Rise to the Challenge or Not Give a Damn:

Differential Performance in High vs. Low Stakes Tests

Yigal Attali
Educational Testing Service
Rosedale Rd.
MS-16-R
Princeton, NJ 08541
USA
Voice: 609-734-1747
Fax: 609-734-1755
e-mail: yattali@ets.org

Zvika Neeman
The Eitan Berglas School of Economics
Tel Aviv University
P.O.B. 39040
Ramat Aviv, Tel Aviv, 69978
ISRAEL
Office: +972-3-6409488
Fax: +972-3-6409908
e-mail: zvika@post.tau.ac.il

Analia Schlosser
The Eitan Berglas School of Economics
Tel Aviv University
P.O.B. 39040
Ramat Aviv, Tel Aviv, 69978
ISRAEL
Office: +972-3-6409064
Cel:+972-54-4902414
Fax: +972-3-6409908
e-mail: analias@post.tau.ac.il

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]

Abstract

We study how different demographic groups respond to incentives by comparing their performance in “high” and “low” stakes situations. The high stakes situation is the GRE examination and the low stakes situation is a voluntary experimental section that examinees solved after the GRE. Males exhibit a larger difference in performance between the high and low stakes examinations than females, and Whites exhibit a larger difference in performance relative to Asians, Blacks, and Hispanics. The larger differential performance between high and low stakes tests among men and whites is partially explained by lower effort invested in the low stake test.

[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 demographic groups respond differently to incentives and cope differently with competitive pressure. Interest in this subject stems from attempts to explain gender, racial, and ethnic differences in human capital accumulation and labor market performance. More practically, interest in the effects of incentives and competitive pressure on performance is also motivated by the increased use of aptitude tests for college admissions and job screening and the growing use of standardized tests for the measurement of school advancement and the assessment of student's learning.

While it is clear that students' motivation affects performance, less attention has been given to differences in test-taking motivation across demographic groups or group differences in response to performance based incentives or what is at stake in a given test. 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).[1] 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 situation is the real GRE examination and the low stakes situation 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 situations and we can compare performance in the exact same task. This allows us to examine differences in performance based on comparisons that involve the same individual rather than

[1] 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/.

comparisons between groups. A third unique feature of our study is the availability of a rich data on individuals' characteristics that includes information on family background, college major and academic performance, and intended graduate field of studies. These comprehensive data allow us to compare individuals of similar academic and family background and examine the persistence of our results across different subgroups. A fourth important advantage 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. 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. This finding is important as it shows that our results are unlikely to be driven by differential selection into the experiment.

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 compared 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.

We find that group differences in performance change between high and low stakes tests appear 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).

We also go a step further and explore various alternative explanations for the differential response to incentives across demographic groups and show that the higher differential performance of males and whites between the high and the low stakes test is partially explained by lower levels of effort exerted by these groups in the low stakes situations relative to women and minorities. We do not find evidence supporting alternative explanations such as test anxiety or stereotype threat.

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, 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.

Figure 1 summarizes our main finding. We ranked individuals according to their performance (from best, ranked “1”, to worst) both in the high stakes test and in the low stakes test. For each examinee we calculated the difference in his/her ranking between the high and the low stakes test. The figure plots the distribution of rank difference by gender and race. Panels (a) and (b) show that females are more likely to move up in their ranking while males are more likely to move down when switching from the high to the low stakes test. On average the men’s ranking declines by 164 and 152 positions in the low stakes test relative to the high stakes test in both the Q-section and the V-section, respectively. In contrast, the women’s ranking improves, on average, by 96 and 74 positions, respectively. The rank change of males and females are statistically different (p-values of Mann-Whitney tests <0.0001) both in the Q- and the V-sections. Similarly, we see in panels (c) and (d) that the relative ranking of whites declines while the relative ranking of minorities improves when switching from the high to the low stakes test in both the Q- and the V-sections. The change in the rank of whites and minorities is statistically different (p-values of Mann-Whitney tests <0.0001).

Our findings that differences in performance between individuals could vary according to the level of incentives and what is at stake suggest that the quality of a match between a worker and a job would not only depend on worker’s ability but also on his/her differential performance according to the incentive scheme attached to the job.

The rest of the paper proceeds as follows. In the next section we review the related literature. In Section 3 we describe the experimental setup and data. We present the empirical framework in Section 4. In Section 5 we present the results and in Section 6 we discuss alternative possible explanations for our findings. Section 7 concludes.

2. Related Literature

The experimental literature in economics contains many examples that demonstrate that incentives affect individuals’ performance. In recent years, much attention has been given to the question of whether response to incentives varies across individuals, with a particular focus on differences by gender. Surprisingly, differences in response to incentives by race and ethnicity received little attention. 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; Niederle et al,. 2008; Dohmen and Falk, 2011, and additional references in the comprehensive review of Niederle and Vesterlund, 2010). Recent studies, however, (e.g., Gunther et al., 2010 and Cotton et al., 2010) find that gender gaps in competitive performance depend crucially on the type of competition and number of interactions. Another recent study also shows that task stereotypes and time constraints can also affect the performance gap between men and women (Shurchkov, forthcoming). A few studies have investigated whether these gender differences are socially constructed or innate (Gneezy et al., 2009, Booth and Nolen, 2011; Booth and Nolen, 2012).

Most of the evidence on gender differences in competitive behavior and response to incentives is based on laboratory experiments. The extension of these findings to real world situations is limited to a small number of recent studies and remains an important empirical open question. Paserman (2010) studies performance of professional tennis players and finds that performance decreases under high competitive pressure but this result is similar for both men and women. Similarly, Lavy (2008) finds no gender differences in performance of high school teachers who participated in a performance-based tournament. On the other hand, in a recent field experiment among administrative job seekers Flory et al. (2010) find that women are less likely to apply to jobs that include performance based payment schemes but this gender gap disappears when the framing of the job is switched from being male- to female-oriented.[2]

An opportunity to observe individuals' performance at different incentive levels occurs in the case of achievement tests in schools and admission tests into universities and colleges. A number of studies within the educational measurement literature demonstrate that high stakes situations induce stronger motivation and higher effort.[3] 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 they differ in their perception of the importance of the test or in their motivation to perform well.[4]

[2] Other studies include Paarsch and Shearer (2007), Jurajda and Munich (forthcoming) and Ors, et al. (2008).

[3] 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 on the effects of incentives and test taking motivation see O'Neil, Surgue, and Baker (1996).

[4] Several studies (e.g., Barres, 2006; Duckworth and 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).

3. Experimental Set-up and Data

We use 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.[5]

Participants in the experiment were randomly assigned into one of four groups: one group was administered a quantitative section (Q-section) with standard time limit (45 minutes), a second group received a verbal section (V-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 one's own achievement in the high stakes section rather than on absolute performance, incentives to perform well in the experimental section were similar for all participants in the experiment.

Appendix Table A1 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 a more heterogeneous population and to control for a similar testing environment. In addition, we want

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

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 received either an extra Q-section (3,922) or an extra V-section (4,310).[6] We select only experiment participants who were randomized into the regular time limit experimental 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.[7]

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 1 compares the characteristics of the full sample of US GRE test takers and the sample of experiment participants.[8] 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.[9]

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 which 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. 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

[6] 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 highly close to that.

[7] 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.

[8] 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.

[9] 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.

distributions within each race/ethnicity. Overall, results presented in Table 1 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.[10]

GRE test takers are required to fill out a form upon registration to the exam. The form collects information on basic background characteristics, college studies, and intended graduate field of studies.[11] Appendix Table A2 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 A2 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" and 28 percent of both groups scored "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 A2 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).

[10] While we do not find differences in observable characteristics, there could still be differences in unobserved characteristics. Nevertheless, for the purpose of our study, we should worry about differential selection into the experiment by unobservables across demographic groups. The fact that we did not find evidence for differential selection across groups according to observables hints that the presence of large differences in selection by unobservables across groups is very unlikely.

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

4. Empirical Framework

To examine the change in individuals' performance between the high and the low stakes test across groups, we estimate the following first difference equation for each of the experimental samples (i.e. individuals randomized to the experimental Q or V section):

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

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* 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.[12] Whites and males are the omitted categories. The coefficients of interest are $\beta_1, \beta_2, \beta_3, \beta_4$ that denote the difference in performance gap between the high and 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.[13]

GRE scores in the quantitative and verbal sections range between 200 and 800, in 10-point increments. To ease the interpretation of the results, we transformed these raw scores into percentile

[12] Race/ethnicity categories in the GRE form are self-exclusive (i.e., it is not possible to check more than one option).

[13] An alternative approach is to estimate a conditional model that regresses the score in the low stakes test on the score in the high stakes test. The score change model described in equation (1) and the conditional regression model both attempt to adjust for baseline outcomes but they answer different questions. The score change model examines how groups, on average, differ in score changes between the high and the low stakes test. The conditional regression model asks whether the score change of an individual who belongs to one group differs from the score change of an individual who belongs to another group under the assumption that the two had come from a population with the same baseline level. The two approaches are expected to provide equivalent answers when the groups have similar baseline outcomes. However, as discussed by Cribbie and Jamieson (2000), when baseline means differ between groups, conditional regression suffers from directional bias. Namely, conditional regression will augment differences when groups start at different levels and then remain parallel or diverge (see Lord's Paradox - Lord, 1967) and will attenuate differences when groups start at different levels and then converge. Because the demographic groups we examine have different baseline GRE performance, we choose to estimate models of score change.

ranks using the GRE official percentile rank tables.[14] All results presented below are based on GRE percentile ranks. As we show later, we obtain similar results when using raw scores or log of raw scores.

5. Results

5.1. Gender Differences in Performance

Table 2 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. Similarly to other comparisons of GRE scores by gender, males outperform females in both the quantitative and verbal sections among the participants in our experiment. On average, Males are placed about 15.3 percentile points 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, which show that males' performance drops by 11.6 percentile points from the high to the low stakes Q-sections while females' performance drops by only 7.1 points. The differential gap in performance between males and females is 4.5 percentile points (s.e.=0.784). That is, a switch from the high to the low stakes situation narrows the gender gap in the quantitative test by about 4.5 percentile points, which is equivalent to 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 background characteristics and academic ability. This finding is important as it suggests that our results are unlikely to be driven by differences in family background and ability.

We also find the same pattern when examining changes in individual's performance between the high and low stakes V-sections. Males' scores drop by 10.2 percentile points, on average, while females' scores drop by a smaller magnitude of 6.1 percentile points. That is, males' scores drop by 4 percentile points (s.e.=0.783) more relative to females. Interestingly, despite the fact that the gender

[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

[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.

gap in the high stakes V-section is smaller than in the Q-section, we find that the differential change in performance between males and females in both sections is of a similar magnitude. However, in this case, the reduction in the gender gap is bigger; namely, the gender gap in verbal scores is reduced by two thirds when moving from the high stakes to the low stakes situation. Note that the largest drop in performance between the high and the low stakes section observed among men is not only evident in absolute terms but also when measured relative to the outcome mean. That is, we see that males' scores drop by 21 percent while females' scores drop by 18 percent in the Q-section. Similarly, we find that males' scores in the V-section drop by 17 percent while females' scores drop by 11 percent.

Table 3 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. Panel A reports results for the Q-section and panel B reports results for the V-section. Rows 1 through 5 in both panels present estimates for the samples stratified by UGPA. As expected, we observe a positive association between UGPA and GRE performance. Students with higher UGPA have higher 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, we observe that the gender gap in performance is narrower in the low stakes section in each of the cells stratified by UGPAs and is even 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).[16] Interestingly, the larger drop in males' performance relative to females is found across all ability levels (see columns 11 and 12) and is evident both in absolute and percentage terms relative to the mean outcome.

The next two rows of Table 3 (in both panels A and B) 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

[16] We use UGPA to stratify the sample by academic ability (instead of using the score in the high stakes section) as this is an independent measure of performance that is not related to the dependent variable.

driven to achievement.[17] 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). It is therefore interesting to examine whether we find the same pattern of gender differences in response to change in incentives in a subsample where selection by gender goes in the opposite direction.

As seen in columns 3 and 4 of table 3, 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 (8.7 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 even narrower (7.1 percentile points) although still significant. 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 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 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.

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 largest drop in males’ performance is evident both in absolute terms and relative to the outcome means in the high stakes test. 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. This

[17] We focus here in a more limited number of fields than the traditional STEM definition (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 STEM fields.

finding is important as it shows that the larger drop in performance among men is found even in subsamples that exhibit no differences in performance in the high stakes test.

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 4 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.[18]

[18] We suspect that the different pattern obtained for Asians and Hispanics in the V-section could be related to language dominance.

Table 5 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). 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). But Hispanics and Asians exhibit a similar drop in performance to that of 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.[19]

Appendix Table A3 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-section 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 that behave similarly to 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

[19] It is worth noting that the conclusions described in this subsection rely on samples that are stratified by gender and race/ethnicity and that are relatively small for Blacks, Hispanics, and Asians.

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, but the difference is not statistically significant.

5.4 Robustness of the Results

In Appendix Table A4 we show that our results are robust to nonlinear transformations of the dependent variable and that they also emerge in additional samples. In the first row of panels A and B, we report differences in performance in the quantitative and verbal sections using raw scores (scaled between 200 and 800). In the second row of each panel, we show differences in performance using the natural logarithm of raw scores in order to measure changes in performance in percentage terms. Both alternative metrics yield results that are equivalent to our main findings. Furthermore, as we show in Figure 1, we obtain the same results when we rely only on the ordinal information embedded in scores. Namely, when we compare examines' change in ranking between the high and the low stakes sections. Overall, these additional results show that our findings are not driven by a specific scale used to measure achievement.

The third row of each panel replicates our main results using the samples of examinees randomized into experimental sections with extended time limit (67.5 minutes for the Q-section and 45 minutes for the V-section). Estimates are similar to our main results showing that our findings are replicable in additional settings.

6. Discussion

The evidence presented above shows that men and Whites exhibit a larger differential performance between high and low stakes tests compared to women and minorities. The larger decline in performance found among men and whites can be due to at least three different reasons: (i) men and Whites simply “don't give a damn” in low stakes situations compared to women and minorities, respectively;[20] (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. We examine below the plausibility of these alternative explanations and discuss some other interpretations. We acknowledge that our data do not

[20] An equivalent argument is that women and minorities are more conscientious and exert relatively more effort in low stakes situations.

allow us to rigorously test the relative contribution of each explanation. Nevertheless, we believe the evidence presented below provides interesting insights that would likely motivate additional research.

6.1 Do Men and Whites Exert Less Effort in Low Stakes Situations?

To examine the likelihood of the first explanation, we would ideally like to measure effort invested in the test. More effort could be exerted by trying harder to solve each question (i.e., investment of more mental energy) or by investment of more time. Figure 2 plots the distribution of time spent by examinees in the experimental Q and V-sections by gender, race, and ethnicity.[21] 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).

Figure 3 exhibits 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. Overall, 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 exerting much effort. We define an indicator of low effort for individuals who invested less than ten minutes in the experimental section. While the ten minutes cutoff is somewhat arbitrary, we choose a time threshold that clearly suggests low effort and cannot be confounded by the ability to solve a test speedily.[22]

We plot in Figure 4 the cumulative test score distribution in the high stake section of subjects who invested low effort in the experimental section and compare it to the equivalent score distribution of subjects who invested some reasonable amount of time in the experimental section. We compare the test score distribution for low effort individuals and the rest within each demographic group and in both the Q- and the V-sections. Each quadrant in the figure refers to a specific demographic group and section. We also report p-values of Kolmogorov-Smirnov tests of equality between the two distributions and p-values of t-tests of equality of means (assuming unequal variances).

[21] Unfortunately, there is no information on time spent on the high stakes sections. Nevertheless, students usually exhaust the time limit.

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

For the quantitative section (panels a through d), we see no differences in the high stakes test score distribution between subjects who invested low effort in the experimental section and those who invested some reasonable amount of time. Indeed we cannot reject the hypothesis of equality of distributions or equality of means for each demographic group. This finding is important as it shows that achievement in the high stakes section is unrelated to effort levels invested in the low stakes section suggesting that subjects were not playing strategically in the low stakes section. This finding also shows that baseline differences in achievement in the high stakes section between demographic groups are unlikely to explain group differences in effort levels.

For the verbal section (panels e through h) we see no differences in test score distributions or means between those who invested low effort and the rest among males. We see some differences in the test score distribution for females (p-value of K-S test=0.04). Nevertheless, differences in the distribution derive from differences in the dispersion around the mean, with a larger variance among those investing low effort. Indeed, we cannot reject the hypothesis of equality of means between the two groups (p-value=0.931). For minorities we see evidence pointing to lower effort levels among those with lower scores in the high stakes section (although the difference in distributions is not statistically significant). These differences go against what we would expect if experiment participants were considering the monetary incentive when deciding about effort levels in the low stakes test.[23] Nevertheless, as discussed above, language difficulties might have affected performance of minorities in the verbal section so we prefer not to put too much weight in the comparison of performance between whites and minorities in this section.

Taken together, the evidence presented in Figure 4, suggests that effort exerted by individuals in the experimental section is not related to performance in the "real" GRE test across all demographic groups in the Q-section and among males, females, and whites in the V-section.

Table 6 reports the characteristics of individuals who invested less than ten minutes in the experimental section. Columns 1 and 2 of the table report the share of examinees who invested less than 10 minutes in the experimental Q- and V-sections stratified by gender, race/ethnicity, academic achievement, and parental education. We also report p-values that test for equality of coefficients between groups.

The results clearly show that males appear to exert less effort in the experimental section compared to females. 17 percent of the males who participated in the Q-experiment spent less than ten

[23] Individuals with higher scores in the high stakes section might find it more difficult to achieve the same result in the low stakes section. So, if they were playing strategically, we would expect to find lower effort levels among high achievers.

minutes trying to solve the experimental section while the equivalent among females is 13 percent. Gender differences are similar for the V-section. It is important to recall that, as shown in Table 1, the share of males and females among experiment participants was equal to their share in the full population of GRE test takers. This suggests 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 exerted according to students' parental education. Although differences are relatively small, it seems that students with more educated parents are more likely to invest less in the exam. In contrast, we find no clear relationship between the likelihood of low effort and students' abilities, neither when defined by students' scores in the high stakes section nor when defined by students' UGPAs. This last finding is important as it shows that the decision to exert low effort in the low stakes section is unrelated to students' ability, suggesting that noncognitive skills are likely to play a more important role in determining performance in low stakes situations. The lack of a relationship between students' 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 who exert very low effort? To examine this, we reproduce our main results of Tables 2 and 5 while limiting the sample to individuals who invested some minimal amount of time in the experimental section. Appendix Table A5 reports differences in performance between the high and the low stakes test for the sample of individuals who spent at least ten minutes in the experimental section. We also re-do the analysis while limiting the sample to individuals who spent more than three minutes in the experimental section as we can see in Figure 2 that some examinees (about 6.7 percent) left the experimental section shortly after it started achieving very low scores. 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. The gap between males and females and between Whites and minorities is also reduced, but to a lower extent, when the sample is limited to those who invested more than three minutes in the experimental section. Nevertheless, we still observe in both cases a larger gap in performance for males and Whites relative to females and minorities.

Finally, in the fourth row of panels A and B in Table A5, we report estimates from a model that uses the full sample and controls for a fourth order polynomial of time invested in the low stakes section. Again, we observe that differences between groups are reduced when accounting for time spent in the experimental section. Nevertheless, we see that the gap in differential performance between males and females and between whites and blacks or Hispanics is still sizable and significant. Note that while we use time invested in the low stakes section as a proxy for effort, we do not observe mental effort, a factor that might explain the remaining differences in performance change between groups.

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.

6.2 Are Women and Minorities More Subject to Stress in High Stakes Situations?

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 a higher level of stress and test anxiety among females and minorities that hinders their performance in high stakes situations. To examine 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 students do 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. We therefore adjust for score volatility and compare the share of examinees who improved their performance across demographic groups.

Columns 1 and 6 of table 7 report the share of examinees who improved their scores in the quantitative and in the verbal experimental sections. To adjust for score improvement due to score volatility and measurement error, we define a score gain for cases where the difference between the low-stakes score and the high-stakes score divided by the conditional standard error of measurement of difference scores is greater than 1.65.[24] Roughly 1.5 percent of examinees have a significant score gain in the experimental Q-section and 5.3 percent in the V-section. Columns 2 through 5 and 7 through 10 report differences in the share of examinees who improve scores by gender and by race/ethnicity. The first row reports raw differences between groups and the second row reports differences after controlling for students' background characteristics. Both sets of estimates show no gender differences in the likelihood of improving performance between the high and the low stakes section. The analysis by race/ethnicity shows no differences in improvement rates between whites and Asians or Hispanics. The comparison between whites and blacks shows no differences for the quantitative section but at positive difference favoring blacks in the verbal section.

We further explore the differential impact of test anxiety across groups using an alternative approach that takes advantage of additional information reported by examinees in the background questionnaire. The questionnaire asked examinees to report the reason(s) for taking the GRE test, allowing them to mark various alternatives. About 7 percent marked "practice" as one of the reasons for taking the exam.[25] If test anxiety hinders performance of females, blacks or Hispanics relative to males or whites in the high stakes section, we expect to find smaller performance gaps between groups among those who are taking the test for practice purposes.[26] To examine this, we regressed test scores in the "real" GRE sections on demographic group indicators, an indicator for practice exam and an interaction between practice exam and demographic group indicators while controlling for examinees' characteristics. To increase precision, we compare gaps in the quantitative and verbal "real" sections using all examinees irrespective of the section into which they were randomized in the experimental section.

We report in Table A6 estimates of the gaps in the "real" GRE sections between males and females and between whites, blacks, Hispanics, and Asians and the interaction terms between practice test and each demographic group. The analysis by gender shows no differences in the gender gap in the "real" GRE section between those taking the test for admission to graduate school and those taking the

[24] We use the conditional standard error of measurement of difference scores reported in Table 6b of the official ETS publication and define an indicator for score improvement following the ETS definition of significant GRE score differences (see ETS, 2007).

[25] The main reasons were of course, admission to graduate school (96%) and graduate department admissions requirement (29%). Other reasons include fellowship/scholarship application requirement (23%), undergraduate program exit requirement (1%), and other (3%). Note that applicants were instructed to select all reasons that apply, so that reasons do not add up to 100. The background questionnaire is filled by examinees before the test so it is not affected by their performance.

[26] Students who took the exam for practice might be different from those who took the exam for university admission. However, for the purpose of our comparison, we only need to assume that selection works in a similar direction for all demographic groups.

test for practice. Similarly, we see no differential gaps in the analysis by race/ethnicity. Namely, all interaction terms between practice test and demographic groups are small and insignificant proving that that gender and race/ethnic gaps in the “real” GRE section are similar between regular examinees and those who are taking the exam for practice.

Taken together, evidence presented in Tables 7 and A6 suggest that test anxiety in the high stakes section is unlikely to explain the smaller change in performance between the high and the low stakes tests observed among females and minorities.

6.3. Are Men and Whites Better Able to Rise to the Challenge in High Stakes Situations?

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 their performance when facing a high stakes or challenging task. This explanation is harder to assess as it is impossible to establish an ability baseline that is independent of performance in a given test of a given stake. It is challenging to even conceive of a thought experiment that could possibly answer this question because performance always depends on the perceived importance of the test.

6.4. Other Explanations

An additional explanation for our results could be that the monetary prize offered to experiment participants had a differential impact on different demographic groups. While this is feasible, we note that the prize consisted of $250 (1.5 times the GRE cost) paid to 100 individuals out of 30,000 experiment participants. Such amount distributed to such a small number of participants seems too low to have a significant differential effect in performance. Alternatively, one can argue that differences in performance in the experimental section could arise from group differences in their opportunity cost of time. However, as shown in Table 1, participation rates in the experiment were similar across demographic groups, suggesting that there were no group differences in the perceived cost or benefit of participating in the experiment.

To further assess the impact of the monetary prize and the opportunity cost of time on performance in the experimental section, we examined the association between the change in performance (from the high to the low stakes section) and earning levels at the state of residence of the examinee. We use two different measures of earnings: median annual earnings of full time workers and median annual earnings of college graduates computed separately by gender and state.[27] If the monetary prize or the opportunity cost of time had any impact on performance at the experimental section, we should expect a smaller reduction in performance in states with lower earnings levels. We report in Appendix Table A7, regression estimates for the association between the change in performance and median earnings for males and females. Columns 1 and 3 report estimates from simple bivariate models and columns 2 and 4 report estimates from regressions that control for examinee characteristics. Overall, we do not find any association between median earnings at the state of residence of the examinee and his/her change in performance suggesting that our main results are unlikely to be explained by a differential impact of the monetary prize or the opportunity cost of time.

An alternative explanation for differential changes in performance could be that performance of females and minorities is lower than expected in the high stakes section due to stereotype threat (e.g. Steele, 1997 and Steel and Aronson, 1995). However, it is unclear why gender and race/ethnicity stereotypes would be more pronounced in the high stakes section. In addition, the fact that we find similar gender differences in both the quantitative and the verbal sections suggest that stereotype threat is unlikely to explain our main results as the theory would predict that women would respond negatively to the quantitative section only. Moreover, stereotype threat theory cannot explain the smaller drop in performance among Asians in the quantitative section.

We further assess the likelihood of stereotype threat explanation by examining the relationship between gender stereotypes in math and verbal achievement at the state of residence of the examinees and the differential change in performance. To proxy for gender stereotypes at the state of residence of the examinee we use the stereotype adherence index developed by Pope and Sydnor (2010) which reflects gender disparities in test scores favoring boys in math and science and favoring girls in reading and was shown by the authors to be positively associated with other measures of gender stereotype attitudes at the state level.[28] We hypothesize that stereotype threat plays a more important role in states with higher values in the stereotype index. Therefore, for our results to be consistent with stereotype threat, we should observe a larger gender differential in the Q-section and a smaller gender

[27] Earnings come from data published by the U.S. Census Bureau based on 5-year average earnings by state and gender from American Community Survey for the years 2005-2008.

[28] Pope and Sydnor (2010) use test score data from the National Assessment of Educational Progress (NAEP) and show that states that have larger gender disparities in stereotypically male-dominated tests of math and science also tend to have larger gender disparities (of the opposite sign) in stereotypically female-dominated tests of reading. The authors develop a state stereotype adherence index that is defined as the average of the male-female ratio in math and science and female-male ratio in reading for the top 5 percent of the students. We use a standardized version of their index to ease the interpretation of regression estimates.

differential in the V-section in states with a higher stereotype index. In Appendix Table A8, we assess this hypothesis by regressing the score difference between the high and the low stakes section on a female indicator, the gender stereotype index and an interaction between these two variables. Consistent with our previous results, we find a negative coefficient for the female indicator, which shows that females have a smaller differential in performance between the high and the low stakes section in both the quantitative and the verbal section. On the other hand, estimates for the interaction term between female and the stereotype index are all small and not significant. Moreover, their sign goes in the opposite direction than would be expected by the stereotype threat theory.

An additional alternative interpretation of our findings could be that group differences in underlying ability might generate differential drop in performance. However, as we note above, we observe the same pattern of gender and race/ethnic differences across different subsamples and even in subsamples that exhibit similar performance in the high or the low stakes section.

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. Furthermore, in the context of aptitude tests, Ackerman and Kanfer (2009) and Liu et al. (2004) show no evidence for a decline in test performance in the longer test conditions.

In the context of our study, the fact that we find similar participation rates in the experiment among males and females and whites, blacks, Hispanics, and Asians, provides further evidence that a differential effect of fatigue is unlikely to explain our findings. Moreover, as shown in Appendix table A4, the fact that we can replicate our results in the samples of students randomized into the extended time limit sections, provides strong evidence that mitigates this concern.

Finally, one could argue that group differences in performance change between the low and the high stakes section can be explained by differences in learning or test familiarization. To assess this conjecture, we took advantage of one additional piece of information at our disposal. The background questionnaire collected information on examinees' preparation methods for the GRE exam (e.g., use of software or books published by the ETS or other providers, coaching courses offered by commercial companies, coaching courses offered by educational institutions, no preparation, etc.). We coded this information in a vector of dummy variables and re-estimated our main models while controlling for these additional covariates. Results of these expanded models are reported in Appendix Table A9 together with results from our main specification. All estimates are highly similar to our main results suggesting that learning or test familiarization cannot explain our findings.

7. 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. We also find no evidence for alternative explanations such as stereotype threat, test practice or differences in examines' alternative cost of time. But 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 able 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). They are 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).[29]

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 similarity 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.

A curious finding that relates to this question is presented in Figure 5, 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 may also have important implications for admission policies that are intended to achieve demographic diversity in educational institutions and the workplace. If different groups perform differently in low and high stakes situations, then policymakers may be able to diversify the population admitted to colleges, universities, specific study fields, and workplaces by gentle manipulation of the stakes of admission exams. There are several different possible mechanisms that may help facilitate such a change. For example, allowing students to retake an admission test and consideration of the average or the maximum score obtained would reduce the stake of any given test.[30]

[29] 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).

[30] 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.

Finally, our results may also have implications for personnel and incentive policies as they suggest that differences in productivity between workers could vary according to the incentive scheme attached to the job and induce workers to self-select into jobs accordingly.

References

Ackerman P. L. and R. Knafer, (2009) “Test Length and Cognitive Fatigue: an Empirical Examination of Effects of Performance and Test-Taker Reactions”, Journal of Experimental Psychology: Applied, Vol. 15 No.2, pp. 163-181.

Barres, B. (2006) “Does Gender Matter?” *Nature*, No. 442, pp. 133-136.

Booth, A. and P. Nolen (2011) “Choosing To Compete: How Different Are Girls and Boys?” Forthcoming, *Journal of Economic Behavior and Organization*.

Booth, A. and P. Nolen (2012) “Gender Differences in Risk Behavior: Does Nurture Matter” Forthcoming, *Economic Journal*.

Borghans, L. H. Meijers, and B. T. Weel (2008), “The Role of Noncognitive Skills in Explaining Cognitive Test Scores” *Economic Inquiry*, Vol. 46 No. 1, pp. 2-12.

Bridgeman B., F. Cline, and J. Hessinger (2004) “Effect of Extra Time on Verbal and Quantitative GRE Scores” *Journal Applied Measurement in Education* No. 17, pp. 25-37.

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

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* No. 33, pp. 609-624.

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

Cotton, C., F. McIntyre, and J. Price (2010) “The Gender Gap Under Pressure: A Detailed Look at Male and Female Performance Differences During Competitions” NBER Working Paper No. 16436.

Cribbie, R. A. and J. Jamieson, (2000) “Structural Equation and the Regression Bias for Measuring Correlates of Change”, Educational and Psychological Measurement, Vol. 60 No. 6, pp. 893-907.

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

Dohmen, T. J. and A. Falk, (2011) “Performance Pay and Multi-Dimensional Sorting: Productivity, Preferences, and Gender”, *American Economic Review*, Vol. 101, No. 2, pp. 493-525.

Duckworth, A. L., and M. E. P. Seligman (2006) "Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores" *Journal of Educational Psychology, Vol. 98 No. 1*, pp. 198-208.

Educational Testing Service, ETS (2007), "Graduate Record Examinations, Guide to the Use of Scores 2007-2008".

Flory, J. A., A. Leibbrandt and J. A. List (2010) "Do Competitive Work Places Deter Female Workers? A Large-Scale Natural Experiment on Gender Differences in Job Entry Decisions" NBER Working Paper No. 16546.

Gneezy, U., K. L. Leonard, and J. A. List (2009) "Gender Differences in Competition: Evidence From a Matrilineal and a Patriarchal Society" *Econometrica* Vol. 77 No. 5, pp. 1637-1664.

Gneezy, U., M. Niederle, and A. Rustichini (2003) "Performance in Competitive Environments: Gender Differences" *Quarterly Journal of Economics*, Vol. 108 No. 3, pp. 1049-1074.

Gneezy, U., and A. Rustichini (2004) "Gender and Competition at a Young Age" *American Economic Review Papers and Proceedings*, Vol. 94 No. 2, pp. 377-381.

Gunther, C., N. A. Ekinci, C. Schwieren, and M. Strobel, (2010) "Women Can't Jump? An Experiment on Competitive Attitudes and Stereotype Threat" *Journal of Economic Behavior and Organization,* Vol. 75 No. 3, pp. 395-401.

Heckman, J. J. and Y. Rubinstein (2001), "The Importance of Noncognitive Skills: Lessons from the GED Testing Program", *American Economic Review Papers and Proceedings*, Vol. 91 No. 2, pp. 145-149.

Heckman, J. J., and F. Cunha (2007), "The Technology of Skill Formation", *American Economic Review*, Vol. 97 No. 2, pp. 31-47.

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

Jurajda, Š. and D. Münich (2011), "Gender Gap in Admission Performance under Competitive Pressure", American Economic Review Papers and Proceedings, May.

Lavy, V. (2008) "Gender Differences in Competitiveness in a Real Workplace: Evidence from Performance-Based Pay Tournaments among Teachers." NBER Working Paper 14338.

Liu J., J. R. Allspach, M. Feigenbaum, H. J. Oh, and N. Burton (2004) "A Study of Fatigue Effects from the New SAT," College Board Research Report No. 2004-2005.

Lord F. M. (1967), "A Paradox in the Interpretation of Group Comparisons", Psychological Bulletin, Vol. 68 No. 5, pp. 304-305.

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.

Niederle, M., and L. Vesterlund (2007) "Do Women Shy Away From Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics*, Vol. 122 No. 3, pp. 1067-1101.

Niederle, 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., B. Sugrue, and E. L. Baker (1996) "Effects of Motivational Interventions on the National Assessment of Educational Progress Mathematics Performance", Educational Assessment, Vol. 2 No. 2, pp. 135-157.

Örs, E., F. Palomino, and E. Peyrache (2008). "Performance Gender-Gap: Does Competition Matter?" CEPR Discussion Paper No. 6891.

Paarsch, H. and Bruce S. Shearer (2007) "Do Women React Differently to Incentives? Evidence from Experimental Data and Payroll Records", European Economic Review, Vol. 51, pp. 1682-1707.

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

Pope, D. G. and J. R. Sydnor (2010), "Geographic Variation in the Gender Differences in Test Scores," *Journal of Economic Perspectives*, Vol. 24 No. 2, pp. 95-108.

Rothstein, J. (2004), "College Performance Predictions and the SAT," *Journal of Econometrics*, Vol. 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.

Segal, C., (2010), “Motivation, Test Scores, and Economic Success”, Working Paper, Universitat Pompeu Fabra.

Shurchkov, Olga. (Forthcoming) “Under pressure: Gender Differences in Output Quality and Quantity under Competition and Time Constraints,” *Journal of the European Economic Association*.

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.

Steele, C. M. and J. Aronson (1995) “Stereotype threat and the intellectual test performance of African Americans”, *Journal of Personality and Social Psychology*, Vol. 69, pp. 797-811.

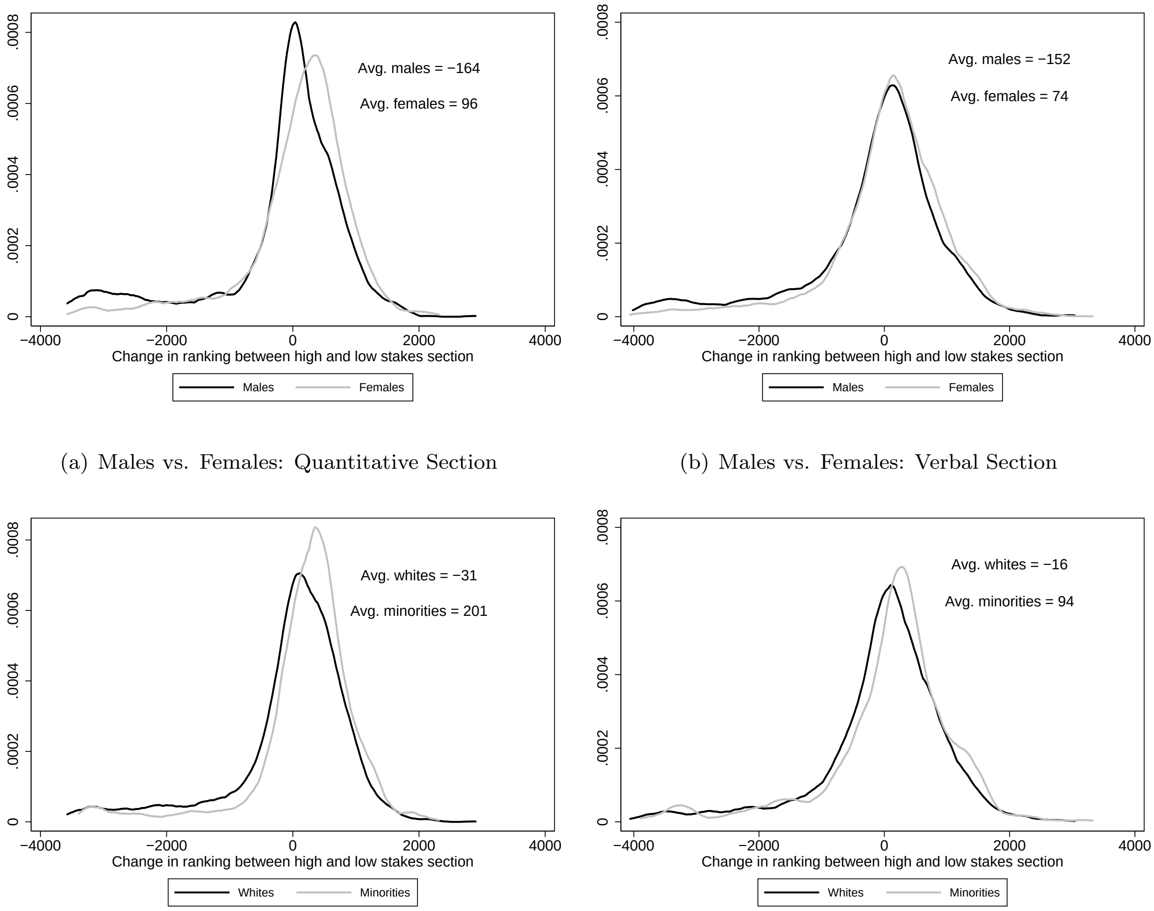
Stevenson, H. W., and J. W. Stigler (1992). The learning gap: Why our schools are failing and what we can learn from Japanese and Chinese education. New York: Simon & Schuster.

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

Willingham, W.W., and N. S. 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.

Figure 1: Difference in Ranking Between High and Low Stakes Test



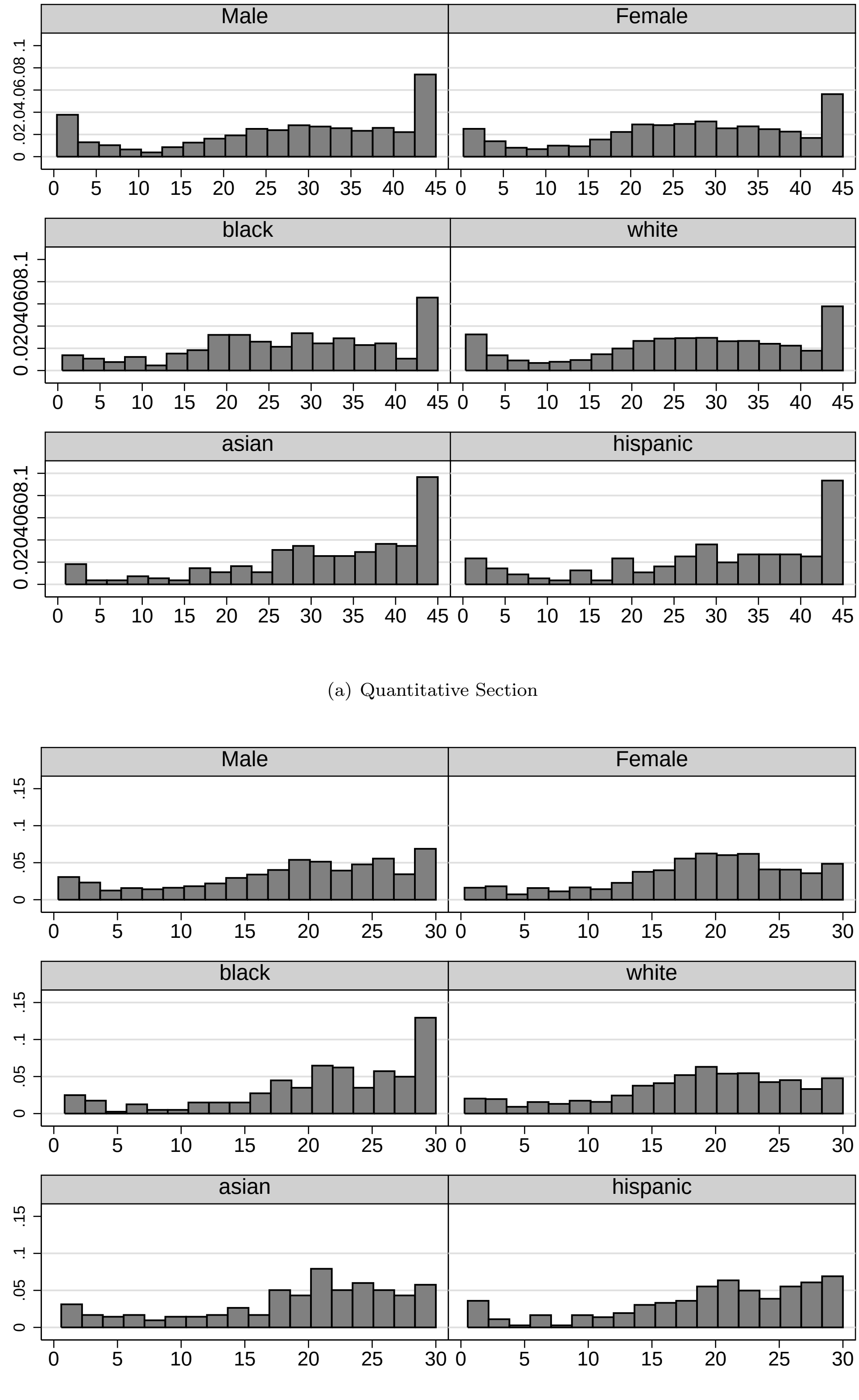
(a) Males vs. Females: Quantitative Section

(b) Males vs. Females: Verbal Section

(c) Whites vs. Minorities: Quantitative Section

(d) Whites vs. Minorities: Verbal Section

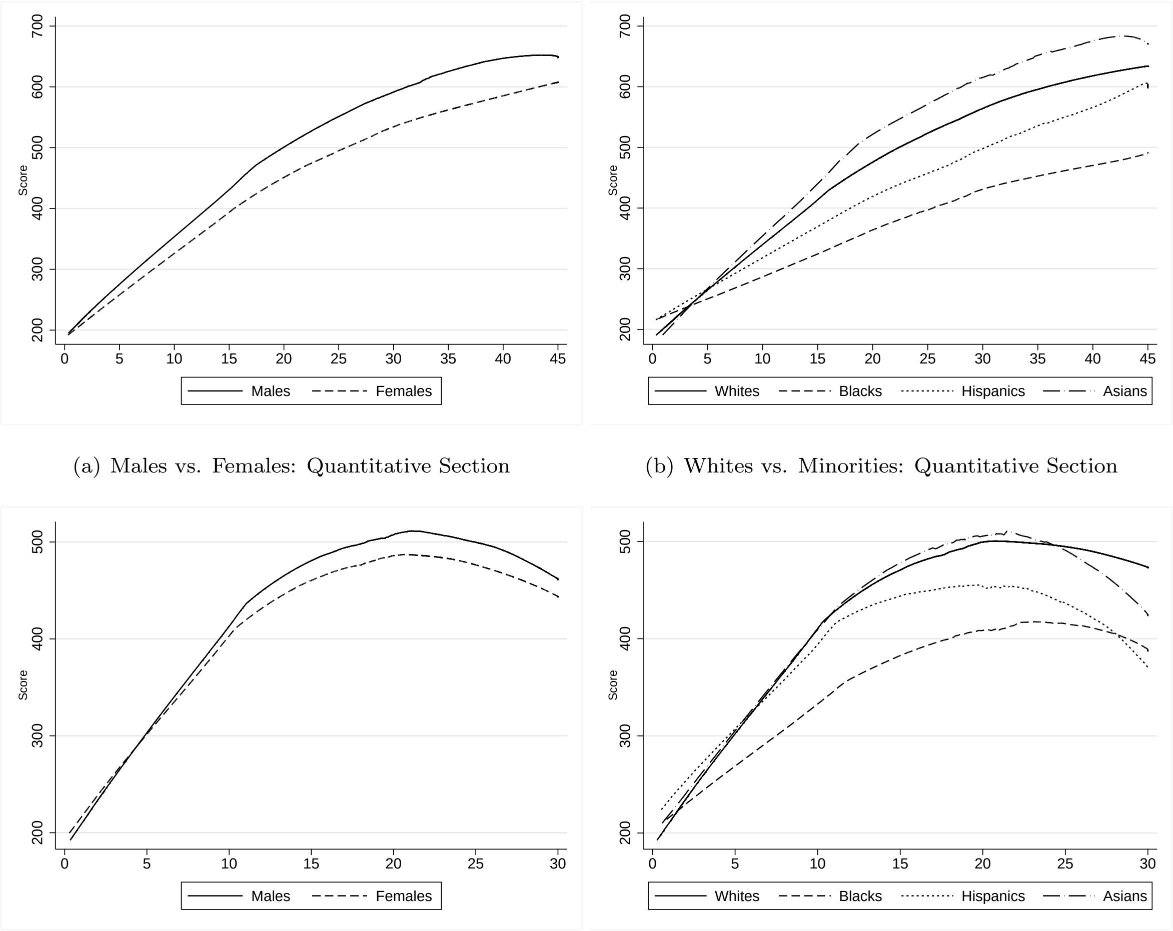
Figure 2: Distribution of Time Invested in the Experimental Section



(a) Quantitative Section

(b) Verbal Section

Figure 3: Relationship Between Time Invested in the Experimental Section and Test Score Achieved in that Section



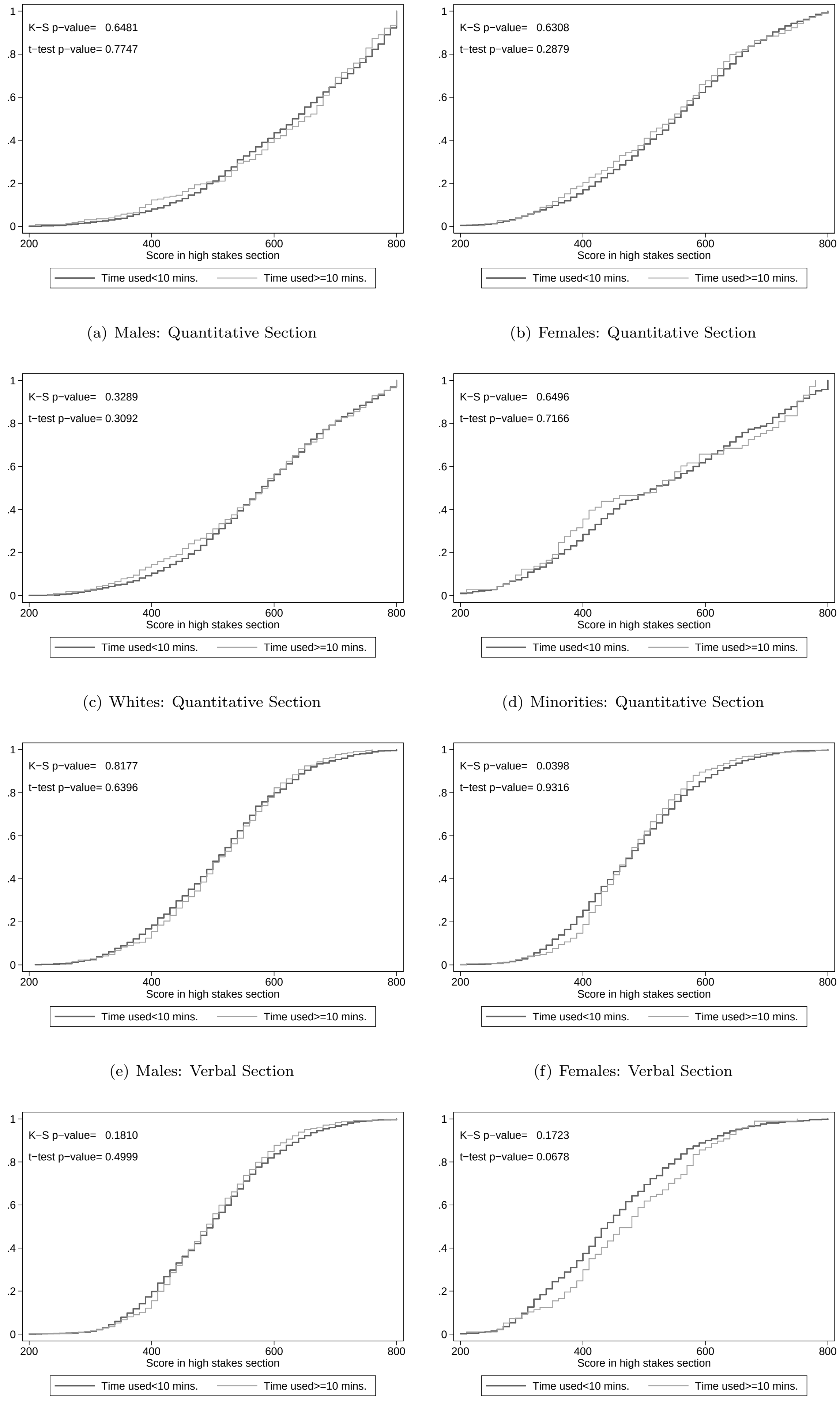
(a) Males vs. Females: Quantitative Section

(b) Whites vs. Minorities: Quantitative Section

(c) Males vs. Females: Verbal Section

(d) Whites vs. Minorities: Verbal Section

Figure 4: CDFs of Test Score in High Stake Section by Effort Invested in Experimental Section



(a) Males: Quantitative Section

(b) Females: Quantitative Section

(c) Whites: Quantitative Section

(d) Minorities: Quantitative Section

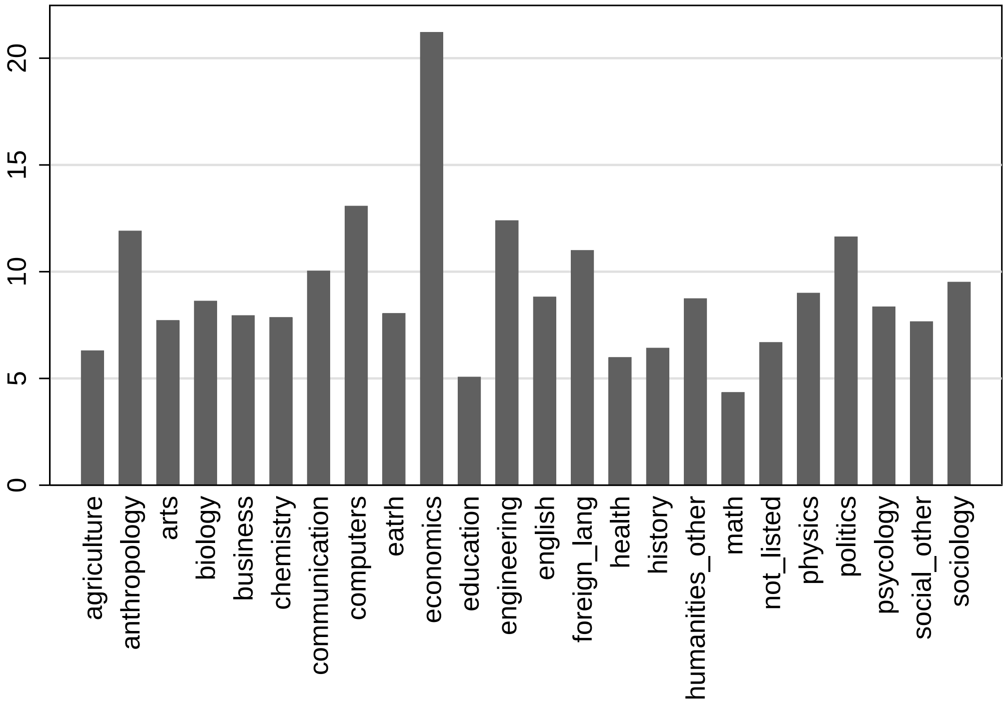
(e) Males: Verbal Section

(f) Females: Verbal Section

(g) Whites: Verbal Section

(h) Minorities: Verbal Section

Figure 5: Performance Gap Between High and Low Stakes Test by Undergraduate Major: Quantitative Section



Notes: The figure plots the performance gap between the high and the low stakes Q-section by subjects' undergraduate major. We include only majors that have at least 30 observations. Test scores are measured in percentile score ranks.

Table 1. Comparison Between Full Population of GRE Test Takers and Experiment Participants

A. By gender

	Males			Females		
		Experiment Participants			Experiment Participants	
	Full Sample	Q section	V section	Full Sample	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		
		Experiment Participants			Experiment Participants			Experiment Participants			Experiment Participants	
	Full Sample	Q section	V section	Full Sample	Q section	V section	Full Sample	Q section	V section	Full Sample	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												
Mean	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 2. Performance in High and Low Stakes Tests by Gender

	Number of Obs.		High Stakes Score			Low Stakes Score			High Stakes - Low Stakes			
	Males	Females	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	1,368	2,553	55.579	40.277	15.302	43.935	33.162	10.773	11.644	7.115	4.529	3.893
			(27.432)	(24.382)	(0.854)	(25.475)	(31.342)	(0.927)	(0.683)	(0.385)	(0.784)	(0.809)
Verbal Section	1,465	2,845	62.902	56.453	6.450	52.481	50.345	2.136	10.421	6.108	4.313	4.041
			(24.959)	(24.538)	(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 3. Performance in High and Low Stakes Tests by Gender and Examinee Characteristics

	Number of Obs.		High Stakes Score			Low Stakes Score			High Stakes - Low Stakes			
	Males	Females	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)	(18.445)	(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)	(19.377)	(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)	(22.755)	(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)	(23.935)	(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)	(23.462)	(1.869)	(27.321)	(34.295)	(2.318)	(1.908)	(0.959)	(2.135)	(2.459)
Undergrad major in	362	132	78.644	69.955	8.689	65.870	63.295	2.575	12.773	6.659	6.114	4.950
Physics, Math, Comp. or Eng.			(17.321)	(23.107)	(1.935)	(27.074)	(31.352)	(3.078)	(1.549)	(2.121)	(2.624)	(2.667)
Grad intended studies in	340	122	77.674	70.574	7.100	65.515	64.369	1.146	12.159	6.205	5.954	4.512
Physics, Math, Comp. or Eng.			(18.191)	(21.707)	(2.024)	(25.909)	(31.265)	(3.161)	(1.596)	(2.167)	(2.689)	(2.846)
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)	(22.986)	(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)	(23.495)	(1.214)	(24.850)	(32.199)	(1.356)	(1.013)	(0.529)	(1.142)	(1.188)
At least some graduate studies or	357	619	63.588	48.724	14.864	49.952	39.069	10.883	13.636	9.654	3.982	2.938
professional degree			(25.921)	(25.125)	(1.689)	(27.697)	(32.106)	(1.953)	(1.455)	(0.929)	(1.725)	(1.824)

Table 3 (cont.). Performance in High and Low Stakes Tests by Gender and Examinee Characteristics

	Number of Obs.		High Stakes Score			Low Stakes Score			High Stakes - Low Stakes			
	Males	Females	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)	(22.205)	(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)	(23.273)	(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)	(23.512)	(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)	(23.097)	(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)	(22.573)	(1.589)	(28.622)	(31.125)	(2.130)	(1.598)	(0.933)	(1.850)	(2.025)
Undergrad major in	388	161	66.781	65.839	0.942	54.036	62.012	-7.976	12.745	3.826	8.919	8.982
Physics, Math, Comp. or Eng.			(24.124)	(25.365)	(2.296)	(25.708)	(31.769)	(2.824)	(1.424)	(1.301)	(1.929)	(2.048)
Grad intended studies in	378	142	66.341	66.056	0.285	53.643	60.535	-6.892	12.698	5.521	7.177	8.237
Physics, Math, Comp. or Eng.			(23.796)	(24.881)	(2.372)	(27.411)	(31.356)	(2.986)	(1.445)	(1.340)	(1.970)	(2.187)
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)	(23.959)	(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)	(23.942)	(1.134)	(27.139)	(30.420)	(1.343)	(1.033)	(0.591)	(1.190)	(1.258)
At least some graduate studies or	376	731	68.830	63.848	4.982	58.495	56.791	1.704	10.335	7.057	3.278	3.825
professional degree			(22.931)	(24.094)	(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 2. 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 4. Performance in High and Low Stakes Tests by Race and Ethnicity

Number of Obs.				High Stakes Score							Low Stakes Score						
W	B	H	A	Whites (W)	Blacks (B)	Hispanics (H)	Asians (A)	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																	
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																	
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 5. Differential Performance Between High and Low Stakes Tests by Race and Ethnicity

High Stakes - Low Stakes				Raw Difference			Controlled Difference		
Whites (W)	Blacks (B)	Hispanics (H)	Asians (A)	W-B	W-H	W-A	W-B	W-H	W-A
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
A. Quantitative Section									
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									
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 report 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 6. Share of Experiment Participants who Spent Less than Ten Minutes in the Experimental Section

Share who spent less than ten minutes among	Q-section (1)	V-section (2)
Gender		
Males	0.167	0.181
Females	0.132	0.138
p-value of difference: Males-Females	*0.0042*	*0.0004*
Race/ethnicity		
Whites	0.152	0.154
Blacks	0.106	0.101
p-value of difference: whites-blacks	*0.0196*	*0.0077*
Hispanics	0.129	0.140
p-value of difference: whites-hispanics	*0.3277*	*0.5581*
Asians	0.071	0.161
p-value of difference: whites-asians	*0.0000*	*0.7901*
Maternal Education		
High School or less	0.134	0.133
College or some college	0.134	0.155
At least some graduate studies or professional degree	0.163	0.157
p-value of difference	*0.1031*	*0.1880*
Paternal Education		
High School or less	0.145	0.136
College or some college	0.130	0.151
At least some graduate studies or professional degree	0.161	0.166
p-value of difference	*0.0606*	*0.0010*
Undergraduate GPA		
C or C-	0.148	0.161
B-	0.120	0.122
B	0.128	0.136
A-	0.159	0.176
A	0.151	0.155
p-value of difference	*0.1242*	*0.0218*
Achievement decile in high stakes test		
1	0.166	0.160
2	0.147	0.092
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
p-value of difference	*0.7360*	*0.0011*
Number of Observations	565	659

Notes: Columns 1 and 2 report the share of examinees that spent less than 10 minutes in the experimental Q or V sections respectively out of their relevant group. The p-values reported in italics test for equality of the coefficients of the different subgroups.

Table 7. Share of Experiment Participants who Improved their Score in the Low Stakes Section Relative to the High Stakes Section

	Q-section					V-section				
	Mean	Males - Females	Whites-Blacks	Whites-Hispanics	Whites-Asians	Mean	Males-Females	Whites-Blacks	Whites-Hispanics	Whites-Asians
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Raw difference	0.015	-0.004	-0.006	-0.018	-0.005	0.053	0.000	-0.038	-0.016	-0.008
		(0.004)	(0.009)	(0.012)	(0.009)		(0.007)	(0.018)	(0.017)	(0.015)
Controlled difference		-0.005	-0.009	-0.020	-0.002		-0.008	-0.033	-0.014	-0.007
		(0.005)	(0.009)	(0.012)	(0.009)		(0.008)	(0.019)	(0.017)	(0.015)

Notes: Columns 1 and 6 report the share of examinees who improved their score in the experimental Q or V sections respectively relative to the real GRE section. A score gain is defined for cases where the score difference between the low and the high stakes section divided by the standard error of measurement of difference in scores is greater than 1.65. Columns 2-5 and 7-10 report differences between males and females and between whites and minorities in the share of examinees who improve their scores. The first row reports raw differences between groups. The second row reports differences between groups after controlling for examinee's covariates detailed in Table 2. Robust standard errors are reported in parenthesis.

Online appendix

Table A1. 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 A2. Descriptive Statistics of Experiment Participants

	Males	Females	Whites	Blacks	Hispanics	Asians
	(1)	(2)	(3)	(4)	(5)	(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				
American Indian or Alaskan Native	0.00	0.01				
Other	0.05	0.04				
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
Missing	0.06	0.05	0.04	0.07	0.04	0.05
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
Missing	0.01	0.01	0.01	0.04	0.02	0.01
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
Missing	0.04	0.04	0.03	0.06	0.07	0.05
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.

Online appendix

Table A3. Performance in High versus Low Stakes Tests by Gender and Race/Ethnicity - Quantitative Section

	High Stakes		Low Stakes		High-Low Stakes		**Controlled Difference**
	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)	(23.342)	(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)	(19.039)	(16.851)	(26.150)	(2.146)	(0.906)	**(2.306)**
Controlled Difference					**5.803**	**3.491**	
(Whites-Blacks)					**(2.385)**	**(1.140)**	
Hispanics	44.022	31.363	38.405	28.748	5.618	2.615	**0.609**
	(27.048)	(22.875)	(23.230)	(29.775)	(2.422)	(1.561)	**(3.301)**
Controlled Difference					**7.539**	**4.182**	
(Whites-Hispanics)					**(2.601)**	**(1.649)**	
Asians	72.167	56.386	66.071	48.671	6.095	7.714	**0.747**
	(23.589)	(26.875)	(29.090)	(29.509)	(2.603)	(1.955)	**(3.919)**
Controlled Difference					**9.412**	**-0.169**	
(Whites-Asians)					**(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 group. Columns 3-4 report mean performance in the low stakes test for each gender-race/ethnicity group. Performance change between the high and the low stakes tests are reported in columns 5 and 6. Controlled differences in performance change between males and females stratified by race/ethnicity are reported in bold in column 7. Controlled differences in performance change between whites and minorities stratified by gender are reported in bold in columns 5 and 6. Test scores are reported in percentile ranks. Standard deviations and robust standard errors are reported in parenthesis.

Online appendix

Table A4. Robustness Check: Differential Performance in High versus Low Stakes Tests

	Difference in individual performance between high and low stake test						Controlled difference between groups			
	Males	Females	Whites	Blacks	Hispanics	Asians	Males-Females	Whites-Blacks	Whites-Hispanics	Whites-Asians
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
A. Quantitative Section										
Raw scores	76.952	51.324	65.601	21.019	32.321	43.616	22.752	31.330	31.697	27.070
	(4.402)	(2.596)	(2.650)	(6.085)	(8.818)	(9.464)	(5.331)	(7.082)	(9.275)	(10.337)
Ln(raw scores)	0.182	0.130	0.161	0.054	0.091	0.098	0.047	0.080	0.068	0.074
	(0.011)	(0.007)	(0.007)	(0.016)	(0.022)	(0.022)	(0.013)	(0.019)	(0.023)	(0.024)
Percentile score ranks	11.560	6.330	8.708	1.585	3.995	10.573	4.360	4.788	3.098	-0.966
Extended time limit sample	(0.733)	(0.421)	(0.421)	(1.071)	(1.611)	(1.789)	(0.910)	(1.194)	(1.720)	(1.868)
B. Verbal Section										
Raw scores	45.993	26.882	34.275	11.935	27.059	39.255	17.660	11.357	3.011	-2.480
	(3.022)	(1.748)	(1.734)	(5.779)	(6.118)	(7.073)	(3.625)	(6.371)	(6.511)	(7.386)
Ln(raw scores)	0.121	0.072	0.091	0.037	0.072	0.103	0.046	0.028	0.009	-0.006
	(0.008)	(0.005)	(0.005)	(0.016)	(0.016)	(0.018)	(0.009)	(0.017)	(0.017)	(0.019)
Percentile score ranks	11.380	4.575	7.428	1.240	3.894	8.008	5.693	3.759	2.227	0.653
Extended time limit sample	(0.748)	(0.413)	(0.423)	(1.101)	(1.539)	(1.814)	(0.883)	(1.240)	(1.620)	(1.857)

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 use difference in raw test scores as a dependent variable. The second row of each panel use difference in the natural logarithm of raw scores as a dependent variable. The last row of each panel uses difference in percentile score ranks (as all our main results tables) but uses the sample of students who got the extra GRE section with an extended time limit. Standard deviations and robust standard errors are reported in parenthesis.

Online appendix

Table A5. Performance Gap Between High and Low Stakes Section by Time Spent in Low Stakes Section

	Controlled difference between groups			
	Males-Females	Whites-Blacks	Whites-Hispanics	Whites-Asians
Sample	(1)	(2)	(3)	(4)
	A. Quantitative Section			
Full	3.893	4.160	5.231	3.292
	(0.809)	(1.016)	(1.416)	(1.693)
Time spent in experimental section ≥ 10 mins.	1.060	2.049	4.368	0.252
	(0.554)	(0.769)	(0.977)	(1.207)
Time spent in experimental section ≥ 3 mins.	2.061	2.987	4.503	1.966
	(0.697)	(0.847)	(1.222)	(1.422)
Full sample - controlling for 4th order polynomial of time spent in experiment	2.819	1.988	2.579	-1.032
	(0.573)	(0.931)	(1.040)	(1.199)
	B. Verbal Section			
Full	4.041	3.080	0.326	-0.747
	(0.818)	(1.459)	(1.543)	(1.700)
Time spent in experimental section ≥ 10 mins.	0.997	2.196	-0.748	-0.076
	(0.555)	(1.100)	(1.240)	(1.214)
Time spent in experimental section ≥ 3 mins.	2.101	4.023	0.128	0.513
	(0.667)	(1.123)	(1.365)	(1.404)
Full sample - controlling for 4th order polynomial of time spent in experiment	2.017	1.039	0.021	-0.759
	(0.551)	(1.070)	(1.273)	(1.132)

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 reproduces results reported in tables 2 and 5. The second row of each panel reports results for the subsample of examinees who spent less than 10 minutes in the experimental section. The third row of each panel reports results for the subsample who spent at least 3 minutes in the experimental section. The fourth row of each panel reports results for the full sample from a model that controls for a 4th order polynomial of time spent in the experimental section. Test scores are reported in percentile ranks. Standard errors are reported in parenthesis.

Online appendix

Table A6. Differential Gap in "Real" GRE Between Students Taking Test for Practice and Other Students

	Gaps by Gender		Gaps by Race/Ethnicity					
	Female	Female x Practice	Black	Black x Practice	Hispanic	Hispanic x Practice	Asian	Asian x Practice
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Quantitative section	-4.361	0.953	-13.803	1.414	-6.646	2.239	2.923	3.765
	(0.570)	(1.909)	(0.856)	(2.927)	(1.068)	(3.788)	(1.286)	(4.781)
Verbal section	-3.960	0.557	-21.450	0.996	-8.969	2.191	0.144	6.032
	(0.615)	(2.247)	(1.265)	(4.071)	(1.325)	(4.659)	(1.307)	(4.960)

Notes: The table reports estimates from a regression of test score in the high stakes section on demographic group indicators and interactions between demographic indicators and an indicator for practice exam. The model controls also for an indicator of practice exam and student's background characteristics detailed in Table 2. Robust standard errors are reported in parenthesis.

Table A7. Associations Between Median Earnings at the Examinee State of Residence and Differential Performance

	Males		Females	
	(1)	(2)	(3)	(4)
A. Quantitative Section				
Median earnings (in thousand dollars)	-0.169 (0.161)	-0.178 (0.168)	0.124 (0.130)	0.055 (0.131)
Median earnings of college graduates (in thousand dollars)	0.049 (0.151)	0.101 (0.149)	0.012 (0.143)	-0.040 (0.134)
B. Verbal Section				
Median earnings (in thousand dollars)	0.211 (0.162)	0.106 (0.182)	0.158 (0.076)	0.093 (0.080)
Median earnings of college graduates (in thousand dollars)	0.191 (0.111)	0.158 (0.129)	0.109 (0.112)	0.055 (0.110)
Controls for examinee's covariates	--	✓	--	✓

Notes: The table reports regression estimates for the coefficient of annual median earnings (in thousand US$) of full time workers or college graduates working full time at the state of residence of the examinee. The dependent variable is the score difference (in percentile points) between the high and the low stakes section. Estimates reported in columns (2) and (4) come from regressions that control for race/ethnicity, mother’s and father’s education, dummies for UGPA, undergraduate major, intended graduate field of studies, and disability status. Standard errors clustered at the state levels are reported in parenthesis.

Table A8. Differential Performance and Stereotype Threat

	High Stakes - Low Stakes			
	Quantitative Section		Verbal Section	
	(1)	(2)	(3)	(4)
Female	-4.434	-3.712	-4.851	-4.651
	(0.692)	(0.829)	(0.749)	(0.779)
State stereotype index	-1.108	-0.903	-0.149	0.001
	(0.651)	(0.626)	(0.702)	(0.667)
Female x State stereotype index	0.606	0.633	-0.796	-0.645
	(0.681)	(0.747)	(0.706)	(0.701)
Controls for examinee's covariates	--	✓	--	✓

Notes: The table reports estimates from models that regress the score difference (in percentile points) between the high and the low stakes section on a female dummy, the gender stereotype index of the state of residence of the examinee and the interaction between these two variables. Estimates reported in columns (2) and (4) come from regressions that control also for race/ethnicity, mother's and father's education, dummies for UGPA, undergraduate major, intended graduate field of studies, and disability status. The gender stereotype index is a standaryzed version of the index developed by Pope and Sydnor (2010) where higher values denote stronger gender stereotypes. Standard errors clustered at the state levels are reported in parenthesis.

Online appendix

Table A9. Differential Gap Between High and Low Stakes Section After Controlling for Test Preparation Methods

	Quantitative Section				Verbal Section			
	Males-Females	Whites-Blacks	Whites-Hispanics	Whites-Asians	Males-Females	Whites-Blacks	Whites-Hispanics	Whites-Asians
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Basic specification	3.893	4.160	5.231	3.292	4.041	3.080	0.326	-0.747
	(0.809)	(1.016)	(1.416)	(1.693)	(0.818)	(1.459)	(1.543)	(1.700)
Controlling for test preparation methods	3.942	3.864	5.284	3.647	4.228	3.310	0.500	-0.378
	(0.817)	(1.031)	(1.428)	(1.677)	(0.825)	(1.455)	(1.552)	(1.701)

Notes: The table reports differences in performance between the high and the low stakes tests by gender and race/ethnicity. Columns 1-4 report differences in the Q-section and columns 5-8 report differences in the V-section. The first row of the table reproduces estimates from the full specification reported in tables 2 and 5. The second row reports results from regressions that control also for indicators for test preparation methods reported by the examinees: no preparation, software from the ETS, books published by the ETS, software from other providers, books from other providers, attended a coaching course offered by a commercial company, attended a coaching course offered by an educational institution, used *ScoreItNow!* online writing practice, used GRE enhanced diagnostic service, other type of preparation. Robust standard errors are reported in parenthesis.