuring knowledge in

[19] 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 capturing whether the given observation has a missing value for the covariate in question. (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.

[20] Analyzing the final grades, note that it was theoretically possible to get a grade 11 in this course and two students indeed received a calculated grade of 10.5 because the homework bonus point was added on top of the two exam grades.

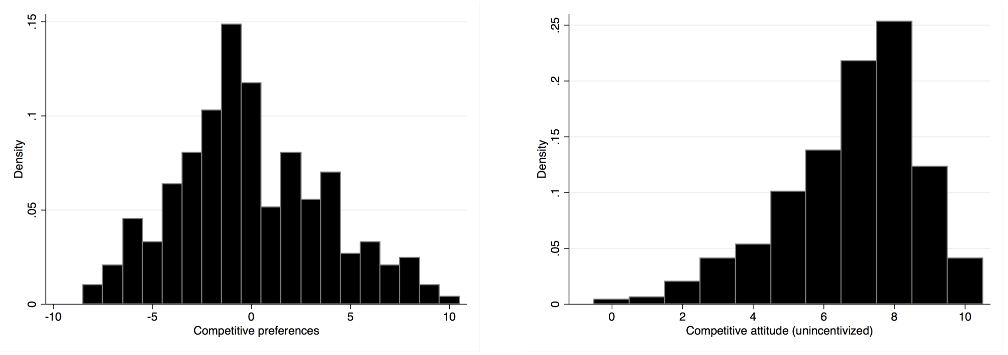
prerequisites was rather low, 4.67 correct answers out of 10 questions, possibly due to the intense time pressure students were subjected to during the test. Scores on this test turn out to be highly significantly correlated with the final grade of the course ($corr = 0.23$), lending credibility to our use of this metric as ability proxy. On average we find students to be overconfident about their test performance according to all the measures we have elicited. In the table we present the *ex ante* relative overconfidence variable (calculated such that a score of zero corresponds to a correct guessed rank and any positive number indicates overconfidence) which has a mean of 18.23 and a standard deviation of 29.65.

As mentioned in the previous section, students' attitudes towards risk and uncertainty were elicited by means of Holt & Laury (2002)-style choice lists. The risk preference measure we report here reflects the difference for each student between the risk neutral and their actual switching point (a score of zero thus indicates risk neutrality, and negative numbers correspond to risk aversion). With an average score of -1.10 we find indication for moderate risk aversion in our sample.[21] This finding is not reflected in students' self-reported risk attitudes where the mean score is 5.41 on a scale running from 0 to 10 (where higher numbers mean greater tolerance for risk).[22]

We discuss students' taste for tournament-style incentives by calculating the variable *competitive preferences*: the difference between respondents' "optimal" (based on their relative performance guess, assuming risk neutrality) and actual switching point in the choice list shown in Figure C3. This measure shows relatively low average competitiveness in our sample (a mean of 0.02 on a scale that runs from -10 to 10 such that positive numbers indicate a taste for competition). Depicting the distribution of the measure, Figure 1a shows that the modal score is -1, suggesting a mild aversion to tournaments. We also report a second measure we call *competitive attitudes*, containing students' self-evaluation of competitiveness in general. Contrary to our results from the incentivized measure, we find that students on average consider themselves competitive: the mean score is 6.79 and the most frequently chosen answer is 8 (as shown in Figure 1b) on a scale where 0 corresponds to "not competitive at all" and 10 to "highly competitive". The two measures of competitiveness are not significantly correlated. While the second method has the advantage that it does not exclusively focus on tournament entry and allows a broader interpretation of competitiveness that potentially fits our setting better, we must emphasize that it is based on an unincentivized question that, unlike the risk elicitation technique of Dohmen *et al.* (2011), has not been validated experimentally.

[21] The ambiguity aversion score has a less straightforward interpretation and is more useful for between-subject comparisons.

[22] The correlation between the two measures of the willingness to take risk is 0.16 and highly significant.



(a) Incentivized, based on switching points

(b) Unincentivized, self-reported measure

Figure 1: Distribution of competitive preferences

On average, students expect to receive a course grade of 7.04, which is slightly higher than the mean of the actual course grades. Students' overconfidence is also reflected in their guessed rank in terms of course grades: on average students expect that out of 100, only 37.37 of their peers will do better than them. Finally, we observe that students have 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 and higher numbers reflect a more positive attitude).

Comparing observable characteristics of students between the two treatment groups (also separately by gender), we find that the randomization was successful (see Table D1 in Appendix D). 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.[23] The table also allows for gender comparisons. We observe that women are more likely than men to follow the international program and are 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. 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 line with results from previous studies (see, e.g. Croson & Gneezy (2009)) we find that men and women differ in their attitudes toward risk, with women being significantly more risk averse than men according to both of our measures. Contrary to our expectations, we find no significant gender differences in competitiveness using the

[23] 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.

incentivized choice between piece rate and tournament (*competitive preferences* variable). However, men rate themselves significantly higher on the self-reported competitiveness scale (*competitive attitudes*) than women.

4 Results

As we explained in Section 2.3, our design allows for two different identification strategies: a between-subject and a within-subject comparison of outcomes under absolute and relative grading. We start by presenting results from the between-subject analysis: we compare student behavior in (preparation for) the midterm exam between the two treatment groups. The outcomes considered are participation in the exams, quality of homework assignments, class attendance, self-reported study time, and exam scores. As we have argued before, due to the random assignment of students to grading schemes a simple comparison of the two groups' results shows us whether students performed differently under relative than under absolute grading. In Section 4.2 we then pool the data from the first and the second half of the course and estimate the impact of relative grading using student and exam fixed effects.

4.1 Between-subject analysis

4.1.1 Selection

Having shown in Section 3 that there are no concerning pre-intervention differences between the treatment groups, we also 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 & Vesterlund (2007) and several replications suggest that even high-ability women are likely to shy away from competition. We would thus expect to see lower midterm show-up among females in the relative grading group. We find no support for this hypothesis in our data: first, the number of non-participants is very low: 16 vs. 30 in the relative and absolute group, respectively. Show-up is thus slightly *higher* under relative grading (a raw difference of 4.9 percentage points in participation rates, significant at the 5% level). Moreover, there is no gender difference in the propensity to participate in the midterm exam among those assigned to relative grading (the show-up rates in this group are 95.5% for female and 93.5% for male students, and a t-test yields a p-value of 0.23). Finally, we have checked that selection does not ruin the balancedness of the two treatment groups. We are thus fairly confident that non-random exam participation does not bias our results.

4.1.2 Preparation behavior

We continue 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, the quality of their homework, their lecture and tutorial attendance or their self-reported study times. Table 5 summarizes our results. Competitive grade incentives had little impact on preparation behavior prior to the midterm exam: students in both treatment 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 four weeks. Relative grading had a positive impact on female students' quality of homework assignments (column (2)): the grade average of the first two assignments was 0.483 points higher (on a scale of 0 to 10) for female students assigned to relative grading.[24]

Table 5: THE EFFECT OF RELATIVE GRADING ON PREPARATION BEFORE THE MIDTERM EXAM.

	hand in HW	*avg. HW grade*	*attendance*	*prep. time*
	(1)	(2)	(3)	(4)
relative	0.451	0.082	0.046	0.080
	(0.302)	(0.136)	(0.157)	(0.094)
female	0.506	-0.186	-0.063	0.289^{**}
	(0.388)	(0.159)	(0.189)	(0.113)
relative*female	0.023	0.483^{**}	-0.020	-0.135
	(0.604)	(0.222)	(0.268)	(0.160)
Demographic controls	✓	✓	✓	✓
Ability controls	✓	✓	✓	✓
Constant	3.053^{*}	6.528^{***}	3.124^{***}	2.124^{***}
	(1.844)	(0.890)	(0.898)	(0.582)
N	527	426	516	384
Pseudo-/Adj. R^2	0.250	0.051	-0.001	0.085

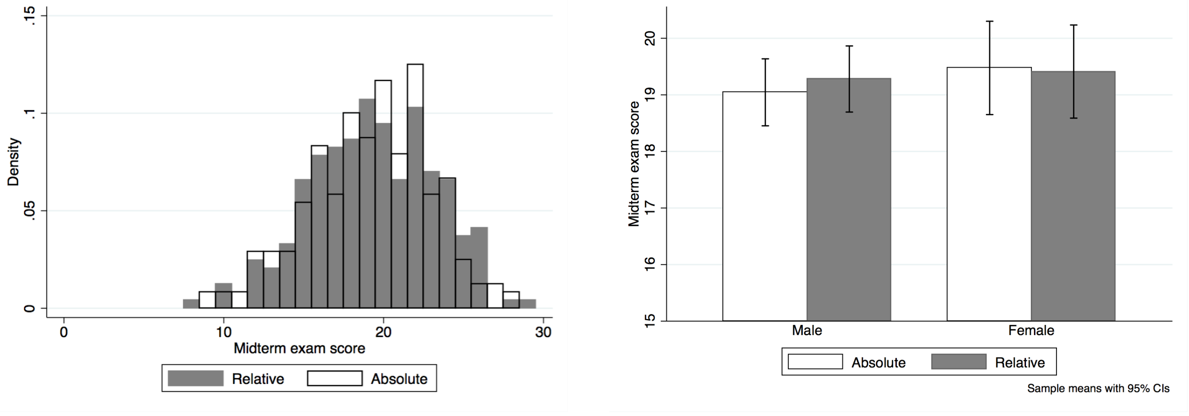
Notes: The table displays estimated coefficients from (1): logistic and (2)-(4): OLS regressions. Dependent variables: (1): hand in HW 1&2, (2): avg. grade HW 1&2, (3): attendance weeks 1-3, (4): self-reported study time. Covariates (1)-(4): int. program, age, Dutch born, Math grades, num. retakes, test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

[24] This result, however, may suffer from endogeneity since solving homework assignments was voluntary. While Column (1) suggests that relative grading did not effect the overall rate of handing in homework, it may have induced different types of students to hand in their work. In the absence of a credible exclusion restriction, we are unable to perform a Heckman selection model to alleviate these concerns, and therefore we treat the result that relative grading improved the quality of female students' assignments with caution.

4.1.3 Midterm exam performance

The next measure we consider is the midterm exam performance. Given that we do not observe consistent differences between the treatment groups in preparation behavior, we do not expect a large impact of relative grading on the midterm exam (designed to measure course-specific knowledge), either. On the other hand, right before starting their exams, students were required to fill out a questionnaire about absolute/relative grading and their treatment assignment (see Section 2.3), reminding them of the scheme applied to their test and potentially influencing their level of concentration, stress or effort exertion during the exam.

A raw comparison of scores reveals no impact of relative grading on midterm exam performance. Out of 30 questions, 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). According to a two-sample t-test with unequal variances, the difference is insignificant (p-value: 0.62).[25] As Figure 2a 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).



(a) Distribution of midterm exam scores

(b) Mean midterm exam scores by gender

Figure 2: The effect of absolute vs. relative grading on midterm exam performance

A simple comparison between the two groups might hide important compositional differences. In particular, as outlined in Section 2.1, our theoretical model predicts that the response to relative grading differs by students' competitiveness. Given the empirical stylized fact of gender differences in response to tournaments, we expect male students to do better and female students do to worse under relative than under absolute grading.

[25] Our sample size allows us to detect an effect size of approx. 0.25 SD, or a 0.97 change in exam scores with an 80% power. Consequently, we are underpowered to identify a score change as small as 0.17, the raw difference between the two sample means. However, we find such a small effect also to be economically insignificant.

Figure 3b, presenting mean midterm scores with 95% confidence intervals by grading scheme and gender, does not support this hypothesis: there seems to be no significant difference in performance between the treatment groups among either male or female students.

The above results are also supported by regression analysis. Results are presented in Table 6: column (1) repeats the finding that there is no overall difference between the scores by grading schemes, while column (2) confirms that the interaction between relative grading and female is also insignificant. Column (3) presents results from a specification that controls for demographic (age, nationality, study program) and ability (average grade from previous math-related courses, share of exam retakes, performance on the online prerequisites test) characteristics. We find that adding covariates largely improves the explanatory power of our model (the adjusted R^2 increases to 0.257). In line with alternative hypothesis, the point estimate for the effect of relative grading is positive and the coefficient associated with *relative*female* is negative; however, both are small and not significantly different from zero.

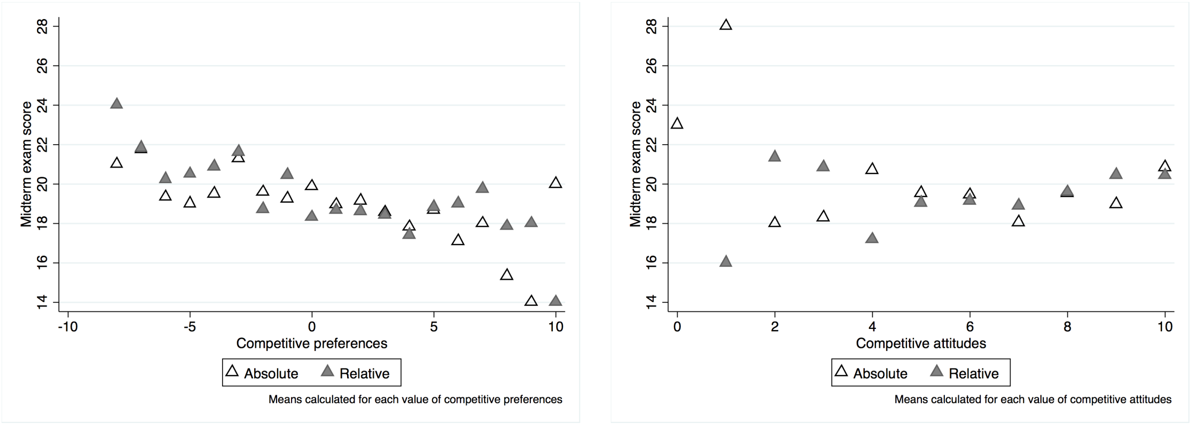
Table 6: THE EFFECT OF RELATIVE GRADING ON MIDTERM SCORES.

midterm score	No covariates	Gender interaction	With covariates
	(1)	(2)	(3)
relative	0.170	0.236	0.455
	(0.346)	(0.428)	(0.219)
female		0.431	0.281
		(0.512)	(0.446)
relative*female		-0.300	-0.746
		(0.722)	(0.625)
Demographic controls			✓
Ability controls			✓
Constant	19.196***	19.045***	14.145***
	(0.245)	(0.303)	(2.345)
N	483	482	482
Adj. R^2	-0.002	-0.004	0.257

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

4.1.4 Heterogeneity by competitiveness and ability

In our analysis so far we relied on students' gender as a proxy for their competitiveness. In the following we try a different approach and directly use the competitiveness values we have elicited in our online incentivized survey. We have two such measures available: *competitive preferences*, based on students' switching point between a piece rate and a tournament remuneration scheme (controlling for confidence), and *competitive attitudes*, based on students' self-assessment of general competitiveness (for a more detailed description of the variables, please refer to Sections 2.5 and 3).



(a) Incentivized

(b) Unincentivized

Figure 3: The effect of absolute vs. relative grading on midterm exam scores, by competitiveness

Figure 3 shows the mean exam score of students for any given level of competitiveness (3a depicts competitive preferences, 3b competitive attitudes), separately by grading schemes. Our alternative hypothesis predicts no relationship between competitiveness and exam scores under absolute grading, while we expect a positive relationship in case of relative grading. Figure 3 does not support our hypothesis: regardless of the competitiveness measure used, we find no evidence for the predicted pattern in our data. A regression analysis provides a similar conclusion: as shown in Table 7, the interaction terms between relative grading and competitive preferences (column (1)) and competitive attitudes (column (2)) are not significantly different from zero.

While our theory does not predict the response to relative grading to differ by ability, the findings of e.g. Paredes (forthcoming) and Müller & Schotter (2010) suggest that reaction to competitive grade incentives could be heterogeneous with respect to students' skills. We analyze the impact of ability using two different measures: *math grades*, i.e. the average grade of all mathematics- or statistics-related exams the student had taken prior to our course, and *test performance*, denoting the score the student obtained on the prerequisites test in the online survey. Columns (3) and (4) of Table 7 suggest that these measures of ability do not seem to influence the reaction to grade incentives, either. 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.[26] Figure D1 in Appendix D, showing midterm scores

[26] Regression outputs available from the authors upon request.

by grading schemes for different ability levels, further supports this conclusion. We find no heterogeneity by risk or ambiguity aversion, grade expectations or attitudes towards the two grading practices.[27]

Table 7: THE IMPACT OF COMPETITIVENESS AND ABILITY ON THE RESPONSE TO RELATIVE GRADING.

	(1) Competitive preferences	(2) Competitive attitudes	(3) Math grades	(4) Test performance
relative	0.145	-0.863	1.330	0.022
	(0.299)	(1.143)	(1.331)	(0.928)
relative * comp. preferences	-0.033			
	(0.080)			
relative * comp. attitudes		0.148		
		(0.161)		
relative * math grade			-0.194	
			(0.217)	
relative * test performance				0.030
				(0.187)
Demographic controls	✓	✓	✓	✓
Ability controls	✓	✓	✓	✓
Constant	14.630***	15.214***	13.650***	14.433***
	(2.326)	(2.467)	(2.470)	(2.404)
N	482	482	482	482
Adjusted R^2	0.268	0.254	0.255	0.254

Notes: The table displays estimated coefficients from OLS regressions. All specifications include the following control variables: female, int. program, age, Dutch born, Math grades, num. retakes, test questions; column (1) also controls for competitive preferences and column (2) for competitive attitudes. In all specifications, indicator variables for missing covariates included. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.2 Within-subject analysis

We continue the analysis by exploiting the within-subject nature of our design. As discussed in Section 2.3, our dataset contains two observations per student for all outcomes of interest: one when the student was exposed to absolute, the other when exposed to relative grading. We can therefore construct a panel dataset and estimate a regression with student fixed effects to filter out individual-specific, time-invariant characteristics. By including exam fixed effects, we can also control for the potential differences in difficulty between the mid- and end-of-term exams. The results of this exercise are presented in Table 8. Our findings are in line with the conclusions of the between-subject analysis: relative grading does not induce women to stay away from the exam (column (1)), and neither preparation effort (columns (2) to (4)) nor exam scores (column (5)) are significantly affected by a change in grading schemes. Contrary to the results of Table 5, we find no impact of relative grading on female homework assignment grades (see column

[27] A previous version of this paper also included heterogeneity analysis by study program, and reported a gender difference in response to competitive grade incentives among students following the international program. This subsample, however, was too small (N=126) and the analysis was underpowered. We therefore decided to omit these results from the current version of the paper.

(3)). We also test for a heterogeneity in response to relative grading by competitiveness or ability, but find no significant effect. Results are presented in Table D2 in Appendix D.

Table 8: THE IMPACT OF RELATIVE GRADING, WITHIN ESTIMATION

	exam show-up	*hand in HW*	*avg. HW grade*	*attendance*	*exam score*
	(1)	(2)	(3)	(4)	(5)
relative	-0.007	0.006	0.196	0.085	-0.233
	(0.012)	(0.024)	(0.134)	(0.093)	(0.252)
relative * female	-0.008	0.042	-0.002	0.025	-0.002
	(0.017)	(0.039)	(0.203)	(0.150)	(0.429)
Student fixed effects	✓	✓	✓	✓	✓
Exam fixed effects	✓	✓	✓	✓	✓
N	1054	1054	834	1032	942
Within-R^2	0.041	0.175	0.265	0.373	0.196

Notes: The table displays estimated coefficients from within estimations. Dependent variables: (1): participate in exam, (2) hand in HW, (3): avg. grade of HW assignments, (4): attendance, (5) exam score. Standard errors are corrected to account for the fact that there are two observations per student, and are reported in parentheses. $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5 Discussion

In this study we set out to test the effectiveness of competitive grade incentives in inducing more preparation and better exam scores. We found no indication that competitive grade incentives induce students in our sample to work harder or perform better. In the following, we discuss potential explanations for this result.

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 to test understanding of the treatments (see Section 2.3) rules out this explanation. Out of the 483 students who participated in the exam, 403 filled out the questionnaire and 84% of them gave the correct answers to questions. When we exclude those who did not fill out the questionnaire or gave 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. We do not believe that this would dilute incentives: according to our theoretical model, students with competitive preferences have a higher return on effort under relative than under absolute grading (because of the extra "rank-utility" they obtain from a high grade when graded on the curve), and this remains true regardless of the future reversal of the grading schemes.

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 persistently low attendance and preparation effort among our participants. Relatedly, as Fryer (2011) argues, even if students are motivated by the incentives, they may not know how to respond productively to them. However, Leuven *et al.* (2010) also study Economics & Business Bachelor students at the University of Amsterdam some years prior to our experiment and find that (high-ability) students are able to improve their performance in response to financial incentives, suggesting that our results are not driven by a lack of knowledge of the education production function.

Our intervention could also be ineffective because rewards are not immediate (see Levitt *et al.* (forthcoming)): students need to invest in preparation long before the gains (in the form of higher grades) are realized. While high discount rates may explain why attendance and homework effort are unchanged, it is not clear whether the null result in the midterm exam can also be attributed to this factor. Students got a powerful reminder of the particular grading scheme they were subject to right at the start of the midterm exam, and still we observe no response in scores to competitive grade incentives, suggesting either that effort provision was unaffected or that test scores are fully determined by prior preparation and can not be improved by trying harder during the exam itself.

We note that our results are consistent with the predictions of our baseline model that assumes students with competitive preferences to derive the same utility from a higher rank under both schemes, regardless of how explicitly relative performance is signaled by the grades. We do not find this assumption realistic, especially given how difficult it is to collect reliable relative performance information under absolute grading in our setting. As explained in Section 2.1, individual grades or cohort averages are not made publicly available, and with a class size of 500 students are unlikely to get a clear picture of their place in the cohort grade distribution just by asking their peers. Moreover, results by Azmat & Iriberri (2010b), Blanes i Vidal & Nossol (2011) and Kosfeld & Neckermann (2011) suggest that explicit information on relative performance induces higher effort provision.

Finally, as a result of the particular way the relative grading curve was fixed, our theoretical model predicts no difference in effort provision between the two grading schemes in case students do not have competitive preferences. A straightforward interpretation of our results is therefore that students in our sample do not gain utility from outperforming their peers. This reasoning is in line with the findings of Buser *et al.* (2014) who study the link between competitiveness and track choices among Dutch high school students and report that students with relatively low levels of competitiveness tend to select into the Economics and Society academic track (high school track choices are strongly correlated with the choice of major in tertiary education in The Netherlands). Still it is not clear why those with higher levels of competitiveness do not respond to the treatment either (see columns (1) and (2) in Table 7 showing that regardless of our measure of competitiveness, the interaction effect between relative grading and taste for competition is not significantly different from zero).

All in all, these considerations prompt us to question the validity of the assumption that "competitive" students in our sample derive rank utility from obtaining higher grades than their peers. In a culture where academic excellence is not valued, it is not obvious that outperforming one's classmates leads to a higher status or to gains in ego utility. As we have argued in the Introduction, Dutch students are characterized by a "just pass" attitude and are unlikely to strive for the highest possible grades. Participants in the EMO course are no exception: out of the 529 students enrolled in the class, only 25 were part of the university's honors program. Similarly, in the online survey at the beginning of the course only 20 students reported an expected final grade of at least 9 (where 10 is the top mark and 5.5 is the lowest requirement to pass). As a result, even students who in other domains exhibit competitive preferences might be unresponsive to tournament-style grading because the prize ('best student', 'top of the class') is not worth enough for them to provide extra effort.

While we have shown that tournament-style grade incentives are not the cure for insufficient effort provision among college students with low levels of ambition, we must emphasize that our result of no gender differences in response to competitive grading does not necessarily generalize to a setting where students value academic excellence more. It is an interesting avenue for future research to focus on more ambitious students and test experimentally whether competitive grading hinders the academic performance of females, especially in more mathematics-related subjects as suggested by Niederle & Vesterlund (2010).

6 Conclusion

This study aims to test a potential remedy for the problem of students' insufficient effort provision. A few recent studies focus on the effectiveness of grade incentives in addressing this issue (Chevalier *et al.*, 2014; Grove & Wasserman, 2006): reassuringly, they find that making assignments count towards one's final grade increases student effort provision. In this paper we test whether the *type* of grade incentive matters by experimentally comparing student effort provision and exam performance under the two most commonly used evaluation schemes: absolute and relative grading. In particular, we test whether relative grading, by conveying explicit information on relative performance, improves student outcomes, and whether male students, generally found to be more competitive, are more responsive to grading on the curve than females. We conduct our study in the Netherlands, a country where the motivation to excel academically is low, and competitive incentives are traditionally not in use in higher education.

Contrary to our expectations, we find no difference in preparation or exam performance under the two grading schemes, nor do we observe heterogeneity in response to the schemes by gender, competitiveness or ability. We argue that our results are not driven by confusion, the specifics of the particular design we used, or a lack of knowledge among students about the education production function. High discount rates and low competitiveness can not fully account for our findings, either. Instead, we believe that students did not respond to competitive grade incentives because outperforming peers academically does not lead to status gains in an environment where the norm is to pass the class with minimal effort provision (*"zesjescultuur"*).

This culture is not unique to our sample or to Dutch universities: according to Brennan *et al.* (2009), more than 60% of students in EU-countries report not to work more than what is required to pass. Such an approach is dangerous if students are overconfident about their performance because even a slightly lower than expected score could lead to failing courses, contributing to longer study durations or leaving the program without a degree. Our study thus highlights the importance of research into incentives in education: in many cases, grades (either criterion- or norm-referenced) alone do not seem to provide sufficient motivation to students, so other measures are required to reduce the social cost of underperformance, increased study times and dropping out.

References

Angrist, Joshua, Lang, Daniel, & Oreopoulos, Philip. 2009. Incentives and Services for College Achievement: Evidence from a Randomized Trial. *American Economic Journal: Applied Economics*, **1**(1), 136–63.

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

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

Barankay, Iwan. 2012. *Rank Incentives: Evidence from a Randomized Workplace Experiment*. Working paper, University of Pennsylvania.

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

Bigoni, Maria, Fort, Margherita, Nardotto, Mattia, & Reggiani, Tommaso G. 2015. Cooperation or Competition? A Field Experiment on Non-monetary Learning Incentives. *The B.E. Journal of Economic Analysis & Policy*, **15**(4), 1753–1792.

Blanes i Vidal, Jordi, & Nossol, Mareike. 2011. Tournaments Without Prizes: Evidence from Personnel Records. *Management Science*, **57**(10), 1721–1736.

Brennan, John, Patel, Kavita, & Tang, Winnie. 2009. *Diversity in the student learning experience and time devoted to study: a comparative analysis of the UK and European evidence*. Report to HEFCE by Centre for Higher Education Research and Information, The Open University.

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

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

Buser, Thomas, Niederle, Muriel, & Oosterbeek, Hessel. 2014. Gender, Competitiveness, and Career Choices. *The Quarterly Journal of Economics*, **129**(3), 1409–1447.

Charness, Gary, & Rabin, Matthew. 2002. Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, **117**(3), pp. 817–869.

Chevalier, Arnaud, Dolton, Peter, & Lührmann, Melanie. 2014 (Oct.). *Making It Count: Evidence from a Field Study on Assessment Rules, Study Incentives and Student Performance.* IZA Discussion Papers 8582. Institute for the Study of Labor (IZA).

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

De Paola, Maria, Gioia, Francesca, & Scoppa, Vincenzo. 2015. Are females scared of competing with males? Results from a field experiment. *Economics of Education Review*, **48**, 117 – 128.

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

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

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

Garibaldi, Pietro, Giavazzi, Francesco, Ichino, Andrea, & Rettore, Enrico. 2012. College Cost and Time to Complete a Degree: Evidence from Tuition Discontinuities. *The Review of Economics and Statistics*, **94**(3), 699–711.

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

Gneezy, Uri, Niederle, Muriel, & Rustichini, Aldo. 2003. Performance In Competitive Environments: Gender Differences. *The Quarterly Journal of Economics*, **118**(3), 1049–1074.

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

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

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

Harbring, Christine, & Irlenbusch, Bernd. 2003. An experimental study on tournament design. *Labour Economics*, **10**(4), 443–464.

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

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

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

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

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

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

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

Kosfeld, Michael, & Neckermann, Susanne. 2011. Getting More Work for Nothing? Symbolic Awards and Worker Performance. *American Economic Journal: Microeconomics*, **3**(3), 86–99.

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

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

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

Levit, Nancy, & Downs, Robert C. 1997. If it Can't Be Lake Woebegone...A Nationwide Survey of Law School Grading and Grade Normalization Practices. *University of Missouri-Kansas City Law Review*, **65**, 819.

Levitt, Steven, List, John, Neckermann, Susanne, & Sadoff, Sally. forthcoming. The Behavioralist Goes to school: Leveraging Behavioral Economics to Improve Educational Performance. *American Economic Journal: Economic Policy.*

Lindo, Jason M., Sanders, Nicholas J., & Oreopoulos, Philip. 2010. Ability, Gender, and Performance Standards: Evidence from Academic Probation. *American Economic Journal: Applied Economics*, **2**(2), 95–117.

Moldovanu, Benny, & Sela, Aner. 2001. The Optimal Allocation of Prizes in Contests. *American Economic Review*, **91**(3), 542–558.

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

Müller, Wieland, & Schotter, Andrew. 2010. Workaholics and Dropouts in Organizations. *Journal of the European Economic Association*, **8**(4), 717–743.

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

Niederle, Muriel, & Vesterlund, Lise. 2010. Explaining the Gender Gap in Math Test Scores: The Role of Competition. *Journal of Economic Perspectives*, **24**(2), 129–44.

Niederle, Muriel, & Vesterlund, Lise. 2011. Gender and competition. *Annual Review of Economics*, **3**(September), 601–630.

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

Onderstal, Sander. 2014. *Economics of Organizations and Markets.* Pearson, Amsterdam.

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

Paredes, Valentina. forthcoming. Grading System and Student Effort. *Education Finance and Policy.*

Petrie, Ragan, & Segal, Carmit. 2014 (November). *Gender Differences in Competitiveness: The Role of Prizes.* GMU Working Papers in Economics No. 14-47.

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

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

Tran, Anh, & Zeckhauser, Richard. 2012. Rank as an inherent incentive: Evidence from a field experiment. *Journal of Public Economics*, **96**(9–10), 645 – 650.

van den Broek, Anja, Wartenbergh, Froukje, Hogeling, Lette, Brukx, Danny, Warps, Jules, Kurver, Bas, & Muskens, Marjolein. 2009. *Studentenmonitor Hoger Onderwijs 2007*. ResearchNed Nijmegen.

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

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 baseline model in which competitive preferences are independent of whether or not students obtain explicit information about their relative performance. Then, we discuss an extension where this assumption relaxed.

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 such as ability, motivation, and competitive preferences. Students' utility is a function $U(e, g, r, \alpha)$ of their effort $e \geq 0$, their exam grade g, their rank r in the grade distribution, 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). Typically, it is assumed that students do not care about their rank in and of itself. The model also captures a setting where students are bound by an effort frontier (which could be modelled by letting $U(e, g, r, \alpha) = -\infty$ 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. In equilibrium, each student maximizes utility by picking

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

where F denotes the equilibrium distribution of grades over the student population, which is assumed to exist and to be consistent with students' optimal effort choices (i.e., it is indeed the equilibrium distribution). Because there is a continuum of students, the individual student's score has no influence on this distribution.

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.

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 and f the density function corresponding to F.

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 $F^{-1} : [0, \overline{\sigma}] \rightarrow [0, 1]$ represent the inverse function of F.

Relative grading

Under relative grading, a student's grade is determined by her rank r in the scoring 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

(2)

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

Observe that maximization problems (1) and (2) coincide so that $e = e^A(\alpha)$ is indeed a best response. ■

Grading schemes and competitive preferences

In this section, we adapt our model by letting rank utility depend on whether or not students obtain explicit information about their rank: the more informative the signal on relative performance, the more utility students gain from a high rank. For ease of exposition, we discuss the extreme case where students only care about their rank if they are explicitly informed about it. It follows from this assumption that the extent to which students care about their rank in the grading distribution depends on the scheme used to determine their grades. In particular, students under absolute grading do not obtain rank utility as they only receive imperfect information about their relative performance. By definition, relative grading provides students with explicit information about their rank as grades in this scheme correspond directly to students' relative performance.

Absolute grading

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

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

Suppose that the students' α's are one-dimensional and distributed according to cumulative distribution function F that 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 $r \in [0,1]$.)

Relative grading

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 preference, 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

(3)

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

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 misreprents 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

$$e(\beta) = \sqrt{2 \textstyle\int_0^{\beta} x dF^{-1}(H(x))}. \tag{4}$$

■

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

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

In this case, relative grading induces that 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 (keeping ability constant).

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 'streching' 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 obtains for the subpopulation with competitive preferences as the next proposition shows.

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

Proof. The result straightforwardly follows from

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

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

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 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.

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:

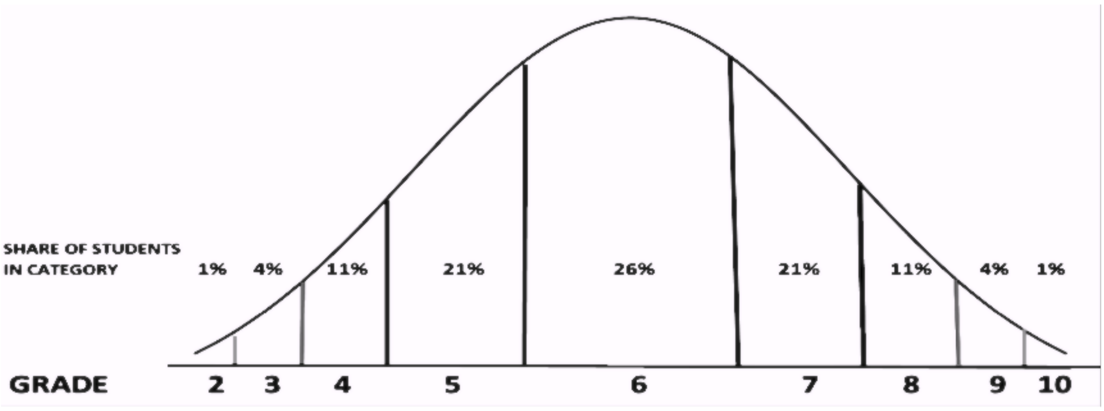
$$\text{Grade exam} = 10 - 0.4^{*}(\text{number of errors})$$

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

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

Relative grading

Under a relative grading scheme, or grading on a curve, students' grades depend on how well they perform in the exams compared to other students taking this course. It is not the individual score, but the students' position in the class score distribution (i.e., the students' rank among all students taking the exam) that determines the exam grade. For this course the curve is fixed so that the average score translates into an exam grade of 6, and the highest performing 1% of students receive a grade 10 while the lowest performing 1% get a grade 2. We illustrate this scheme by the figure and the table below:



Relative rank	**Grade**
(calculated from the top)	
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

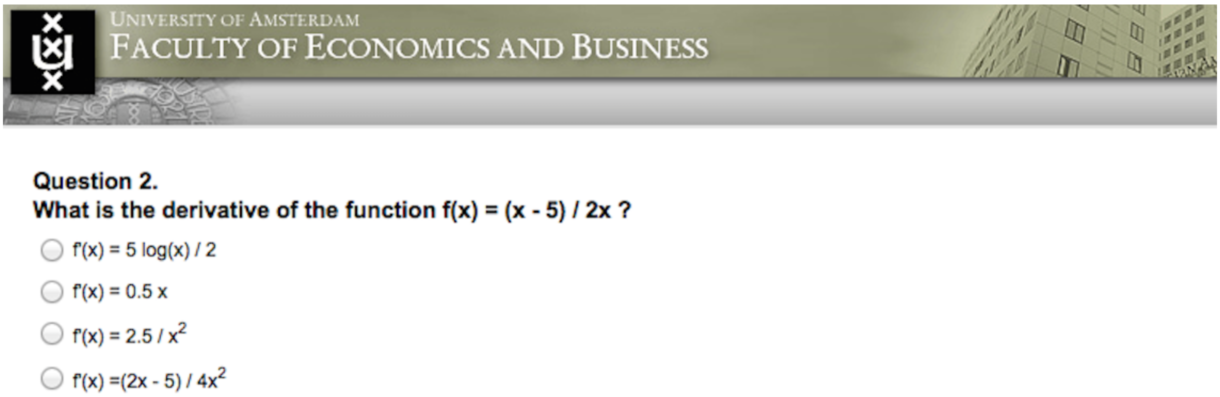
In order to compare the two grading schemes, we will randomly divide all students into two grading groups: the blue group and the yellow group. Students in the two groups will take exams of the same difficulty level but will face different grading schemes:

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.

C Screenshots from the survey



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

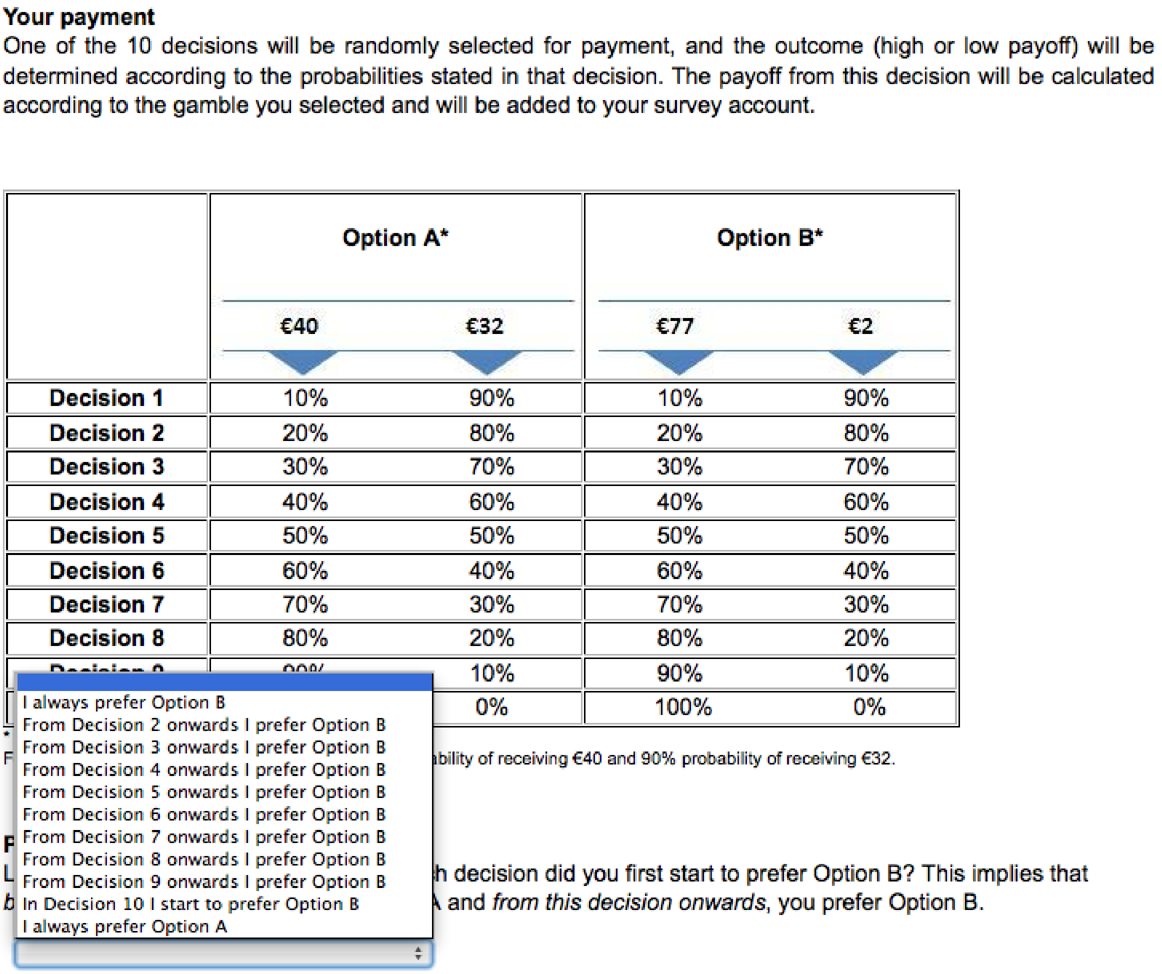
○ $f'(x) = 5 \log(x) / 2$

○ $f'(x) = 0.5 x$

○ $f'(x) = 2.5 / x^2$

○ $f'(x) = (2x - 5) / 4x^2$

Figure C1: Example of a multiple-choice test question



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.

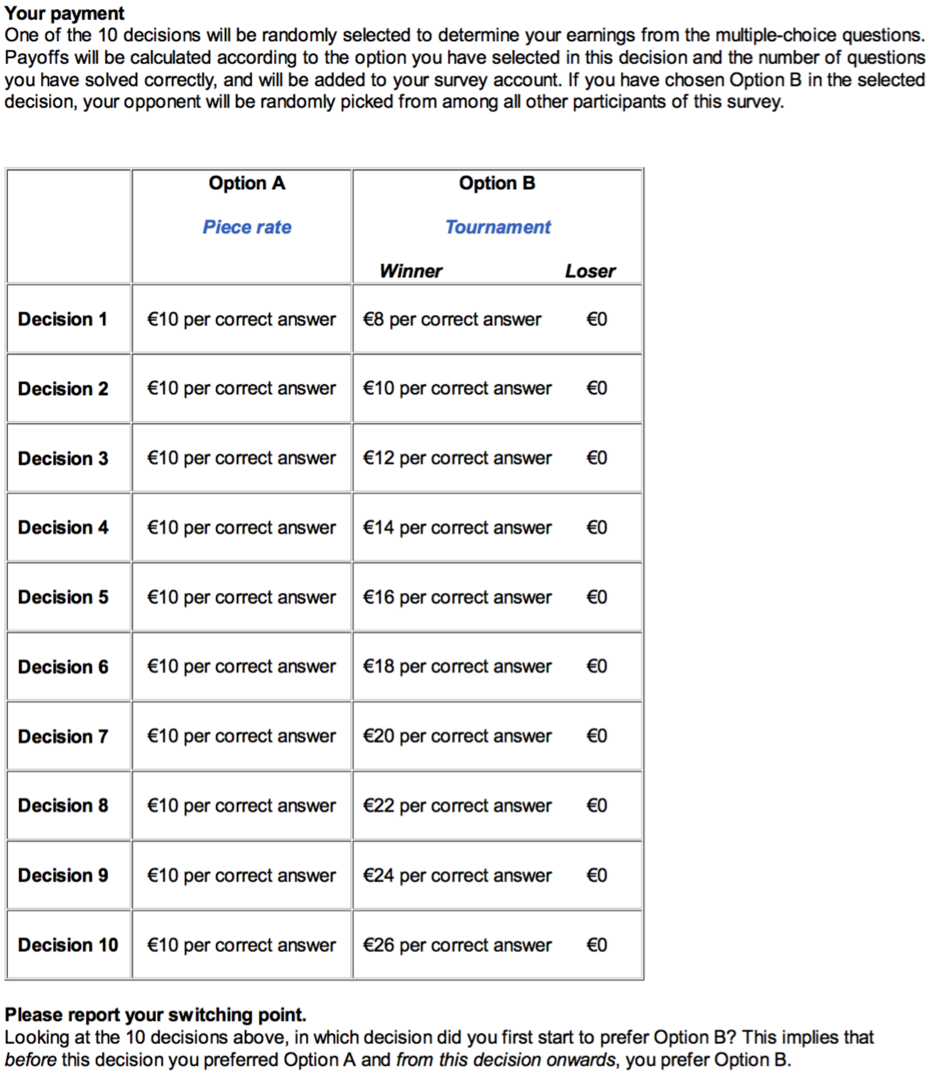
	Option A*		Option B*	
	€40	€32	€77	€2
Decision 1	10%	90%	10%	90%
Decision 2	20%	80%	20%	80%
Decision 3	30%	70%	30%	70%
Decision 4	40%	60%	40%	60%
Decision 5	50%	50%	50%	50%
Decision 6	60%	40%	60%	40%
Decision 7	70%	30%	70%	30%
Decision 8	80%	20%	80%	20%
Decision 9	90%	10%	90%	10%
		0%	100%	0%

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

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

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

Figure C2: Eliciting risk preferences



Your payment
One of the 10 decisions will be randomly selected to determine your earnings from the multiple-choice questions. Payoffs will be calculated according to the option you have selected in this decision and the number of questions you have solved correctly, and will be added to your survey account. If you have chosen Option B in the selected decision, your opponent will be randomly picked from among all other participants of this survey.

	Option A	**Option B**	
	Piece rate	***Tournament***	
		Winner	***Loser***
Decision 1	€10 per correct answer	€8 per correct answer	€0
Decision 2	€10 per correct answer	€10 per correct answer	€0
Decision 3	€10 per correct answer	€12 per correct answer	€0
Decision 4	€10 per correct answer	€14 per correct answer	€0
Decision 5	€10 per correct answer	€16 per correct answer	€0
Decision 6	€10 per correct answer	€18 per correct answer	€0
Decision 7	€10 per correct answer	€20 per correct answer	€0
Decision 8	€10 per correct answer	€22 per correct answer	€0
Decision 9	€10 per correct answer	€24 per correct answer	€0
Decision 10	€10 per correct answer	€26 per correct answer	€0

Please report your switching point.
Looking at the 10 decisions above, in which decision did you first start to prefer Option B? This implies that *before* this decision you preferred Option A and *from this decision onwards*, you prefer Option B.

Figure C3: Eliciting the willingness to compete

D Additional tables and figures

Table D1: COMPARISON OF MEANS BETWEEN TREATMENT GROUPS, BY GENDER.

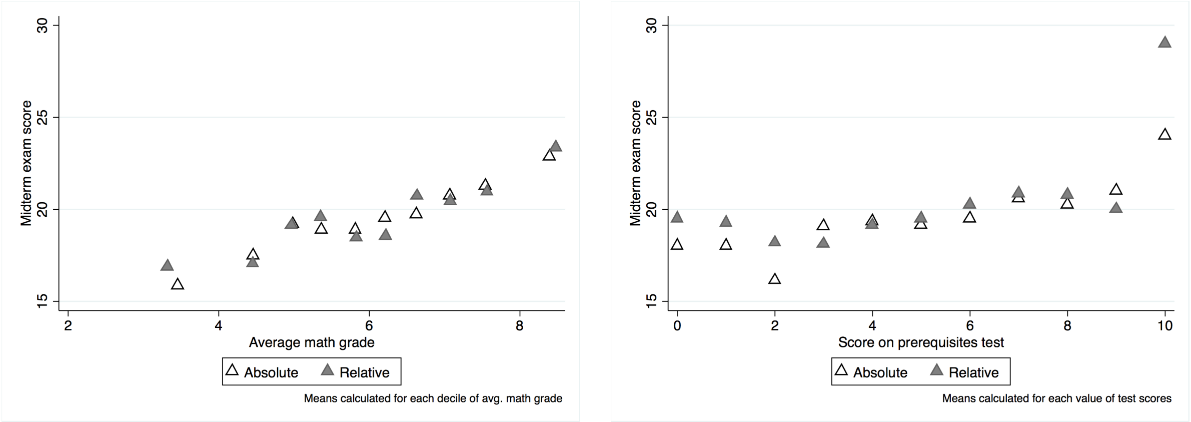
	MEN			WOMEN			GENDER
	BLUE	YELLOW	Diff.	BLUE	YELLOW	Diff.	Diff.
DEMOGRAPHICS							
int. program	0.219	0.179		0.337	0.360		***
	(0.031)	(0.030)		(0.050)	(0.051)		
age	20.951	20.795		20.774	20.759		
	(0.175)	(0.142)		(0.270)	(0.208)		
Dutch born	0.779	0.812		0.659	0.644		***
	(0.032)	(0.031)		(0.050)	(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 preferences	-0.881	-0.903		-1.541	-1.391		***
	(0.160)	(0.145)		(0.195)	(0.204)		
risk attitudes	5.736	5.452		5.024	5.149		***
	(0.165)	(0.152)		(0.211)	(0.203)		
ambiguity aversion	6.830	6.200	*	6.341	6.276		
	(0.250)	(0.253)		(0.340)	(0.335)		
competitive preferences	-0.132	0.033		0.365	-0.391		
	(0.273)	(0.311)		(0.441)	(0.416)		
competitive attitudes	7.264	6.994		6.153	6.230		***
	(0.138)	(0.146)		(0.220)	(0.217)		
expected course 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: Column (3) indicates significance of difference between the two treatment groups among male, column (6) among female students. Column (7) reports significant differences in means between the two genders. Differences of means compared by two-sample t-test with unequal variances. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D2: THE IMPACT OF RELATIVE GRADING BY COMPETITIVENESS AND ABILITY, WITHIN ESTIMATION

	(1)	(2)	(3)	(4)
	Competitive preference	Competitive attitude	Math grades	Test performance
relative	-0.241	0.226	-0.492	-0.684
	(0.204)	(0.644)	(0.539)	(0.548)
relative * comp. preferences	-0.047			
	(0.052)			
relative * comp. attitudes		-0.069		
		(0.091)		
relative * math grade			0.047	
			(0.089)	
relative * test performance				0.098
				(0.112)
Student fixed effects	✓	✓	✓	✓
Exam fixed effects	✓	✓	✓	✓
N	942	942	942	942
Within-R^2	0.197	0.196	0.196	0.197

Notes: The table displays estimated coefficients from within estimations. Dependent variable: exam score. Standard errors are corrected to account for the fact that there are two observations per student, and are reported in parentheses. $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$



(a) Average math grades

(b) Score on prerequisites test

Figure D1: The effect of absolute vs. relative grading on midterm exam scores, by ability