Rise to the Challenge or Not Give a Damn:

Differential Performance in High vs. Low Stakes Tests

Yigal Attali,[1] Zvika Neeman,[2] and Analia Schlosser[3]

May 2010

PRELIMINARY AND INCOMPLETE

Abstract. We compare performance in the GRE examination in "high" and "low" stakes situations. The high stakes test is the real GRE examination and the low stakes test is a voluntary experimental section of the GRE examination that examinees were invited to take immediately after they finished the real GRE exam. We show that males exhibit a larger difference in performance between the high and low stakes examinations than females, and that Whites exhibit a larger difference in performance between the high and low stakes examinations relative to Asians, Blacks, and Hispanics. The possible reasons as well as the implications of our findings are discussed.

[1] ETS, Princeton, NJ.
[2] Eitan Berglas School of Economics, Tel Aviv University.
[3] Eitan Berglas School of Economics, Tel Aviv University.

1. Introduction

Recently, there has been much interest in the question of whether different groups respond differently to incentives and cope differently with competitive pressure. Economists' interest in this subject stems from attempts to explain gender, racial, and ethnic differences in labor market performance and in human capital accumulation.

Outside economics, the increased use of aptitude tests for college admissions (NACAC, 2008) and the proliferation of standardized tests for school accountability purposes and assessment of student's learning (Koretz, 2008), has generated a vivid debate on what test scores measure and how they are influenced by noncognitive skills such as test taking motivation, attentiveness, self discipline, and effort. While it is clear that students' motivation affects performance, less attention has been given to differences in test-taking motivation across groups or group differences in response to performance based incentives. Rather, it has been implicitly assumed that all groups have the same level of motivation and exert equal effort when facing a test of a given stake.

In this paper, we examine whether individuals respond differently to incentives by analyzing their performance in the Graduate Record Examination General Test (GRE).[4] We examine differences in response to incentives between males and females as well as differences among Whites, Asians, Blacks, and Hispanics. Specifically, we compare performance in the GRE examination in "high" and "low" stakes situations. The high stakes test is the real GRE examination and the low stakes test is a voluntary experimental section of the GRE test that examinees were invited to take immediately after they finished the real GRE examination.

A unique characteristic of our study is that we observe individuals' performance in a "real" high stakes situation that has important implications for success in life. This feature distinguishes our work from most of the literature, which is usually based on controlled experiments that require individuals to perform tasks that might not bear directly on their everyday life, and that manipulate the stakes, degree of competitiveness, or incentive levels in somewhat artificial ways. A second distinctive feature of our research is that we are able to observe performance of the same individual in high and low stakes

[4] The GRE test is a commercially-run psychometric examination that is part of the requirements for admission into most graduate programs in arts and sciences schools and departments in the US and other English speaking countries. Each year, more than 600,000 prospective graduate school applicants from approximately 230 countries take the GRE General Test. The exam measures verbal reasoning, quantitative reasoning, critical thinking, and analytical writing skills that have been acquired over a long period of time and that are not related to any specific field of study. For more information see ETS website: http://www.ets.org/gre/general/about/.

situations and that we can measure performance in the exact same task. This allows us to examine differences in performance based on within individual comparisons rather than comparisons between groups. The third unique feature of our study is that we are able to observe the selection of individuals into the experiment and examine the extent of differential selection within and across groups. Interestingly, we do not find any evidence of differential selection into the experiment, neither according to gender, race or ethnicity, nor according to individual's scores in the "real" GRE exam.

Our results show that males exhibit a larger difference in performance between the high and low stakes GRE test than females, and that Whites exhibit a larger difference in performance between the high and low stakes GRE test relative to Asians, Blacks, and Hispanics. A direct consequence of our findings is that test score gaps between males and females or between Whites and Blacks or Hispanics are larger in a high stakes test than in a low stakes test, while the test score gap between Asians and Whites is larger in the low stakes test.

Interestingly, we find that this differential performance between high and low stakes tests across groups appears across all ability levels (proxied by undergraduate GPA), family backgrounds (measured by mother's education), and even among students with similar orientation towards math and sciences (identified by their undergraduate major or intended graduate filed of studies).

Our findings imply that inference of ability from cognitive test scores is not straightforward. Test performance depends on the perceived significance or importance of the exam. Moreover, it appears that variations in the perceived importance of the test generate different changes in performance across gender, racial, and ethnic groups. Therefore, the perceived importance of a test can significantly affect the ranking of individuals by performance and may have important implications for the analysis of performance gaps by gender, race, and ethnicity.

2. Related Literature

The experimental literature in economics has long documented that individuals' performance is affected by incentives. In recent years, much attention has been given to whether response to incentives varies across individuals, with a particular focus on differences by gender. Surprisingly, little attention has been given to differences in response to incentives by race and ethnicity. A number of studies have shown that men are more willing to self-select into competitive environments relative to women and outperform women in mixed gender competitions (see, e.g. Datta Gupta et al., 2005; Gneezy et al.,

2003, Gneezy and Rustichini, 2004; Niederle and Vesterlund, 2007; Nierdele et al,. 2008; and additional references in the comprehensive review of Nierdele and Vesterlund, 2010). A few studies have investigated whether these gender differences are socially constructed or innate (Gneezy et al., 2008, Booth and Nolen, 2009). With the exception of Paserman (2010), these studies are based on laboratory experiments.

A natural situation where individuals' performance is observed at different incentive levels is in achievement tests in schools, qualifying exams, and university admissions. A number of studies within the educational measurement literature demonstrate that high stakes situations affect test scores by increasing motivation and effort.[5] However, high stakes also increase test anxiety and so might harm performance (Cassaday and Johnson, 2002). Performance in tests is also affected by noncognitive skills as shown by Heckman and Rubinstein (2001), Cunha and Heckman (2007), Borghans et al. (2008), and Segal (2009). Therefore, individuals with similar cognitive skills might obtain different scores in aptitude tests if their perceived importance of the test differs or if they are less motivated to perform well.[6]

3. Experiment Set-up and Data

We re-analyze data from a previous study conducted by Bridgeman et al. (2004), whose purpose was to examine the effect of time limits on performance in the GRE Computer Adaptive Test (CAT) examination. All examinees who took the GRE CAT General Test during October-November 2001 were invited to participate in an experiment that would require them to take an additional test section. GRE examinees who agreed participate in the experiment were promised a monetary reward if they perform well compared to their performance in the real examination.[7]

[5] For example, Cole et al. (2008) show that students' effort is positively related to their self reports about the interest, usefulness, and importance of the test; and that effort is, in turn, positively related to performance. For a review of the literature and studies on the effects of incentives and test taking motivation on NAEP performance see O'Neil, Surgue, and Baker (1996).

[6] Several studies (e.g., Barres, 2006; Duckworth & Seligman, 2006; and the references therein) suggest that girls outperform boys in school because they are more serious, diligent, studious, and self disciplined than boys. Other important noncognitive dimensions that affect test performance are discussed by the literature on stereotype threat that suggests that performance of a group is likely to be affected by exposure to stereotypes that characterize the group (see Steele, 1997; Steele and Aronson, 1995; and Spencer et al., 1999).

[7] Specifically, at the end of the regular test, a screen appeared that invited voluntary participation in a research project. The instructions stated "It is important for our research that you try to do your best in this section. The sum of $250 will be awarded to each of 100 individuals testing from September 1 to October 31. These awards will recognize the efforts of the 100 test takers who score the highest on questions in the research section relative to how well they did on the preceding sections. In this way, test takers at all ability levels will be eligible for the

Experiment participants were randomly assigned to one of four groups: one group was administered a quantitative section with standard time limit (45 minutes), a second group received a verbal section with standard time limit (30 minutes), the third group received a quantitative section with extended time limit (68 minutes) and the fourth group received a verbal section with extended time limit (45 minutes). The research sections were taken from regular CAT pools (over 300 items each) that did not overlap with the pools used for the real examination. The only difference between the research section and the real sections was the appearance of a screen that indicated that performance on the research section did not contribute to the examinee's official test score. We therefore consider performance in the real section to be performance in a high stakes situation and performance in the experimental section to be performance in a low stakes situation. Even though a monetary reward based on performance was offered to those who participated in the experiment, it is clear that success in the experimental section was less significant to examinees and involved less pressure. More importantly, since the monetary reward was conditional on performance relative to own achievement in the high stakes section rather than on absolute performance, incentives to perform well in the experimental section were identical for all participants in the experiment.

Table 1 shows details of the construction process of our analysis sample. From a total of 81,231 GRE examinees in all centers (including overseas), 46,038 were US citizens that took the GRE test in centers located in the US. We focus on US citizens tested in the US to avoid dealing with more heterogeneous populations and to control for a similar testing environment. In addition, we want to abstract from differences in performance that are due to language difficulties. 15,945 out of the 46,038 US examinees agreed to participate in the experiment. About half of them (8,232) were randomized into the regular time limit sections receiving either an extra Q-section (3,922) or an extra V-section (4,310).[8] We select only experiment participants who were randomized into the regular time limit experimental

award. Award recipients will be notified by mail." See Bridgeman et al. (2004) for more details about the experiment design and execution.

[8] Since the experimental sections were randomized among the full sample of experiment participants, which included all students (US and international) tested in all centers across the globe, the proportion of US participants assigned to each section is not exactly 50 percent but is extremely close to that.

groups because we are interested in examining differences in performance in the exact same task that differs only by the stake examinees associated with it.[9]

A unique feature of our research design that distinguishes our study from most of the experimental literature is that we are able to identify and characterize the experiment participants out of the full population of interest (i.e., GRE examinees in our case). Table 2 compares the characteristics of the full sample of US GRE test takers and the sample of experiment participants.[10] Interestingly, the two populations are virtually identical in terms of proportions of females, males, and minorities. For example, women comprise 66 percent of the full population of US domestic examinees while the share of women among those who agreed to participate in the Q or the V section was 65 and 66 respectively. Likewise, whites make up about 78 percent of GRE US domestic examinees and they are equally represented among experiment participants. The shares of Blacks, Hispanics, and Asians range between 6 and 5.5 percent in both the full sample and the sample of experiment participants.[11]

Not only are the different subgroups of interest (males, females, Whites, and minorities) equally represented among experiment participants, but we also observe that experiment participants have similar GRE test scores relative to the full population from where they were drawn. For example, males are located, on average, at the 56 percentile rank of the Q-score distribution, which is equal to the average performance of experiments participants in the Q or the V sections. The median score (57 percentile rank) and standard deviation (27 points) are also identical for the full sample of GRE US male test takers, the sample of experiment participants randomized to the Q-section, and the sample of experiment participants randomized to the V-section. The test score distribution of female GRE test takers is also identical to that of female experiment participants. We observe also the same result when comparing test score distributions within each race/ethnicity. Overall, results presented in Table 2 show that there is no differential selection into the experiment according to gender, race/ethnicity or GRE test scores. Moreover, we do not find any evidence of differential selection within each gender or race/ethnic group.

[9] One limitation of our study is that we were not able to randomize the order of the tests, so that all examinees received the low stakes test after the high stakes test. As we discuss later, we believe this constraint does not affect our main results or interpretation.

[10] Due to data restrictions we cannot compare experiment participants to non-participants as we received the data on experiment participants and the data on the full population of GRE examinees in two separate datasets that lacked individual identifiers.

[11] Reported proportions by race/ethnicity do not add up to one since the following additional groups are not reported in the table: American Indian, Alaskan, and examinees with missing race/ethnicity.

GRE test takers are required to fill a form upon registration to the exam. The form collects information on basic background characteristics, college studies, and intended graduate field of studies.[12] Table 3 reports descriptive statistics of these background characteristics for the sample of experiment participants stratified by gender, race, and ethnicity. Note that the comparisons presented here are across the population of GRE test takers, which is a selected sample of college students, and therefore they do not represent group differences across the population of college students but rather differences across college students who intend to pursue graduate studies.

Averages reported in columns 2 and 3 of Table 2 show that Males and Females seem to come from similar family backgrounds as denoted by both mother's and father's educational levels and by the proportion of native English speakers (about 92 percent). Females and males have also similar distributions of undergraduate GPA (UGPA). For example, 19 percent of males and 19 percent of females have an UGPA that is equal to "A". Nevertheless, males are more likely to come from undergraduate majors in math, computer science, physics or engineering and they are also more likely to intend to pursue graduate studies in these fields (26 percent for males versus 5 percent for females).

Columns 3 through 6 in Table 3 report descriptive statistics of the analysis sample stratified by race/ethnicity. Maternal education is similar among Whites and Asians but Asians are more likely to have a father with at least some graduate studies or a professional degree relative to Whites (45 versus 35 percent). Hispanics and Blacks come from less educated families. Asians are less likely to be native English speakers (86 percent) relative to Whites (93 percent), Blacks (95 percent), and Hispanics (90 percent). In terms of undergraduate achievement, we observe that Whites and Asians have similar UGPAs distributions but Hispanics and Blacks have, on average, lower UGPAs. Asians are more likely to do math, science, and engineering either as an undergraduate major or as an intended field of graduate studies (30 percent) relative to Whites (11 percent), Blacks (8 percent), or Hispanics (12 percent).

4. Empirical Framework

To compare between subject's performance in the high and the low stakes test, we estimate the following first difference equation for each of the experimental samples (i.e. those randomized to an experimental Q or V section):

$$(1)\ Y_{iHS} - Y_{iLS} = \beta_0 + \beta_1 Female_i + \beta_2 Black_i + \beta_3 Hispanic_i + \beta_4 Asian_i + \beta_5 Other_i + x_i'\gamma + u_i$$

[12] Unfortunately, we obtained the background information on experiment participants only.

Where Y_{iHS} denotes the test score of individual *i* in the high stakes section; Y_{iLS} is the test score of individual *i* in the low stakes section; x_i is vector of individual characteristics that includes the following covariates: mother's and father's education, dummies for UGPA, undergraduate major, intended graduate field of studies, and disability status. *Female*, *Black*, *Hispanic*, *Asian*, and *Other* are dummy variables for the gender and race/ethnicity of the examinee.[13] Whites and males are the omitted categories. The coefficients of interest are $\beta_1 - \beta_4$ that denote the difference in performance gap between the high to the low stakes test of the relevant group (Females or Blacks/Hispanics/Asian) relative to the omitted category (Males or Whites). To simplify the interpretation, we reverse the sign of the coefficients and report in all tables differences between males and females and differences between Whites and Blacks/ Hispanics/Asians.

Note that by using a first difference specification, we are differencing out an individual's fixed effect that accounts for all factors that affect examinee's performance in both the low stakes and the high stakes test. By including a vector of covariates we allow for individual's characteristics to affect the change in performance between the high and low stakes situation.

GRE scores in the quantitative and verbal sections range between 200 to 800, in 10-point increments. To ease the interpretation of the results, we transformed these raw scores into percentile ranks using the GRE official percentile rank tables.[14] All results presented below are based on GRE percentile ranks. We obtained similar results when using raw scores or logs of raw scores.

5. Results

5.1. Gender Differences in Performance

Table 4 exhibits examinees' performance in the high and low stakes test by section and gender. Columns 3 through 5 report performance in the high stakes section. As common in comparisons of GRE scores by gender, we also find among our experiment participants a gender gap in performance favoring males in both the quantitative and verbal sections. On average, Males are placed about 15.3 percentile points

[13] Race/ethnicity categories in the GRE form are self-exclusive and it is not possible to check more than one option.
[14] For more information regarding on the interpretation of GRE scores, exam administration and validity see "Guide to GRE Scores" available online at the ETS website: http://www.ets.org/Media/Tests/GRE/pdf/gre_0910_guide.pdf

higher in the test score distribution of the Q-section relative to females. The gender gap in the V-section is smaller but still sizable, with males scoring about 6.4 percentile points higher than females.[15]

Students' performance in the low stakes section is reported in columns 6-8. On average, performance in the low stakes section is lower than in the high stakes section. Interestingly, the test score gender gap is narrower in the low stakes section but is still significant (10.7 percentile points in the Q-section and 2 percentile points in the V-section). The reduction of the gender gap in the low stakes section suggests a differential drop in performance between the high and low stakes section between males and females. This is reported in columns 9 and 10 that show that males' performance between the high and low stakes Q-sections drops by 11.6 percentile points while females' performance drops by only 7.1 points. The differential gap in performance between males and females is of 4.5 percentile points (s.e.=0.784). That is, moving from the high to the low stakes situation narrows the gender gap in the quantitative test by about 4.5 percentile points, which is a 30 percent drop in the gender gap of the high stakes test. The differential change in performance remains almost unchanged after controlling for individual's characteristics.

We also find the same pattern when examining changes in individual's performance between the high and low stakes V-sections. Men's scores drop by 10.2 percentile points, on average, while females' scores drop by a smaller magnitude, 6.1 percentile points. That is, Males' scores drop by 4 percentile points (s.e.=0.783) more relative to females. The differential change in performance between males and females in the V-section is similar to what we found in Q-section. Nevertheless, in this case, it can be translated into a larger reduction in the gender gap: the gender gap in verbal scores is reduced by two thirds when moving from the high stakes to the low stakes situation.

Table 5 reports the gender gap in students' performance in high and low stakes tests for different subsamples stratified by undergraduate GPA (UGPA), student's major, intended field of graduate studies, and mother's education. As expected, we observe a positive association between UGPA and GRE performance. Students with higher UGPA have higher test scores in both the high and the low stakes sections of the quantitative and verbal exams. Males' advantage in the high stakes test appears across all cells of the UGPA distribution both in the quantitative and the verbal sections. Again in this case, the gender gap in performance is narrower in the low stakes section for all cells stratified by

[15] Note that percentile scores of males and females do not add to 100 since they are constructed using the official GRE tables, which include also international examinees.

UGPAs and it even becomes insignificant when comparing performance in the V-section between male and female students with an UGPA of A, A- or B-.

We see in columns 9 and 10 of the table that all students, regardless of their academic ability (proxied by UGPA), exhibit a significant drop in performance between the high and the low stakes sections (both the quantitative and the verbal). Interestingly, the larger drop in males' performance relative to females is found across all ability levels (see columns 11 and 12). Note that the larger drop in males' performance is evident both in absolute and percentage terms relative to the mean outcome.

The next two rows of Table 5 report the gender gap in performance for the sample of students who majored in math, computer science, Physics or engineering or who intend to pursue graduate studies in one of these fields (to simplify the discussion we will call them math and science students). We focus on these students to target a population of females that is expected to be highly selected, with a strong academic orientation towards math and science, and perhaps also more driven to achievement.[16] While females represent the majority among the full population of GRE examinees (65 percent) they are certainly a minority among math and science students (26 percent).

Achievement in the GRE Q-section is much higher among math and science students relative to the sample average and even relative to those students whose UGPA is an "A". Math and science students also attain higher scores in the V section relative to the sample average but they score slightly lower compared to those students with an "A" UGPA. As expected, the gender gap in the high stakes Q-section among math and science students is smaller (7 percentile points) than the gender gap in the full sample (15.3 percentile points), although we still observe that males have higher achievement than females. The gender gap among those who intend to pursue graduate studies in these fields is a bit wider (8.7 percentile points) but still significantly smaller than the gap observed for the full sample. Finally, there is no gender gap achievement in the V high stakes section in the subsamples of math and science students.

Achievement of math and science students in the Q low stakes section is lower than in the high stakes section but these students still perform better relative to other students in the low stakes section. Consistent with our previous results, the gender gap in Q performance among math and science

[16] Note that in contrast to most of the literature that focus on gender differences in academic and career orientations towards Science, Technology, Engineering, and Math (STEM) we focus here in a more limited number of fields (e.g., we exclude biology) to select those fields that are predominately populated by males. Our results do not change when using the broader definition of STEMs fields.

students is narrower in the low stakes section relative to the high stakes section and in this case, it is even insignificant. The pattern for the V section is similar, but in this case, we observe for the first time that math and science females actually outperform their male counterparts in the low stakes V-section with an average achievement that is about 7 to 8 percentile points higher.

Overall, if we compare performance of male and female examinees with academic orientations in math and science, we reach the conclusion that there are no gender differences in average performance if we consider low stakes scores but that males outperform females by about 7-8 percentile points if we focus instead on high stakes scores. Likewise, a comparison of performance in a verbal high stakes examination among math and science students reveals no gender gap in performance while inspection of gender differences in performance in a low stakes verbal exam shows that females outperform males by about 7-8 percentile points.

A direct corollary of these results that is consistent with our previous findings is that even in this subsample of high achieving students, there is a drop in performance between the high and the low stakes test that is larger for males (who reduce their performance by about 12-13 percentile points in both subjects) relative to females (who reduce their performance by 6-7 percentile points in the Q section and by 4-5 percentile points in the V section). The gender differences in relative performance in these subsamples of high achieving students is of about 5 percentile points in the Q section and 8 percentile points in the V sections. Both gaps are statistically significant and do not change much after controlling for examinees' observed characteristics.

We also look at gender gaps within groups stratified by mother's education. Our interest was to examine whether female examinees whose mothers attended graduate school would behave more like males and exhibit a larger gap in performance between the high and low stakes situation. Interestingly, the gender gap in relative performance between high and low stakes test appears across all levels of maternal education in both the quantitative and the verbal sections.

5.2. Differences in Performance by Race/Ethnicity

Table 6 reports differences in performance among Black, Hispanic, and Asian students relative to White students in the high and low stakes sections. Asians have the highest achievements among all ethnic/racial groups in the high stakes Q-section. Their test scores are about 15 percentile points above Whites. Hispanics lag behind Whites by an average of 10.6 percentile points. Q-scores of Blacks are

lower and they are placed, on average, about 25 percentile points below Whites in the test score distribution.

Average performance of all race/ethnic groups is lower in the low stakes test, but the drop in performance differs for each group. As a result, test score gaps between groups differ in the low and the high stakes test. For example, the score gap between Whites and Blacks shrinks from 25 to 19 percentile points when comparing between the high versus low stakes Q-section. Likewise, the gap between Whites and Hispanics shrinks from 10.6 to 5 percentile points while the gap between Asians and Whites widens a bit (from 15.3 to 17.6 percentile points in favor of Asians).

The results for the V-section are similar to those described for the Q-section when comparing performance of Whites versus Blacks. Again in this case, the test score gap between Whites and Blacks narrows from 23.2 percentile points to 17.7 percentile points when comparing between performances in the high versus low stakes section of the V-test. Contrasts between Whites and Hispanics or Asians differ in the Q and in the V sections. First, we observe that while Asians outperform Whites in the Q-sections they perform similarly to Whites in the V-sections. Second, we observe that the score gaps between Whites and Asians or Whites and Hispanics are similar in the high and in the low stakes V-sections.[17]

Table 7 reports change in performance between the high and the low stakes test for Whites, Blacks, Hispanics, and Asians and the raw and controlled differences between Whites and each of these groups. Whites exhibit the largest drop in performance between the high and the low stakes Q-section. Whites' performance drops by 9.4 percentile points, while that of Asians drops by 7 percentile points, Blacks' performance drops by 3 percentile points, and Hispanics' performance drops by 3.8 percentile points. Differences between Whites and each of the minority groups are all significant. The controlled difference between Whites and Blacks, after accounting for individual's characteristics, is of 4.16 percentile points (s.e.=1.02). The equivalent difference between Whites and Hispanics is 5.23 (s.e.=1.42) and the difference between Whites and Asians is 3.29 (s.e.=1.70).

[17] We suspected that the different pattern obtained for Asians and Hispanics in the V-section could be related to language dominance. We therefore replicated the analysis while limiting the sample to students who selected English as their best language for communication. Results, not reported here but available upon request, were similar to those obtained for the full sample and did not support our hypothesis. Still, we believe that Asian and Hispanic students that chose English as their best language for communication might have a more limited vocabulary relative to Whites, a fact that could affect their performance in the V-section.

In the verbal section, the performance drop from the high to the low stakes section is larger for Whites than for Blacks (7.8 percentile points versus 2.3 percentile points). In contrast, Hispanics and Asians seem to exhibit a similar drop in performance compared to Whites.

5.3 Within Race/Ethnicity and Gender Differences in Performance

Results presented above showed that males and Whites exhibit the largest drop in performance between the high and the low stakes tests compared to females and minorities. We check here for gender and race/ethnicity interactions by examining whether differences between males and females appear across all race/ethnic groups and whether differences between Whites and minorities show up for males and for females.[18]

Table 8 reports differences in performance between males and females within each race/ethnicity group as well as differences between Whites and minorities for males and females separately. The table also reports performance in the high and low stakes section for each gender and ethnicity/race group. We focus here in the Q sections as we think performance is less influenced by language constraints among Hispanics and Asians. The results show that White males have the largest differential performance between the high and the low stakes test compared to Black, Asian, and, Hispanic males. We obtain a similar result for females with the exception of Asian females, where we observe that they behave in the same way as White females.

Comparisons between males and females within each race/ethnicity group reveal that males exhibit a larger drop in performance relative to females among Whites, Blacks, and Hispanics although differences between genders are only statistically significant among Whites. In contrast, we observe no gender differences among Asians. Asian males and females have an average drop in performance between the high and the low stakes test of 6 and 7 percentile points respectively. In fact, the drop observed among females is even larger than the drop observed among males, albeit the difference is not statistically significant.

[18] It is worth noting that conclusions drawn from this subsection should be taken with caution as the sample sizes stratified by gender and race/ethnicity are relatively small for Blacks, Hispanics, and Asians.

5.4 Rising to the Challenge or not Giving a Damn?

The evidence presented above shows that men and Whites exhibit a larger differential performance between high and low stakes tests relative to women and minorities. The following three alternative arguments could explain our results: (i) men and Whites simply "don't give a damn" in low stakes situations compared to women and minorities, respectively; (ii) women and minorities find it relatively more difficult to deal with high stakes and stressful situations; and (iii) men and Whites are more capable of "rising to the challenge" in high stakes situations compared to women and minorities, respectively.[19]

To examine the likelihood of the first explanation, we would ideally like to measure effort invested in the test. Effort could be manifested by trying harder to solve each question (i.e., investing more mental energy) or by investing more time. Figure 1 plots the distribution of time spent by examinees in the experimental Q and V-sections by gender, race, and ethnicity. We learn from the figure that there is a significant variation in time invested in the experimental section. Some examinees spent very little time and some of them exhausted the time limit (45 minutes for the Q-section and 30 minutes for the V-section).

We examine in Figure 2 the relationship between achievement in the experimental section and time invested in that section for males, females, Whites, Blacks, Hispanics, and Asians. The figure shows that achievement increases with time invested in the quantitative section for all gender, racial, and ethnic groups. The relationship between time invested and performance in the verbal section is also positive at the lower values of the distribution but switches signs after about 20 minutes. It is clear from the figures that it is not possible to get a high score without investing some minimal amount of time. We therefore conclude that subjects who invested very little time were obviously not investing enough effort.

Table 9 reports the characteristics of individuals that invested less than ten minutes in the experimental section. While the ten minutes cutoff is somewhat arbitrary in determining who invests

[19] Because examinees participated in the experimental section after they completed the real GRE examination, it is also possible that our results are due to the fact that women and minorities are less fatigued by the GRE examination than men and Whites, respectively. This argument seems unlikely as it goes against recent psychological and medical literature that claims that, if anything, females appear to exhibit a higher level of fatigue after performance of cognitive tasks (see, e.g., Yoon et al., 2009). In addition, we are not aware of any studies that show that Whites exhibit a higher level of fatigue in response to cognitive tasks relative to Blacks, Hispanics, or Asians.

low effort, we choose a time threshold that clearly suggests low effort and cannot be confounded by ability to solve a test speedily.[20] Columns 1 and 4 of the table report the share of examinees who invested less than 10 minutes in the experimental Q and V sections. Columns 3 and 6 report the p-values for F-tests that check for equality of proportions between groups.

The results clearly show that males appear to invest less effort in the experimental section compared to females. 17 percent of the males who participated in the Q-experiment spent less than ten minutes trying to solve the experimental section while the equivalent among women is 13 percent. Gender differences are similar for the V-section. It is important to recall that, as shown in Table 2, the share of males and females among experiment participants was equal to their share in the full population of GRE test takers; so that gender differences in effort among experiment participants cannot be attributed to a differential selection into the experiment. Statistics by race/ethnicity show that Whites are more likely to invest low effort relative to Blacks and Asians. Whites also appear to invest less effort than Hispanics, although differences in this case are not statistically significant.

The stratification of the sample by background characteristics and achievement shows some interesting patterns. First, we observe some differences in effort invested according to student's parental education. Although differences are relatively small, it seems that students with more educated parents are more likely to invest low effort. In contrast, we find no clear relationship between the likelihood of investing low effort and student's ability, neither when defined by student's score in the high stakes section nor when defined by student's UGPA. This last finding is important as it shows that the decision to invest low effort in the low stakes section is unrelated to student's ability, suggesting that noncognitive abilities are likely to play a more important role in determining performance in low stakes situations. The lack of a relationship between student's ability and effort invested in the low stakes section suggests also that our previous results on differential gaps in performance by gender, race, and ethnicity are unlikely to be explained by ability differences between groups.

Are all gender, racial, and ethnic differences in the performance gap between the low and the high stakes test explained by a larger share of males/Whites investing very low effort? To examine this, we reproduce our main results of Tables 4 and 7 while limiting the sample to individuals who invested at least ten minutes in the experimental section. Appendix Table A1 reports differences in performance

[20] Participants who invested less than 10 minutes in the experimental Q-section were all located below the 58th percentile of the test score distribution. 94% of all those who spent less than 10 minutes in the V-section were also located below the 58th percentile.

between the high and the low stakes test for the sample of individuals who spent at least ten minutes in the experimental section. Panel A reports results for the Q-section and Panel B reports results for the V-section. To facilitate comparison, we reproduce the results for the full sample of experiment participants in the first row of each panel. Our results show that differences between males and females and between Whites and minorities are reduced when the sample is limited to those who invested at least ten minutes in the experimental section. Nevertheless, we still observe a larger gap in performance for males and Whites relative to females and minorities.

To summarize, evidence on time invested in the experimental section suggests that the larger gap in performance between the high and the low stakes section found among men and Whites can be partly explained by a lower level of effort exerted by these groups in the low stakes section.

As noted above, a second possible explanation for the larger gap in performance between the high and the low stakes section among men and Whites could be due to a higher level of stress and test anxiety among females and minorities that hinders their performance in high stakes situations. To examine the merit of this explanation we inspect the distribution of changes in performance between the high and the low stakes test. Although most individuals have lower test scores in the low stakes section, we find that some of them improve their performance. This improvement can be due to usual volatility or measurement error in test scores, due to learning or increased familiarity with the test, or due to a lower level of stress and anxiety involved in the low stakes test.

Table 10 reports the proportion of males, females, Whites, Black, Hispanics, and Asians who exhibit an increase in performance between the high and the low stakes section. Nearly 30 percent of examinees improve their score in the low stakes section. Interestingly, females and minorities are more likely to improve their performance in the low stakes section relative to males and Whites respectively. This result could be due to a higher level of stress in the high stakes section among females and minorities or could just be a statistical artifact induced by the lower average performance of females and minorities in the high stakes section. Figure 3 shows that once we compare examinees located at the same decile of the score distribution in the high stakes exam, there are no differences in the share of those improving their test score by gender, race or ethnicity. Overall, males and females who are located at the same decile of the test score distribution in the high stakes test are equally likely to improve their performance. The same is true for Whites and minorities. These findings suggest that test

anxiety and stress are unlikely to explain the smaller gap in performance between the high and low stakes test among females and minorities.

Finally, a third possible explanation for the larger gap in performance between the high and low stakes test among men and whites is that these groups are more able to boost performance when facing a high stakes or challenging task. This explanation is harder to assess as it is not possible to establish an ability baseline that is independent of performance in a given test of a given stake.

6. Discussion and Conclusions

In this study we examine the differential performance of females, males, Whites, and minorities in low and high stakes situations by contrasting performance of GRE examinees in the real and in an experimental section of the test. As opposed to the majority of previous studies in this subject, we are able to examine achievement in a real high stakes situation and look at changes in performance at the individual level in the exact same task in a low stakes condition. Our results show that males and Whites have the highest differential change in performance relative to females, Asians, Hispanics, and Blacks.

We show that the larger differential performance observed among males and Whites is at least partially due to the fact that these two groups invest relatively less effort in low stakes exams. We did not find empirical support for the hypothesis that the smaller differential among females and minorities is due to higher levels of stress and test anxiety among these groups that hinder their performance in high stakes situations. Finally, we cannot rule out the possibility that part of the differences between genders and race/ethnicity are due to the fact that males and Whites are better capable to top themselves in high stakes situations.

Our findings indicate that men and Whites who perform well in high stake tests might not perform as well in ordinary assignments, and that women and minorities who do not perform so well in high stake tests may do relatively better in ordinary, day to day, assignments. Another implication of our findings is that test score gaps between males and females or between Whites and minorities might vary according to the stakes of the test, as each group appears to respond differently to level of stakes. Therefore, it is important to consider the stakes of a test and the differential performance of each group according to the stakes level when analyzing test score gaps.

Our findings are consistent with evidence that shows that girls' performance in low stakes examinations, such as NAEP, is equal or better than boys' performance, while boys outscore girls in high

stakes tests such as SAT, ACT, and GRE (Hill et al., 2010). This is also consistent with the findings that standardized tests usually underpredict college and graduate school performance for women and overpredict performance for men (see, e.g., Willingham and Cole, 1997 and Rothstein, 2004).[21]

It is interesting to try to determine to what extent differences in performance between high and low stakes situations are socially constructed or innate. While this question is beyond the scope of the current study, we speculate that the lack of differences between Asian males and females suggests that part of the source for the gender differences observed among other ethnic and racial groups might be explained by acquired rather than innate skills. This is also consistent with Stevenson and Stigler (1992) who claim that in cultures that produce a large number of math and science graduates, especially women in South and East Asian cultures, the basis of success is generally attributed to effort rather than to inherent ability.

More findings on this issue are presented in Figure 4, where we plot differences in achievement between the high and the low stakes Q-section by students' undergraduate major. Interestingly, those who exhibit the largest gap in achievement between the high and the low stakes section are economics majors. This finding could be either a result of self-selection into economic majors or skills acquired during undergraduate studies. Be that as it may, it is consistent with Rubinstein (2006) who finds that economics majors have a much stronger tendency to maximize profits relative to other undergraduate majors.

Our results suggest that policy makers may be able to diversify the population of students who are admitted to colleges, universities, and graduate schools, and who study math and science, by gentle manipulation of the stakes of the exams that these students are required to take prior to being admitted into their respective programs. There are several different possible mechanisms that may help facilitate such a change. For example, allowing students to retake a test while considering the average or the maximum score would reduce the stake of any given test.[22]

[21] Our findings also suggest the same pattern for Whites compared to minorities, but this is not the case in practice (see, e.g., Mattern et al., 2008), presumably because the lower performance of minority students in college can be explained by other factors such as their relatively disadvantaged background (Rothstein, 2004).

[22] Indeed, this type of policy has recently been adopted by many colleges in their undergraduate admission policies. The new policy, known as Score Choice, gives students the option to choose the SAT scores by test date and subject test to be sent to colleges (CollegeBoard, 2009). While this policy is expected to lower the level of stress and stakes of any given test, Vigdor and Clotfelter (2001) show that that minority students are less likely to retake the SAT, a fact that offsets the possible benefits of this policy.

References

Barres, B. (2006): "Does Gender Matter?," *Nature*, 442, 133-136.

Booth, A. and P. Nolen (2009) "Choosing To Compete: How Different Are Girls and Boys?" Working Paper.

Bridgeman B., F. Cline, and J. Hessinger (2004) "Effect of Extra Time on Verbal and Quantitative GRE Scores," *Journal Applied Measurement in Education* 17, 25-37.

Cassaday, J. C., & Johnson, R. E. (2002). "Cognitive Test Anxiety and Academic Performance," *Contemporary Educational Psychology*, 27, 270–295.

Clark, M. J., & Grandy, J. (1984). "Sex differences in the academic performance of Scholastic Aptitude Test takers" (College Board Rep. No. 84 – 8; Educational Testing Science Research Rep. No. 84 – 43). New York: College Entrance Examination Board.

Cole, J. S., D. A. Bergin, and T. A. Whittaker (2008) "Predicting student achievement for low stakes test with effort and task value," *Contemporary Educational Psychology* 33, 609-624.

CollegeBoard (2009). SAT Score-Use Practices by Participating Institution.

Datta Gupta, N., A. Poulsen, and M.-C. Villeval (2005): "Male and Female Competitive Behavior: Experimental Evidence," IZA Discussion Paper No. 1833.

Eckel, C., and P. Grossman (2002): *Differences in the Economic Decisions of Men and Women: Experimental Evidence.* Elsevier Science B.V./North-Holland, Amsterdam, The Netherlands.

Eckel, C. C., and P. J. Grossman (2008): "Sex and Risk: Experimental Evidence," in *Handbook of Experimental Economics Results*, ed. by C. Plott, and V. Smith. Elsevier Science, Amsterdam, The Netherlands, Forthcoming.

Gneezy, U., K. L. Leonard, and J. A. List (2008): "Gender Differences in Competition: Evidence From a Matrilineal and a Patriarchal Society," NBER Working Paper 13727.

Gneezy, U., M. Niederle, and A. Rustichini (2003): "Performance in Competitive Environments: Gender Differences," *Quarterly Journal of Economics*, pp. 1049-1074.

Gneezy, U., and A. Rustichini (2004): "Gender and Competition at a Young Age," *American Economic Review Papers and Proceedings*, pp. 377-381.

Hill C., C. Corbett, and Rose A., 2010, “Why So Few? Women in Science, Technology, Engineering and Mathematics”, American Association of University Women (AAUW).

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. (1984). “Methodological, statistical, and ethical issues in the study of bias in psychological tests.” In C. R. Reynolds & R. T. Brown (Eds.), Perspectives on bias in mental testing (pp. 41–99). New York: Plenum Press.

Koretz, D. (2008) “The Evolution of American Testing”, Chapter 2 in Measuring Up: What Educational Testing Really Tells Us. Harvard University Press, Cambridge Massachusetts.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Lemann, Nicholas (1999) The Big Test: The Secret History of the American Meritocracy. New York: Farrar, Straus and Giroux.

Mattern, K. D., B. F. Patterson, E. J. Shaw, J. L. Kobrin, and S. M. Barbuti (2008) “Differential Validity and Prediction of the SAT” College Board Research Report No. 2008-4, The College Board, New York, 2008.

The National Association for College Admission Counseling (NACAC) 2008. Report of the Commission on the Use of Standardized Tests in Undergraduate Admission.

National Science Board. (2008). Science and engineering indicators 2008 (Volume 1, NSB 08-01; Volume 2, NSB 08-01A). Arlington, VA: National Science Foundation.

Niederle, M., and L. Vesterlund (2007). “Do Women Shy Away From Competition? Do Men Compete Too Much?,” *Quarterly Journal of Economics*, pp. 1067-1101.

Nierdele, M., C. Segal, and L. Vesterlund (2008). “How Costly is Diversity? Affirmative Action in Light of Gender Differences in Competitiveness”, NBER Working Paper No. 13922.

Niederle, M., and L. Vesterlund (2010). “Explaining the Gender Gap in Math Test Scores,” *Journal of Economic Perspectives*, Vol. 24, No. 2, pp. 124-144.

O’Neil, H. F., Sugrue, B., and Baker, E. L. (1996) “Effects of Motivational Interventions on the National Assessment of Educational Progress Mathematics Performance”, Educational Assessment, 2(2), pp. 135-157.

Paserman, D. (2010) “Gender Differences in Performance in Competitive Environments: Evidence from Professional Tennis Players.” Discussion Paper, Boston University.

Rothstein, J. (2004), “College Performance Predictions and the SAT,” *Journal of Econometrics*, 121. No.1-2, pp. 123-144.

Rubinstein, A. (2006), “A Skeptic’s Comment on the Study of Economics”, *Economic Journal*, Vol. 116, C1-C9.

Spencer, S., Steele, C. M., and D. M. Quinn, (1999), “Stereotype Threat and Women’s Math Performance,” *Journal of Experimental Social Psychology*, Vol. 35 No. 1, pp. 4-28.

Steele, C. M., (1997), “A threat in the Air: How Stereotypes Shape the Intellectual Identities and Performance of Women and African-Americans”, American Psychologist, Vol. 52 No. 6, pp. 613-629.

Stevenson, H. W., & Stigler, J. W. (1992). The learning gap: Why our schools are failing and what we can learn from Japanese and Chinese education. New York: Simon & Schuster.

Stricker, L. J., Rock, D. A., & Burton, N. W. (1993). “Sex differences in predictions of college grades from Scholastic Aptitude Test scores.” *Journal of Educational Psychology*,85, 710 –718.

Vigdor, J. L. and C. T. Clotfelter (2001), “Retaking the SAT,” Journal of Human Resources, Vol. 38, pp. 1-33.

Willingham, W.W., & Cole, N.S. (1997). Gender and fair assessment. Mahwah, NJ: Erlbaum.

Yoon T., M. Keller, B. Schinder De-Lap, A. Harkins, R. Lepers, and S. Hunter, (2009) “Sex Differences in Response to Cognitive Stress During a Fatiguing Contraction”, Journal of Applied Physiology, Vol. 107, pp. 1486-1496.

Table 1. Sample Selection Process

		Gender			Race/ethnicity				
	Total	Males	Females	Missing	Whites	Blacks	Hispanics	Asians	Other/ Missing
Population (all GRE tested 9/1/2001-10/31/2001)	81,231	34,723	41,617	4,891					
US citizens tested in the US	46,038	15,749	30,160	129	36,042	2,877	2,400	2,584	2,135
Experiment participants (total)	29,962	13,359	14,803	1,800					
US citizens tested in the US	15,945	5,486	10,458	1	12,374	1,024	850	982	715
Participants in regular time limit experiment	8,232	2,834	5,398	0	6,407	513	445	479	388
Participants in Q section	3,922	1,369	2,553		3,027	265	224	224	182
Participants in V section	4,310	1,465	2,845		3,380	248	221	255	206

Notes: The table reports the process we followed to select our analysis samples.

Table 2. Comparison Between Full Population of GRE Test Takers and Experiment Participants

A. By gender

	Males			Females		
	Full Sample	Participants		Full Sample	Participants	
		Q section	V section		Q section	V section
N	15,749	1,369	1,465	30,160	2,553	2,845
Share	0.34	0.35	0.34	0.66	0.65	0.66
Quantitative score						
Mean	55.8	55.6	56.8	40.7	40.3	41.2
S.D	26.7	27.4	27.0	23.9	24.4	23.9
Median	57	57	57	39	39	39
Verbal score						
Mean	64.1	62.4	62.9	57.0	56.2	56.5
S.D	24.5	25.0	25.0	24.8	25.0	24.5
Median	67	67	67	57	57	57

B. By Race/Ethnicity

	Whites			Blacks			Hispanics			Asians		
	Full Sample	Participants		Full Sample	Participants		Full Sample	Participants		Full Sample	Participants	
		Q section	V section		Q section	V section		Q section	V section		Q section	V section
N	36042	3027	3380	2877	265	248	2400	224	221	2584	224	255
Share	0.783	0.772	0.784	0.062	0.068	0.058	0.052	0.057	0.051	0.056	0.057	0.059
Quantitative score												
	46.8	47.0	47.4	24.6	21.9	24.7	36.5	36.4	38.4	63.0	62.3	64.3
S.D	25.0	25.5	25.2	21.8	21.8	21.2	24.9	25.3	26.1	25.4	26.8	24.9
Median	44	44	48	18	13	18	31	31	35	66	66	71
Verbal score												
Mean	61.5	60.6	60.5	37.8	35.7	37.4	47.6	48.8	48.7	62.0	61.5	60.8
S.D	23.6	23.8	23.7	24.1	23.2	24.2	26.0	26.8	26.2	26.8	27.1	26.8
Median	62	62	62	35	29	35	46	46	52	67	62	62

Notes: The table reports students performance (in percentile score ranks) of the full sample of GRE test takers and performance of experiment participants stratified by gender and race/ethnicity. The samples are restricted to US citizens tested in the US.

Table 3. Descriptive Statistics of Experiment Participants

	Males (1)	Females (2)	Whites (3)	Blacks (4)	Hispanics (5)	Asians (6)
Females			0.66	0.74	0.65	0.63
Race/Ethnicity						
Whites	0.78	0.78				
Blacks	0.05	0.07				
Hispanics	0.06	0.05				
Asians	0.06	0.06				
Mother's Education						
High School or less	0.23	0.22	0.21	0.33	0.40	0.24
College or some college	0.45	0.48	0.48	0.41	0.37	0.46
At least some graduate studies or professional degree	0.26	0.25	0.26	0.19	0.19	0.25
Father's Education						
High School or less	0.21	0.23	0.20	0.43	0.40	0.15
College or some college	0.40	0.44	0.44	0.38	0.33	0.39
At least some graduate studies or professional degree	0.37	0.32	0.35	0.16	0.25	0.45
Native English speaker	0.93	0.92	0.93	0.95	0.90	0.86
Undergraduate GPA						
C or C-	0.07	0.05	0.05	0.20	0.08	0.05
B-	0.11	0.10	0.10	0.18	0.13	0.07
B	0.30	0.33	0.32	0.36	0.37	0.36
A-	0.28	0.28	0.30	0.13	0.23	0.30
A	0.19	0.19	0.21	0.07	0.13	0.18
Undergraduate major in Physics, Math, Comp. Science or Engineering	0.26	0.05	0.12	0.10	0.12	0.31
Grad. intended studies in Physics, Math, Comp. Science or Engineering	0.25	0.05	0.11	0.07	0.13	0.30

Notes: The table reports descriptive statistics of participants in the regular time limit experiment. The samples are restricted to US citizens tested in the US.

Table 4. Performance in High and Low Stakes Tests by Gender

			High Stakes Score			Low Stakes Score			High Stakes - Low Stakes			
	N Males	N Fem.	Males	Females	Diff.	Males	Females	Diff.	Males	Females	Raw Diff.	Controlled Diff.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Quantitative Section	1368	2553	55.579	40.277	15.302	43.935	33.162	10.773	11.644	7.115	4.529	3.893
			(27.432)	(27.432)	(0.854)	(25.475)	(31.342)	(0.927)	(0.683)	(0.385)	(0.784)	(0.809)
Verbal Section	1465	2845	62.902	56.453	6.450	52.481	50.345	2.136	10.421	6.108	4.313	4.041
			(24.959)	(24.959)	(0.794)	(27.649)	(30.534)	(0.922)	(0.673)	(0.400)	(0.783)	(0.818)

Notes: The table reports students test scores in the high (columns 3-4) and the low stakes sections (columns 6-7) of the GRE test. Columns 5 and 8 report test scores gaps between males and females in the high and the low stakes section of the exam respectively. Columns 9 and 10 report differences in individual's performance between the high and the low stakes section. Column 11 reports the differential change in performance between males and females (col. 9 - col. 10). Column 12 reports the controlled difference between males and females after accounting for the following individual covariates: mother's and father's education, dummies for race/ethnicity, UGPA, undergraduate major, intended graduate field of studies, and disability status. Test scores are reported in percentile ranks. Robust standard deviations and standard errors of the differences are reported in parenthesis. Sample sizes are reported in columns 1 and 2.

Table 5. Performance in High and Low Stakes Tests by Gender and Examinee Characteristics

			High Stakes Score			Low Stakes Score			High Stakes - Low Stakes			
	N Males	N Fem.	Males	Females	Diff.	Males	Females	Diff.	Males	Females	Raw Diff.	Controlled Diff.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
A. Quantitative Section												
Undergraduate GPA												
C or C-	102	134	39.784	21.157	18.628	30.461	18.590	11.871	9.324	2.567	6.756	6.738
			(24.462)	(24.462)	(2.793)	(17.397)	(25.557)	(2.800)	(1.947)	(0.851)	(2.124)	(2.197)
B-	144	266	43.028	28.267	14.761	34.458	24.034	10.425	8.569	4.233	4.336	3.822
			(25.528)	(25.528)	(2.248)	(19.386)	(26.841)	(2.306)	(1.939)	(0.837)	(2.111)	(2.317)
B	426	855	48.962	36.063	12.899	38.418	29.958	8.460	10.545	6.105	4.439	3.182
			(25.942)	(25.942)	(1.415)	(23.056)	(28.660)	(1.486)	(1.152)	(0.613)	(1.305)	(1.346)
A-	393	717	63.237	46.815	16.422	51.438	37.756	13.682	11.799	9.059	2.740	3.360
			(24.906)	(24.906)	(1.524)	(27.150)	(31.765)	(1.812)	(1.273)	(0.823)	(1.516)	(1.596)
A	251	490	69.821	50.700	19.121	53.801	42.382	11.419	16.020	8.318	7.702	8.309
			(25.227)	(25.227)	(1.869)	(27.321)	(34.295)	(2.318)	(1.908)	(0.959)	(2.135)	(2.459)
Undergrad major in Physics, Math, Comp. or Eng.	340	122	77.674	70.574	7.100	65.515	64.369	1.146	12.159	6.205	5.954	4.512
			(18.191)	(18.191)	(2.024)	(25.909)	(31.265)	(3.161)	(1.596)	(2.167)	(2.689)	(2.846)
Grad intended studies in Physics, Math, Comp. or Eng.	362	132	78.644	69.955	8.689	65.870	63.295	2.575	12.773	6.659	6.114	4.950
			(17.321)	(17.321)	(1.935)	(27.074)	(31.352)	(3.078)	(1.549)	(2.121)	(2.624)	(2.667)
Maternal Education												
High School or less	320	582	43.903	32.973	10.931	35.581	27.038	8.543	8.322	5.935	2.387	2.132
			(26.374)	(26.374)	(1.687)	(23.117)	(27.255)	(1.716)	(1.235)	(0.672)	(1.405)	(1.475)
College or some college	621	1228	58.097	39.965	18.132	46.018	33.800	12.218	12.079	6.165	5.914	5.846
			(26.830)	(26.830)	(1.214)	(24.850)	(32.199)	(1.356)	(1.013)	(0.529)	(1.142)	(1.188)
At least some graduate studies or professional degree	357	619	63.588	48.724	14.864	49.952	39.069	10.883	13.636	9.654	3.982	2.938
			(25.921)	(25.921)	(1.689)	(27.697)	(32.106)	(1.953)	(1.455)	(0.929)	(1.725)	(1.824)

Table 5 (cont.). Performance in High and Low Stakes Tests by Gender and Examinee Characteristics

			High Stakes Score			Low Stakes Score			High Stakes - Low Stakes			
	N Males	N Fem.	Males	Females	Diff.	Males	Females	Diff.	Males	Females	Raw Diff.	Controlled Diff.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
B. Verbal Section												
Undergraduate GPA												
C or C-	106	161	48.689	38.441	10.248	43.208	35.435	7.773	5.481	3.006	2.475	2.451
			(23.915)	(23.915)	(2.864)	(24.116)	(26.541)	(3.140)	(2.036)	(1.513)	(2.536)	(3.374)
B-	167	275	53.695	47.949	5.746	46.144	44.447	1.696	7.551	3.502	4.049	3.261
			(26.025)	(26.025)	(2.389)	(25.274)	(27.002)	(2.545)	(1.719)	(1.129)	(2.056)	(2.401)
B	436	945	58.690	51.935	6.755	50.197	46.309	3.888	8.493	5.626	2.867	2.898
			(23.905)	(23.905)	(1.368)	(25.740)	(29.117)	(1.555)	(1.165)	(0.664)	(1.340)	(1.378)
A-	405	799	68.225	62.016	6.208	54.138	55.253	-1.115	14.086	6.763	7.323	7.228
			(22.888)	(22.888)	(1.405)	(27.634)	(32.032)	(1.780)	(1.391)	(0.793)	(1.600)	(1.676)
A	292	560	74.137	66.366	7.771	61.709	58.664	3.045	12.428	7.702	4.726	4.417
			(20.914)	(20.914)	(1.589)	(28.622)	(31.125)	(2.130)	(1.598)	(0.933)	(1.850)	(2.025)
Undergrad major in Physics, Math, Comp. or Eng.	378	142	66.341	66.056	0.285	53.643	60.535	-6.892	12.698	5.521	7.177	8.237
			(23.796)	(23.796)	(2.372)	(27.411)	(31.356)	(2.986)	(1.445)	(1.340)	(1.970)	(2.187)
Grad intended studies in Physics, Math, Comp. or Eng.	388	161	66.781	65.839	0.942	54.036	62.012	-7.976	12.745	3.826	8.919	8.982
			(24.124)	(24.124)	(2.296)	(25.708)	(31.769)	(2.824)	(1.424)	(1.301)	(1.929)	(2.048)
Maternal Education												
High School or less	344	628	54.302	49.244	5.059	45.959	45.051	0.908	8.343	4.193	4.150	3.924
			(26.892)	(26.892)	(1.679)	(25.717)	(29.148)	(1.810)	(1.305)	(0.745)	(1.502)	(1.556)
College or some college	658	1354	64.114	56.078	8.036	53.157	49.908	3.249	10.957	6.171	4.787	5.193
			(23.671)	(23.671)	(1.134)	(27.139)	(30.420)	(1.343)	(1.033)	(0.591)	(1.190)	(1.258)
At least some graduate studies or professional degree	376	731	68.830	63.848	4.982	58.495	56.791	1.704	10.335	7.057	3.278	3.825
			(22.931)	(22.931)	(1.504)	(28.787)	(30.521)	(1.865)	(1.318)	(0.827)	(1.556)	(1.640)

Notes: The table reports gender differences in performance in the low and the high stakes sections of the GRE test for different subsamples. Panel A reports results for experiment participants in the Q-Section Panel B reports results for experiment participants in the V-Section. Controlled differences in column 12 include the covariates detailed in Table 4. Test scores are reported in percentile ranks. Robust standard deviations and standard errors of the differences are reported in parenthesis. Sample sizes are reported in columns 1 and 2.

Table 6. Performance in High and Low Stakes Tests by Race and Ethnicity

					High Stakes Score							Low Stakes Score						
	N W	N B	N H	N A	Whites	Blacks	Hispanics	Asians	W-B	W-H	W-A	Whites	Blacks	Hispanics	Asians	W-B	W-H	W-A
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
A. Quantitative Section																		
Full sample	3,026	265	224	224	46.99	21.85	36.39	62.30	25.13	10.59	-15.32	37.55	18.90	32.58	55.20	18.65	4.97	-17.64
					(25.46)	(21.80)	(25.33)	(26.76)	(1.62)	(1.75)	(1.75)	(27.78)	(19.72)	(26.39)	(30.38)	(1.75)	(1.90)	(1.90)
B. Verbal Section																		
Full sample	3,380	248	221	255	60.55	37.37	48.73	60.84	23.18	11.82	-0.30	52.79	35.08	42.22	51.78	17.71	10.57	1.01
					(23.69)	(24.23)	(26.20)	(26.85)	(1.58)	(1.67)	(1.56)	(28.17)	(24.08)	(27.87)	(31.42)	(1.85)	(1.95)	(1.83)

Notes: The table reports students performance in the high and the low stakes sections stratified by race/ethnicity. Columns 9-11 report test score gaps in the high stakes section between Whites and Blacks/Hispanics/Asians respectively. Columns 16-18 report test score gaps in the high stakes section between Whites and Blacks/Hispanics/Asians respectively. Test scores are reported in percentile ranks. Robust standard deviations and standard errors of the differences are reported in parenthesis. Sample sizes for each race/ethnicity group are reported in columns 1-4.

Table 7. Differential Performance Between High and Low Stakes Tests by Race and Ethnicity

	High Stakes - Low Stakes				Raw Gap Relative to Whites			Controlled Gap Relative to Whites		
	Whites	Blacks	Hispanics	Asians	Blacks	Hispanics	Asians	Blacks	Hispanics	Asians
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
A. Quantitative Section										
Full sample	9.431	2.951	3.808	7.107	6.480	5.623	2.323	4.160	5.231	3.292
	(0.399)	(0.863)	(1.346)	(1.561)	(0.949)	(1.402)	(1.609)	(1.016)	(1.416)	(1.693)
B. Verbal Section										
Full sample	7.755	2.282	6.511	9.067	5.473	1.244	-1.312	3.080	0.326	-0.747
	(0.390)	(1.316)	(1.457)	(1.625)	(1.371)	(1.506)	(1.669)	(1.459)	(1.543)	(1.700)

Notes: Columns 1-4 differences in performance between the high and the low stakes section by race/ethnicity. Columns 5-7 report the differential change in performance between Whites and Blacks/Hispanics/Asians respectively. Columns 8-10 report controlled differences between Whites and Blacks/Hispanics/Asians respectively after accounting for the following individual covariates: mother's and father's education, dummy for female, dummies for UGPA, undergraduate major, intended graduate field of studies, and disability status. Test scores are reported in percentile ranks. Robust standard errors are reported in parenthesis.

Table 8. Performance in High versus Low Stakes Tests by Gender and Race/Ethnicity - Quantitative Section

	High Stakes		Low Stakes		High-Low Stakes		**Controlled Diff.**
	Males	Females	Males	Females	Males	Females	**(Males-Females)**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Whites	56.701	41.800	43.914	34.161	12.787	7.639	**4.767**
	(26.403)	(26.403)	(25.179)	(31.132)	(0.793)	(0.437)	**(0.937)**
Blacks	28.769	19.605	24.215	17.175	4.554	2.430	**0.475**
	(27.739)	(27.739)	(16.851)	(26.150)	(2.146)	(0.906)	**(2.306)**
Controlled Diff. (Whites-Blacks)					**5.803**	**3.491**	
					(2.385)	**(1.140)**	
Hispanics	44.022	31.363	38.405	28.748	5.618	2.615	**0.609**
	(27.048)	(27.048)	(23.230)	(29.775)	(2.422)	(1.561)	**(3.301)**
Controlled Diff. (Whites-Hispanics)					**7.539**	**4.182**	
					(2.601)	**(1.649)**	
Asians	72.167	56.386	66.071	48.671	6.095	7.714	**0.747**
	(23.589)	(23.589)	(29.090)	(29.509)	(2.603)	(1.955)	**(3.919)**
Controlled Diff. (Whites-Asians)					**9.412**	**-0.169**	
					(2.942)	**(2.052)**	

Notes: The table reports test scores in the Q-section of the GRE exam. Columns 1-2 report mean performance in the high stakes test for each gender-race/ethnicity cell. Columns 3-4 report mean performance in the low stakes test for each gender-race/ethnicity cell. Differences in performance between the high and the low stakes tests are reported in columns 5-7. Test scores are reported in percentile ranks. Standard deviations and robust standard errors are reported in parenthesis.

Table 9. Share of Experiment Participants who Spent Less than Ten Minutes in the Experimental Section

	Q-section			V-section		
	Share	H0	P-value	Share	H0	P-value
Share who spent less than ten minutes among	(1)	(2)	(3)	(4)	(5)	(6)
Gender						
Males	0.167			0.181		
Females	0.132	H0: Males=Females	0.0042	0.138	H0: Males=Females	0.0004
Race/ethnicity						
Whites	0.152			0.154		
Blacks	0.106	H0: Blacks=Whites	0.0196	0.101	H0: Blacks=Whites	0.0077
Hispanics	0.129	H0: Hispanics=Whites	0.3277	0.140	H0: Hispanics=Whites	0.5581
Asians	0.071	H0: Asians=Whites	0.0000	0.161	H0: Asians=Whites	0.7901
Maternal Education						
High School or less	0.134			0.133		
College or some college	0.134	H0: =High School or less	0.9990	0.155	H0: =High School or less	0.0990
At least some graduate studies or professional degree	0.163	H0: =High School or less	0.0790	0.157	H0: =High School or less	0.1130
Paternal Education						
High School or less	0.145			0.136		
College or some college	0.130	H0: =High School or less	0.2920	0.151	H0: =High School or less	0.2600
At least some graduate studies or professional degree	0.161	H0: =High School or less	0.3180	0.166	H0: =High School or less	0.0370
Undergraduate GPA						
C or C-	0.148	H0: No differences by		0.161	H0: No differences by	0.0218
B-	0.120	UGPA	0.1242	0.122	UGPA	
B	0.128			0.136		
A-	0.159			0.176		
A	0.151			0.155		
Achievement decile in high stakes test						
1	0.166	H0: No differences by		0.160	H0: No differences by	
2	0.147	achievement deciles	0.7360	0.092	achievement deciles	0.0011
3	0.128			0.103		
4	0.128			0.152		
5	0.153			0.174		
6	0.150			0.177		
7	0.132			0.170		
8	0.137			0.147		
9	0.166			0.169		
10	0.137			0.133		
Number of Observations	565			659		

Notes: Columns 1 and 4 report the share of examinees that spent less than 10 minutes in the experimental Q or V sections respectively out of their relevant group. Columns 3 and 6 report p-values that test for equality of the coefficients specified in columns 2 and 5 respectively.

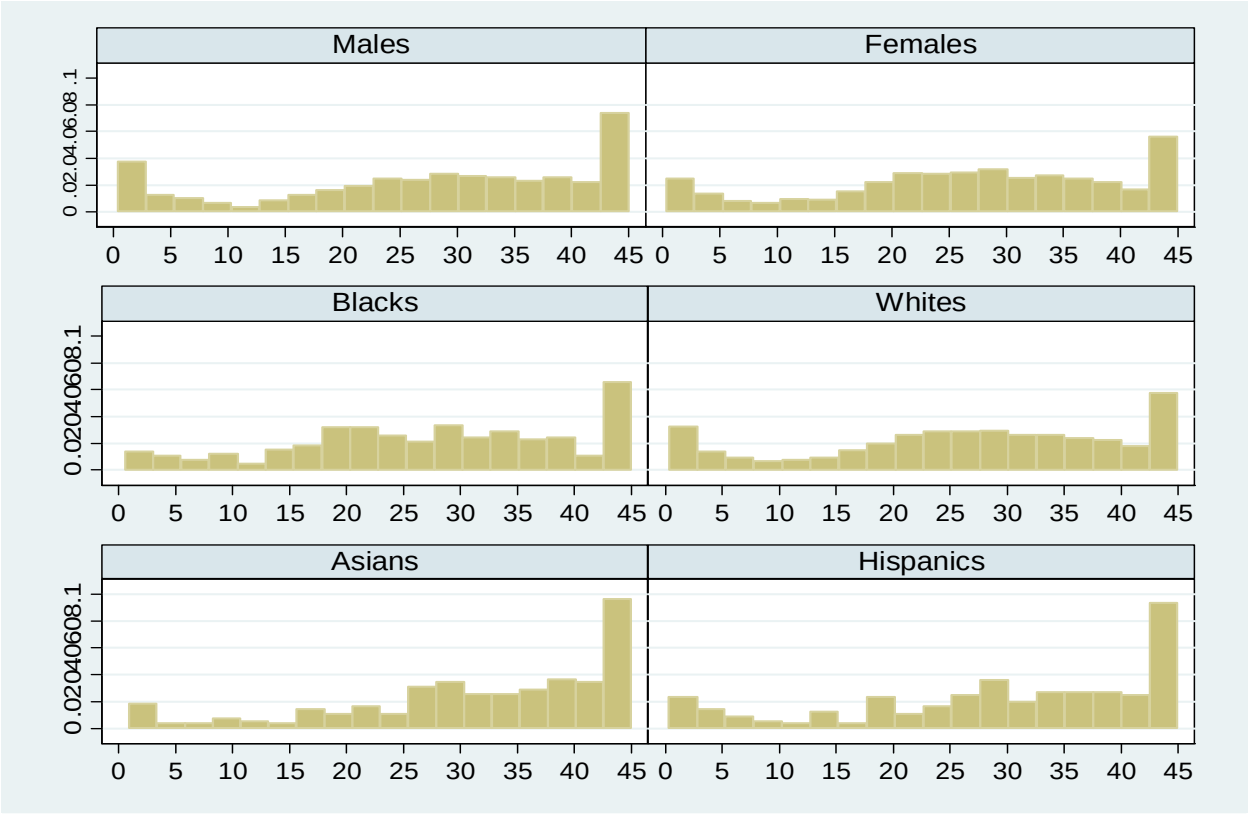
Table 10. Share of Experiment Participants who Improved their Score in the Low Stakes Section Relative to the High Stakes Section

	Q-section			V-section		
	Share	H0	P-value	Share	H0	P-value
Share who improved their score among	(1)	(2)	(3)	(4)	(5)	(6)
Gender						
Males	0.267			0.309		
Females	0.316	H0: Males=Females	0.0010	0.342	H0: Males=Females	0.0240
Race/ethnicity						
Whites	0.286			0.326		
Blacks	0.347	H0: Blacks=Whites	0.0450	0.423	H0: Blacks=Whites	0.0030
Hispanics	0.429	H0: Hispanics=Whites	0.0000	0.303	H0: Hispanics=Whites	0.4690
Asians	0.339	H0: Asians=Whites	0.1050	0.345	H0: Asians=Whites	0.5430
Number of Observations	1172			1426		

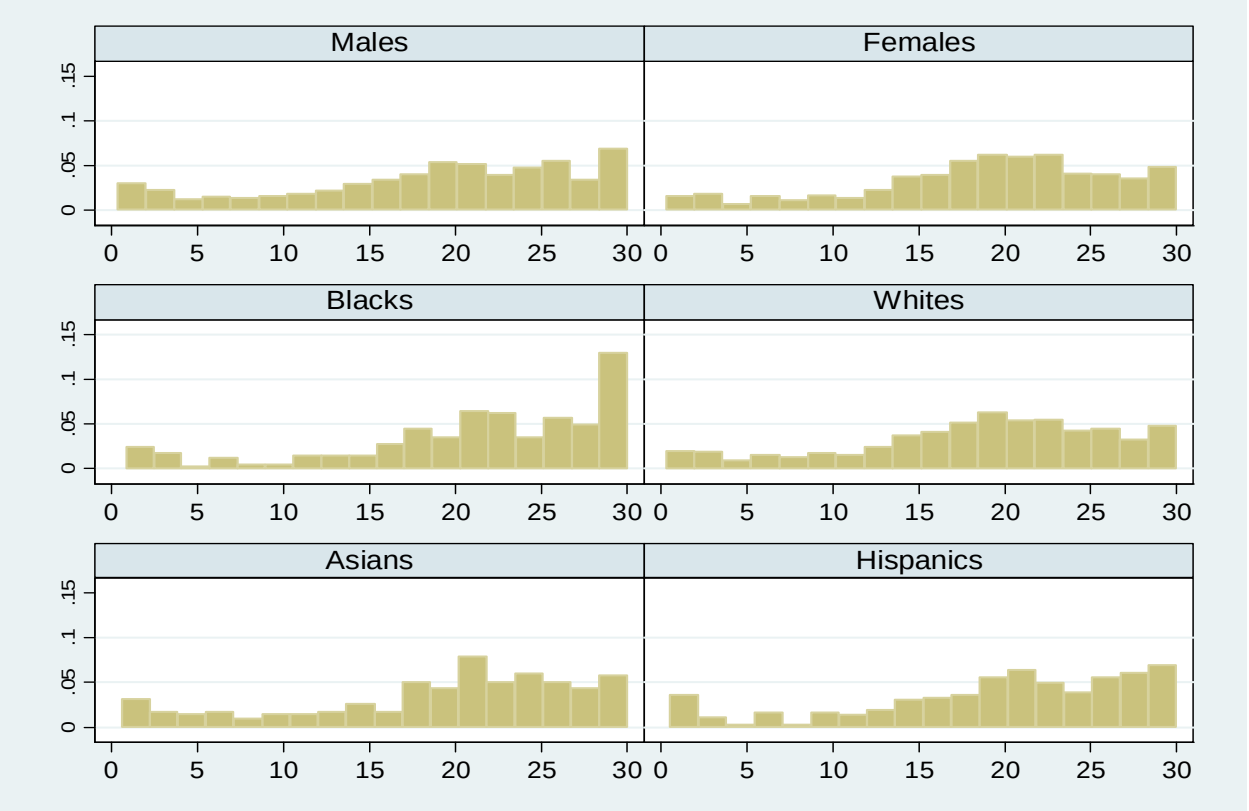
Notes: Columns 1 and 4 report the share of examinees who improved their score in the experimental Q or V sections respectively relative to the real GRE section. Columns 3 and 6 report p-values that test for equality of the coefficients specified in columns 2 and 5 respectively.

Figure 1. Distribution of Time Invested in Experimental Section

A. Quantitative Section



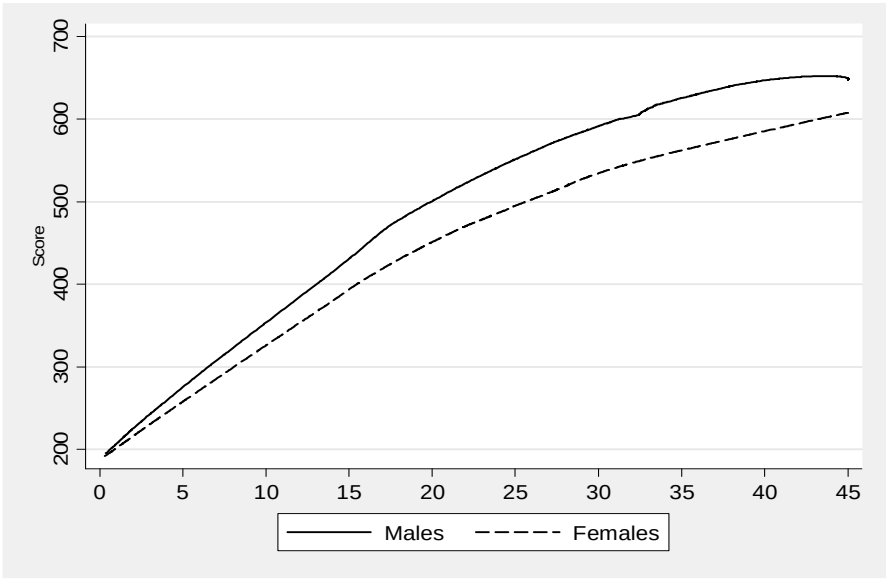
B. Verbal Section

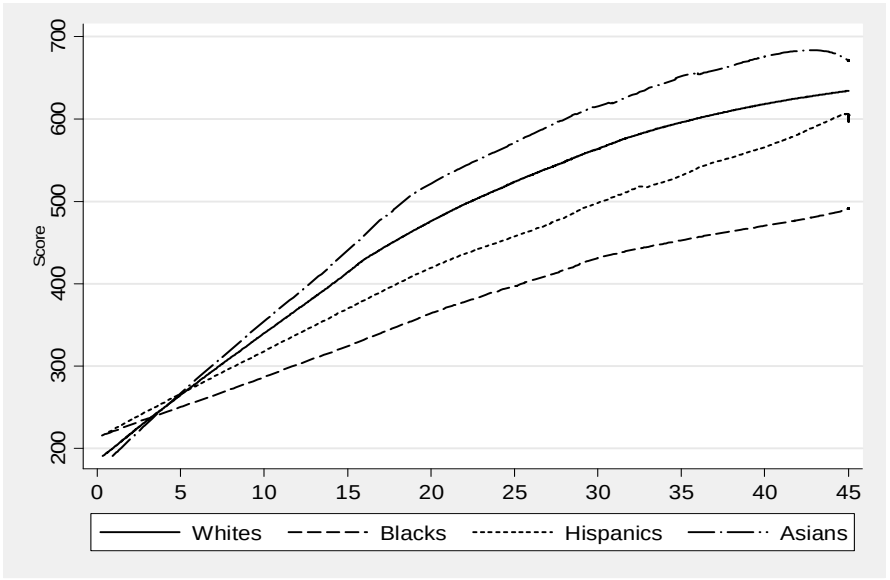


Notes: The figures plot the distribution of time (in minutes) spent in the experimental section by Gender and race/ethnicity. Panel A plots distributions for the Q-section and Panel B plots distributions for the V-section.

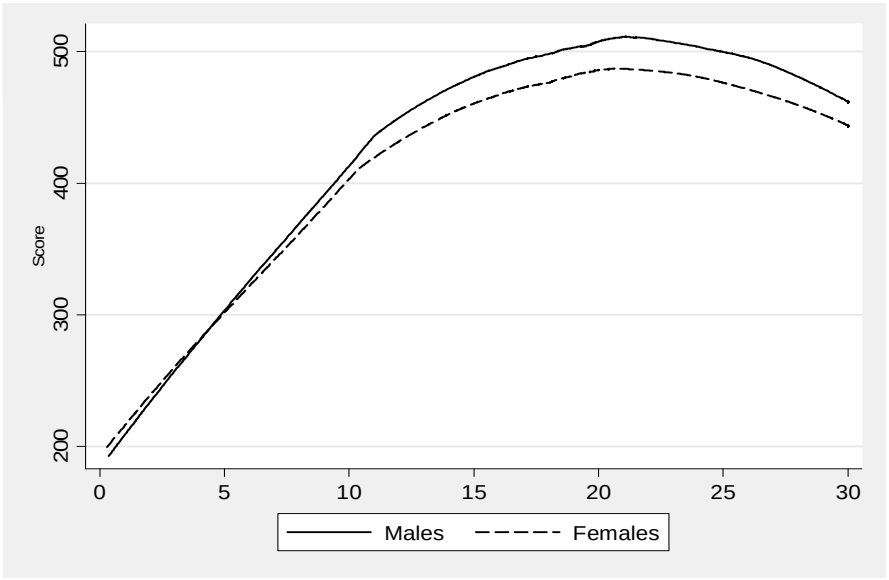
Figure 2. Relationship Between Time Invested in the Experimental Section and Test Score Achieved in that Section

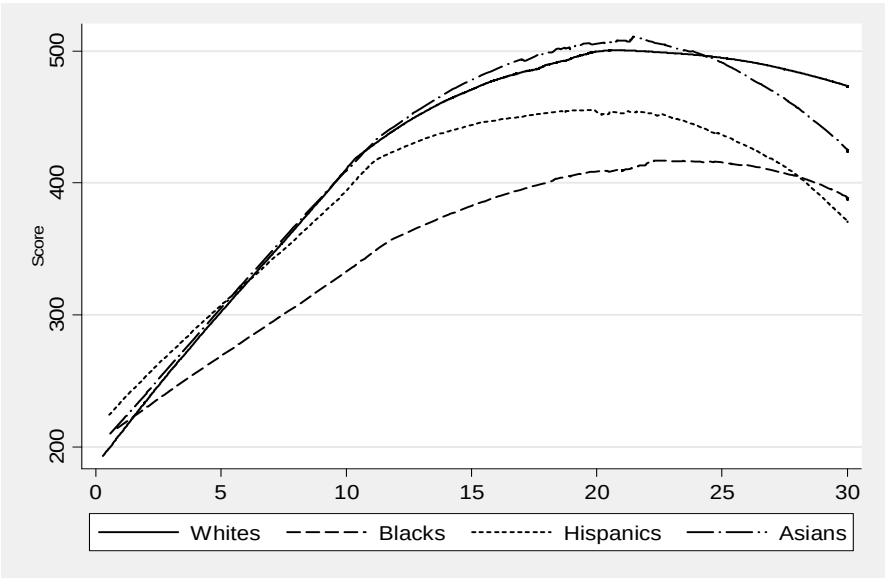
A. Quantitative Section





B. Verbal Section

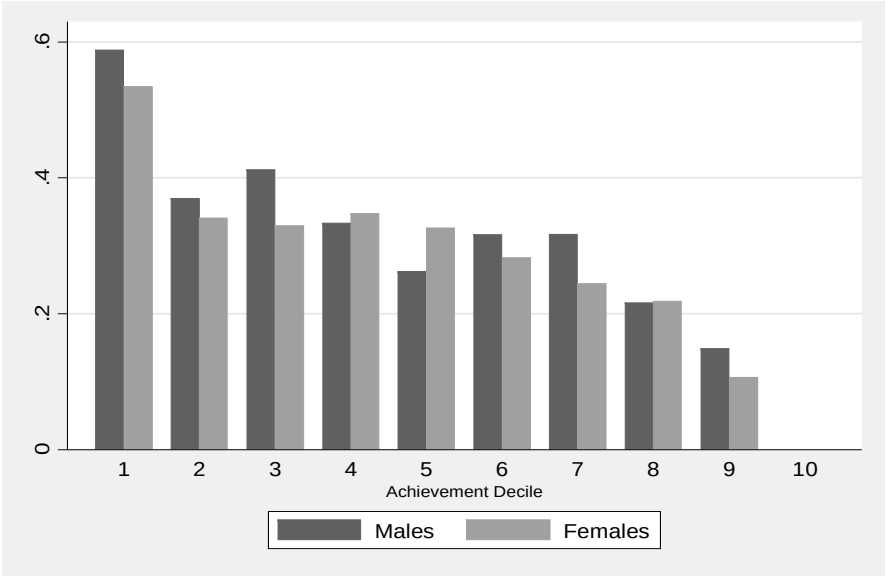


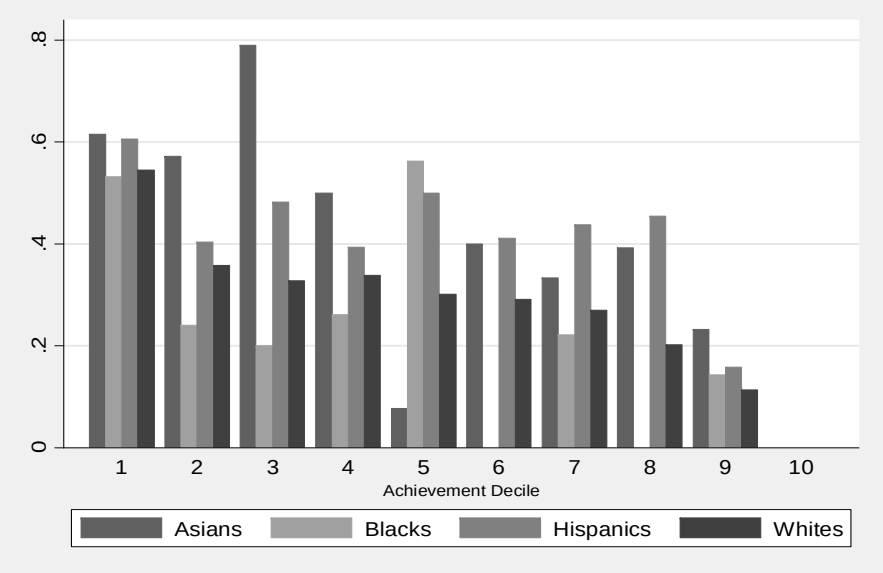


Notes: The figures plots local weighted regressions of score in the experimental section on time invested in that section. Plots are stratified by gender and race/ethnicity. Panel A plots figures for the Q-section and Panel B plots figures for the V-section.

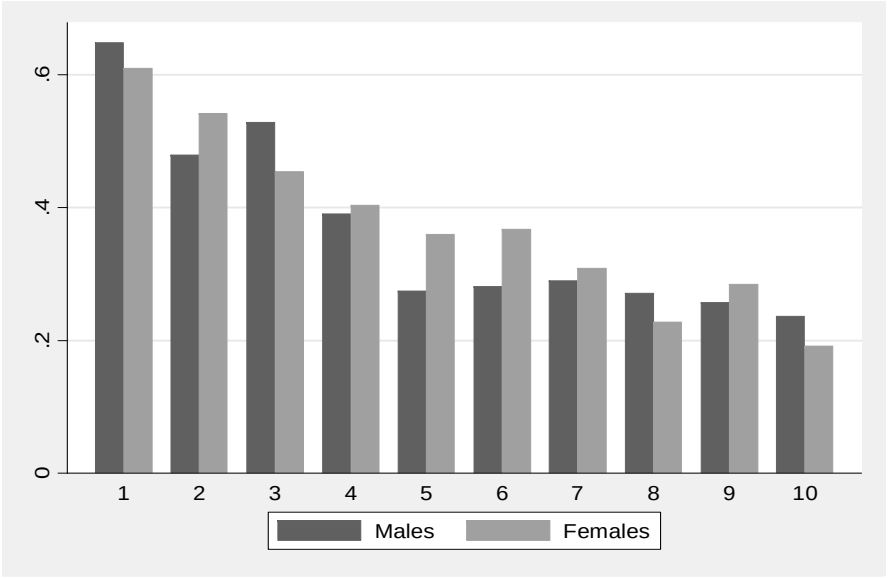
Figure 3. Share of Examinees who Improved their Score in the Experimental Section by Achievement Decile in Real GRE Section

A. Quantitative Section





B. Verbal Section



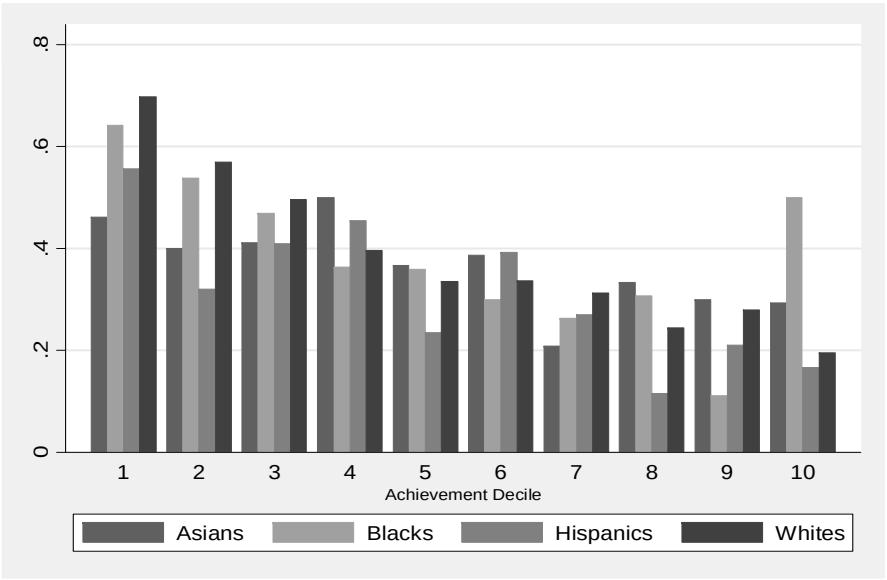
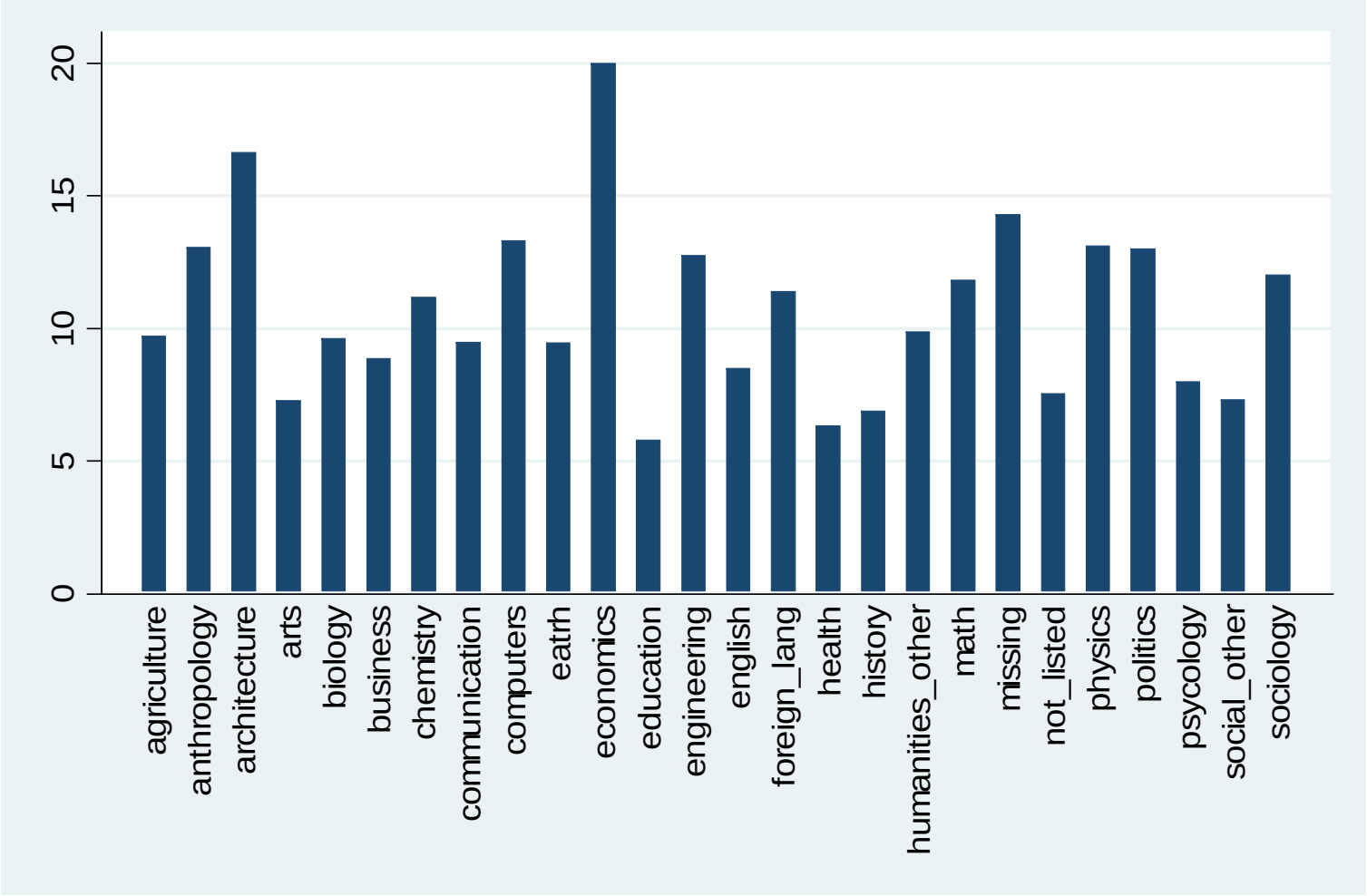


Figure 4. Performance Gap Between High and Low Stakes Test by Undegraduate Major: Q Section



Notes: The figure reports differences in performance between the high and the low stakes examination for students stratified by their undergraduate major.

Table A1. Performance Gap Between High and Low Stakes Section by Time Spent in Low Stakes Section

	Difference in individual performance between high and low stake test						Controlled difference between groups			
Sample	Males (1)	Females (2)	Whites (3)	Blacks (4)	Hispanics (5)	Asians (6)	Males-Females (7)	Whites-Blacks (8)	Whites-Hispanics (9)	Whites-Asians (10)
					A. Quantitative Section					
Full	11.644	7.115	9.431	2.951	3.808	7.107	3.893	4.160	5.231	3.292
	(0.683)	(0.385)	(0.399)	(0.863)	(1.346)	(1.561)	(0.809)	(1.016)	(1.416)	(1.693)
Time spent in experimental section ≥ 10 mins.	3.414	2.606	3.264	0.949	-1.236	2.832	1.060	2.049	4.368	0.252
	(0.421)	(0.289)	(0.271)	(0.644)	(0.895)	(1.153)	(0.554)	(0.769)	(0.977)	(1.207)
					B. Verbal Section					
Full	10.421	6.108	7.755	2.282	6.511	9.067	4.041	3.080	0.326	-0.747
	(0.673)	(0.400)	(0.390)	(1.316)	(1.457)	(1.625)	(0.818)	(1.459)	(1.543)	(1.700)
Time spent in experimental section ≥ 10 mins.	1.927	0.848	1.337	-1.529	1.879	1.607	0.997	2.196	-0.748	-0.076
	(0.429)	(0.287)	(0.260)	(1.015)	(1.199)	(1.153)	(0.555)	(1.100)	(1.240)	(1.214)

Notes: The table reports differences in performance between the high and the low stakes tests by gender and race/ethnicity. Panel A reports differences in the Q-section and panel B reports differences in the V-section. The first row of each panel reproduce results reported in tables 4 and 7. The second row of each panel reports results for the subsample of examinees who spent less than 10 minutes in the experimental section. Test scores are reported in percentile ranks. Standard deviations and robust standard errors are reported in parenthesis.