The Gender Gap Cracks Under Pressure: A Detailed Look at Male and Female Performance Differences During Competitions

Christopher Cotton
Frank McIntyre
Joseph Price*

July 19, 2010

Abstract

Using data from multiple-period math competitions, we show that males outperform females of similar ability during the first period. However, the male advantage is not found in any subsequent period of competition, or even after a two-week break from competition. Some evidence suggests that males may actually perform worse than females in later periods. The analysis considers various experimental treatments and finds that the existence of gender differences depends crucially on the design of the competition and the task at hand. Even when the male advantage does exist, it does not persist beyond the initial period of competition. (*JEL* J16, J24)

Keywords: competitiveness; gender differences; field experiment

*Cotton: Department of Economics, University of Miami, Coral Gables, Florida 33146; cotton@miami.edu. McIntyre and Price: Department of Economics, Brigham Young University, Provo, Utah 84602; mcintyre@byu.edu, joe_price@byu.edu.

1 Introduction

In their influential paper, Gneezy et al. (2003) show that males and females respond differently to competition. The authors conduct experiments in which college students are paid to solve mazes, either on their own or in competition with others. They show that competition causes males to increase their performance relative to females. Gneezy and Rustichini (2004) study footraces between fourth graders and find a similar result: boys respond favorably to competition, while girls do not. It has been argued that these results help explain the gender gap in achievement and pay in the work place, higher education, and other settings.

Our paper works to provide a better understanding of gender differences in response to competition. Its main goal is to determine how well the earlier results hold up against changes in experiment design and participant experience. At the heart of our analysis are a series of in-classroom experiments that we conducted with 505 primary school students. In each of the 24 classrooms that participated, students were randomly paired with an opponent and then competed against their opponent to complete an age-appropriate math quiz as quickly and accurately as possible. We then repeated the process, rematching opponents and assigning a new set of questions. Each classroom participated in up to five sequential rounds of competition, resulting in 2171 total individual-period level observations.

Our experimental approach has four primary advantages over past analyses. First, where past experiments identify gender differences in a single competitive interaction, we observe how the results change over sequential periods of competition as participants warm up and gain experience. Second, we are able to run a variety of treatments in which we vary the design of the competition. Third, we focus on math competitions, which are academic in nature and clearly relate to one's ability to succeed in academic and professional settings. Fourth, we have data on state assessment test scores for our participants, giving us a formal measure of ability that is absent from earlier analyses. Since the questions used in our competition also come from the state assessments, we can directly compare how a given student's behavior changes as the environment changes. Not only are test scores an outcome of substantial interest in the education literature, they also are a relevant input to explaining why males and females with similar academic qualifications may experience different levels of success in a competitive workplace.

In the most common of our experimental treatments—the "race treatment"—the math competitions were framed as races. Participants were told to complete as many questions as possible before the five-minute time limit. The winner was the one who solved the most questions. If someone finished the quiz before the time limit, then the quicker time won in the event of a tie. The race results provide significant insight into the gender differences

in reaction to competition. In the first period of competition, males perform significantly better than females of the same ability. This result is consistent with the previous literature (e.g., Gneezy et al. 2003, Gneezy and Rustichini 2004), which is not surprising since the first period of competition in our experiments closely resembles the one-time competitions in past papers. The main contribution of the paper comes when we look at later periods of competition. In the second period of competition onward, we find absolutely no evidence that males perform better than females of similar ability. Although gender has a significant effect on performance in the initial period of competition, the male advantage vanishes almost immediately. In fact, we find some evidence that males perform *worse* than females in later periods.

A closer look at performance in the math races reveals some additional insight into the results. First, we show that the initial male advantage is more likely due to males overperforming in the first period relative to their trend in later rounds, rather than due to female underperformance. Second, we subdivide the race treatment data by quiz length to analyze separately short-quiz competitions in which many participants finished the quiz early, and long-quiz competitions in which few participants submitted answers before the time limit. All subdivisions of the data told the same story: an initial male advantage disappears before the second period of competition.

After presenting the results for the race treatment, we present three smaller experimental treatments designed to provide a better understanding of the gender differences. The "not-a-race treatment" was identical to the race treatment except that the competition was not framed as a race. That is, participants did not benefit from submitting answers before the time limit, and they were explicitly told that the competition was "not a race." It remained clear to the participants, however, that they were competing against their assigned opponent to answer the most questions correctly. The "reading-race treatment" was identical to the race treatment, except that each quiz was made up of reading (i.e., language arts) questions rather than math questions. Finally, for the "repeat treatment" we returned to some of the race treatment classrooms two weeks after first running the experiments and reran the same experiment with new questions.

In these alternative treatments we don't even find evidence of a first-period male advantage. In both the reading-race and not-a-race treatments, we find no evidence of the gender effect in any period of competition, including the first. This suggests that both the task at hand and the design of the competition affect the initial gender effect. The result for the reading treatment is merely suggestive, as we do not have the same ability measures as we do for the math treatments. The result for the not-a-race treatment is relatively robust, suggesting that males initially react more positively than females to competition when they

view the competition as a race, but not otherwise. Taken together, these alternative treatments show that the existence of the initial gender effect depends crucially on the nature of the competition. Furthermore, in the treatments where we find an initial male advantage, the advantage is always short-lived, vanishing after the first period.

The repeat treatment speaks to the importance of the results for the literature. If once the male advantage vanishes it is gone forever, then there is little reason to think that the initial advantage could have any substantial effect on long-run achievement. On the other hand, if the male advantage reappears at the beginning of every new competition, then it may still help explain long-run achievement differences between males and females in competitive environments. When we return to the classrooms two weeks after first conducting the experiment and re-run the competitions, we find no significant evidence of a male advantage in any period of competition including the first. Although the relatively small sample size in this treatment prevents us from being certain of these results, the evidence points to the male advantage not returning even weeks after the initial competition.

Throughout the paper, the evidence calls into question the claim that performance differences in competition may explain long-run achievement differences between males and females. We show that the existence of the male advantage is highly dependent on the type of competition. It exists in math races, but does not exist if the races involve reading questions, or if the competition is not framed as a race. Very few competitive interactions in the workplace are viewed as races. Furthermore, the male advantage (when it does exist) vanishes completely after the initial period of competition and we see no evidence that it returns even after a two-week break from competition. Because of this, we see little reason to believe that the male advantage, identified in the first period of our math races and in other one-time competitions, could drive significant differences in long-run achievement. However, the same behavioral trait that causes males to increase their initial performance during a competition—whether it is increased initial excitement about competing or an initial increase in testosterone in the face of competition—may also make it more likely that males choose to compete in the first place. This alternative explanation of long-run achievement differences is consistent with Niederle and Vesterlund (2007), who show that males are more likely to enter competition compared with similarly able females.

2 Literature Review

A number of past articles assess whether males and females react differently to competition. Gneezy et al. (2003) run a series of experiments on college students and show that males respond more favorably than females to competition when solving mazes. Gneezy and Rus-

tichini (2004) produce similar results when they have primary school students run footraces. Günther et al. (in press, 2010) also identify a male advantage in maze competitions, but then find no significant difference between male and female performance in competitions involving word games. In these papers' experiments, the competitions lasted only one period, and were framed as races. Our results are consistent with these earlier findings. We identify a male advantage in the first period of the math races, but do not find any evidence of a male advantage in the first round of our reading races. In none of these earlier experiments did competitors participate in more than one competition.

Another branch of this literature considers the impact of opponent gender on performance. Inzlicht and Ben-Zeev (2000) find evidence that one's performance on difficult verbal and math tests may depend on the gender composition of the group of people sitting in close proximity, even when they are not directly competing. Antonovics et al. (2009) showed that males were more likely to answer trivia questions correctly when a larger fraction of their competitors were female. Price (2008) shows that competitive funding can affect time to candidacy in graduate school, with both males and females responding more positively to the competition when more of their peers are female. In the present analysis, we focus on whether males perform better than females in competitive environments; we are less concerned with whether performance depends on the opponent's gender. We do, however, find suggestive evidence that males perform slightly better when competing against females.

Compared to much of the previous literature, our analysis benefits from having access to state assessment test scores. Most other papers that measure gender differences randomly divide their subject pool into two groups for each gender. One of the groups race to complete a task, and the other group completes the same task in a non-competitive setting. The analyses then compare the distribution of performance in each group to see if there are significant differences. In our analysis, we are able to directly control for each participant's past performance on the math section of state assessment tests. Furthermore, our focus on math competitions lends real world appeal to the analysis, as math ability has a significant affect on career success (Joensen and Nielsen (2009)).

Other articles consider whether males and females have different preferences for competition. Niederle and Vesterlund (2007) and Wozniak (2009) show that, given a choice, males are more likely to compete than females. Similarly, Sapienza et al. (2009) claim that at Northwestern University, 36% of female versus 57% of male MBA students choose competitive finance careers.[1] We find these results particularly interesting in the light of our

[1]Booth and Nolen (2009) and Gneezy et al. (2009) show that preferences towards competition may be due to past exposure and experience. Specifically, Booth and Nolen (2009) show that females who attend all girls schools are more likely to choose competition (even competition against males) than are females in coed schools. Gneezy et al. (2009) show that in a matrilineal society, women prefer competition more than

analysis. Although we find no evidence of long-run performance differences between males and females within competition, males do perform better in the very first period of competition. This initial performance boost experienced by the males may be related to their eagerness to compete in the first place.

3 Research Design

3.1 Experiments and Test Score Data

Working with school officials and teachers, we went into 24 elementary school classrooms to run a series of in-classroom, curriculum-based competitions. In each period of competition, student participants were randomly paired with another classmate. The students were given a quiz with questions selected from past state-assessment tests for the appropriate grade level. They had five minutes to answer as many of the questions as they could. At the end of five minutes, a winner was determined for each pair of students, and the winner received two raffle tickets. If a pair of students tied, each received one raffle ticket. In each classroom, we repeated this process up to five times, each time pairing students with new opponents. There was a minimum amount of delay between each competition, with the entire activity taking about an hour. At the end of the final competition, we randomly selected three raffle tickets, and the students who won the raffle in each classroom received a candy bar or other prize.

Our tournaments took place a few weeks before the state assessment tests and were used as a way of preparing the students for those tests. In total, we worked with 657 elementary school students, including 86 third graders, 297 fourth graders, and 122 sixth graders. Twenty-one of our classrooms participated in five periods of competition. Due to time constraints, one classroom participated in only four periods, and two classrooms participated in three periods. We returned to four of the classrooms two weeks after first conducting the experiment and reran the competitions a second time. 239 of the participants are female, and 266 are male. The school districts in which we conducted the experiments had little ethnic or racial diversity (approximately 90 percent of the area residents are white, non-Hispanic), so we are not concerned about the racial mix of the competitions. In total, we have 2171 observations.

In addition to the data collected during the experiments, the school district provided us with the previous year's state assessment test scores for each of the participants in our math

males. Kleinjans (2009) presents evidence that differences in taste for competition may help explain some of the sorting of males and females into different professions.

competitions. The state where the school district is located uses its own criterion-referenced test for its end-of-year assessment. All students are required to participate, and the tests are not timed. Access to this test score data is one of the primary advantages our analyses has over past competition experiments.

We first describe the race treatment with which we begin the analysis. Then, we describe three alternative treatments that we use later in the analysis.

Race Treatment

After being introduced to the students by the teacher, we read the students the rules of the competition. The description was thorough, informing them of the number of rounds of competition, the number and origins of questions, procedure for determining their opponent each period, the rules for determining each period's winner, and the raffle and prize structure. After explaining the rules, we answered any questions about rules, prizes, or procedure.

In each period of competition, the students were randomly assigned an opponent, then raced against their opponent to complete as many math questions as possible within a five-minute time limit. The questions in the standard treatment were selected from previous year state assessment tests for math, and each period's quiz consisted of 5, 10, or 15 questions. Within each classroom, the length of the quiz was held constant. The questions given in each round were randomly selected and differed between classrooms, but were the same for all students within a class. Five sample questions for fourth-grade participants are provided in the appendix. In each two-person competition, the participant who answered the most questions within the time limit won. Participants had the option to submit their answers before the time limit. If both competitors had the same number of correct answers, then the one who submitted his or her answers first won. The winner received two raffle tickets. If both had the same number correct and no one submitted answers early, then the participants tied, and each received one raffle ticket. After the final period of competition, three raffle tickets were randomly selected and the students with those tickets received a candy bar or other prize.

Eight of the classrooms in the math races involved the participants moving to sit next to their opponent in each round. In the four other classrooms, participants stayed in their own desk the whole time but were told who they were competing against in each round. We find no evidence that participants performed differently based on whether or not they were sitting next to their opponent.

In total, the race treatment was conducted in 12 classrooms with a total of 253 students. The quizzes were five questions long in four classrooms, 15 questions long in three classrooms, and 10 questions long in five of the classrooms. Section 4.2 considers the impact, if any, that

the quiz length has on the results. To be able to compare performance across treatments and quiz lengths, the scores are normalized by round and class to be mean zero and standard deviation one. We present the average normalized and unnormalized scores for all treatments in Tables 1 and 2.

Alternative Treatments

Following the results for the race treatment, we present results from three smaller experimental treatments. First, a not-a-race treatment was identical to the race treatment with one major exception: we did not frame the competition as a race. Although participants knew they were competing to perform better than their opponent on a quiz, we never referred to the competition as a race. Consistent with this change, we also eliminated the faster-finish tie-breaking rule; participants could finish their quizzes early, but doing so did not provide a competitive advantage. In total, we conducted the not-a-race experiment in six classrooms with a total of 122 students. Second, a reading-race treatment was identical to the race treatment except that each period's quiz was comprised of questions selected form the reading section of the state assessment test. In total, we conducted the reading treatment in six classrooms with a total of 130 students. Both the not-a-race and reading treatments involved 10-question quizzes. Third, a repeat treatment involved returning to four of the five-question math race classrooms two weeks after first conducting the experiments and re-running the competitions with new questions and new opponent matching.

The presentation of our results follows the same outline as our investigation. We first present results from the race treatment. Then, we consider the alternative treatments, and what they can tell us about gender differences in response to competition.

3.2 Empirical Strategy

Let $y_{i,r}$ be student i's score in round r, where all scores are normalized by round and class to be mean 0 and standard deviation 1.[2] We run regressions of the following form:

$$y_{i,r} = \alpha + \beta\theta_{i,r} + \delta G_{i,-i,r} + \epsilon_{i,r},$$

where $\theta_{i,r}$ is a measure of the student's innate ability and $G_{i,-i,r}$ is a scalar or vector of dummy variables that captures gender effects based on the student's gender and potentially

[2]We normalize scores for the regressions to make results easily comparable to the education research on tests scores and to remove potential classroom idiosyncrasies. Regressions using percentage correct produce the same substantive results.

that of his or her opponent. In the simplest specification, $G_{i,-i,r}$ is a dummy variable for boys. We also consider specifications that interact gender and opponent gender.

We observe two variables related to participant ability, $\theta_{i,r}$. The first is simply the average performance of the student in the other rounds of competition, which we can write as $\bar{y}_{i,-r}$. On its own, $\bar{y}_{i,-r}$ would be problematic for two reasons. First, $\bar{y}_{i,-r}$ has a fair bit of measurement error in it, biasing β down and thus meaning we would have incorrectly controlled for innate ability. Second, student performance in the other rounds will vary both because of students' innate ability and due to their competitive ability or preferences. In which case it will control for competitive differences across gender, biasing δ towards zero. We deal with both these problems with our second ability measure—the student's score on the prior year's state assessment. Since we use state assessment questions in our competitions, these are an ideal measure of how the student performs on the same material, but in a relatively non-competitive environment.

Because measurement error in the two variables is uncorrelated, we use the state assessment score as an instrument for the student's average performance in the other rounds. The state assessment score only contains variation in ability from a non-competitive environment, so it is a valid instrument for innate ability uncontaminated by direct competitive pressure. We use the average in-competition score to purge any measurement error concerns. β then consistently estimates how innate ability helps students do better in competitive environments.[3]

If there were no measurement error in the state assessment, we could use OLS regressions with the state assessment as the sole measure of ability, $\theta_{i,r}$. Since this reduced form is a more transparent methodology, Section 4.6 presents the results for the OLS analysis, which provides similar results as the IV analysis.

4 Results

4.1 Race Treatment

Table 3 presents the results for the race treatment using state assessment scores as the instrument for ability as described above. Regressions (1) and (2) use data from the first period of competition; (1) controls for the participant's own gender, while (2) also controls for opponent gender. Regressions (3) and (4) do the same for the second period of competition,

[3]In unreported results we allowed β to vary by gender in our baseline race treatment. We saw no evidence that such an interaction was important as the estimated β was the same for males and females. Furthermore, allowing the interaction had no effect on our point estimates or standard errors for δ.

and regressions (5) and (6) do the same for the combined data from periods two through five. (Tables 1 and 2 report on average scores by round for all rounds.)

4.1.1 First Period of Competition

We first identify significant gender differences in performance during the first period of competition. This gender gap is consistent with the literature that looks at one-period competitions.

In Table 3, the first period results clearly show that in competition, males perform significantly better than females of similar ability. In the first regression, the male coefficient is a highly significant 0.34. This means that if we take a male and female with identical past test scores, we expect the male to score 0.34 standard deviations higher than the female in the first period of our math competitions. The male coefficient is significantly different from zero, with p-value of 0.004.

When controlling for opponent gender in regression (2), it remains clear that males perform better during the first round of competition compared with similarly able females. Opponent gender, however, does not have a significant impact on performance. Both male coefficients (i.e., MvM and MvF) are significantly different from the FvM baseline (p-values of 0.01 and 0.03), and MvM is also significantly different from FvF at the ten percent level (p-value = 0.06). Although we cannot reject equality between the MvF and FvF coefficients (p-value=0.17), the coefficient is consistent with a male advantage. Males may perform somewhat better when competing against other males, but this difference is far from significant (p-value = 0.56). Similarly, females may perform slightly better when competing against other females, but the difference is also not significant (p-value = 0.59). Therefore, although females perform significantly worse than males, there is no evidence that first-period performance is influenced by the gender of one's opponent.

These results are consistent with the main findings in past research: in the first period of competition, males perform significantly better than females of similar ability.

4.1.2 Multiple Periods of Competition

After the first round of competition, the male advantage disappears. In the second period of competition and in all subsequent periods, we find no evidence that males perform better than females of similar ability. Even more surprisingly, in later periods of competition, males may actually perform *worse* than females of similar ability.

Regressions (3) and (4) in Table 3 provide results for the second period of competition. Comparing regression (3) to regression (1), a few things are apparent. Most importantly, the

male advantage vanishes by the second period of competition, with the second-period male coefficient equal to -0.01. Equality between the first- and second-period male coefficients is rejected with a p-value of 0.05. Additionally, the ability coefficient and the R-squared are both larger in the second-period regression. That is, in the second period of competition, gender no longer matters and ability becomes a better predictor of performance.[4] When the second period analysis controls for opponent gender in regression (4), the results are similar.[5]

Later periods of competition look a lot like the second period of competition. Instead of presenting separate results for each period of competition, we pool the data from the later periods of competition in regressions (5) and (6). The results when we pool the data are similar to the results from the second period by itself. They suggest that males perform significantly worse than females in later rounds of competition. After the first round of competition, males tend to perform 0.10 standard deviations worse than females of similar ability, which is not only statistically different from zero (p-value = 0.014) but also from the first round gender effect (p-value = 0.007)

A few other patterns in the multiple round data are worth mentioning. Table 2 shows that female performance drops slightly between the first and second periods of competition, then improves over the later rounds of competition. Male performance falls drastically between the first and second rounds of competition, then increases steadily over the later rounds of competition. For females, performance in the third through fifth periods is significantly better than performance in either of the first two periods (p-value < 0.001). We can reject the hypothesis that male performance in the first two periods is the same (p-value < 0.001). Male performance in the second period is also significantly worse than male performance in the fifth period (p-value = 0.001). That is, male performance decreases significantly after the first period, and then—like female performance—gradually improves over the later periods of competition. Taken together, these results suggest that the male advantage in the first round of competition is due to males overperforming compared to their trend in later periods, rather than from females underperforming.

These results are quite surprising. Although we find a male advantage in the initial period of competition, the male advantage promptly disappears. In later rounds of competition, we find no evidence that males outperform females of similar ability. In fact, we present some evidence that males perform worse than females in later periods of competition. These findings suggest that the gender differences in reaction to competition may not be as robust

[4]The difference between the round one and round two ability coefficients is insignificant with a p-value equal to 0.3. As we'll see below, though, the regression grouping all the later rounds has a much more precise estimate that allows us to reject that round one has the same ability coefficient as other rounds.

[5]In round two, males now tend to perform worse when competing against another male than when competing against a female; however, this difference between the MvF and MvM coefficients is not significant.

as previously thought. However, before drawing conclusions about the importance of our findings, we provide work to better understand our results in the following subsections.

4.2 Impact of Quiz Length

The race treatment includes quizzes of different length. Four of the classrooms were given five questions in each period, five of the classrooms were given 10 questions each period, and three of the classrooms were given 15 questions each period. In the 15-question classrooms 21 percent of the competitions had at least one participant finish early; this rises to 34 percent in the 10-question classrooms and to 100 percent in the five-question classrooms. With shorter quizzes, the students could more easily finish the questions within the allotted time and so were much more likely to ring in early.

Table 4 seperates our analysis of the math races by quiz length and round. It shows that for all quiz lengths, males tend to perform better than females in the first round but not in later rounds. The first-period male coefficient is largest in the competitions with 15 questions; although none of the first period male coefficients are significantly different from one another. When we subdivide the standard treatment, the smaller subsamples are less precise, such that the first period male coefficient is only significantly different from zero in the 15-question classification (p-value = 0.038). We cannot reject that the male coefficients are zero in the five- and ten-question classifications as the p-values are 0.334 and 0.134, respectively. However, the magnitude of the coefficients is consistent with the first-round gender difference found elsewhere in our analysis. Furthermore, if we combine the data from the five- and ten-question classifications, the first-round male coefficient becomes 0.27, which is significantly different than zero with a p-value of 0.052. These results show that the gender difference does not persist beyond the first round for any quiz length.[6]

One possible explanation of the changing gender gap is that students may perceive that early completion was not as frequent as they thought it would be. If, after one round, they decided that the competition was not really much of a race, they might change behavior to a non-competitive mode where no gender gap exists. In that case we would expect to see a gender gap in all rounds for the five-question subsample, where some student in each pair always buzzed in early, along with the first round of the longer question classifications. Table 4 suggests that this"revised beliefs" explanation does not fit the data, as the gender premium disappears just as readily in the five-question subsample as in the others. Indeed we can formally reject the hypothesis that the later rounds of the five question subsample

[6]We easily reject that all six gender coefficients are the same (p-value = 0.02). We can further reject the three-restriction test that the first round coefficients are separately each equal to their later round counterparts (p-value = 0.08).

exhibit the same gender effect as the first round of all three classifications (p-value = 0.01).

4.3 Not-A-Race Treatment

The not-a-race treatment is identical to the race treatment except that the competition was not framed as a race, and there was no benefit to finishing the quiz quicker than one's opponent. Participants were told that they would be rewarded for getting more answers correct than their opponent. Although it was clear that the competition involved answering as many questions correctly as possible in a limited amount of time, the participants were explicitly told that the quiz was "not a race." That is, the participants understood that they were competing against their opponents, but we made an effort to downplay the racing aspect of the competition.

Regressions (1) and (2) in Table 5 provide the results for this treatment. Surprisingly, we find no evidence of gender differences in any period of competition, including the first. This finding is particularly interesting since Section 4.2 found a significant first-round gender effect that persisted even with long quizzes that made the time limit binding. Since few participants finished the long quizzes early, the only substantial difference between the long-quiz race treatment and the not-a-race treatment was whether the competition was framed as a race. The results imply that competition does not generally result in a male advantage. Rather, it is perception about the type of competition (e.g., whether it is a race) that causes the gender differences.[7] In our setting, when the participants perceive the competition as a race, the gender differences exist; when they are told that the competition is not a race, the initial male advantage does not exist.

Unlike in the race treatment, the ability measure in the not-a-race treatment is not increasing across rounds of competition. This means that ability is just as good of a predictor of performance in the first period as it is in later periods. The R-squared is also much higher in this regression than the previous ones—close to 0.45 compared to 0.15 to 0.20 in the race treatment—suggesting that there is less variance in performance when the competition is not viewed as a race. This is consistent with the idea that the perception of racing may affect nerves, making performance less predictable.

[7] We reject with a p-value of 0.04 that the first round gender effect here is the same as the one in the race treatment. We can further reject with a p-value of 0.055 that these not-a-race results are the same as the 15 question subsample of the race treatment.

4.4 Reading-Race Treatment

The reading treatment is similar to the standard treatment except that quizzes are made up of reading questions, rather than math questions. A weakness of the analysis here is that the reading treatment was conducted in a different (but neighboring) school district from the math treatments.[8] Because of this, we were unable to get individual-level assessment test scores for the participants in this treatment, and are therefore unable to control for ability with this treatment. We do, however, observe the distribution of reading state assessment scores by gender, which allows us to compare the relative distribution of male and female scores in our experiment with the distribution of scores on the assessment test.

Regressions (3) and (4) in Table 5 present the results from the reading treatment. These regressions, which do not control for ability, show that in the first period of competition, males tend to perform 0.17 standard deviations lower than the typical female, and in later rounds of competition, this male disadvantage is 0.19 standard deviations. The difference between the first-round and later-round male advantage is highly insignificant (p-value = 0.91).[9]

These results show that any gender differences in the reading treatment do not change across periods of competition. They do not, however, rule out the possibility that there exist persistent gender differences in response to reading competitions. To rule this out, we compare the distribution of male and female normalized scores in the reading treatment with the distribution of normalized scores from the reading section of the state assessment tests for a sample of students for whom we do have data. In the noncompetitive sample, the female advantage is slightly more pronounced, at 0.27, but we are unable to reject the null hypothesis that both competitive and noncompetitive scores have the same mean (p-value = 0.92).

We find no evidence of any gender differences in response to competition when using reading questions. Males consistently perform worse than females, and the performance gap appears consistent with the male-female performance gap seen on the reading portion of state assessment tests. Combined with the results from Section 4.3, this suggests that both the design of the competition (e.g., whether it is a race), and the task (e.g., math versus reading quizzes) affect whether there exist initial gender differences in response to competition.

[8] There is very little difference in the school population compared to that used for the math treatments.

[9] We can reject that the first round gender premium is the same as what we found in the math treatment (p-value = 0.02). In unreported results, we controlled for reading ability using the student's reading scores in the other rounds of the competition, which will have the consistency problems we discussed in Section 3.2. Nevertheless, we again saw no significant difference between male performance in the first round and male performance in later rounds (p-value = 0.70).

4.5 Repeat Treatment

The analysis above provides substantial evidence that the male competitive advantage depends on the nature of the competition, and when it does appear, it vanishes almost immediately. If once the male advantage vanishes it is gone forever, then there is little reason to think that the initial advantage could have any substantial effect on long-run achievement. On the other hand, if the male advantage reappears at the beginning of every new competition, then it may still help explain long-run achievement differences between males and females in competitive environments. To test this, we return to four of the standard treatment classrooms two weeks after first conducting the experiments and reran the same experiment with new partner matching and new questions.

Columns (5) and (6) in Table 5 provide the results for this repeat treatment. We see no significant evidence of a gender difference in competition in any period. Although the first-period male coefficient (i.e., 0.10) is positive, we cannot reject equality with zero (p-value = 0.610). Furthermore, it is less than half the value of the first-period male coefficient in any of our race treatment regressions. For later periods, the male coefficient falls to 0.00. These results are consistent with the male advantage not reappearing for at least a couple weeks after the initial period of competition.

We hesitate to push this conclusion too far, however, as we are also unable to reject equality between the first-round male coefficient in the repeat treatment and the same coefficient in the initial race treatment (p-value = 0.29). We can't reject, and it seems perfectly plausible, that there may be an attenuated gender gap in the first round of the repeat treatment. Unfortunately it is difficult to pick up such fine gradations in experimental data. Future research may provide either sufficient data or a more powerful statistical test for exploring this particular effect.

4.6 OLS Analysis

Until now, the analysis presents the results from the IV analysis. We prefer the IV methodology's robustness to measurement error in our ability variable, as discussed in Section 3.2. We also recognize that an OLS analysis is more straightforward. Table 6 provides the male coefficients for the various regressions when we use OLS. The results are substantively unchanged from the IV analysis above, although the unreported ability coefficients are all noticeably lower, presumably due to the downward bias of measurement error.

5 Discussion

In the first period of our math races, we identify a significant male advantage compared to females of the same math ability, a result that is consistent with the literature. Our analysis, however, did not stop there. We repeated the initial competitions multiple times, and ran alternative treatments in which we changed the task and rules of competition. These additional rounds of competition and alternative treatments greatly improve our understanding of gender differences in reaction to competition.

The main findings of our analysis are as follows. Males perform significantly better than females in the first period of math races, but the male advantage quickly disappears and is not found in any subsequent period of competition. Some evidence suggests that males may perform worse than females of similar ability in later periods. After a two-week break, we find no evidence that the initial male advantage returns. Furthermore, the initial gender difference only appears when we frame the competition as a race; it does not appear when we tell participants that the competition is "not a race." We also find no evidence of a male advantage when the competitions involve reading rather than math questions. These findings suggest that the existence of an initial male advantage depends crucially on the design of the competition and the task at hand, and when the male advantage does exist it does not persist beyond the initial period of competition.

All of these results call into question the argument that the male advantage in reaction to competition may drive long-run achievement differences in the workplace. Workplace competitions are rarely perceived as races, and success usually depends on performance across many periods—two aspects of competition that, according to our analysis, minimize the male advantage. Gender differences in long-run career outcomes (Bertrand and Hallock (2001)) may still be driven by differences in taste for competition. If males become more excited about the prospect of competition (or experience an initial increase in testosterone, etc.), they are both more likely to choose to compete in the first place (Niederle and Vesterlund (2007)), and will put in more effort at the beginning of competition before the excitement of competition wears off.

We recognize a number of questions that should be addressed by future research. For example, although our focus on math competitions makes sense when looking to explain career outcomes (as math ability has been linked to career success, e.g., Joensen and Nielsen (2009)), we are uncertain whether the male advantage disappears as quickly in other settings such as footraces.[10] One may also ask whether an initial male advantage vanishes as quickly

[10]It is possible that males perform better in the first period because they view the competition as a race, but then their excitement subsides after the first period because they start to think of the competition as a series of quizzes rather than a series of races. We maintain that our use of math competitions for the analysis

when there is more at stake (e.g., a promotion rather than a candy bar), or with participants of other ages (e.g., professionals rather than students).

References

Antonovics, Kate, Peter Arcidiacono, and Randall Walsh, "The Effects of Gender Interactions in the Lab and Field," *Review of Economics and Statistics*, 2009, *91* (1), 152–162.

Bertrand, Marianne and Kevin Hallock, "The Gender Gap in Top Corporate Jobs," *Industrial and Labor Relations Review*, 2001, *55* (1), 3–21.

Booth, Alison L. and Patrick J. Nolen, "Choosing to Compete: How Different Are Girls and Boys?," February 2009. IZA Discussion Paper Series, No. 4027.

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

_ , Kenneth L. Leonard, and John A. List, "Gender Differences in Competition: Evidence from a Matrilineal and a Patriarchal Society," *Econometrica*, 2009, *77* (5), 1637–1664.

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

Günther, Christina, Neslihan Arslan Ekinci, Christiane Schwieren, and Martin Strobel, "Women can't jump? An experiment on competitive attitudes and stereotype threat," *Journal of Economic Behavior and Organization*, in press, 2010.

Inzlicht, Michael and Talia Ben-Zeev, "A Threatening Intellectual Environment: Why Women Are Susceptible to Experiencing Problem-Solving Deficits in the Presence of Males," *Psychological Science*, 2000, *11* (5), 365–371.

Joensen, Juanna Schrter and Helena Skyt Nielsen, "Is there a Causal Effect of High School Math on Labor Market Outcomes?," *Journal of Human Resources*, 2009, *44* (1), 171–198.

Kleinjans, Kristin J., "Do gender differences in prefereces from competition matter for occupational expectations?," *Journal of Economic Psychology*, 2009, *30* (5), 701–710.

was the right choice, however, as workplace and academic competitions tend to have more in common with the math competitions than footraces.

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

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

Sapienza, Paola, Luigi Zingales, and Dario Maestripieri, "Gender differences in financial risk aversion and career choices are affected by testosterone," *Proceedings of the National Academy of Sciences*, 2009. http://www.pnas.org/cgi/doi/10.1073/pnas.0907352106.

Wozniak, David, "Choices About Competition: Differences by gender and hormonal fluctuations, and the role of relative performance feedback," December 2009. University of Oregon working paper.

Sample questions for 4th graders

1. Jenny is building a chest that is 5 feet long, 2 feet wide, and 2 feet high. Which is the volume of the chest?

   (a) 32 cubic feet

   (b) 28 cubic inches

   (c) 20 cubic feet

   (d) 18 cubic feet

2. Joe, Adam, and Dan each bought 9 balloons. What is the total number of balloons the 3 friends bought?

   (a) 3

   (b) 12

   (c) 27

   (d) 36

3. Which number is a prime number?

   (a) 121

   (b) 81

   (c) 31

   (d) 12

4. Which of the following can be used to represent the length of a piece of string?

   (a) Ounces

   (b) Square inches

   (c) Inches

   (d) Square feet

5. What is the difference?

$$\begin{array}{r} 4,104,183 \\ \underline{-1,893,214} \end{array}$$

   (a) 2,200,969

   (b) 2,210,969

   (c) 2,791,171

   (d) 3,731,171

Table 1: Normalized Scores By Round For All Treatments

	Round				
	1	2	3	4	5
Race Treatment					
Male	0.15	0.05	-0.01	0.07	0.02
	[0.08]	[0.08]	[0.08]	[0.09]	[0.09]
Female	-0.22	-0.04	0.02	-0.07	-0.01
	[0.09]	[0.09]	[0.09]	[0.10]	[0.10]
Not-a-Race Treatment					
Male	0.08	0.03	0.09	0.02	0.03
	[0.11]	[0.10]	[0.10]	[0.10]	[0.11]
Female	-0.02	0.01	-0.02	0.00	-0.01
	[0.09]	[0.10]	[0.10]	[0.10]	[0.09]
Reading Race Treatment					
Male	-0.08	-0.03	-0.12	-0.08	-0.05
	[0.12]	[0.13]	[0.12]	[0.13]	[0.13]
Female	0.09	0.09	0.16	0.10	0.12
	[0.13]	[0.11]	[0.12]	[0.11]	[0.11]
Repeat Treatment					
Male	0.05	-0.02	0.05	-0.07	0.06
	[0.13]	[0.12]	[0.14]	[0.15]	[0.13]
Female	-0.02	0.07	-0.07	0.11	-0.03
	[0.20]	[0.21]	[0.19]	[0.17]	[0.18]

Cells report the average normalized score by round, gender, and treatment. Scores are normalized by treatment/classroom/round to be mean 0 with standard deviation 1. Standard errors are in brackets. Sample Sizes are given in the text.

Table 2: Fraction Correct By Round For All Treatments

	Round				
	1	2	3	4	5
Race Treatment					
Male	0.57	0.49	0.53	0.57	0.61
	[0.02]	[0.03]	[0.02]	[0.02]	[0.03]
Female	0.47	0.44	0.52	0.54	0.59
	[0.03]	[0.02]	[0.03]	[0.02]	[0.03]
Not-a-Race Treatment					
Male	0.64	0.64	0.66	0.68	0.68
	[0.02]	[0.02]	[0.02]	[0.02]	[0.03]
Female	0.63	0.64	0.65	0.68	0.67
	[0.02]	[0.02]	[0.02]	[0.02]	[0.03]
Reading Race Treatment					
Male	0.51	0.53	0.58	0.53	0.52
	[0.03]	[0.03]	[0.03]	[0.03]	[0.03]
Female	0.58	0.57	0.65	0.60	0.60
	[0.04]	[0.04]	[0.03]	[0.04]	[0.04]
Repeat Treatment					
Male	0.78	0.64	0.68	0.52	0.65
	[0.03]	[0.03]	[0.04]	[0.04]	[0.03]
Female	0.76	0.64	0.66	0.54	0.64
	[0.05]	[0.04]	[0.05]	[0.05]	[0.03]

Cells report the average fraction correct by round, gender, and treatment. Standard errors are in brackets.

Table 3: Instrumental Variables Results of Gender Gap in the Race Treatment

	Round 1		Round 2		Rounds 2-5	
	(1)	(2)	(3)	(4)	(5)	(6)
Male:	0.34***		-0.01		-0.10**	
	[0.117]		[0.105]		[0.040]	
Opp. Male (MvM)		0.42***		-0.05		-0.16**
		[0.163]		[0.129]		[0.066]
Opp. Female (MvF)		0.33**		0.05		-0.11
		[0.155]		[0.150]		[0.076]
Female: Opp. Female (FvF)		0.09		0.03		-0.08
		[0.174]		[0.163]		[0.094]
Ability	0.60***	0.60***	0.84***	0.84***	1.16***	1.17***
	[0.165]	[0.165]	[0.145]	[0.144]	[0.062]	[0.062]
Constant	-0.21**	-0.25**	0.02	0.01	0.05*	0.09*
	[0.086]	[0.114]	[0.079]	[0.098]	[0.028]	[0.052]
R	0.160	0.162	0.289	0.291	0.176	0.177
Observations	253		249		905	

The dependent variable is a student's normalized score in one round of competition. Robust standard errors in brackets are clustered by student for multiple round regressions. *** p<0.01, ** p<0.05, * p<0.1. The excluded group is females in columns (1), (3), and (5) and females vs. males (FvM) in the other three columns. Ability is the average performance in other rounds, instrumented with the student's prior year state assessment score. The F-statistic on the excluded instrument is always above 80.

Table 4: Gender Gap by Quiz Length

	5 Questions		10 Questions		15 Questions	
	Round 1	Rounds 2-5	Round 1	Rounds 2-5	Round 1	Rounds 2-5
	(1)	(2)	(3)	(4)	(5)	(6)
Male	0.21	-0.04	0.27	-0.12	0.48**	-0.09
	[0.217]	[0.057]	[0.180]	[0.076]	[0.230]	[0.085]
Ability	0.89***	1.01***	0.52***	1.22***	0.35	1.32***
	[0.224]	[0.061]	[0.290]	[0.152]	[0.365]	[0.154]
Constant	-0.09	0.02	-0.18	0.07	-0.31**	0.06
	[0.185]	[0.047]	[0.127]	[0.054]	[0.142]	[0.044]
R^2	0.127	0.167	0.178	0.100	0.139	0.254
Obs.	86	345	97	286	70	274

The dependent variable is a student's normalized score in one round of competition. Robust standard errors in brackets are clustered by student for multiple round regressions. *** p<0.01, ** p<0.05, * p<0.1. The excluded group is females. Ability is the average performance in other rounds, instrumented with the student's prior year state assessment score. The F-statistic on the excluded instrument is always above 15.

Table 5: Gender Gap for Alternative Treatments

	Not-a-Race		Reading		Repeat	
	Round 1	Rounds 2-5	Round 1	Rounds 2-5	Round 1	Rounds 2-5
	(1)	(2)	(3)	(4)	(5)	(6)
Male	-0.02	-0.00	-0.17	-0.19	0.10	0.00
	[0.131]	[0.035]	[0.174]	[0.130]	[0.196]	[0.052]
Ability	0.95***	1.06***			1.16***	0.94***
	[0.104]	[0.030]			[0.294]	[0.070]
Constant	-0.01	0.01	0.09	0.12	-0.09	-0.01
	[0.085]	[0.022]	[0.130]	[0.081]	[0.142]	[0.039]
R^2	0.447	0.494	0.007	0.009	0.059	0.034
Obs.	122	491	130	516	80	320

The dependent variable is a student's normalized score in one round of competition. Robust standard errors in brackets are clustered by student for multiple round regressions. *** p<0.01, ** p<0.05, * p<0.1. The excluded group is females. Ability is the average performance in other rounds, instrumented with the student's prior year state assessment score. The F-statistic on the excluded instrument is always above 30.

Table 6: Male Coefficients from OLS Regression

	Round 1	Rounds 2-5
Race Treatment	0.37***	0.07
	[0.121]	[0.077]
5 Questions Subsample	0.32	0.15
	[0.212]	[0.118]
10 Questions Subsample	0.29	- 0.01
	[0.199]	[0.142]
15 Questions Subsample	0.48*	0.02
	[0.25]	[0.151]
Not a Race Treatment	-0.01***	0.01
	[0.152]	[0.112]
Reading Treatment	-0.17	-0.19
	[0.175]	[0.130]
Repeat Treatment	0.03	-0.05
	[0.204]	[0.117]

The dependent variable is a student's normalized score in one round of competition. Robust standard errors in brackets are clustered by student for multiple round regressions. *** p<0.01, ** p<0.05, * p<0.1. Each coefficient is from a separate regression and reports the male difference from the female baseline. Regressions control for ability using the student's prior year state assessment score, normalized to be mean 0 with standard deviation 1.