Gender differences and the effect of facing harder competition

June Park John
Graduate School of Education
520 Galvez Mall, Stanford, CA 94305
Stanford University
juneparkjohn@stanford.edu

Abstract:

Gender differences in competition have been demonstrated in a variety of contexts, yet it remains unclear how people respond to competitors they perceive to be hard or easy, and whether gender differences exist in this response. I run an experiment in eighteen public high school classrooms to study the effect of competing in a math task against different levels of competitors. I exploit natural sorting within grade levels in Malaysian public schools to randomly assign competitors of different perceived difficulty levels. Using a standard competition measure, males are significantly more competitive than females. However, when students face harder competitors, males respond by lowering performance while the performance of females does not vary significantly by level of competition.

Keywords: Gender differences; Competition; Gender performance; Tournament; Piece-rate; Information

JEL Codes: I20, J16, J24.

** I am especially grateful to Muriel Niederle for her mentorship and guidance. I thank Doug Bernheim, Al Roth, Charlie Sprenger, Erin Fahle, and participants in the behavioral economics and CEPA seminars for their valuable feedback. I thank the editor and referees who provided helpful comments. I am grateful to the staff and students at the five schools. Mary John and K.J. John provided access for all the schools in the study. Johann John, Kimberly Gan and Sheng Wei Chiam provided excellent research assistance. The research reported here was supported by the Freeman Spogli Institute for International Studies through the Mentored Global Research Fellowship and the Institute of Education Sciences, U.S. Department of Education, through Grant R305B090016 to the Board of Trustees of the Leland Stanford Junior University. The opinions expressed are those of the author and do not represent views of the Institutes or the U.S. Department of Education.*

1. Introduction

Many studies have shown that females are less competitive than males in stereotypically male tasks (see Niederle and Vesterlund, 2011 for review), which explains some of the gender differences in later education and career outcomes (Almås et al., 2016; Buser et al., 2014; Buser et al., 2017; Ors et al., 2013; Zhang, 2013). One important aspect of competition is the perceived difficulty of the competitors: people may react differently in competition when facing easier or harder opponents. Gender differences in these reactions can help explain dynamics of competition and inform policy decisions about the characteristics of competitions in schools or the workplace. Existing research on the perceived difficulty of the competition primarily relies on information provided in a laboratory context which may have limited applicability in the field. In the current study, I exploit natural sorting within grade levels to randomly assign competitors of different perceived difficulty levels to examine the effect of facing harder competitors by gender in addition to replicating the standard gender gap on a math task in Malaysian public schools.

Gender gaps in competition have been categorized by both choice and performance. Females are shown to be less likely than males to choose into competition, a well-established finding in the literature (Niederle and Vesterlund, 2007). Recent research explores how factors such as task or information affect this gender gap (see Niederle, 2016 for review). There is less consistent evidence, however, of gender differences in performance in competitive environments. A seminal paper finds that females perform worse than males when solving puzzles under a competitive incentive scheme, although there is no difference in performance under a non-competitive incentive scheme (Gneezy et al., 2003). Other studies use similar designs and puzzle tasks with similar results (Datta Gupta et al., 2013; Günther et al., 2010). Niederle et al. (2013) finds that males outperform females in math tasks under competition. However, other studies show no gender differences in performance under either non-competitive or competitive incentives in math tasks (Ertac and Szentes, 2011; Niederle and Vesterlund, 2007; Wozniak et al., 2014).

The literature indicates that gender differences in competitive performance cannot be simply explained by differential ability, which has shifted some recent literature to study how features of competition may differentially affect males' and females' performance. One aspect of

competition is how people respond to harder or easier competitors and whether there are gender differences in these responses, the focus of the current study.

Prior research has examined reactions to different levels of competition by providing information or relative feedback during competition[1] in a laboratory environment (Buser, 2016; Cason et al., 2010; Eriksson et al., 2009; Ertac and Szentes, 2011; Gill and Prowse, 2014; Kuhnen and Tymula, 2011; Wozniak et al., 2014), with one recent study conducted in a field setting (Wozniak et al., 2016). In these studies, information about either random competitors or deliberately lower- or higher-performing competitors is given to subjects prior to subsequent competition decisions and performance.

Rational behavior predicts that people would be more reluctant to enter into competition against more difficult competition. Cason et al. (2010) created groups of relatively weaker, stronger, or superstar competition and the study finds that, as expected, the fraction of entry into a tournament is highest against the weaker group and lowest against the superstar group. No breakdown by gender is provided, although there is some indication of gender differences--females under-enter a proportional pay tournament given their expected payout, with no gender difference in under- or over-entry for the winner-take-all tournament. A clear gender difference in choice of competition is demonstrated in an unpublished study by Niederle and Yestrumskas (2008), which shows that females choose a less difficult and less lucrative task than males; however, both genders receive lower payout than if they had optimally chosen their task difficulty.

There is consistent evidence that information about target or relative score provided to subjects decreases or even eliminates the gender gap in entry into competition (Ertac and Szentes, 2011; Wozniak et al., 2014), although Wozniak et al. (2016) finds a persistent gender gap in competition entry among low-ability participants even after information is provided. However, the effect of information on gender differences in performance is less clear.

When subjects must compete, there are mixed results in reactions to information about competitors. Eriksson et al. (2009) finds that feedback on relative performance does not significantly change performance. The study reports positive peer effects in tournaments;

[1] The following discussion of existing literature focuses on studies that involve competition in a math-related task and explore gender differences, although Gill and Prowse use a slider task specifically designed to measure effort (Gill and Prowse, 2014). Other studies examine how information affects performance without any differences in incentives and will not be discussed (e.g. Azmat and Iriberri, 2010).

frontrunners do not slack off and underdogs rarely quit, although continuous feedback reduces the quality but not quantity of effort for underdogs. However, Gill and Prowse (2014) finds that subjects reduce effort after a loss, although males reduce effort only after failing to win large prizes. Buser (2016) shows somewhat different results depending on gender. Buser created three groups based on random pairing in a first round winner-take-all tournament: winners, losers, and those who receive scores, which he refers to as the no information group. Losers from the first round seek harder challenges, are less successful in the challenges and overall make less money in the second round compared to the winners. While there are no gender differences in average outcomes, such as the challenge level selected or performance in the challenge, males react to losing by becoming more challenge-seeking than winners and females react by lowering their performance.

The findings in these previous studies are contingent on random or contrived information about competitors to elicit a reaction from subjects. Although there is a range in the type of information provided, from relative scores to more direct messages of winning or losing, the explicit information acts as a treatment. The use of explicit information may contribute to results in the previous studies-- a study shows that the possibility of receiving feedback induces subjects to work harder even when they are not compensated for the extra effort, which demonstrates how responsive subjects can be to explicit information (Kuhnen and Tymula, 2011).

I focus on the effect of competitor level on competition performance, a relatively less understood aspect of gender differences in competition. I explore reactions to a subtler but realistic scenario of the perception of competitor difficulty, since people often compete with incomplete information about their competitors. For example, students may not know their rankings in class prior to taking a test; even if these rankings are known from a prior test, they do not perfectly transfer to another subject or even another test in the same subject. Despite this uncertainty, students must perform on assignments or tests. Thus, it is important to explore how a noisier yet realistic signal of competitor difficulty affects performance in competition. Although the context is essentially a lab-in-field environment rather than an actual school competition, the school setting allows students to compete against meaningful categories of competitors instead of relying on artificial competitors created by researchers.

By closely following Buser et al.'s (2014) protocol used in secondary schools in the Netherlands, the current study also provides evidence for replicability of findings in a different

context. Cultural context is demonstrated to play a role in gender differences in competition (Gneezy et al., 2009), although not necessarily in expected ways (Cárdenas et al., 2012); thus, it is important to acknowledge potential cultural influences on these differences. Nearly all of the studies use university subject pools in Western countries. To the author's knowledge, this is the first such experiment performed in a Muslim country and one of few performed in Asia. While this paper highlights several differences in the Science, Technology, Engineering & Math (STEM) and gender context particular to Malaysia, the findings are suggestive of gender stereotypes and differences in competition in STEM generally found in the literature.

The results of this study demonstrate that in a context where the standard gender difference in competition entry exists, males appear to be affected by the level of competition while females are not. When students face harder competitors, males respond by lowering performance while the performance of females does not vary significantly by level of competition. These somewhat surprising findings suggest that policies that require females to enter into more difficult competitive situations may not be detrimental to their performance in these situations.

The rest of the article proceeds as follows. Section 2 provides an overview of the study details, including context, data collection procedures and study design. The results from the study are detailed in Section 3. First, I provide descriptive analyses of the behavioral characteristics and other control variables used in later analyses. Then, I provide the analysis of the standard gender differences in competition (same-class competitions). Lastly, I provide an analysis of the response to different levels of competition (other-class competitions). Section 4 discusses potential mechanisms of these findings. Section 5 concludes.

2. Study overview

2.1 Context

Gender differences in competition appear to exist at a young age (Eccles et al., 1993; Gneezy and Rustichini, 2004; Harbaugh et al., 2002; Sutter and Glätzle-Rützler, 2014). These early differences may affect the trajectories of individuals' future decisions and outcomes. To understand competition phenomena in a relevant setting, this study uses a sample of high school students prior to any academic tracking.

This study takes place in public schools in Malaysia, a multicultural developing country in Southeast Asia with a majority Muslim population. Malaysia is a useful context for this study for several reasons. First, the informal but widespread ranking system within grades in public schools provides a unique opportunity to exogenously vary the level of competitor within classrooms, which will be discussed further in Section 2.2. Second, the STEM context in Malaysia appears to favor females compared to the populations used in prior studies, although standard male stereotypes of STEM seem to persist. Several studies view stereotypes associated with tasks as potential explanations for gender differences in math task competitions (Dreber et al., 2014; Grosse and Riener, 2010; Günther et al., 2010; Kamas and Preston, 2010; Shurchkov, 2012), thus any competitive differences found in the Malaysian context could help bring insight into whether gender differences in competition are similar in an environment with greater female STEM participation.

The Malaysian education system consists of six years of primary school and five years of secondary school; during the last two years of secondary school, or upper secondary school[2], students are placed into academic tracks with different associated prestige: the arts track (less prestigious) and the science track (more prestigious). Although there is no official tracking policy prior to the last two years of secondary school, many secondary schools use unofficial methods[3] of ranking and sorting students into classrooms within grade levels. Enrollment in preschool, primary school and secondary school is gender-balanced (49%-50% of enrollment is female). However, there are differences in gender proportions in the upper secondary school academic tracks. In upper secondary school, females constitute about half (47-49%) of the arts stream and the majority (about 58-59%) of students in the science streams[4]. Thus, there are more females than males in the most prestigious science track at the upper secondary level (Ministry of Education Malaysia, 2014). A similar gender distribution is found in the lower secondary Form 3[5] classes in this study, prior to the official academic tracking (see Section 2.2 for details).

The female advantage continues in tertiary education. Malaysia has a slightly lower ratio than the U.S. of females to males in tertiary education, although in both countries, females make

[2] Form 4 & 5 are known as upper secondary and are equivalent to grades 10 & 11.
[3] For example, sorting students into classrooms based solely on overall test scores.
[4] Science and arts streams are the two most common streams; some schools offer "sub-science" or "sub-arts" as well.
[5] Equivalent to grade 9.

up the majority of tertiary students (Malaysia: 1.21 to US: 1.36). However, nearly half of Malaysian female students (46%) versus less than a third of U.S. female students (30%) major in STEM fields (World Economic Forum, 2014). In fact, Malaysian females make up the majority of entrants, enrollments and graduates in most fields of study in the public universities including about two-thirds of graduates in Science, Mathematics and Computer; the only field in which females are a minority is Engineering, Manufacturing and Construction (Ministry of Education Malaysia, 2015). A qualitative study of the University of Malaya's[6] Computer Science and Information Technology department reveals that the majority of faculty, heads of department and dean were women in 2001 (Mellström, 2009). Mellström hypothesizes that computer science professions may be considered more suitable for females because of the office rather than field nature of the work; however, labor market data is limited such that it is not possible to identify the percentages of women in these fields.

Thus, females in Malaysia appear to face a more positive STEM climate in education than in many other countries. Nevertheless, gendered stereotypes for STEM and reading exist (see Section 3.1). Furthermore, prevailing gender norms may discourage females from being too "aggressive", which could influence gender responses to competition (Curriculum Development Division, 2016). These features demonstrate that multiple components of culture create a complex atmosphere that may affect gender dynamics in competition.

2.2 Data collection

This experiment was conducted in public secondary schools in one school district in Selangor, the largest and most urban state in Malaysia. I invited co-educational secondary schools in this district to participate in this study, asking for one classroom period of time; five schools agreed to participate. All schools in this study sort students into classes within grades by prior achievement, a widespread practice in Malaysia, and have a minimum of five classes in Form 3[7] to ensure sufficient variation in competition levels. Three to five classes from Form 3 were selected from each school to participate. The data collection was conducted over the span of one month, from July-August 2015. For a given school, the experiments in different classrooms[8] were conducted during the same day and often at the same time. Not every

[6] Malaysia's oldest and most prestigious public university.
[7] 9th grade equivalent; last year of lower secondary school and prior to academic track specializations.
[8] The experiment for one classroom at one school was conducted about three weeks after the rest of the classrooms at that school because of scheduling problems.

classroom in Form 3 in a school participated, experiments were often conducted at the same time within a school, and the bulk of the classroom experiments in the entire sample was conducted within one week, so there is little reason to worry that students knew about the experiment and strategized prior to participating. Students were paid two weeks after the experiment through sealed envelopes; there was no fixed participation fee and the average payout was RM10.26[9], with a minimum of RM0 and maximum of RM71.

Four of the five schools provided administrative information including student gender and midterm grades (the most recent official grades). The study was conducted during regular classroom instruction time in eighteen classrooms[10]. Each school engaged in some form of classroom rankings such that the classrooms were ordered according to student achievement, prior to official academic tracking practices at the end of Form 3. Students are well aware of this ranking, similar to how students in other countries such as the U.S. are aware of being in advanced or remedial classes. For example, in three of the five schools, classes are named in alphabetical order from top to bottom class. The top class, bottom class, and one to three middle-ranked classes in Form 3 of each school participated in this study. There were 562 secondary school students in Form 3 who participated in this study, but one student was dropped because there was no gender information available, leaving a sample of 561 students (290 males and 271 females). In the sample, females make up 40% of the bottom classes, 48% of the middle classes and 54% of the top classes[11]. The analyses of the effect of facing a different level of competition (i.e., easier or harder competition) are limited to the sample of middle classes (266 students), which were oversampled for this purpose.

The schools in this study represent over a fifth of the 24 public co-educational secondary schools[12] in the district. Although they may not be representative of the country as a whole, the schools appear to be similar on average to Malaysian public secondary schools. The average classroom size in the schools in the sample is 35.28, similar to the national average lower

[9] Currency was given in Malaysian Ringgit (MYR), which has a similar purchasing power to USD although the exchange rate was roughly 4 MYR:1 USD in summer 2015.

[10] One additional classroom was dropped due to technical problems.

[11] Post hoc ANOVA comparisons using the Sidak ($p=0.036$), Bonferroni ($p=0.036$), Scheffe ($p=0.043$) and Tukey ($p=0.032$) methods indicate that only the bottom and top classes have a statistically significant different proportion of females at the $p<0.10$ significance level.

[12] Most students in Malaysia attend co-educational schools. Wiseman (2008) finds that 14.67% of schools (indexed by 8th grade math classrooms) were sex-segregated, which was not statistically different from the international mean of 18.94%.

secondary classroom size of 34 (Ministry of Education Malaysia, 2014). Females make up 48% of the sample, similar to the national percentage of 50% (2015 data) in Form 3 (Ministry of Education Malaysia, 2015).

2.3 Study design

The objective of this experiment is to measure the rates of entering a competition when competing against classroom peers, and, in a subsequent round, to measure differences in performance when forced to compete against students from another higher- or lower-ranked class in the same grade and school.

The experiment has four rounds of tests with varying incentive structures followed by a survey, similar to the design first used in Niederle and Vesterlund (2007). The test instrument for each round was a five-minute math test with 40 double digit multiplication questions, which is a slightly longer and more difficult task than the one used by Niederle and Vesterlund, in order to enable more variance in scores due to an additional incentivized round in this study. This task was designed to measure the level of effort, not mathematical knowledge or attitudes. None of the questions repeat in the study and all numbers with zeroes were removed in order to keep the level of difficulty comparable across each test. There were no penalties for incorrect answers. Students were not allowed to use calculators but were given pieces of scratch paper to solve problems on. Directions about the specific incentive system of the round's test were read out loud in Malay, the language of instruction, prior to each test. All documents were given in both English and Malay. Students were told not to speak during the duration of the study, and had to place their pens down and stand up when the end of each test was announced. Furthermore, students were informed that only 1 out of the 4 rounds of tests would be compensated, randomly chosen at the end of the session, in order to avoid hedging and to encourage each student to try his/her best during each round. Thus, at the end of each session, a representative from the class picked a ball numbered from 1 to 4 out of an opaque bag to choose which round was paid out for that entire class.

Test 1 was scored according to a piece-rate incentive; for this test, students were paid RM0.50 per each correct answer. Test 2 was scored according to a winner-take-all tournament incentive (i.e. a competitive incentive). For this test, students were told they would be competing against 3 other randomly selected students (4 students per group) from their class. If they

obtained the highest score (i.e. first place), they received a payment of RM2 per each correct answer, but if they did not obtain the highest score, they received nothing[13].

Prior to Test 3, students were given the choice of how they wanted to be compensated for the third test. Each student chose between one of the prior two incentive schemes, marked the choice on a form, then inserted the form into an envelope. Students were informed prior to decision-making that if they chose the winner-take-all tournament incentive, they would compete against a new set of three randomly selected competitors' scores from Test 2 so they could be competing against any of their classmates, not just those who chose the tournament incentive (see Niederle and Vesterlund, 2007). Test 3 proceeded after every student selected a choice and put away the form in an envelope.

Prior to Test 4, students were given slips of paper that informed them which class they would be competing against in the fourth test. Thus, in the fourth round of the study, students were told they would be competing in a winner-take-all tournament, competing against three randomly selected students from the other class, under the same incentive structure as Test 2 but using only Test 4 scores. In each class, students were randomly assigned to one of two other classes in their grade (e.g. bottom or top class if the student were in a middle class); classes were referred to by their official school names with no explicit reference to positioning within the grade level. However, as described earlier, students are well aware of the implicit differences between classes.

After Test 4, students completed a survey which included incentivized questions on levels of confidence and risk aversion, in addition to non-incentivized questions about their attitudes, opinions and family background. Students never received information about their scores during the experiment. Students could estimate how they had performed only after they were given their payments, a couple weeks after the experiment had been completed.

3. Results

3.1 Same-class competition analysis

The following section presents results from the first three rounds of the study, which replicates the design from Buser et al. (2014). First, I provide the descriptive results of the

[13] Ties were awarded the same rank, and then skipped the next number of ranking (Stata's egen rank, field option).

performance, competition choice, behavioral and other individual characteristics of students. I then present the regression results that confirm the gender gap in competition.

There is no gender gap in performance for this multiplication task, whether students are under piece-rate or tournament incentive against their classroom peers. A table of descriptive characteristics shows the performance and competition choice prior to Test 3, when all students are under the same incentive structures (Table 1). Although it is not a focus of this paper, there is evidence that the sorting mechanism into classrooms by student prior achievement resulted in classes with overall differences in student performance, which is an important component of the analyses of performance against other classes. The average number of questions correct for the first test, under the piece-rate incentive, is 10.141 although this varies between 5.937 in the bottom classes to 12.432 in the top classes. The average number of questions correct for the second test, under the winner-takes-all tournament incentive, is significantly higher at 12.041, ranging from 7.746 in the bottom classes to 14.444 in the top classes[14]. Overall, females appear to outperform males on these first two tests, though these gender differences disappear when taking into account the class level and corresponding differences in gender distribution across class levels. Thus, it is established that there are no gender differences in performance under either of the incentives for this task.

[Table 1 about here]

Furthermore, both genders increase performance under the competition incentive. The different incentive structures between Test 1 and Test 2 affects both genders; the average number of answers correct between Test 1 and Test 2 statistically significantly increases for both males and females (Appendix A-1). This increase could indicate learning with successive tests (discussed further in Section 3.2); however, a recent study finds that the order of piece-rate and tournament rounds does not significantly affect the difference in performance under the two incentives in a similar experiment (Wozniak et al., 2016). Therefore, we can interpret the positive increase as the response to competition.

[14] The numbers of correct answers for both Test 1 and Test 2 are different between all three class levels according to the analysis of variance comparisons, which indicates that student ability in these tasks has been appropriately sorted by class levels (ANOVA analyses available upon request).

Unlike performance on the tests, there is a clear difference in the rates at which males and females choose competition, both overall and at each class level. Overall, less than a third of students (29.6%) choose competition for the incentive structure of Test 3. Females choose into competition at almost half the rate of males, with an average of 20.7% of females versus 37.9% of males choosing competition, with the greatest difference in the top classes (18.5% of females versus 46.8% of males).

The choice into competition for Test 3 does not appear to incentivize students to perform better than those who did not choose into competition for Test 3. There is no difference in the increase in number of correct answers from Test 2 to Test 3 for those who chose competition and those who chose piece-rate (Table 2). This can indicate either insensitivity to the choice, or poor measurement of effort (e.g. ceiling effects) on performance. Subsequent increased performance on Test 4 discussed in Section 3.2 implies that students did not respond to choice, rather than the task failing to measure changes in effort.

[Table 2 about here]

Other factors such as confidence, risk-aversion, academic performance, attitudes and expectations towards math/science, and socio-economic status may be influential in students' choice of competition. A summary of student behavioral and personal characteristics is shown in Table 3 (Appendix A-2 for detail). There are several characteristics that differ by gender.

Males are more confident and over-confident than females in competitions against their own class. Confidence is measured by two questions on the survey, similar to what is used in Niederle and Vesterlund (2007). These questions ask what rank (1-first place to 4-last place) students think they had achieved for the two forced competition rounds, Test 2 (against own class) and Test 4 (against other class). Students received RM1 per correct answer for these questions. Overconfidence is defined as the difference between actual rank[15] and guessed rank, with a range of -3 to 3. This measure provides the student's level of confidence for the particular task rather than a more generalized measure (e.g. soliciting student perceptions about class rank).

[15] Actual rank is constructed from 1000 simulations of random draws of 3 other students from the appropriate class against a given student's score; the modal value was selected as actual rank.

The average guessed rank of males against their own class is 2.441 versus 2.715 for females ($p=0.001$); thus, males guessed that they obtained a better rank than females guessed. After accounting for actual ranks, females are under-confident while males' guessed ranks are closer to their actual ranks (slightly under-confident against their own class and slightly over-confident against another class).

It appears that males are more accurate in their rankings, although both males and females appear less confident about winning than other studies have found (e.g. Niederle and Vesterlund, 2007). However, the male percentage is roughly in line with what was found in a sample of similarly-aged students (Buser et al., 2014). About 21% of males and 9% of females believe that they won the tournament in Test 2 ($p<0.001$), while 30% of males and 24% of females actually win the tournament, with no significant difference.

Males are more risk-seeking than females according to both risk measures in this study. Risk preference is measured in two ways on the survey. First, students answered an incentivized question based on a modified question used by Eckel and Grossman (2002) that asked them to choose between an option with 100% certainty (RM2) or one of four 50/50 lottery options based on a flip of a coin at the end of the study: RM3 or RM1.50, RM4 or RM1, RM5 or RM0.50 or RM6 or RM0. The coin was flipped in front of the classroom at the end of the study and the individual's choice was paid out with the rest of his/her earnings. Second, students answered a non-incentivized risk preference question taken from the 2004 wave of the German Socio-Economic Panel Study following Dohmen and Falk (2011), which finds that this question predicts incentivized lottery choices. The question is: How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks? Check ONE box on the scale, where the value 0 means: 'unwilling to take risks' and the value 10 means: 'fully prepared to take risks'. Males choose a more risky lottery option and also choose a higher level of risk to describe themselves. In this sample, the correlation between these two measures is 0.243 overall, 0.208 for males and 0.230 for females ($p<0.001$ in both cases).

Females and males perform similarly on their school math midterm grades[16]; there is no gender difference (Appendix A-3 for detail). However, there is a significant female advantage for overall midterm grades: females have a 5 percentage point higher overall midterm grade than

[16] Administrative grade data was obtained from four out of the five schools.

males (57.436 versus 52.414, p=0.005). Despite this academic context, the student survey responses show that male-favoring stereotypes exist for math and science and female-favoring stereotypes exist for reading, similar to Western stereotypes (Appendix A-4 for detail).

Females and males have similar levels of enjoyment of math; 74.3% of males and 69.7% of females agree or strongly agree that they like math (no significant difference) although a higher percentage of males than females like science while a higher percentage of females than males like reading (p=0.015, p<0.001 respectively). In addition, a higher percentage of males believe they are good at math; almost half of males (47.2%) versus a little over a third of females (36.8%) agree or strongly agree that they are good at math (p=0.014). A similar pattern follows for science although it is reversed for reading; over three-quarters of females (77.2%) versus two-thirds of males (67.5%) think they are good at reading (p=0.010).

The science and math fields are most prestigious; 71.4% of all students rate the Science track as the best academic track in upper secondary school, with no statistically significant gender differences. A marginally higher percentage of males than females think that they will end up in the Science track in the next academic year, 47.6% versus 40.6% (p=0.097). On average, students believe that boys are better at math and science while girls are better at reading; males tend to rate boys as better in each of these subjects (Appendix A-4 for detail).

There do not appear to be gender differences in socioeconomic status (SES), using parental education as a proxy. On average, 45.1% of students' fathers and 36.7% of students' mothers hold at least bachelor's degrees (Appendix A-5).

[Table 3 about here]

Given that these variables may contribute to an individual's decision to enter into competition, it is important to control for these variables when determining whether there is a gender difference in competitiveness; that is, choosing competition for Test 3. The measure of competitiveness in this paper is similar to the measure first used in Niederle and Vesterlund (2007). Student choice of whether to enter into competition or piece-rate compensation prior to Test 3, controlling for other variables, is used as the measure of competitiveness (choosing competition is used interchangeably with choosing the tournament incentive for Test 3).

When controlling for only the score on the piece-rate test (Test 1) and the difference between the tournament and piece-rate scores (Test 2-Test 1), females are 17.3 percentage points less likely than males are to choose competition (Table 4, Model 1). When adding in the level of overconfidence, the difference decreases to 14.9 percentage points, which is different from the coefficient in Model 1 at the p=0.005 level[17] (Model 2). This difference remains largely stable when adding in both measures of risk preferences (Model 3), and is not significantly different from Model 2. When student attitudes and SES are added, the gender gap is 13.9 percentage points, although none of the coefficients for these characteristics appear to influence competition entry (Model 4). Lastly, although one school did not provide midterm scores, the gender gap remains when including math and overall midterm grades in addition to all the other covariates (Model 5).

[Table 4 about here]

Similar results hold for the previous models when this school is excluded from the analyses (Appendix A-6) or when session fixed effects are used instead of class fixed effects to account for simultaneous experimental sessions (Appendix A-7). Thus, the gender gap is still significant although the power from the reduced sample size is lower, and is very similar to the gap found in a similar age sample of ninth-grade students in the Netherlands (Buser et al., 2014).

Secondary students in Malaysia show the standard gender gap in choosing competition that has been demonstrated in many different contexts. When only controlling for previous performance, the gender gap is 17.3 percentage points. The gender gap is reduced a total of about 20% when controlling for confidence, risk preferences, student attitudes about math and socioeconomic status, but females are still 13.9 percentage points less likely than males to choose competition ($p<0.05$).

3.2 Other-class competition analysis

[17] Comparisons of the coefficient for female use seemingly unrelated estimations with clustered standard errors (not exact standard errors from main analyses, since Stata's suest command does not accept xtreg models). The coefficient for "Female" in Model 1 is significantly different from the coefficients in Models 2-4 at the $p<0.10$ level.

The previous analysis confirms that the standard gender gap in choosing into math competition exists for this sample of secondary school students. This section focuses on the novel contribution of this paper: how students react to different levels of competition. I present several descriptive findings of the difference in performance when facing different competitors. I then present the experimental results in addition to exploring heterogeneity in these results and whether changes in questions answered or accuracy led to these results.

The sample for the following analyses is restricted to the middle-ranked (middle) classes so that there are both easier (bottom class) and harder (top class) competitors. There are 266 students in 8 middle classes (137 male and 129 female), which represents a little less than half the number of students in the original sample. As described in Section 2.3, students in the middle classes were randomized to compete against either the top ranked class or the bottom ranked class in the same grade and school, although classes were only named by their official titles as to not directly prime students to the level of their competitors. Students received a slip of paper informing them which class their competitors would come from, and were told to put the slip of paper in an envelope and not talk so that treatment assignments remained concealed.

As in the overall sample, there is a general upward trend in the number of correct answers in successive tests, which suggests that learning[18] could play a role in the observed scores (Table 5). This brings up concerns about whether the observed scores reflect learning or ability rather than the effort put into the task. The randomization should alleviate these concerns for this last round, unless learning or ability is not balanced within genders across treatment groups. The randomization produced balanced groups competing against higher and lower competitors across all observable baseline characteristics (gender, math midterm score and overall midterm score). In addition, most student characteristics measured prior to treatment are balanced across groups, including scores on Test 2, Test 3, the difference between Test 2 and Test 1, and the competition choice. Treatment assignment predicts the score on Test 1 at the 10% significance level, although

[18] A limitation of this study is the difficulty in separating out learning effects and response to incentives, given that the order of the rounds remained constant in order to replicate the Niederle and Vesterlund (2007) experiment to determine gender differences in competition. Cotton and colleagues show that repeated competition eliminates the gender gap in performance in their study (Cotton et al., 2013). The results from the current study show some indication that genders may perform differently in successive competitions. The average scores increase from Test 1 to Test 2 for both genders, for only females from Test 2 to Test 3, and then do not increase from Test 3 to Test 4 for either gender. However, there is no indication that males lower their performance during successive rounds, unlike what Cotton and colleagues find.

there is no significant correlation between treatment and Test 1 score within gender (Appendix A-8). The following analyses control for Test 1 score, difference between Test 1 and Test 2 score, and competition choice as robustness checks.

[Table 5 about here]

Although the upward trend in scores on successive tests is clear in the treatment against the bottom class, it is less apparent for those who competed against the top class. However, the incentives between the third and fourth test vary by student choice thus it is most relevant to compare results from Test 4 against Test 2.

In the following analysis, the primary variable of interest is the difference between performance in Test 2 and Test 4. Similar variables are constructed for the difference between total number of questions answered and the difference in accuracy of answers, which are used to explore the main results. Thus, a student's performance against another class (Test 4) is compared against performance against a student's own class (Test 2). This within-subject design allows us to see the effect of a different level of competitor using each subject's baseline value (i.e. performance on Test 2). The average value of the difference in the number of correct answers from Test 2 to Test 4 is 1.34 with a standard deviation of 2.90 and a range of -7 to 10. As Figure 1 shows, there is no gender difference in the change in performance when the competitors are from the bottom class. Both genders perform about 1.5 questions better. However, when matched against competitors from the top class, females increase the number of correct answers by significantly more than males, 1.806 correct answers compared to 0.507 correct answers (Appendix A-9, p=0.017).

[Figure 1 about here]

Since treatment is randomized within class, the following equation can be used to determine the effect of the treatment.

$$y_{ij} = \Gamma_j + \beta_1 Treatment_{ij} + \beta_2 Female_{ij} + \beta_3 (Treatment * Female)_{ij} + \theta \mathrm{X_{ij}} + \epsilon_{ij}$$

where:

y_{ij} is the difference in number of correct answers between other and own class (Test 4 - Test 2) for student i in class j

Γ_j is the class fixed effects

Treatment is 1 if assigned to the top class and 0 if assigned to the bottom class for student i in class j

Female is 1 if female and 0 if male for student i in class j

*Treatment * Female* is 1 if student i in class j is assigned to the top class and is female; 0 otherwise. This represents the gender difference in the effect of treatment on the difference of performance between other and own class

X_{ij} is a vector of student attributes

The regressions in Table 6 show the effects of competing against the top class (competition against bottom class as reference group), relative to competing against the student's own class. Since the treatments were randomly assigned, the estimates of the effect of the treatment can be directly interpreted. Baseline covariates are included in subsequent models, which lowers the precision of the estimates (Columns 2-3). The regressions are also performed separately for males (Columns 4-6) and females (Columns 7-9).

The effect of facing the top class versus the bottom class is about one question less, -1.029 ($p<0.05$) (Table 6, Column 1). However, the interaction effect of being female and facing the top class is positive and similar in magnitude to this negative effect, 1.184 ($p<0.10$). When adding in baseline variables including Test 1 performance, response to competition incentive (difference in Test 1 and Test 2 performance), and competition choice, the pattern remains similar; there is a stable negative main effect although precision decreases so that the female interaction effect is not statistically significantly different from zero (Table 6, Column 3).

[Table 6 about here]

The gender difference in response to harder competition is clearer when examining the regression results separately by gender (Table 6, Columns 4-9). The effect of facing the top class instead of the bottom class is consistently negative and close to 1 question for males after

controlling for behavior and performance from prior rounds[19], ranging from -0.996 to -0.893 (Table 6, Columns 4-6). On the other hand, females do not seem affected by facing the top class as opposed to the bottom class; the effect is not statistically different from zero (Table 6, Columns 7-9). These findings indicate that males are negatively affected by facing a difficult competitor while females are not. Qualitatively similar results hold when the whole sample of students is included and treatment is defined as competing against any higher class (Appendix A-10), session fixed effects are used (Appendix A-11) or absolute score on Test 4 is used controlling for Test 2 performance and other variables (Appendix A-12). Males perform worse when competing against the top class rather than the bottom class, even after controlling for prior performance and competitive behavior, while there is no evidence that females perform differently according to the level of their competitors.

To explore these results, I examine heterogeneity in the sample in addition to whether the effects are due to differential numbers of questions answered or a change in the accuracy of answers.

An important characteristic of this sample is the variance in performance both within schools (e.g. average scores in middle classes compared to top classes) and across schools. All previous results include class fixed effects, which help capture this heterogeneity. However, it is also instructive to view these results in a more easily comparable manner such as the chance of winning against the top class. The chance of winning against the top class conditional on the number of correct answers varies by school; for example, with 18 correct answers, a student in a middle class at School 4 has an 83% chance of winning, while a student in a middle class at School 5 has a 9% chance of winning (Table 7). When the chances of winning are used as controls instead of the numbers of answers correct, the effects of facing harder competition remain negative for males and null for females (Appendix A-13).

[Table 7 about here]

These effects of facing more difficult competitors appear to differ along the distribution of baseline performance by gender. For males, the difference between Test 2 and Test 4 score is

[19] When the baseline variables are added in models 3, 6 and 9, the difference between Test 1 and Test 2 performance (T-PR) shows a consistently large negative coefficient, which could possibly be due to ceiling effects.

greatest at the best and worst quintiles of the baseline (Test 1) performance distribution (Figure 2). Males at the best and worst quintiles who face the top class perform about two questions worse than males who face the bottom class. Females in the top two quintiles perform similarly when facing either the top or bottom class, although females in the bottom two quintiles who face the top class appear to perform a little better than those who face the bottom class. Overall, it appears that males from the top and bottom of the performance distributions respond most to the level of competition.

The change in performance from Test 2 to Test 4 could be due to a combination of the quantity and accuracy of answered questions. For example, individuals can obtain a higher score by answering more questions with the same (or lower) level of accuracy or by answering the same number (or fewer) of questions with higher accuracy. It appears that competitor difficulty has no effect on the number of questions answered; there is a negative effect for males that is not significant after controlling for prior number of questions answered and competitive behavior (Table 8).

[Table 8 about here]

However, males but not females are less accurate when facing more difficult competitors; the difference between females and males when facing harder competition is about 5 percentage points and significant at the 5% level (Table 9, column 3). After controlling for prior accuracy and competitive behavior, the accuracy of males who face harder competitors is a little over 3 percentage points (significant at the 10% level) less than the accuracy of males who face easier competitors (Table 9, column 6). Thus, it appears that males change the quality (accuracy) of performance rather than the quantity of effort against more difficult competition.\

[Table 9 about here]

4 Discussion

This study shows the robustness of the gender gap in competition. Overall, females choose into competition at about half the rate of males—20.7% versus 37.9%. After controlling for student performance, confidence, risk preferences, and other student characteristics, females still have a 13.9 percentage point lower probability of choosing into competition less than males. This gender gap is very similar to what is found in the Netherlands with a similar age group and experiment protocol, although the overall rates of competition are lower in Malaysia.

There is another gender gap that emerges when facing different levels of competitors. The performance of females is not affected by facing harder competitors. However, males perform almost one question worse when facing competitors from the top class (about one-third of a standard deviation) than when competing against the bottom class. It appears that accuracy decreases for males when facing the top class compared to the bottom class. There may be several explanations for the gender difference in performance against harder competitors, such as the gender composition of groups, differential expectations when facing different classes or changes in the chance of winning or expected earnings.

One possible explanation for these results may be the gender composition of the competitor groups. Existing research indicates that the gender composition of competitors can affect performance in competitions (Booth and Nolen, 2012; De Paola et al., 2015; Gneezy et al., 2003; Kuhnen and Tymula, 2011). Thus, the perceived gender composition of the competitors could also play a role in these results. As noted in Section 2.2, there is a higher proportion of females in the top classes than in the middle or bottom classes, although the difference is not statistically significant between the top and middle classes, which is the relevant comparison in these analyses. The range in female composition of the top class across the five schools in the study is reasonably small, from 48.48% to 60.71%. These factors make it unlikely that the female composition of the top classes affected results.

These results could also be explained by different expectations between genders when competing against harder or easier competition, and a corresponding differential change in effort. For example, Kuhnen and Tymula (2011) use gender composition of the group as a proxy for perceived difficulty of competitor and find that females have lower output, worse expected rank and worse actual rank with more males in their group while males are not affected by the gender composition of the group. However, gender composition of the group may be an inappropriate proxy for perceived difficulty of competitors. It is worth noting that they observe that males expect better rankings than females (similar to this study) yet males also outperform females (different from this study).

I use a similar task but more clearly designated groups of easier or harder competitors and find that expectations of males rather than that of females appear to be affected. There are no gender differences in the actual rankings in either treatment condition, although both genders guess a better rank when competing against the bottom class (Table 10). These rankings also

confirm that the difficulty levels of competitors are appropriately categorized; students in the sample have a 55% chance of winning the tournament against the bottom class and a 16% chance of winning against the top class, with no gender difference. However, males guess they are a better rank than females do and are more overconfident when facing the bottom class (p-values 0.019 and 0.061, respectively). There are no gender differences in guessed rank or overconfidence when facing the top class, although males are slightly overconfident and females are under-confident. Since baseline measures of confidence against different classes were not elicited in this study in order to prevent priming, it is not possible to distinguish whether the treatment of facing more difficult competition changed male and female priors about their performance differentially. Nevertheless, these ex-post elicited measures of confidence could indicate a possible mechanism difference between genders; that is, males may lower performance because they expect to do worse against harder competition (on par with females' confidence), relative to their confidence against easier competition (more confident than females).

[Table 10 about here]

Finally, there is a negative effect on the chance of winning (Table 11) and expected earnings (Table 12) when facing harder competition for both females and males. The relatively lower performance of males when facing harder competition does not appear to result in a lower chance of winning or decreased expected earnings for males. Thus, the lower performance of males could reflect greater efficiency (e.g. lower performance for the same financial outcomes).

[Table 11 about here]

[Table 12 about here]

The gender difference in performance under more difficult competition is somewhat surprising, given findings from previous literature which generally show an equal response if not female disadvantage when encountering difficult competition. For example, Eriksson et al. (2009) finds that relative information does not affect performance, Gill and Prowse (2014) finds that both genders lower performance after a loss and Buser (2016) finds that females lower their performance after a loss but males do not.

However, this study design does not depend on explicit information, as previous studies have used, but a more realistic yet less certain competitive situation. The experiment exploited pre-existing differences in levels of competitors without an explicit message about relative

position, which could affect the dynamics in competition. There is suggestive evidence that males may have lowered expectations when facing harder competition, although the gender gap in the effect of facing harder competition on performance does not appear to extend to a gender difference in the chance of winning or expected earnings.

5 Conclusion

This paper presents experimental evidence that females and males have different reactions to more difficult competitors—males lower their performance while females' performance does not change. In addition, it appears that standard gender differences in competitive behavior apply even within a STEM context with more female participation. Given the similar gender gaps in competition choice, it is reasonable to believe that these findings about reacting to harder competition apply in broader contexts.

The results from this study confirm the gender gap in choosing into competition in a math task similar to those that have been linked to future educational choices. Although several previous studies have found that females perform worse than males in competition, the current study adds to the body of literature that finds no gender difference in competitive performance. Furthermore, the within-subject study design shows a gender difference in the response to harder or easier competition.

These findings have implications for policies designed to attract females into more competitive environments. Existing research clearly indicates that, when given a choice, females choose into competition less than males do. There are many situations in which people face competition choices, such as which courses to take in school or which jobs to apply for. Early decisions could have lasting consequences; for example, there may be prerequisite courses for certain majors which are required to pursue certain occupations (e.g. advanced math/science courses required for engineering degrees to become an engineer). If females differentially decline to enter into competition early, gender gaps may widen over time as fewer opportunities remain open.

However, it appears that females may not be negatively affected by the level of competition once they are in a more competitive situation. Thus, if females do not perform worse in more competitive environments even when they do not choose into these environments, perhaps policies can be designed to compel people into more difficult competitive environments.

For example, schools could require more advanced STEM courses or companies could provide mandatory leadership programs, which would require females who may not otherwise choose those programs to participate in them. Then, they may thrive in the more competitive environment. On the other hand, it is important to ensure that males do not perform worse in these more demanding situations where there could be negative outcomes from lowered performance. The results of this study are found in a sample of students in middle-ranked classes with no gender differences in performance, thus these proposed policies may not apply among high or low performance individuals or when gender differences in performance exist. These policies also do not address other barriers such as chilly climates that females face in competitive environments.

Future research could look at the generalizability of and possible mechanisms underlying the results. This study was conducted among secondary students in middle-ranked classes in an Asian country; it would be illuminating to see whether the results hold among different ages, performance levels or cultural contexts. In addition to addressing generalizability, future studies can examine more deeply the potential mechanisms for these results, such as a differential change in expectations when facing different levels of competition. Other possibilities from the psychology literature could be differences in persistence or grit; for example, females may be grittier than males in learning environments. Thus, even if females would not choose more competitive environments, they could persist and succeed in them. Understanding these mechanisms could help design policies that could result in greater participation and performance in environments with more difficult competition.

References

Almås, I., Cappelen, A. W., Salvanes, K. G., Sørensen, E. Ø., & Tungodden, B. (2016). What Explains the Gender Gap in College Track Dropout? Experimental and Administrative Evidence. *American Economic Review*, 106(5), 296–302.

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

Booth, A., & Nolen, P. (2012). Choosing to compete: How different are girls and boys? *Journal of Economic Behavior & Organization*, 81(2), 542–555.

Buser, T. (2016). The Impact of Losing in a Competition on the Willingness to Seek Further Challenges. Forthcoming in *Management Science*.

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

Buser, T., Peter, N., & Wolter, S. C. (2017). Gender, Competitiveness, and Study Choices in High School: Evidence from Switzerland. *American Economic Review*, 107(5), 125–130.

Cárdenas, J.-C., Dreber, A., von Essen, E., & Ranehill, E. (2012). Gender differences in competitiveness and risk taking: Comparing children in Colombia and Sweden. *Journal of Economic Behavior & Organization*, 83(1), 11–23.

Cason, T. N., Masters, W. A., & Sheremeta, R. M. (2010). Entry into winner-take-all and proportional-prize contests: An experimental study. *Journal of Public Economics*, *94*(9–10), 604–611.

Cotton, C., McIntyre, F., & Price, J. (2013). Gender differences in repeated competition: Evidence from school math contests. *Journal of Economic Behavior & Organization*, *86*(C), 52–66.

Curriculum Development Division, Ministry of Education Malaysia. 2016. Sharing Malaysian Experience in Participation of Girls in STEM Education. Geneva, Switzerland, UNESCO International Bureau of Education (IBE).

Datta Gupta, N., Poulsen, A., & Villeval, M. C. (2013). Gender Matching and Competitiveness: Experimental Evidence. *Economic Inquiry*, 51(1), 816–835.

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

Dohmen, T., & Falk, A. (2011). Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender. *American Economic Review*, *101*(2), 556–90.

Dreber, A., Essen, E., & Ranehill, E. (2014). Gender and competition in adolescence: task matters. *Experimental Economics*, *17*(1), 154–172.

Eccles, J., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and gender differences in children's self-and task perceptions during elementary school. *Child Development*, 64(3), 830–847.

Eckel, C. C., & Grossman, P. J. (2002). Sex Differences and Statistical Stereotyping in Attitudes Toward Financial Risk. *Evolution and Human Behavior*, *23*(4), 281–295.

Eriksson, T., Poulsen, A., & Villeval, M. C. (2009). Feedback and incentives: Experimental evidence. *Labour Economics*, *16*(6), 679–688.

Ertac, S., & Szentes, B. (2011). The Effect of Information on Gender Differences in Competitiveness: Experimental Evidence (Koç University-TUSIAD Economic Research Forum Working Paper No. 1104). Koc University-TUSIAD Economic Research Forum.

Gill, D., & Prowse, V. (2014). Gender differences and dynamics in competition: The role of luck. *Quantitative Economics*, *5*(2), 351–376.

Gneezy, U., Leonard, K. L., & List, J. A. (2009). Gender Differences in Competition: Evidence From a Matrilineal and a Patriarchal Society. *Econometrica*, 77(5), 1637–1664.

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in Competitive Environments: Gender Differences. *The Quarterly Journal of Economics*, *118*(3), 1049–1074.

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

Grosse, N. D., & Riener, G. (2010). Explaining Gender Differences in Competitiveness: Gender-Task Stereotypes (Jena Economic Research Paper No. 2010-017). Friedrich-Schiller-University Jena, Max-Planck-Institute of Economics.

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

Harbaugh, W., Krause, K., & Vesterlund, L. (2002). Risk Attitudes of Children and Adults: Choices Over Small and Large Probability Gains and Losses. *Experimental Economics*, 5(1), 53–84.

Kamas, L., & Preston, A. (2010). Are Women Really Less Competitive Than Men? Working Paper, Santa Clara University.

Kuhnen, C. M., & Tymula, A. (2011). Feedback, Self-Esteem, and Performance in Organizations. *Management Science*, 58(1), 94–113.

Mellström, U. (2009). The Intersection of Gender, Race and Cultural Boundaries, or Why is Computer Science in Malaysia Dominated by Women? *Social Studies of Science*, *39*(6), 885–907.

Ministry of Education Malaysia. (2014). Quick Facts 2014. Retrieved 19 February 2016 from http://www.moe.gov.my/cms/upload_files/publicationfile/2014/pubfile_file_002100.pdf

Ministry of Education Malaysia. (2015). Quick Facts 2015. Retrieved 19 February 2016 from http://www.moe.gov.my/cms/upload_files/publicationfile/2015/pubfile_file_002101.pdf

Niederle, M. (2016). Gender. In J. Kagel & A. E. Roth (Eds.), *Handbook of Experimental Economics* (2nd ed., pp. 481–553). Princeton University Press.

Niederle, M., Segal, C., & Vesterlund, L. (2013). How costly is diversity? Affirmative action in light of gender differences in competitiveness. *Management Science*, *59*(1), 1–16.

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

Niederle, M., & Vesterlund, L. (2011). Gender and Competition. *Annual Review in Economics*, *3*, 601–630.

Niederle, M., & Yestrumskas, A. H. (2008). Gender Differences in Seeking Challenges: The Role of Institutions (Working Paper No. 13922). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w13922

Ors, E., Palomino, F. & Peyrache, E. (2013). Performance Gender Gap: Does Competition Matter? *Journal of Labor Economics*, 31(3), 443–499.

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

Sutter, M., & Glätzle-Rützler, D. (2014). Gender Differences in the Willingness to Compete Emerge Early in Life and Persist. *Management Science*, *61*(10), 2339–23354.

Wiseman, A. W. (2008). A culture of (in) equality?: A cross-national study of gender parity and gender segregation in national school systems. *Research in Comparative and International Education*, *3*(2), 179–201.

World Economic Forum. (2014). The Global Gender Gap Report 2014. Retrieved from http://reports.weforum.org/global-gender-gap-report-2014/

Wozniak, D., Harbaugh, W. T., & Mayr, U. (2014). The Menstrual Cycle and Performance

Feedback Alter Gender Differences in Competitive Choices. *Journal of Labor Economics*, *32*(1), 161 – 198.

Wozniak, D., Harbaugh, W. T., & Mayr, U. (2016). The Effect of Feedback on Gender Differences in Competitive Choices. (SSRN Scholarly Paper No. ID 1976073).

Zhang, Y. J. (2013). Can Experimental Economics Explain Competitive Behavior Outside the Lab? (SSRN Scholarly Paper No. ID 2292929). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2292929

Table 1. Descriptive statistics of number of correct answers and competition choice, by class level.

Variable	Class level	Overall	Male	Female	Diff	p-value
Test 1 (Piece-Rate)	Overall	10.141	9.693	10.620	-0.927	0.040
	Bottom	5.937	5.908	5.980	-0.072	0.948
	Middle	10.677	10.307	11.070	-0.763	0.173
	Top	12.432	12.338	12.511	-0.173	0.847
Test 2 (Tournament)	Overall	12.041	11.710	12.395	-0.684	0.082
	Bottom	7.746	7.789	7.680	0.109	0.785
	Middle	12.549	12.482	12.620	-0.138	0.650
	Top	14.444	14.208	14.641	-0.434	0.354
T-PR	Overall	1.900	2.017	1.775	0.242	0.578
	Bottom	1.810	1.882	1.700	0.182	0.469
	Middle	1.872	2.175	1.550	0.625	0.205
	Top	2.012	1.870	2.130	-0.260	0.378
Competition choice	Overall	0.296	0.379	0.207	0.173	<0.001
	Bottom	0.325	0.395	0.220	0.175	0.041
	Middle	0.271	0.321	0.217	0.104	0.057
	Top	0.314	0.468	0.185	0.283	<0.001

Number of observations are from the whole sample: 561 overall, with 290 males and 271 females overall. The gender breakdown is: 76 males and 50 females in the bottom classes; 137 males and 129 females in the middle classes; 77 males and 92 females in the top classes. T-PR is the difference between number correct on the tournament (Test 2) versus piece-rate test (Test 1). Competition choice is the proportion that chose the tournament rather than the piece-rate incentive. P-values are from Mann-Whitney *U* tests.

Table 2. Change in number of correct answers between Test 2 and Test 3.

	Overall		Chose Piece-rate		Chose Competition			
	Mean	N	Mean	N	Mean	N	Diff	p-value
Overall	0.720	561	0.681	395	0.813	166	-0.132	0.518
Male	0.638	290	0.600	180	0.700	110	-0.100	0.771
Female	0.808	271	0.749	215	1.036	56	-0.287	0.385

Differences are calculated by student (Test 3-Test 2). P-values are from Mann-Whitney *U* tests.

Table 3. Descriptive statistics of student characteristics.

	Overall		Male		Female			
Variable	**Mean**	**N**	**Mean**	**N**	**Mean**	**N**	**Diff**	**p-value**
Confidence								
Guessed Rank Test 2	2.573	560	2.441	290	2.715	270	-0.273	0.001
Guessed Rank Test 4	2.455	560	2.360	289	2.557	271	-0.197	0.028
Overconfidence Test 2	-0.221	560	-0.097	290	-0.356	270	0.259	0.016
Overconfidence Test 4	-0.136	560	0.042	289	-0.325	271	0.366	<0.001
Risk								
Incentivized risk scale (1-5; 5 most risky)	2.588	561	3.103	290	2.037	271	1.067	<0.001
Non-incentivized risk scale (0-10; 10 most risky)	6.161	559	6.410	288	5.897	271	0.513	0.001
Midterm scores								
Math	49.842	463	49.457	230	50.223	233	-0.767	0.852
Overall GPA	54.936	432	52.414	215	57.436	217	-5.022	0.005
Attitudes and Beliefs								
Like Math	0.721	555	0.743	284	0.697	271	0.046	0.232
Like Science	0.770	556	0.812	287	0.725	269	0.087	0.015
Like Reading	0.752	537	0.647	275	0.863	262	-0.215	<0.001
Good at Math	0.422	552	0.472	286	0.368	266	0.104	0.014
Good at Science	0.410	554	0.455	286	0.362	268	0.093	0.027
Good at Reading	0.722	554	0.675	286	0.772	268	-0.098	0.010
Rank Science 1	0.714	532	0.722	270	0.706	262	0.016	0.681
Guess Science Stream	0.442	559	0.476	288	0.406	271	0.070	0.097
Stereotype views								
Gender better at math (-1 to 1)	-0.220	549	-0.270	282	-0.169	267	-0.101	0.053
Gender better at science (-1 to 1)	-0.160	550	-0.236	284	-0.079	266	-0.157	0.003
Gender better at reading (-1 to 1)	0.376	553	0.320	284	0.435	269	-0.115	0.048
Socioeconomic status								
Father is college grad	0.451	552	0.483	286	0.417	266	0.065	0.124
Mother is college grad	0.367	551	0.384	284	0.348	267	0.035	0.388

Guess Rank ranges from 1 to 4 (1 is the best rank and 4 is the worst rank). Overconfidence is calculated as Actual-Guessed rank (actual rank based on modal rank in 1,000 simulations). Midterm scores are available for 4 schools, and are on a scale of 0-100. Attitudes and beliefs are based on dichotomized variables where 1=yes/agree and 0=no/disagree, except for "Gender better at" questions which are coded -1 (Boys are better) 0 (Both are equally as good) and 1 (Girls are better). Socioeconomic status are dichotomized variables for each parent holding at least a bachelor's degree. P-values are from Mann-Whitney *U* tests.

Table 4. Models for tournament entry (Competitiveness).

	Model 1	Model 2	Model 3	Model 4	Model 5
Female	-0.173**	-0.149**	-0.145**	-0.139**	-0.150*
	(0.048)	(0.047)	(0.044)	(0.047)	(0.056)
Num. Correct-Test 1	0.018**	0.026***	0.024***	0.026**	0.028**
	(0.005)	(0.006)	(0.006)	(0.007)	(0.008)
T-PR	0.009	0.018*	0.015*	0.019**	0.019*
	(0.008)	(0.007)	(0.006)	(0.006)	(0.007)
Overconfidence Test 2		0.060**	0.051*	0.053**	0.067*
		(0.017)	(0.018)	(0.018)	(0.022)
Nonincentivized risk			0.025+	0.024*	0.037**
			(0.012)	(0.011)	(0.011)
Incentivized risk			-0.004	-0.004	-0.020*
			(0.012)	(0.013)	(0.009)
Math stereotype				-0.013	-0.036
				(0.038)	(0.045)
Likes math				0.006	-0.010
				(0.037)	(0.050)
Thinks is good at math				-0.010	-0.018
				(0.044)	(0.058)
Expects science stream				0.035	0.074
				(0.042)	(0.042)
Father is college grad				-0.033	-0.056
				(0.060)	(0.055)
Mother is college grad				0.009	-0.017
				(0.059)	(0.070)
Midterm math score					-0.000
					(0.002)
Midterm overall score					-0.005
					(0.004)
Observations	561	560	558	524	409

All models provide OLS linear probability results that include class fixed effects. T-PR is the difference between number correct on the tournament (Test 2) versus piece-rate test (Test 1). Overconfidence Test 2 is measured as the difference between Actual and Guessed rank on Test 2. Nonincentivized risk is a scale from 0 to 10 (10 is most risky). Incentivized risk is the choice between a certain option or set of lotteries, ranging from 1 to 5 (5 is most risky). Robust standard errors are provided in parentheses. Significance levels are set at + $p<0.10$ * $p<0.05$ ** $p<0.01$ *** $p<0.001$.

Table 5. Number of correct answers, by treatment condition.

				Treatment Condition					
		Overall		Bottom Class		Top Class			
		Mean	N	Mean	N	Mean	N	Diff	p-value
Overall	Test 1	10.677	266	10.977	133	10.376	133	0.602	0.228
	Test 2	12.549	266	12.714	133	12.383	133	0.331	0.567
	Test 3	13.429	266	13.669	133	13.188	133	0.481	0.387
	Test 4	13.891	266	14.286	133	13.496	133	0.789	0.253
Males	Test 1	10.307	137	10.848	66	9.803	71	1.046	0.134
	Test 2	12.482	137	12.636	66	12.338	71	0.298	0.725
	Test 3	13.263	137	13.652	66	12.901	71	0.750	0.436
	Test 4	13.467	137	14.136	66	12.845	71	1.291	0.270
Females	Test 1	11.070	129	11.104	67	11.032	62	0.072	0.870
	Test 2	12.620	129	12.791	67	12.435	62	0.356	0.623
	Test 3	13.605	129	13.687	67	13.516	62	0.170	0.631
	Test 4	14.341	129	14.433	67	14.242	62	0.191	0.627

Analyses are limited to the sample of students in the middle classes. P-values are from Mann-Whitney *U* tests.

Table 6. Change in number of correct answers between Test 2 and Test 4 due to level of competition.

	All			Male			Female		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Versus top class	-1.029*	-1.043*	-0.815+	-0.996*	-1.008*	-0.893+	0.181	0.179	0.059
	(0.342)	(0.348)	(0.421)	(0.326)	(0.325)	(0.392)	(0.603)	(0.622)	(0.626)
Female	-0.002	-0.053	-0.112						
	(0.564)	(0.576)	(0.615)						
Female * Vs top class	1.184+	1.197+	0.864						
	(0.596)	(0.618)	(0.756)						
Competition		-0.448	-0.305		-0.380	-0.237		-0.722	-0.598
		(0.257)	(0.244)		(0.417)	(0.362)		(0.563)	(0.583)
Test 1			-0.047			-0.080+			0.016
			(0.035)			(0.037)			(0.049)
T-PR			-0.403***			-0.284**			-0.485***
			(0.045)			(0.066)			(0.079)
Observations	266	266	266	137	137	137	129	129	129

Analyses are limited to the sample of students in the middle classes. All models provide OLS linear probability results that include class fixed effects. Competition is the competition choice prior to Round 3. Test 1 is the number of correct answers on Test 1 (Piece-Rate). T-PR is the difference between number correct on the tournament (Test 2) versus piece-rate test (Test 1). Robust standard errors are provided in parentheses. Significance levels are set at: + p<0.10 * p<0.05 ** p<0.01 *** p<0.001.

Table 7. Chance of winning in Test 4 against top class, by school.

Questions	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**25**	**29**
School 1	0.2	0.6	1.2	1.4	5.7	10.8	12.4	17.5	30.5	50.1	71.3	91.5	93.2	-	-	-
School 2	0.5	0.5	1	1.1	3.1	-	20.3	23.4	39.2	55.6	-	-	81.5	-	-	-
School 3	0.3	0.3	-	-	-	3.4	-	-	24.5	-	53.8	-	-	-	-	-
School 4	1.7	-	21.5	-	-	53.1	66.7	-	83.1	-	-	-	-	-	-	-
School 5	0	0	0	0.8	1	2.4	3	7	9.4	13.7	23.7	39.4	-	52.6	61.8	100

Analyses only include the sample of students in the middle classes who face the top class. The chance of winning in Test 4 is the chance of getting 1st place in a group of 4 total competitors: the individual and 3 competitors from the top class at the same school (percentages are obtained by simulating 1,000 random draws of groups of 3 competitors from the top class for each individual).

Table 8. Change in number of answered questions between Test 2 and Test 4 due to level of competition.

	All			Male			Female		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Versus top class	-0.645*	-0.637*	-0.465	-0.616*	-0.610*	-0.448	-0.146	-0.146	-0.234
	(0.238)	(0.238)	(0.254)	(0.250)	(0.252)	(0.256)	(0.423)	(0.421)	(0.455)
Female	0.062	0.088	0.085						
	(0.369)	(0.373)	(0.400)						
Female * Vs top class	0.474	0.468	0.212						
	(0.404)	(0.393)	(0.539)						
Competition		0.223	0.174		0.182	0.165		0.268	0.060
		(0.224)	(0.210)		(0.173)	(0.226)		(0.480)	(0.461)
Test 1			0.029			-0.012			0.096*
			(0.032)			(0.036)			(0.038)
T-PR			-0.235***			-0.296***			-0.144+
			(0.036)			(0.042)			(0.070)
Observations	266	266	266	137	137	137	129	129	129

Analyses are limited to the sample of students in the middle classes. All models provide OLS linear probability results that include class fixed effects. Competition is the competition choice prior to Round 3. Test 1 is the number of total (incorrect + correct) answers on Test 1 (Piece-Rate). T-PR is the difference between number of total answers on the tournament (Test 2) versus piece-rate test (Test 1). Robust standard errors are provided in parentheses. Significance levels are set at: + p<0.10 * p<0.05 ** p<0.01 *** p<0.001.

Table 9. Change in accuracy between Test 2 and Test 4 due to level of competition.

	All			Male			Female		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Versus top class	-0.067+	-0.067+	-0.042*	-0.066+	-0.066+	-0.034+	0.020	0.019	0.008
	(0.035)	(0.035)	(0.015)	(0.034)	(0.034)	(0.015)	(0.021)	(0.022)	(0.016)
Female	-0.014	-0.017	-0.013						
	(0.033)	(0.033)	(0.022)						
Female * Vs top class	0.086+	0.086+	0.051*						
	(0.042)	(0.042)	(0.019)						
Competition		-0.023	-0.020		-0.004	-0.013		-0.062*	-0.037+
		(0.024)	(0.015)		(0.039)	(0.027)		(0.022)	(0.019)
Test 1			-0.519***			-0.592***			-0.566***
			(0.050)			(0.100)			(0.070)
T-PR			-0.875***			-1.021***			-0.785***
			(0.095)			(0.164)			(0.067)
Observations	266	266	266	137	137	137	129	129	129

Analyses are limited to the sample of students in the middle classes. All models provide OLS linear probability results that include class fixed effects. Competition is the competition choice prior to Round 3. Test 1 is the percentage of correct answers on Test 1 (Piece-Rate). T-PR is the difference between percentages of correct answers on the tournament (Test 2) versus piece-rate test (Test 1). Robust standard errors are provided in parentheses. Significance levels are set at: + p<0.10 * p<0.05 ** p<0.01 *** p<0.001.

Table 10. Confidence on Test 4 by treatment and gender.

		Overall		Male		Female			
Variable	**Treatment**	**Mean**	**N**	**Mean**	**N**	**Mean**	**N**	**Diff**	**p-value**
Actual Rank	Bottom class	1.541	133	1.500	66	1.582	67	-0.082	0.830
	Top class	2.932	133	2.944	71	2.919	62	0.024	0.761
Guessed Rank	Bottom class	1.962	133	1.758	66	2.164	67	-0.407	0.019
	Top class	3.000	133	2.930	71	3.081	62	-0.151	0.685
Overconfidence	Bottom class	-0.421	133	-0.258	66	-0.582	67	0.325	0.061
	Top class	-0.068	133	0.014	71	-0.161	62	0.175	0.579
Probability of win	Bottom class	55.024	133	56.114	66	53.951	67	2.162	0.601
	Top class	15.794	133	15.437	71	16.203	62	-0.767	0.474

Analyses are limited to the sample of students in the middle classes. Actual rank is based on modal rank on Test 4 based on 1,000 simulations. Guessed rank is from the survey question asking students to guess their rank. Overconfidence is the difference between Actual and Guessed rank. Probability of win is calculated as the percentage of wins (i.e. rank 1) based on the 1,000 simulations. P-values are from Mann-Whitney *U* tests.

Table 11. Change in chance of winning Test 4 due to level of competition.

	All			Male			Female		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Versus top class	-40.047***	-39.864***	-36.768**	-39.707***	-39.569***	-37.446***	-38.539***	-38.522***	-37.635***
	(5.087)	(5.168)	(6.887)	(5.105)	(5.153)	(6.726)	(5.871)	(5.721)	(4.387)
Female	-1.425	-0.776	-0.689						
	(4.320)	(4.377)	(4.060)						
Female * Vs top class	1.247	1.088	-1.011						
	(4.680)	(4.698)	(5.095)						
Competition		5.652*	-3.551		4.093	-2.671		8.357	-5.595
		(2.108)	(2.080)		(4.977)	(2.155)		(4.794)	(7.064)
Test 1			4.479***			4.078***			5.306***
			(0.263)			(0.341)			(0.441)
T-PR			2.185***			2.865**			1.816*
			(0.309)			(0.651)			(0.625)
Observations	266	266	266	137	137	137	129	129	129

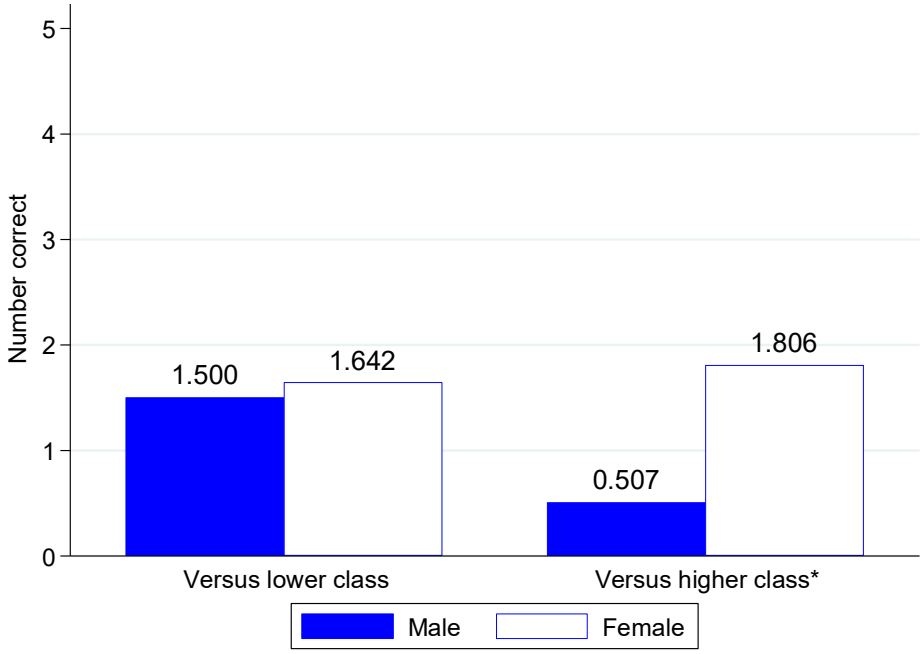
Analyses are limited to the sample of students in the middle classes. All models provide OLS linear probability results that include class fixed effects. The chance of winning in Test 4 is the chance of getting 1st place in a group of 4 total competitors: the individual and 3 competitors from the other class (percentages are obtained by simulating 1,000 random draws of groups of 3 competitors for each individual, by class). Competition is the competition choice prior to Round 3. Test 1 is the number of correct answers on Test 1 (Piece-Rate). T-PR is the difference between number correct on the tournament (Test 2) versus piece-rate test (Test 1). Robust standard errors are provided in parentheses. Significance levels are set at: + p<0.10 * p<0.05 ** p<0.01 *** p<0.001.

Table 12. Change in expected earnings in Test 4 due to level of competition.

	All			Male			Female		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Versus top class	-12.006***	-11.907***	-10.512**	-11.836***	-11.755***	-10.772**	-12.041***	-12.032***	-11.570***
	(1.573)	(1.632)	(2.225)	(1.560)	(1.584)	(2.115)	(2.153)	(2.118)	(1.393)
Female	-0.181	0.169	0.229						
	(1.913)	(1.921)	(1.785)						
Female * Vs top class	-0.151	-0.236	-1.125						
	(2.175)	(2.161)	(2.185)						
Competition		3.054*	-1.354		2.424	-0.828		4.260*	-2.409
		(1.080)	(0.965)		(2.362)	(1.165)		(1.357)	(2.857)
Test 1			2.138***			1.959***			2.515***
			(0.215)			(0.241)			(0.320)
T-PR			1.164**			1.430**			1.029*
			(0.270)			(0.407)			(0.340)
Observations	266	266	266	137	137	137	129	129	129

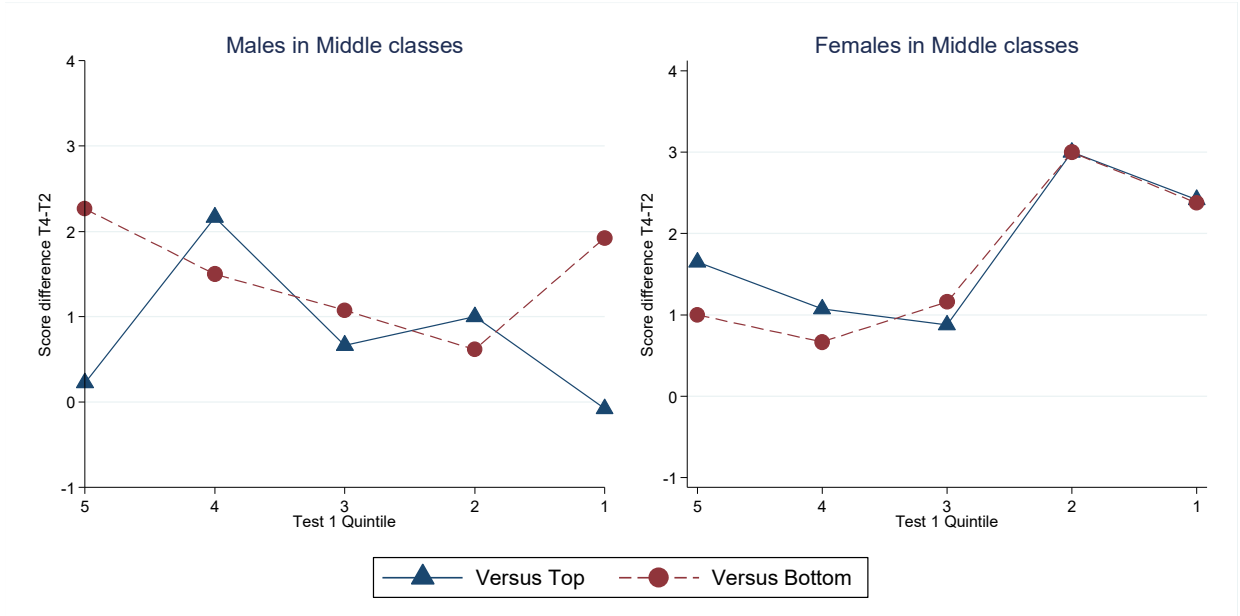
Analyses are limited to the sample of students in the middle classes. All models provide OLS linear probability results that include class fixed effects. The expected earnings in Test 4 is the chance of getting 1st place in the group of 4 multiplied by 2 (percentages are obtained by simulating 1,000 random draws of groups of 3 competitors for each individual, by class). Competition is the competition choice prior to Round 3. Test 1 is the number of correct answers on Test 1 (Piece-Rate). T-PR is the difference between number correct on the tournament (Test 2) versus piece-rate test (Test 1). Robust standard errors are provided in parentheses. Significance levels are set at: + p<0.10 * p<0.05 ** p<0.01 *** p<0.001.

Figure 1: Change in number of correct answers between Test 2 and Test 4, by treatment and gender



Note: Significance levels are set at: + p<0.10 * p<0.05 ** p<0.01 *** p<0.001.

Figure 2. Change in number of correct answers between Test 2 and Test 4, by treatment and gender and initial performance quintile



Note: Quintiles calculated within each class. 1 is best and 5 is worst.

A-1. Average difference in number of correct answers between tests.

		Overall		Male		Female	
	Class level	Diff	p-value	Diff	p-value	Diff	p-value
Test 1 to Test 2	All	1.900	<0.001	2.017	<0.001	1.775	<0.001
	Bottom	1.810	0.012	1.882	0.043	1.700	0.142
	Middle	1.872	<0.001	2.175	0.001	1.550	0.003
	Top	2.012	<0.001	1.87	0.037	2.130	0.002
Test 2 to Test 3	All	0.720	0.010	0.638	0.153	0.808	0.026
	Bottom	0.214	0.747	0.105	0.945	0.380	0.734
	Middle	0.880	0.017	0.781	0.204	0.984	0.031
	Top	0.846	0.055	0.909	0.217	0.793	0.148
Test 3 to Test 4	All	0.451	0.203	0.345	0.569	0.565	0.163
	Bottom	0.492	0.657	0.539	0.637	0.420	0.895
	Middle	0.462	0.308	0.204	0.783	0.736	0.192
	Top	0.402	0.483	0.403	0.751	0.402	0.409
Test 2 to Test 4	All	1.171	<0.001	0.983	0.045	1.373	0.001
	Bottom	0.706	0.451	0.645	0.565	0.800	0.599
	Middle	1.342	0.001	0.985	0.114	1.721	0.002
	Top	1.249	0.012	1.312	0.147	1.196	0.029

This table reports the differences in number correct. Number of observations are from the whole sample: 561 overall, with 290 males and 271 females overall. The gender breakdown is: 76 males and 50 females in the bottom classes; 137 males and 129 females in the middle classes; 77 males and 92 females in the top classes. P-values are from Mann-Whitney *U* tests for the difference in number correct between tests.

A-2. Student behavioral characteristics, by gender and class level.

		Overall			Male		Female			
Variable	**Class level**	**Mean**	**N**	**SD**	**Mean**	**N**	**Mean**	**N**	**Diff**	**p-value**
Guess Rank Test 2	Overall	2.573	560	0.941	2.441	290	2.715	270	-0.273	0.001
	Bottom	2.698	126	0.998	2.632	76	2.800	50	-0.168	0.502
	Middle	2.515	266	0.933	2.358	137	2.682	129	-0.325	0.005
	Top	2.571	168	0.906	2.403	77	2.714	91	-0.312	0.030
Guess Rank Test 4	Overall	2.455	560	1.072	2.36	289	2.557	271	-0.197	0.028
	Bottom	3.048	126	0.954	2.934	76	3.220	50	-0.286	0.088
	Middle	2.481	266	1.068	2.365	137	2.605	129	-0.240	0.072
	Top	1.970	168	0.925	1.776	76	2.130	92	-0.354	0.011
Overconfidence Test 2 (Actual-Guessed rank)	Overall	-0.221	560	1.159	-0.097	290	-0.356	270	0.259	0.016
	Bottom	-0.484	126	1.129	-0.539	76	-0.400	50	-0.139	0.627
	Middle	-0.139	266	1.185	0.051	137	-0.341	129	0.392	0.011
	Top	-0.155	168	1.116	0.078	77	-0.352	91	0.430	0.036
Overconfidence Test 4 (Actual-Guessed rank)	Overall	-0.136	560	1.136	0.042	289	-0.325	271	0.366	<0.001
	Bottom	0.310	126	1.196	0.408	76	0.160	50	0.248	0.180
	Middle	-0.244	266	1.128	-0.117	137	-0.380	129	0.263	0.078
	Top	-0.298	168	1.018	-0.039	76	-0.511	92	0.471	0.001
Incentivized risk scale (1-5; 5 most risky)	Overall	2.588	561	1.573	3.103	290	2.037	271	1.067	<0.001
	Bottom	2.317	126	1.505	2.763	76	1.640	50	1.123	0.001
	Middle	2.613	266	1.555	3.066	137	2.132	129	0.934	<0.001
	Top	2.751	169	1.632	3.506	77	2.120	92	1.387	<0.001
Non-incentivized risk scale (0-10; 10 most risky)	Overall	6.161	559	2.150	6.410	288	5.897	271	0.513	0.001
	Bottom	5.427	124	2.292	5.824	74	4.840	50	0.984	0.006
	Middle	6.308	266	2.049	6.489	137	6.116	129	0.373	0.048
	Top	6.467	169	2.087	6.831	77	6.163	92	0.668	0.028

Guess Rank ranges from 1 to 4 (1 is the best rank and 4 is the worst rank). Overconfidence is calculated as Actual-Guessed rank (actual rank based on modal rank in 1,000 simulations). P-values are from Mann-Whitney *U* tests.

A-3. Student midterm scores, by gender and class level.

		Overall			Male		Female			
Variable	**Class level**	**Mean**	**N**	**SD**	**Mean**	**N**	**Mean**	**N**	**Diff**	**p-value**
Math	Overall	49.842	463	23.379	49.457	230	50.223	233	-0.767	0.852
	Bottom	27.064	94	19.407	25.204	54	29.575	40	-4.371	0.363
	Middle	48.043	234	18.009	50.534	116	45.593	118	4.941	0.045
	Top	68.822	135	18.073	69.200	60	68.520	75	0.680	0.629
Malay	Overall	53.620	463	19.600	49.700	230	57.489	233	-7.789	<0.001
	Bottom	31.638	94	18.686	28.833	54	35.425	40	-6.592	0.138
	Middle	53.667	234	14.681	51.009	116	56.280	118	-5.271	0.004
	Top	68.844	135	11.615	65.950	60	71.160	75	-5.210	0.014
English	Overall	63.641	462	17.896	62.000	230	65.267	232	-3.267	0.125
	Bottom	41.462	93	16.731	39.259	54	44.513	39	-5.254	0.188
	Middle	63.889	234	12.823	64.483	116	63.305	118	1.178	0.264
	Top	78.489	135	7.753	77.667	60	79.147	75	-1.480	0.167
Overall	Overall	54.936	432	17.403	52.414	215	57.436	217	-5.022	0.005
	Bottom	32.739	94	12.150	30.973	54	35.123	40	-4.150	0.200
	Middle	54.637	203	11.165	54.089	101	55.179	102	-1.091	0.526
	Top	70.843	135	9.114	68.892	60	72.404	75	-3.512	0.060

Midterm scores are available for 4 schools, and are on a scale of 0 to 100. P-values are from Mann-Whitney *U* tests.

A-4. Student opinions and stereotypes, by gender and class level.

Variable	Class level	Overall			Male		Female			
		Mean	N	SD	Mean	N	Mean	N	Diff	p-value
Like Math	Overall	0.721	555	0.449	0.743	284	0.697	271	0.046	0.232
	Bottom	0.585	123	0.495	0.548	73	0.640	50	-0.092	0.311
	Middle	0.736	265	0.442	0.787	136	0.682	129	0.105	0.054
	Top	0.796	167	0.404	0.853	75	0.750	92	0.103	0.100
Like Science	Overall	0.770	556	0.421	0.812	287	0.725	269	0.087	0.015
	Bottom	0.637	124	0.483	0.635	74	0.640	50	-0.005	0.956
	Middle	0.795	264	0.404	0.853	136	0.734	128	0.119	0.017
	Top	0.827	168	0.379	0.909	77	0.758	91	0.151	0.010
Like Reading	Overall	0.752	537	0.432	0.647	275	0.863	262	-0.215	<0.001
	Bottom	0.648	122	0.480	0.514	72	0.840	50	-0.326	<0.001
	Middle	0.789	251	0.409	0.714	126	0.864	125	-0.150	0.004
	Top	0.774	164	0.419	0.662	77	0.874	87	-0.211	0.001
Good at Math	Overall	0.422	552	0.494	0.472	286	0.368	266	0.104	0.014
	Bottom	0.240	121	0.429	0.233	73	0.250	48	-0.017	0.830
	Middle	0.392	263	0.489	0.485	136	0.291	127	0.194	0.001
	Top	0.601	168	0.491	0.675	77	0.538	91	0.137	0.072
Good at Science	Overall	0.410	554	0.492	0.455	286	0.362	268	0.093	0.027
	Bottom	0.295	122	0.458	0.274	73	0.327	49	-0.053	0.534
	Middle	0.407	263	0.492	0.500	136	0.307	127	0.193	0.001
	Top	0.497	169	0.501	0.545	77	0.457	92	0.089	0.251
Good at Reading	Overall	0.722	554	0.448	0.675	286	0.772	268	-0.098	0.010
	Bottom	0.677	124	0.469	0.622	74	0.760	50	-0.138	0.107
	Middle	0.695	262	0.461	0.667	135	0.724	127	-0.058	0.311
	Top	0.798	168	0.403	0.740	77	0.846	91	-0.106	0.090
Rank Science 1	Overall	0.714	532	0.452	0.722	270	0.706	262	0.016	0.681
	Bottom	0.567	104	0.498	0.567	60	0.568	44	-0.002	0.988
	Middle	0.695	262	0.461	0.716	134	0.672	128	0.045	0.435
	Top	0.837	166	0.370	0.855	76	0.822	90	0.033	0.567
Guess Science Stream	Overall	0.442	559	0.497	0.476	288	0.406	271	0.070	0.097
	Bottom	0.208	125	0.408	0.267	75	0.120	50	0.147	0.049
	Middle	0.406	266	0.492	0.453	137	0.357	129	0.096	0.112
	Top	0.673	168	0.471	0.724	76	0.630	92	0.093	0.201

Gender better at math								
Overall	-0.220	549	-0.270	282	-0.169	267	-0.101	0.053
Bottom	-0.169	118	-0.157	70	-0.188	48	0.030	0.811
Middle	-0.229	262	-0.259	135	-0.197	127	-0.062	0.415
Top	-0.243	169	-0.390	77	-0.120	92	-0.270	0.003
Gender better at reading								
Overall	0.376	553	0.320	284	0.435	269	-0.115	0.048
Bottom	0.248	121	0.236	72	0.265	49	-0.029	0.974
Middle	0.420	264	0.378	135	0.465	129	-0.087	0.289
Top	0.399	168	0.299	77	0.484	91	-0.185	0.040
Gender better at science								
Overall	-0.160	550	-0.236	284	-0.079	266	-0.157	0.003
Bottom	-0.092	120	-0.194	72	0.063	48	-0.257	0.041
Middle	-0.206	262	-0.237	135	-0.173	127	-0.064	0.433
Top	-0.137	168	-0.273	77	-0.022	91	-0.251	0.007

Attitudes and beliefs are based on dichotomized variables where 1=yes/agree and 0=no/disagree, except for "Gender better at" questions which are coded -1 (Boys are better) 0 (Both are equally as good) and 1 (Girls are better). P-values are from Mann-Whitney *U* tests.

A-5. Descriptive statistics of student characteristics, by gender and class level.

		Overall			Male		Female			
Variable	**Class level**	**Mean**	**N**	**SD**	**Mean**	**N**	**Mean**	**N**	**Diff**	**p-value**
Female	Overall	0.483	561	0.500						
	Bottom	0.397	126	0.491						
	Middle	0.485	266	0.501						
	Top	0.544	169	0.500						
Father is college grad	Overall	0.451	552	0.498	0.483	286	0.417	266	0.065	0.124
	Bottom	0.276	123	0.449	0.297	74	0.245	49	0.052	0.526
	Middle	0.462	262	0.499	0.518	137	0.400	125	0.118	0.056
	Top	0.563	167	0.498	0.600	75	0.533	92	0.067	0.384
Mother is college grad	Overall	0.367	551	0.482	0.384	284	0.348	267	0.035	0.388
	Bottom	0.281	121	0.451	0.292	72	0.265	49	0.026	0.752
	Middle	0.341	264	0.475	0.372	137	0.307	127	0.065	0.265
	Top	0.470	166	0.501	0.493	75	0.451	91	0.043	0.584

P-values are based on Mann-Whitney *U* tests.

A-6. Models for tournament entry (Competitiveness), excluding school without administrative records.

	Model 1	Model 2	Model 3	Model 4	Model 5
Female	-0.170**	-0.144*	-0.136*	-0.142*	-0.150*
	(0.048)	(0.048)	(0.047)	(0.048)	(0.056)
Num. Correct Test 1	0.018*	0.026**	0.023**	0.027**	0.028**
	(0.006)	(0.007)	(0.007)	(0.008)	(0.008)
T-PR	0.007	0.018*	0.014+	0.019*	0.019*
	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)
Overconfidence Test 2		0.064**	0.053*	0.058*	0.067*
		(0.020)	(0.021)	(0.020)	(0.022)
Nonincentivized risk			0.038**	0.037**	0.037**
			(0.010)	(0.010)	(0.011)
Incentivized risk			-0.005	-0.010	-0.020*
			(0.012)	(0.013)	(0.009)
Math stereotype				-0.038	-0.036
				(0.043)	(0.045)
Likes math				-0.009	-0.010
				(0.046)	(0.050)
Thinks is good at math				-0.036	-0.018
				(0.051)	(0.058)
Expects science stream				0.069	0.074
				(0.042)	(0.042)
Father is college grad				-0.084	-0.056
				(0.058)	(0.055)
Mother is college grad				0.003	-0.017
				(0.070)	(0.070)
Midterm math score					-0.000
					(0.002)
Midterm overall score					-0.005
					(0.004)
Observations	464	463	462	439	409

All models provide OLS linear probability results that include session fixed effects (13 session vs 18 classes). T-PR is the difference between number correct on the tournament (Test 2) versus piece-rate test (Test 1). Overconfidence Test 2 is measured as the difference between Actual and Guessed rank on Test 2. Nonincentivized risk is a scale from 0 to 10 (10 is most risky). Incentivized risk is the choice between a certain option or set of lotteries, ranging from 1 to 5 (5 is most risky). Robust standard errors are provided in parentheses. Significance levels are set at + p<0.10 * p<0.05 ** p<0.01 *** p<0.001.

A-7. Models for tournament entry (Competitiveness), clustered by session.

	Model 1	Model 2	Model 3	Model 4	Model 5
Female	-0.174**	-0.154**	-0.153**	-0.152**	-0.151*
	(0.047)	(0.047)	(0.042)	(0.041)	(0.058)
Num. Correct Test 1	0.017***	0.022***	0.020***	0.022***	0.028***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)
T-PR	0.008	0.015+	0.012	0.016+	0.018+
	(0.010)	(0.008)	(0.007)	(0.009)	(0.009)
Overconfidence Test 2		0.051**	0.042*	0.046**	0.062**
		(0.013)	(0.017)	(0.014)	(0.018)
Nonincentivized risk			0.025+	0.024+	0.037*
			(0.014)	(0.013)	(0.012)
Incentivized risk			-0.007	-0.007	-0.023*
			(0.012)	(0.014)	(0.008)
Math stereotype				-0.000	-0.026
				(0.025)	(0.029)
Likes math				-0.004	-0.021
				(0.024)	(0.033)
Thinks is good at math				-0.005	-0.003
				(0.057)	(0.058)
Expects science stream				-0.002	0.058
				(0.032)	(0.034)
Father is college grad				-0.041	-0.061
				(0.084)	(0.064)
Mother is college grad				0.015	-0.004
				(0.050)	(0.059)
Midterm math score					0.000
					(0.001)
Midterm overall score					-0.006*
					(0.002)
Observations	561	560	558	524	409

All models provide OLS linear probability results that include session fixed effects (13 session vs 18 classes). T-PR is the difference between number correct on the tournament (Test 2) versus piece-rate test (Test 1). Overconfidence Test 2 is measured as the difference between Actual and Guessed rank on Test 2. Nonincentivized risk is a scale from 0 to 10 (10 is most risky). Incentivized risk is the choice between a certain option or set of lotteries, ranging from 1 to 5 (5 is most risky). Robust standard errors are provided in parentheses. Significance levels are set at + $p<0.10$ * $p<0.05$ ** $p<0.01$ *** $p<0.001$.

A-8. Balance check of covariates for middle classes.

	All			Males			Females		
Variable	**Coeff**	**SE**	**Obs**	**Coeff**	**SE**	**Obs**	**Coeff**	**SE**	**Obs**
Female	-0.038	0.033	266						
Math midterm score	0.282	2.399	234	-3.071	2.100	116	3.563	2.991	118
Overall midterm score	0.610	1.320	203	-0.103	1.762	101	1.530	1.545	102
Test 1 (Piece-Rate)	-0.626+	0.290	266	-1.058	0.651	137	-0.086	0.570	129
Test 2 (Tournament)	-0.373	0.359	266	-0.372	0.532	137	-0.338	0.394	129
Test 3	-0.513	0.539	266	-0.793	0.518	137	-0.026	0.797	129
Tournament-Piece Rate	0.252	0.337	266	0.686	0.409	137	-0.251	0.477	129
Competition choice	-0.015	0.031	266	-0.034	0.050	137	-0.002	0.054	129

Analyses are limited to the sample of students in the middle classes. This table presents results of regressions of the covariates on treatment, for the overall sample and then by gender. Each row represents a regression. All regressions use class fixed effects. Robust standard errors are provided in parentheses. Significance levels are set at + p<0.10 * p<0.05 ** p<0.01 *** p<0.001

A-9. Change in number of correct answers between Test 2 and Test 4 by treatment in middle classes.

	Overall			Males		Females			
Class Level	**Mean**	**N**	**SD**	**Mean**	**N**	**Mean**	**N**	**Diff**	**p-value**
Versus lower	1.571	133	2.827	1.500	66	1.642	67	-0.142	0.849
Versus higher	1.113	133	2.972	0.507	71	1.806	62	-1.299	0.017

Analyses are limited to the sample of students in the middle classes. Differences are calculated by individual (Test 4-Test 2). P-values are from Mann-Whitney *U* tests.

A-10. Change in number of correct answers between Test 2 and Test 4 due to level of competition, using whole school sample.

$$y_{ij} = \Gamma_j + \beta_1 Vshigher_{ij} + \beta_2 Female_{ij} + \beta_3 (Vshigher * Female)_{ij} + \theta \mathrm{X_{ij}} + \epsilon_{ij}$$

y_{ij} = Difference in Number of Correct Answers between Other and Own class (Test 4 - Test 2) for student i in class j.
Γ_j is the class fixed effects.
$Vshigher$ is 1 if assigned higher class and 0 if assigned lower class for student i in class j. This means that for all the bottom classes & half of middle classes, Treatment=1.
Female is 1 if female and 0 if male for student i in class j.
*Female * Vshigher* is 1 if subject is assigned to higher class & is female; 0 otherwise.
$\mathrm{X_{ij}}$ is vector of student attributes.

	All			**Male**			**Female**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Vs higher	-0.793*	-0.794*	-0.571	-0.996**	-0.991**	-0.814*	0.181	0.180	0.060
	(0.373)	(0.376)	(0.373)	(0.313)	(0.311)	(0.381)	(0.579)	(0.590)	(0.584)
Female	-0.083	-0.088	-0.054						
	(0.328)	(0.337)	(0.363)						
Female * Vs higher	0.672	0.674	0.356						
	(0.438)	(0.441)	(0.481)						
Competition		-0.021	-0.014		0.143	0.120		-0.473	-0.227
		(0.280)	(0.237)		(0.377)	(0.328)		(0.437)	(0.400)
Test 1			-0.032			-0.030			-0.064
			(0.021)			(0.031)			(0.042)
T-PR			-0.374***			-0.305***			-0.456***
			(0.041)			(0.047)			(0.066)
Observations	561	561	561	290	290	290	271	271	271

All models provide OLS linear probability results that include class fixed effects. The whole school sample is used; thus those who received the treatment "Vshigher" are half the students in the middle classes and all the students in the bottom classes, which is not random. Competition is the competition choice prior to Round 3. Test 1 is the number of correct answers in Test 1 (Piece-Rate). T-PR is the difference between number correct on the tournament (Test 2) versus piece-rate test (Test 1). Robust standard errors are provided in parentheses. Significance levels are set as: + p<0.10 * p<0.05 ** p<0.01 *** p<0.001.

A-11. Change in number of correct answers between Test 2 and Test 4 due to level of competition, clustered by session.

	All			Male			Female		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Versus top class	-1.032*	-1.046*	-0.816	-1.013*	-1.025*	-0.896+	0.177	0.175	0.051
	(0.417)	(0.419)	(0.464)	(0.404)	(0.402)	(0.444)	(0.674)	(0.693)	(0.698)
Female	-0.005	-0.056	-0.116						
	(0.544)	(0.550)	(0.612)						
Female * Vs top class	1.189+	1.201	0.866						
	(0.602)	(0.625)	(0.760)						
Competition		-0.452	-0.317		-0.341	-0.223		-0.645	-0.488
		(0.240)	(0.227)		(0.419)	(0.366)		(0.549)	(0.572)
Test 1			-0.043			-0.064			0.000
			(0.030)			(0.035)			(0.040)
T-PR			-0.401***			-0.282**			-0.496***
			(0.044)			(0.053)			(0.082)
Observations	266	266	266	137	137	137	129	129	129

Analyses are limited to the sample of students in the middle classes. All models provide OLS linear probability results that include session fixed effects (7 sessions vs 8 classes). Competition is the competition choice prior to Round 3. Test 1 is the number of correct answers on Test 1 (Piece-Rate). T-PR is the difference between number correct on the tournament (Test 2) versus piece-rate test (Test 1). Robust standard errors are provided in parentheses. Significance levels are set at: + p<0.10 * p<0.05 ** p<0.01 *** p<0.001.

A-12. Number of correct answers on Test 4 due to level of competition.

	All			Male			Female		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Versus top class	-1.438*	-1.384+	-0.815+	-1.368+	-1.325+	-0.893+	-0.157	-0.153	0.059
	(0.596)	(0.616)	(0.421)	(0.590)	(0.605)	(0.392)	(0.715)	(0.694)	(0.626)
Female	-0.343	-0.151	-0.112						
	(0.738)	(0.763)	(0.615)						
Female * Vs top class	1.234	1.187	0.864						
	(0.826)	(0.750)	(0.756)						
Competition		1.671*	-0.305		1.293	-0.237		2.132**	-0.598
		(0.580)	(0.244)		(0.975)	(0.362)		(0.459)	(0.583)
Test 2			0.953***			0.920***			1.016***
			(0.035)			(0.037)			(0.049)
T-PR			-0.356***			-0.205*			-0.502***
			(0.043)			(0.065)			(0.091)
Observations	266	266	266	137	137	137	129	129	129

Analyses are limited to the sample of students in the middle classes. All models provide OLS linear probability results that include class fixed effects. Competition is the competition choice prior to Round 3. Test 2 is the number of correct answers in Test 2 (Tournament). T-PR is the difference between number correct on the tournament (Test 2) versus piece-rate test (Test 1). Robust standard errors are provided in parentheses. Significance levels are set at: + p<0.10 * p<0.05 ** p<0.01 *** p<0.001.

A-13. Change in number of correct answers between Test 2 and Test 4 due to level of competition, controlling for chance of winning.

	All			Male			Female		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Versus top class	-1.029*	-1.043*	-0.971*	-0.996*	-1.008*	-0.993*	0.181	0.179	0.100
	(0.342)	(0.348)	(0.393)	(0.326)	(0.325)	(0.364)	(0.603)	(0.622)	(0.680)
Female	-0.002	-0.053	-0.049						
	(0.564)	(0.576)	(0.595)						
Female * Vs top class	1.184+	1.197+	1.055						
	(0.596)	(0.618)	(0.780)						
Competition		-0.448	-0.279		-0.380	-0.168		-0.722	-0.808
		(0.257)	(0.280)		(0.417)	(0.426)		(0.563)	(0.516)
Chance win T1			0.000			-0.008			0.010
			(0.007)			(0.007)			(0.011)
Chance T2-T1			-0.032**			-0.019			-0.037*
			(0.009)			(0.012)			(0.013)
Observations	266	266	266	137	137	137	129	129	129

Analyses are limited to the sample of students in the middle classes. All models provide OLS linear probability results that include class fixed effects. Competition is the competition choice prior to Round 3. Chance winning T1 is the chance of getting 1st place if Test 1 were a tournament with groups of 4 competitors (percentages obtained by simulating 1,000 draws of groups of 3 competitors for each individual, by class). Chance T2-T1 is the difference in the chances of winning in Test 1 and Test 2. Robust standard errors are provided in parentheses. Significance levels are set at: + p<0.10 * p<0.05 ** p<0.01 *** p<0.001.