THE ECONOMIC JOURNAL

The Economic Journal, **129** (*October*), 2916–2948 DOI: 10.1093/ej/uez015
Advance Access Publication Date: 16 May 2019

DIFFERENTIAL PERFORMANCE IN HIGH VERSUS LOW STAKES TESTS: EVIDENCE FROM THE GRE TEST*

Analia Schlosser, Zvika Neeman and Yigal Attali

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 Graduate Record Examination (GRE), and the low stakes situation is a voluntary experimental section of the GRE. We find that males exhibit a larger drop in performance between the high and low stakes examinations than females, and that whites exhibit a larger drop in performance than minorities. Differences between high and low stakes tests are partly explained by the fact that males and whites exert lower effort in low stakes tests compared with females and minorities.

Recently, there has been much interest in the question of whether different demographic groups respond differently to incentives and competitive pressure. Interest in this subject stems from attempts to explain gender, racial and ethnic differences in human capital accumulation and labour market performance, and is further motivated by the increased use of aptitude tests for college admissions and job screening and by the growing use of standardised tests for the assessment of students' learning. While it is clear that motivation affects performance, less attention has been given to demographic group differences in response to performance-based incentives.

In this paper, we examine whether individuals respond differently to incentives by analysing their performance in the Graduate Record Examination (GRE) General Test.[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 part in 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 and that is administered to a very large and easily characterisable population, namely the population of applicants to graduate programs in arts and sciences in the United States. 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, that manipulate the stakes,

* Corresponding author: Analia Schlosser, Berglas School of Economics, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel. Email: analias@tauex.tau.ac.il.

This paper was received on 6 March 2017 and accepted on 20 August 2018. The Editor was Kjell Salvanes.

We are grateful for comments received at the SOLE meeting, 'Discrimination at Work' and 'Frontiers in Economics of Education' workshops, and seminar participants at the Federal Reserve Bank of Chicago, CESifo, Copenhagen Business School, Norwegian Business School, University of Zurich, Bar Ilan University, Ben Gurion University, and University of Haifa. This research was supported by the Israeli Science Foundation(grant No. 1035/12).

[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 in the United States 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 the ETS website: http://www.ets.org/gre/general/about/.

degree of competitiveness, or incentive levels in somewhat artificial ways, and where stakes are not as high as in real-life important events. A second distinctive feature of our research is that we are able to observe the performance of the same individual in high and low stakes situations that involve the exact same task. A third unique feature of our study is the availability of 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 backgrounds and to 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. Notably, 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.

Our results show that males exhibit a larger difference in performance between the high and low stakes GRE than females, and that whites exhibit a larger difference in performance between the high and low stakes GRE test than 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. Specifically, while males outperform females in the high stakes quantitative section of the GRE by 0.55 standard deviations (SD), the gender gap in performance in the low stakes section is only 0.30 SD. Similarly, males' advantage in the high stakes verbal section is 0.26 SD, while the gender gap in the low stakes section is only 0.07 SD. Whites outperform blacks and Hispanics in the high stakes quantitative section by 1.1 SD and 0.42 SD, respectively, but the gaps are significantly reduced in the low stakes section, to 0.63 and 0.14 SD. This pattern is reversed for Asians because they outperform whites by 0.51 SD in the high stakes quantitative section, so that the gap increases to 0.55 SD in the low stakes section. These group differences in performance between high and low stakes tests appear across all undergraduate GPA levels, 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 field of studies).

We 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 compared with women and minorities, respectively. 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: differences in the perceived importance of the 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. The results from our paper have two main implications:

(*i*) stakes have to be taken into account when analysing performance gaps between groups;
(*ii*) some groups are driven mostly by incentives, while other groups exert high effort even if stakes are low or 'nearly zero'.

While these two implications do not, in themselves, amount to direct policy recommendations, they are nevertheless highly relevant for policy. For example, they imply that any analysis of gender or race test score gaps, or studies that examine the effect of a specific educational

intervention by gender or race, should consider the stakes of the test involved in order to interpret the results and effectiveness of the intervention. In addition, our results highlight the fact that university or job admission policies that use standardised aptitude tests should consider that such tests measure only performance under a high stakes setup and are less informative about individuals' performance in low stakes or zero stakes situations, which may be as important at the university or job.

Most of the experimental literature about gender differences in performance focuses on a comparison of performance between a competitive setting, where the best performer receives a higher payment, and a non-competitive environment, where subjects are paid according to their own performance (using a piece-rate schedule). A common finding in these studies is that while the performance of men improves under competition, women's performance is unchanged or even declines slightly (see, e.g., Gneezy *et al*., 2003; Gneezy and Rustichini, 2004). A second finding is that women 'shy away from competition'. Namely, given the choice, women prefer to be compensated according to a non-competitive piece-rate compensation schedule over participation in competitive tournaments (see, e.g., Datta Gupta *et al*., 2005; Niederle and Vesterlund, 2007; Dohmen and Falk, 2011).

There are several variations and extensions to these studies that examine whether the results vary by: (a) the gender composition of the group involved in the tournament; (b) the type of task involved (tasks requiring effort vs. skills, or tasks where males or females have a stereotypical or real advantage); (c) the information provided about own and others' performance during the experiment; (d) the use of priming; (e) letting participants choose the gender of their competitors; (f) manipulating the risk associated with the payments; and (g) the number of iterations involved. For recent reviews of this literature, see Croson and Gneezy (2009), Azmat and Petrongolo (2014), and Niederle (2016).

Our paper differs from these previous studies in several aspects: first, we compare performance between a high stakes setting that has important consequences for life and a task that has almost zero stakes. In a sense, this is more similar to a comparison between performance under a piece-rate and a flat-rate payment scheme. Second, even though GRE scores are also reported in percentiles, the exam is not presented as a direct tournament between subjects (certainly not among those tested on a specific date and in a specific test centre).[2] Accordingly, the focus of our study is not a comparison between a competitive and a non-competitive environment but rather a contrast between a high stakes and a very low stakes setting. As our results show, males invest less effort than females when stakes are low. We therefore add new insights to the experimental literature cited above by suggesting that gender differences found in these lab experiments may significantly understate differences in important real-life situations given that the stakes levels of lab experiments are relatively low.

Evidence on gender differences in real world situations is limited to a small number of recent studies and remains an important empirical open question. Paserman (2010) studied the performance of professional tennis players and found that performance decreases under highly competitive pressure, but this result is similar for both men and women. Similarly, Lavy (2008) found no gender differences in the performance of high school teachers who participated in a performance-based tournament. On the other hand, in a field experiment among administrative job seekers, Flory *et al*. (2010) found that women are indeed less likely to apply for jobs that

[2] While GRE test scores are relative to other students, the competition between students is less salient on the day of the exam as the pool of competitors is very large and not directly visible or known ex ante to GRE test-takers.

include performance-based payment schemes, but that this gender gap disappears when the framing of the job is switched from being male- to female-oriented.[3]

A number of studies within the educational measurement literature demonstrate that high stakes situations induce stronger motivation and higher effort.[4] However, high stakes also increase test anxiety and so might harm performance (Cassaday and Johnson, 2002). Indeed, Ariely *et al*. (2009) found that strong incentives can lead to 'choking under pressure' in both cognitive and physical tasks, although they did not find gender differences. Performance in tests is also affected by non-cognitive skills, as shown by Heckman and Rubinstein (2001), Cunha and Heckman (2007), Borghans *et al*. (2008), and Segal (2010).[5]

Levitt *et al*. (2016) examined how timing, type of rewards and framing of rewards affect performance in a series of field experiments involving primary and secondary school students in Chicago. They report that, in most cases, boys were more likely to respond to incentives than girls were. Azmat *et al*. (2019) is the closest paper to ours. These authors exploited the variation in the stakes of tests administered to students attending a Spanish private school and showed that the performance of female students declines as the stakes become higher, while males' performance improves. Their finding is consistent with ours, but we examine the performance of a much larger population (GRE test-takers) and show gender differences in response to incentives across a wide range of students' background characteristics, fields of study and ability levels. In addition, we are able to explore the role played by students' effort in explaining our findings, and rule out some alternative explanations (including females choking under pressure). Our study also expands the literature by examining differential performance by race and ethnicity. To the best of our knowledge, no other study has examined differences in response to incentives among ethnic groups.

Our paper is also related to Babcock *et al*. (2017), who find that women, more than men, volunteer, are asked to volunteer, and accept requests to volunteer for 'low promotability' tasks. Their results suggest that women's higher tendency to volunteer seems to be shaped by women's beliefs rather than preferences. Accordingly, these authors suggest several alternative assignment schemes to reduce the gender gap in participation in low stakes activities such as turn-taking or random assignment.

In our study, the decision to participate in the low stakes task, which is analogus to 'volunteering', does not generate a group benefit as in Babcock *et al*. However, we examine not just willingness to participate in the low stakes task, but also effort exerted conditional upon participation. That is, our setting contains both the binary decision of whether to volunteer or not, and a continuous decision with respect to how much effort to exert after volunteering. Our results show that while men and women are equally likely to volunteer, the performance of men is significantly lower. Our results therefore suggest that even if men and women are randomly assigned to participate in a certain committee, women might invest more time and effort condi-

[3] Other studies that compare gender performance by degree of competitiveness include Jurajda and Munich (2011) and Ors *et al*. (2008).

[4] 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 *et al*. (1996).

[5] Several studies (see e.g., 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 non-cognitive dimensions that affect test performance are discussed by the literature on stereotype threat that suggests that the performance of a group is likely to be affected by exposure to stereotypes that characterise the group (see Steele, 1997; Steele and Aronson, 1995; Spencer *et al*., 1999).

tional on participation. Consequently, a random assignment mechanism might not overcome the problem of inequality in investment in '*low promotability*' tasks.

The rest of the paper proceeds as follows. In the next section we describe the experimental setup and data. In Section 2, we present the empirical framework. In Section 3 we present the results, and in Section 4 we discuss alternative explanations for our findings as well as other related observations. Section 5 concludes.

1. Experimental Setup 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. At the end of the regular test, a screen appeared that invited examinees to voluntarily participate in a research project that would require them to take an additional test section for experimental purposes.[6] GRE examinees who agreed to participate in the experiment were promised a monetary reward if they performed well compared with their performance in the real examination.[7]

Participants in the experiment were randomly assigned into one of four groups: one group was administered a quantitative section (Q-section) with a standard time limit (45 minutes), a second group was administered a verbal section (V-section) with a standard time limit (30 minutes), the third group was administered a quantitative section with an extended time limit (68 minutes) and the fourth group was administered a verbal section with an 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 experimental section and the real sections was the appearance of a screen that indicated that performance on the experimental 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 (or almost zero 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 centres (including overseas), 46,038 were U.S. citizens who took the GRE test in centres located in the United States. We focus on U.S. citizens tested in the United States to avoid dealing with a more heterogeneous population and to control for a similar testing environment. In addition, we want to abstract from differences in performance

[6] Students saw their score in the regular test only after the experimental section. They were never told their score in the experimental section.

[7] Specifically, 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 recognise 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.

that are due to language difficulties. A total of 15,945 out of the 46,038 U.S. examinees agreed to participate in the experiment. About half of them (8,232) were randomised into the regular time-limit sections and were administered either an extra Q-section (3,922) or an extra V-section (4,310).[8] We select only experiment participants who were randomised 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 that examinees associate with it.[9]

A unique feature of our research design that distinguishes our study from most of the experimental literature is that we are able to identify and characterise 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 U.S. GRE test-takers and the sample of experiment participants.[10] The two populations are virtually identical in terms of proportions of females, males and minorities. For example, women constitute 66% of the full population of U.S. 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% of GRE U.S. domestic examinees and they are similarly represented among experiment participants. The shares of blacks, Hispanics and Asians range between 6% and 5.5% in both the full sample and the sample of experiment participants.[11]

Participants in the experiment also have similar GRE test scores to those in the full relevant sub-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 male participants in the experiment. The median score (57 percentile rank) and standard deviation (27 points) are also identical for the full sample of GRE U.S. male test-takers, the sample of experiment participants randomised to the Q-section, and the sample of experiment participants randomised to the V-section. The test score distribution of female GRE test-takers is also identical to that of female experiment participants. We observe also the same result when comparing test-score distributions within each race/ethnicity. Overall, the results presented in Table 1 show that there is no differential selection into the experiment according to gender, race/ethnicity or GRE test scores; nor do we find any evidence of differential selection within each gender or race/ethnic group.[12]

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.[13] Appendix Table A2 reports the descriptive statistics of these background characteristics

[8] Since the experimental sections were randomised among the full sample of experiment participants, which included all students (U.S. and international) tested in all centres around the world, the proportion of U.S. participants assigned to each section is not exactly 50%.

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

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

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

[12] 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 suggests that the presence of large differences in selection by unobservables across groups is very unlikely.

[13] We obtained the complete background information on experiment participants only, so we only analyse selection in the experiment according to gender, race, ethnicity and GRE scores in the high stakes section.

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

	A. By gender					
	Males			*Females*		
	Full sample	*Experiment participants*		*Full sample*	*Experiment participants*	
		Q-section	*V-section*		*Q-section*	*V-section*
N	15,749	1,369	1,465	30,160	2,553	2,845
Share	0.34	0.35	0.34	0.66	0.65	0.66
Quantitative score						
Mean	55.8	55.6	56.8	40.7	40.3	41.2
SD	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
SD	24.5	25.0	25.0	24.8	25.0	24.5
Median	67	67	67	57	57	57

	B. By race/ethnicity											
	Whites			*Blacks*			*Hispanics*			*Asians*		
	Full sample	*Experiment participants*		*Full sample*	*Experiment participants*		*Full sample*	*Experiment participants*		*Full sample*	*Experiment participants*	
		Q-section	*V-section*		*Q-section*	*V-section*		*Q-section*	*V-section*		*Q-section*	*V-section*
N	36,042	3,027	3,380	2,877	265	248	2,400	224	221	2,584	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
SD	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
SD	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 the performance of experiment participants stratified by gender and race/ethnicity. The samples are restricted to U.S. citizens tested in the United States.

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 come from similar family backgrounds, as measured by both mother's and father's educational levels and by the proportion of native English speakers. Females and males also have similar distributions of undergraduate GPA (UGPA). 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% for males versus 5% for females).

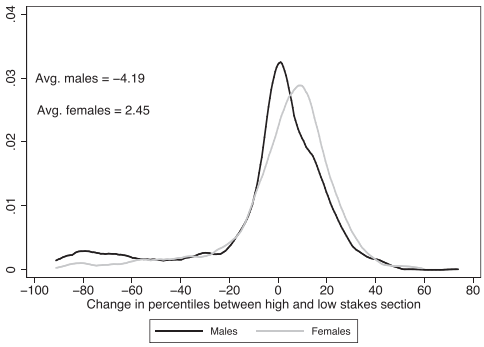
Columns 3 through 6 in Table A2 report descriptive statistics of the analysis sample stratified by race/ethnicity. Maternal education is similar among whites and Asians, but Asians are more likely to have a father with at least some graduate studies or a professional degree relative to whites (45% versus 35%). Hispanics and blacks come from less educated families. Asians are less likely to be native English speakers (86%) relative to whites (93%), blacks (95%) and Hispanics (90%). In terms of undergraduate achievement, we observe that whites and Asians have similar UGPA 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%) relative to whites (11%), blacks (8%) or Hispanics (12%).

2. Empirical Framework

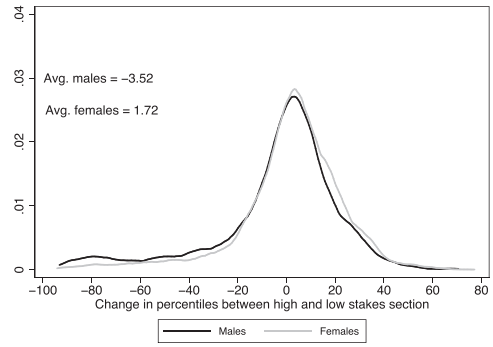
Our main objective is to examine how the performance of different demographic groups changes as a function of the stakes of the test (high stakes: real GRE exam; low stakes: experimental section). We summarise our main finding in Figure 1, using an ordinal metric, which is free of the specific scale of test scores. We ranked individuals according to their performance in each test and plot the rank change distribution (in percentile points) between the high and low stakes test by gender and race for each test. Panels (a) and (b) show that men's ranking declines by 4 percentile points in the low stakes test relative to the high stakes test, while women's ranking improves by 2 percentile points. Panels (c) and (d) show that the ranking of whites declines while the ranking of minorities improves when switching from the high to the low stakes test in both the Q- and the V-section. Focusing on the Q-section, which is less likely to be affected by language problems of minorities, we see that whites' ranking declined by almost 1 percentile point while that of minorities improved by about 5 percentile points.[14] The rank changes between men and women and between whites and minorities are statistically different (p-values of Mann–Whitney tests < 0.0001).

We now turn to measure individuals' change in performance using a simple regression model to control for additional characteristics of individuals and quantify the average change in performance between the high and low stakes test for each group. We estimate the following first difference equation for each of the experimental samples (i.e., individuals randomised to the

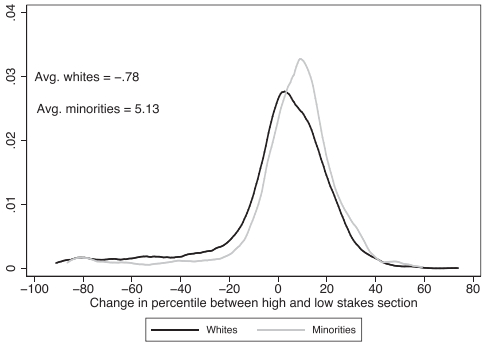
[14] Minorities include Asians, Hispanics and blacks. We excluded students who defined themselves as American Indian or Alaskan Native (43) or other race (271).



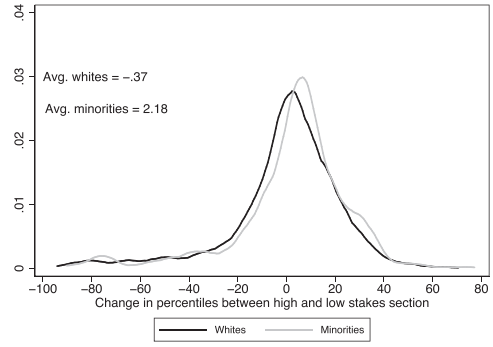
(a) Males vs. Females: Quantitative Section



(b) Males vs. Females: Verbal Section



(c) Whites vs. Minorities: Quantitative Section



(d) Whites vs. Minorities: Verbal Section

Fig. 1. *Difference in Ranking Between High and Low Stakes Test.*
Notes: The figure shows the difference in ranking between the low and the high stakes exam by gender and race ethnicity.

experimental Q- or V- section): [15]

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

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.[16] 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

[15] Note that at that time, there was only one Q/V section. The high stakes GRE score was based on all items in that section.

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

Table 2. *Performance in GRE Test by Gender, Race and Ethnicity.*

	Males (M) (1)	Females (F) (2)	M-F (3)	Whites (W) (4)	Blacks (B) (5)	Hispanics (H) (6)	Asians (A) (7)	W-B (8)	W-H (9)	W-A (10)
(a) High stakes score										
Quantitative section	55.58	40.28	15.30	46.99	21.85	36.39	62.30	25.13	10.59	− 15.32
	(27.43)	(24.38)	(0.85)	(25.46)	(21.80)	(25.33)	(26.76)	(1.62)	(1.75)	(1.75)
Number of observations	1368	2553		3026	265	224	224			
Verbal section	62.90	56.45	6.45	60.55	37.37	48.73	60.84	23.18	11.82	− 0.30
	(24.96)	(24.54)	(0.79)	(23.69)	(24.23)	(26.20)	(26.85)	(1.58)	(1.67)	(1.56)
Number of observations	1465	2845		3380	248	221	255			
(b) Low stakes score										
Quantitative section	43.93	33.16	10.77	37.55	18.90	32.58	55.20	18.65	4.97	− 17.64
	(31.34)	(25.48)	(0.93)	(27.78)	(19.72)	(26.39)	(30.38)	(1.75)	(1.90)	(1.90)
Number of observations	1368	2553		3026	265	224	224			
Verbal section	52.48	50.34	2.14	52.79	35.08	42.22	51.78	17.71	10.57	1.01
	(30.53)	(27.65)	(0.92)	(28.17)	(24.08)	(27.87)	(31.42)	(1.85)	(1.95)	(1.83)
Number of observations	1465	2845		3380	248	221	255			

Notes: The table reports students' test scores in the high stakes and low stakes sections of the GRE and the gaps between males and females, and whites and minorities. Test scores are reported in percentile ranks. Standard deviations are reported in parenthesis.

or blacks/Hispanics/Asians) relative to the omitted category (males or whites). To simplify the exposition, 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 an individual's characteristics to affect the change in performance between the high and low stakes situation.[17]

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 ranks using the GRE official percentile rank tables.[18] All results presented below are based on GRE percentile ranks. As we show below, we obtain similar results when using raw scores, log of raw scores or z-scores.

3. Results

3.1. *Differences in Performance by Gender, Race and Ethnicity*

Panel (a) of Table 2 exhibits examinees' performance in the high stakes test for males, females, whites, blacks, Hispanics, and Asians and the gaps between groups.[19] Similar to other compar-

[17] 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 augments differences when groups start at different levels and then remain parallel or diverge (see Lord's Paradox—Lord, 1967) and attenuates 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.

[18] For more information regarding on the interpretation of GRE scores, exam administration and validity, see Educational Testing Service (2007).

[19] The percentile scores of males and females do not add to 100 since they are constructed using the official GRE tables, which include international examinees and are based on several years of data.

isons 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 6.5 percentile points higher than females. Asians have the highest achievements among all ethnic/racial groups in the Q-section. Their test scores are about 15 percentile points above those of 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. In the verbal section, whites outperform Asians, although the difference between groups is not statistically significant. The gap between whites and blacks is a bit smaller (23 percentile points), while the gap between whites and Hispanics is about 12 percentile points. With the exception of whites versus Asians in the verbal section, all gaps between groups in the high stakes section are statistically significant.

Panel (b) of Table 2 reports students' performance in the experimental section and gaps by gender and race/ethnicity. On average, performance in the low stakes test is lower than in the high stakes test for all groups. Notably, gaps between males and females or whites and blacks or Hispanics are narrower in the experimental section (even though they are still statistically significant). For example, the score gap between males and females shrinks from 15 to 11 percentile points in the Q-section and from 7 to 2 percentile points in the V-section. The score gap between whites and blacks shrinks from 25 to 19 percentile points in the Q-section and from 23 to 18 in the V-section, and the gap between whites and Hispanics shrinks from 11 to 5 percentile points in the Q-section and from 12 to 11 percentile points in the V-section. The gap between Asians and whites in the Q-section widens between the high and the low stakes test (from 15 to 18 percentile points) because Asians outperform whites in this exam.

Table 3 reports the change in performance between the high and the low stakes section for each demographic group (first row of each panel) and the difference (second and third rows) in the drop in performance between males and females or between whites and blacks/Hispanics/Asians. 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 gap in the drop in performance between males and females is significant and stands at 4.5 percentile points (SE = 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 (although it is still significant), which is equivalent to a 30% 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 achievement. This finding is important as it suggests that our results are unlikely to be driven by differences in family background and academic achievement.

We also find a similar gender gap in the V-section. Males' scores drop by 10.4 percentile points, on average, while females' scores drop by a smaller magnitude of 6.1 percentile points. That is, males' scores drop by 4.3 percentile points (SE = 0.783) more relative to females. Note that the proportional drop in males' performance is also larger than females'. Namely, males' scores drop by 21% while females' scores drop by 18% in the Q-section. Similarly, we find that males' scores in the V-section drop by 17% while females' scores drop by 11%.

The stratification by race/ethnicity shows that 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 Asians' performance drops by 7 percentile points, blacks' performance drops by 3 percentile points and Hispanics' performance drops by 3.8 percentile points. Differences in

Table 3. *Difference in Performance Between High and Low Stakes Test by Gender, Race and Ethnicity.*

	Males (1)	Females (2)	Whites (3)	Blacks (4)	Hispanics (5)	Asians (6)
(a) Quantitative section						
High stakes − low stakes	11.644	7.115	9.431	2.951	3.808	7.107
	(0.683)	(0.385)	(0.399)	(0.863)	(1.346)	(1.561)
Raw difference between males and		4.529		6.480	5.623	2.323
females or whites and minority group		(0.784)		(0.949)	(1.402)	(1.609)
Controlled difference		3.905		4.276	5.205	3.145
		(0.820)		(1.050)	(1.402)	(1.701)
(b) Verbal section						
High stakes – low stakes	10.421	6.108	7.755	2.282	6.511	9.067
	(0.673)	(0.400)	(0.390)	(1.316)	(1.457)	(1.625)
Raw difference between males and		4.313		5.473	1.244	− 1.312
females or whites and minority group		(0.783)		(1.371)	(1.506)	(1.669)
Controlled difference		3.577		3.150	0.629	− 0.555
		(0.821)		(1.472)	(1.533)	(1.706)

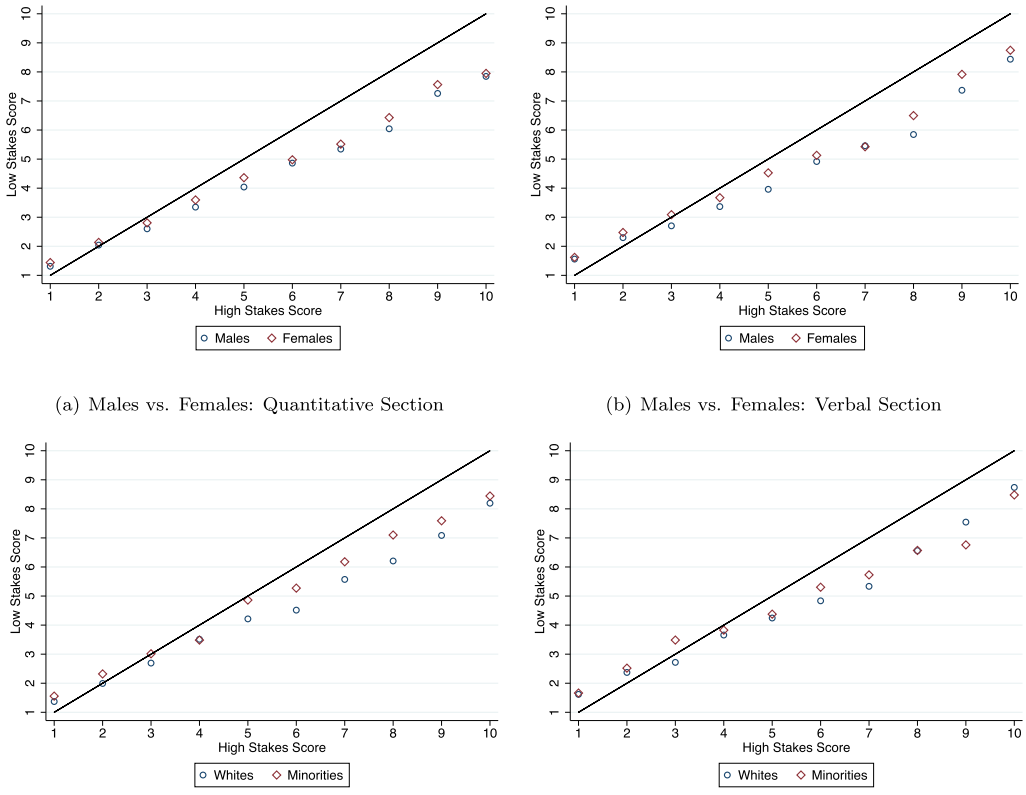
Notes: The first row of each panel reports differences in individual's performance between the high and the low stakes section of the GRE by gender, race and ethnicity. The second row of each panel reports the differences in the drop in performance between males and females or whites and blacks/Hispanics/Asians. The third row of each panel reports differences between groups controlling for the following individual covariates: mother's and father's education, indicators for gender or 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 parentheses.

the performance drop 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 4.3 percentile points (SE = 1.05). The equivalent difference between whites and Hispanics is 5.21 (SE = 1.40) and the difference between whites and Asians is 3.2 (SE = 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. We suspect that the different pattern obtained for Asians and Hispanics in the V-section could be related to language dominance.

Overall, the evidence presented in Table 3 shows that males and whites exhibit the largest drop in performance between the high and the low stakes tests compared with females and minorities. Our results are robust to non-linear transformations and alternative definitions of the dependent variable, as reported in Appendix Table A3. 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 the third row, we report results based on z-scores.[20] All alternative metrics yield results that are equivalent to our main findings: males' drop in performance between the high and low stakes section is 5% or 0.17 SD larger than the drop of females; whites' drop in performance in the Q-section is 8% or 0.23 SD larger than the drop of blacks; 7% or 0.23 SD larger than the drop of Hispanics and 7% or 0.19 SD larger than the drop of Asians. These additional results show that our findings are not driven by a specific scale used to measure achievement. Furthermore, as we show in Figure 1, we obtain the same results when we rely only on the ordinal information embedded in scores.

The fourth row of each panel in Table A3 replicates our main results using the samples of examinees randomised into experimental sections with extended time limits (67.5 minutes for the

[20] Z-scores are computed using the mean and standard deviation of the high stakes test.



(a) Males vs. Females: Quantitative Section

(b) Males vs. Females: Verbal Section

(c) Whites vs. Minorities: Quantitative Section

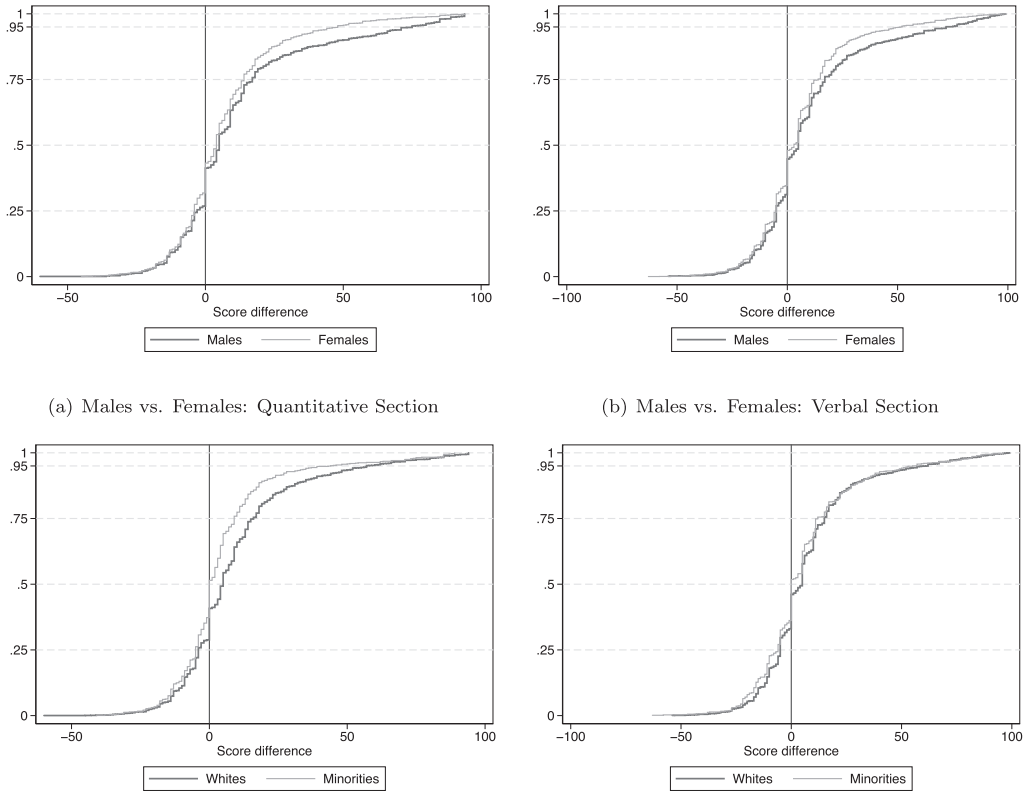
(d) Whites vs. Minorities: Verbal Section

Fig. 2. *Score Distribution in High and Low Stakes Test.*
Notes: The figure shows the low stakes score as a function of the high stakes score. Scores for each gender, race and ethnicity are mapped into deciles using the distribution of the high stakes score of each group.

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. In addition, they demonstrate that our results are not sensitive to time constraints or to differential responses by gender or ethnicity to the length of the exam.

We also examine how the change in performance varies by students' performance in the high stakes exam. To examine this issue, we divide the high stakes score distribution for each group into deciles and define for each individual his/her score decile in the high and low stakes section. We plot in Figure 2 the average score decile of the low stakes section as a function of the score decile in the high stakes section by gender and race. Overall, with the exception of those located at the bottom of the test score distribution in the high stakes section, there is a similar drop in performance (in percentage terms) in all parts of the high stakes score distribution, with males having a larger drop relative to females and whites having a larger drop in performance relative to minorities.

Another relevant question is whether the results are driven by just a small group of males or whites that experience a large performance drop or by most individuals who belong to those



(a) Males vs. Females: Quantitative Section

(b) Males vs. Females: Verbal Section

(c) Whites vs. Minorities: Quantitative Section

(d) Whites vs. Minorities: Verbal Section

Fig. 3. *Distribution of Score Difference.*
Notes: The figure shows the CDF of the difference in score (measured in percentiles) between the high and low stakes section by gender/race and section.

demographic groups. Figure 3 plots the CDF of the difference in score (measured in percentiles) between the high and low stakes section by gender/race and section. For most individuals the change in performance is of a few percentile points, but males have a larger drop in performance than females. In addition, a larger proportion of males has a substantial drop in performance relative to females. The same pattern is observed for whites versus minorities in the Q-section: whites have a larger drop in performance relative to minorities and those who have a very large drop in performance are disproportionally represented by whites.

We further explore this issue by re-estimating our main model after dropping from each demographic group those individuals with the largest drop in performance (i.e., those in the top 10-percentile distribution of the performance change in their demographic group). Results from this subsample (reported in the last row of Appendix Table A3) show that differences between demographic groups in performance change are very similar to differences observed for the full sample. Again, males and whites have a larger drop in performance relative to females and minorities. This implies that the results are not driven only by a few extreme values of a specific demographic group.

Table 4. *Performance in High versus Low Stakes Tests by Gender and Race/Ethnicity—Quantitative Section.*

	High stakes		Low stakes		High − low stakes		**Controlled difference**
	Males (1)	*Females* (2)	*Males* (3)	*Females* (4)	*Males* (5)	*Females* (6)	***(Males − Females)*** (7)
Whites	56.701	41.800	43.914	34.161	12.787	7.639	**4.904**
	(26.403)	(23.342)	(25.179)	(31.132)	(0.793)	(0.437)	**(0.945)**
Blacks	28.769	19.605	24.215	17.175	4.554	2.430	**0.186**
	(27.739)	(19.039)	(16.851)	(26.150)	(2.146)	(0.906)	**(2.531)**
Controlled difference (Whites − blacks)					**5.485**	**3.568**	
					(2.405)	**(1.153)**	
Hispanics	44.022	31.363	38.405	28.748	5.618	2.615	**0.181**
	(27.048)	(22.875)	(23.230)	(29.775)	(2.422)	(1.561)	**(3.502)**
Controlled difference (Whites − Hispanics)					**7.464**	**4.071**	
					(2.608)	**(1.663)**	
Asians	72.167	56.386	66.071	48.671	6.095	7.714	**− 1.307**
	(23.589)	(26.875)	(29.090)	(29.509)	(2.603)	(1.955)	**(4.678)**
Controlled difference (Whites − Asians)					**9.266**	**− 0.399**	
					(2.955)	**(2.055)**	

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. Test scores are reported in percentile ranks. Standard deviations and robust standard errors are reported in parentheses.

3.2. *Within Race/Ethnicity and Gender Differences in Performance*

We check for gender and race/ethnicity interactions by examining whether differences between males and females appear across all racial/ethnic groups and whether differences between whites and minorities show up for males and for females.[21]

Table 4 reports performance in the high and low stakes section for each gender and ethnicity/race as well as differences in performance between males and females within each race/ethnicity and between whites and minorities for males and females separately. We focus on the Q-section, as 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, who behave similarly to white females.

Comparisons between males and females within each racial/ethnic group reveal that males exhibit a larger drop in performance than females among whites, blacks and Hispanics, although differences between genders are only statistically significant among whites. In contrast, we observe no gender differences among Asians. In fact, the drop observed among females is even larger than the drop observed among males, although the difference is not statistically significant.

3.3. *Heterogeneous Effects*

Table 5 reports the gender gap in students' performance in high and low stakes tests for different subsamples stratified by undergraduate GPA (UGPA), student's major, intended field of graduate studies and mother's education. We focus on gender gap and not on gap by race/ethnicity, since

[21] 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, so the results should be taken with caution.

Table 5. *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 (1)	*Females* (2)	*Males* (3)	*Females* (4)	*Diff.* (5)	*Males* (6)	*Females* (7)	*Diff.* (8)	*Males* (9)	*Females* (10)	*Raw diff.* (11)	*Controlled diff.* (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	7.103
			(24.462)	(18.445)	(2.793)	(17.397)	(25.557)	(2.800)	(1.947)	(0.851)	(2.124)	(2.320)
B−	144	266	43.028	28.267	14.761	34.458	24.034	10.425	8.569	4.233	4.336	2.295
			(25.528)	(19.377)	(2.248)	(19.386)	(26.841)	(2.306)	(1.939)	(0.837)	(2.111)	(2.294)
B	426	855	48.962	36.063	12.899	38.418	29.958	8.460	10.545	6.105	4.439	3.492
			(25.942)	(22.755)	(1.415)	(23.056)	(28.660)	(1.486)	(1.152)	(0.613)	(1.305)	(1.375)
A−	393	717	63.237	46.815	16.422	51.438	37.756	13.682	11.799	9.059	2.740	3.109
			(24.906)	(23.935)	(1.524)	(27.150)	(31.765)	(1.812)	(1.273)	(0.823)	(1.516)	(1.641)
A	251	490	69.821	50.700	19.121	53.801	42.382	11.419	16.020	8.318	7.702	7.980
			(25.227)	(23.462)	(1.869)	(27.321)	(34.295)	(2.318)	(1.908)	(0.959)	(2.135)	(2.529)
Undergrad major in physics, math, comp. or eng.	362	132	78.644	69.955	8.689	65.870	63.295	2.575	12.773	6.659	6.114	4.244
			(17.321)	(23.107)	(1.935)	(27.074)	(31.352)	(3.078)	(1.549)	(2.121)	(2.624)	(2.829)
Grad intended studies in physics, math, comp. or eng.	340	122	77.674	70.574	7.100	65.515	64.369	1.146	12.159	6.205	5.954	4.457
			(18.191)	(21.707)	(2.024)	(25.909)	(31.265)	(3.161)	(1.596)	(2.167)	(2.689)	(2.875)
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	2.091
			(26.374)	(22.986)	(1.687)	(23.117)	(27.255)	(1.716)	(1.235)	(0.672)	(1.405)	(1.497)
College or some college	621	1228	58.097	39.965	18.132	46.018	33.800	12.218	12.079	6.165	5.914	5.732
			(26.830)	(23.495)	(1.214)	(24.850)	(32.199)	(1.356)	(1.013)	(0.529)	(1.142)	(1.218)
At least some graduate studies or professional degree	357	619	63.588	48.724	14.864	49.952	39.069	10.883	13.636	9.654	3.982	2.829
			(25.921)	(25.125)	(1.689)	(27.697)	(32.106)	(1.953)	(1.455)	(0.929)	(1.725)	(1.879)

Table 5. *Continued*

	Number of obs.		High stakes score			Low stakes score			High stakes − low stakes			
	Males (1)	*Females* (2)	*Males* (3)	*Females* (4)	*Diff.* (5)	*Males* (6)	*Females* (7)	*Diff.* (8)	*Males* (9)	*Females* (10)	*Raw diff.* (11)	*Controlled diff.* (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	1.121
			(23.915)	(22.205)	(2.864)	(24.116)	(26.541)	(3.140)	(2.036)	(1.513)	(2.536)	(3.641)
B−	167	275	53.695	47.949	5.746	46.144	44.447	1.696	7.551	3.502	4.049	0.677
			(26.025)	(23.273)	(2.389)	(25.274)	(27.002)	(2.545)	(1.719)	(1.129)	(2.056)	(2.583)
B	436	945	58.690	51.935	6.755	50.197	46.309	3.888	8.493	5.626	2.867	2.514
			(23.905)	(23.512)	(1.368)	(25.740)	(29.117)	(1.555)	(1.165)	(0.664)	(1.340)	(1.392)
A−	405	799	68.225	62.016	6.208	54.138	55.253	−1.115	14.086	6.763	7.323	7.098
			(22.888)	(23.097)	(1.405)	(27.634)	(32.032)	(1.780)	(1.391)	(0.793)	(1.600)	(1.738)
A	292	560	74.137	66.366	7.771	61.709	58.664	3.045	12.428	7.702	4.726	3.388
			(20.914)	(22.573)	(1.589)	(28.622)	(31.125)	(2.130)	(1.598)	(0.933)	(1.850)	(2.064)
Undergrad major in physics, math, comp. or eng.	388	161	66.781	65.839	0.942	54.036	62.012	−7.976	12.745	3.826	8.919	7.547
			(24.124)	(25.365)	(2.296)	(25.708)	(31.769)	(2.824)	(1.424)	(1.301)	(1.929)	(2.063)
Grad intended studies in physics, math, comp. or eng.	378	142	66.341	66.056	0.285	53.643	60.535	−6.892	12.698	5.521	7.177	7.506
			(23.796)	(24.881)	(2.372)	(27.411)	(31.356)	(2.986)	(1.445)	(1.340)	(1.970)	(2.135)
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	4.197
			(26.892)	(23.959)	(1.679)	(25.717)	(29.148)	(1.810)	(1.305)	(0.745)	(1.502)	(1.611)
College or some college	658	1354	64.114	56.078	8.036	53.157	49.908	3.249	10.957	6.171	4.787	4.750
			(23.671)	(23.942)	(1.134)	(27.139)	(30.420)	(1.343)	(1.033)	(0.591)	(1.190)	(1.281)
At least some graduate studies or professional degree	376	731	68.830	63.848	4.982	58.495	56.791	1.704	10.335	7.057	3.278	3.614
			(22.931)	(24.094)	(1.504)	(28.787)	(30.521)	(1.865)	(1.318)	(0.827)	(1.556)	(1.702)

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 parentheses. Sample sizes are reported in columns 1 and 2.

subgroups are too small for that stratification. 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, 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 in the verbal section. 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 UGPA, exhibit a significant drop in performance between the high and the low stakes sections (both the quantitative and the verbal).[22] Males' performance drop is larger than females' drop across all levels of UGPA (see columns 11 and 12), and is evident both in absolute and in percentage terms.

The next two rows of Table 5 (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.[23] While females represent the majority among the full population of GRE examinees (65%), they are a minority among math and science students (26%). It is therefore interesting to examine whether we find the same results in a subsample where selection by gender goes in the opposite direction.

As seen in columns 3 and 4 of Table 5, achievement in the GRE Q-section is much higher among math and science students relative to the full sample 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 full sample, but they score slightly lower compared with those students with an 'A' UGPA. 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. In contrast, there is no gender gap achievement in the high stakes V-section in the subsamples of math and science students.

Achievement of math and science students in the low stakes Q-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 is even insignificant. The pattern for the V-section is similar, with math and science females even outperforming their male counterparts in the low stakes V-section.

Even in this subsample of math and science students, the drop in performance between the high and the low stakes test is larger for males (who reduce their performance by about 12–13 percentile points in both subjects) compared with 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 larger drop in males' performance is evident both in absolute terms and relative to the outcome means

[22] We use UGPA to stratify the sample (instead of using the score in the high stakes section) because it provides a measure of students' performance that is taken independently and before the realisation of the dependent variable.

[23] We focus here on a more limited number of fields than the traditional STEM (science, technology, engineering and mathematics) 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.

in the high stakes test. The gender differences in relative performance in these subsamples are 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 finding is important because it shows that the larger drop in performance among males is found even in subsamples that exhibit no differences in performance in the high stakes test.

We also looked at gender gaps within groups stratified by mother's education. We were curious to check 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 situations. This turned out not to be the case. The gender gap in relative performance between high and low stakes tests appears across all levels of maternal education in both the quantitative and the verbal section.

4. Discussion

The evidence presented above shows that males and whites exhibit a larger difference in performance between high and low stakes tests compared with females and minorities. The larger decline in performance found among males and whites could be due to at least distinct two reasons: (*i*) males and whites do not exert as much effort in low stakes situations compared with females and minorities, respectively; *(ii)* females and minorities find it relatively more difficult to deal with high stakes and stressful situations.[24] We examine below the plausibility of these alternative explanations and discuss some other interpretations. We acknowledge that our data do not allow us to rigorously test the relative contribution of each explanation. Nevertheless, we believe that the evidence presented below provides interesting directions for further research.

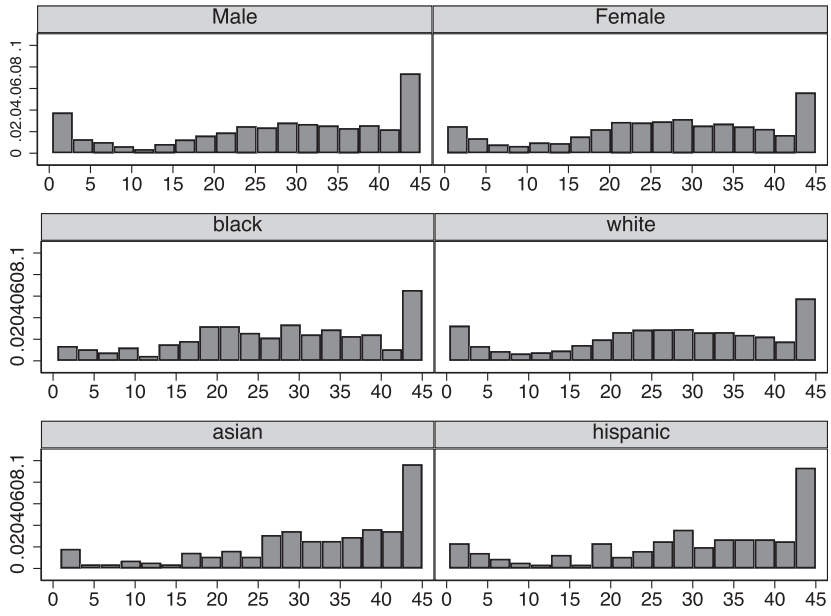
4.1. *Do Males and Whites Exert Less Effort in Low Stakes Situations?*

To examine the likelihood of the first explanation, we would ideally like to measure the 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 4 plots the distribution of time spent by examinees in the experimental Q- and V-sections by gender, race and ethnicity.[25] The figure shows that there is a significant variation in time invested in the experimental section. Some examinees spent very little time and some exhausted the time limit (45 minutes for the Q-section and 30 minutes for the V-section).

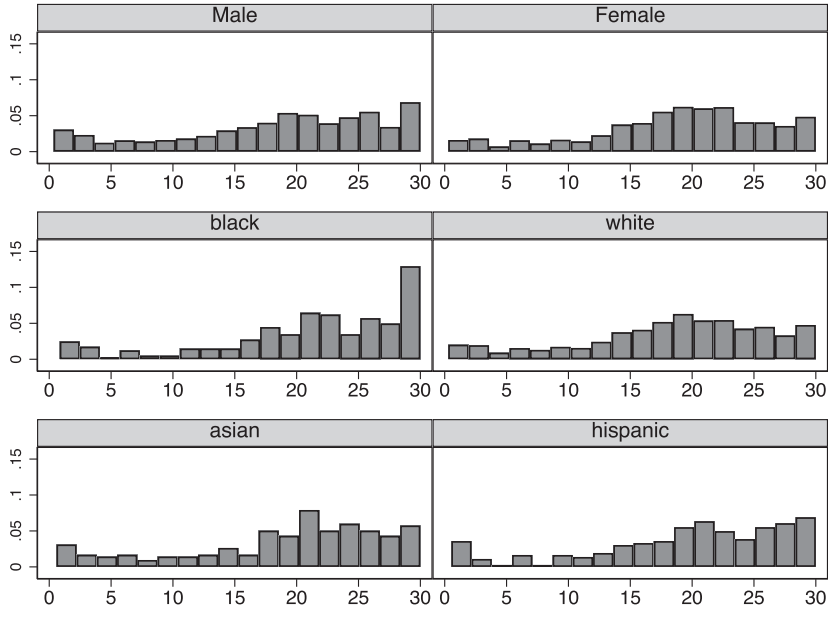
Figure 5 exhibits the relationship between achievement in the experimental section and the time invested in that section for males, females, whites, blacks, Hispanics and Asians. The figure shows that achievement increases with time invested in the quantitative section for all gender, racial and ethnic groups. The relationship between time invested and performance in the verbal section is also positive at the lower values of the distribution, but switches sign after about 20 minutes. Overall, it is clear from the figures that it is impossible to receive a high score without investing some minimal amount of time. We therefore conclude that subjects who invested very

[24] Alternatively, males and whites are arguably better able to boost their performance when stakes are high or the task is challenging. 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.

[25] Unfortunately, there is no information on time spent in the real GRE test. However, students usually exhaust the time limit.

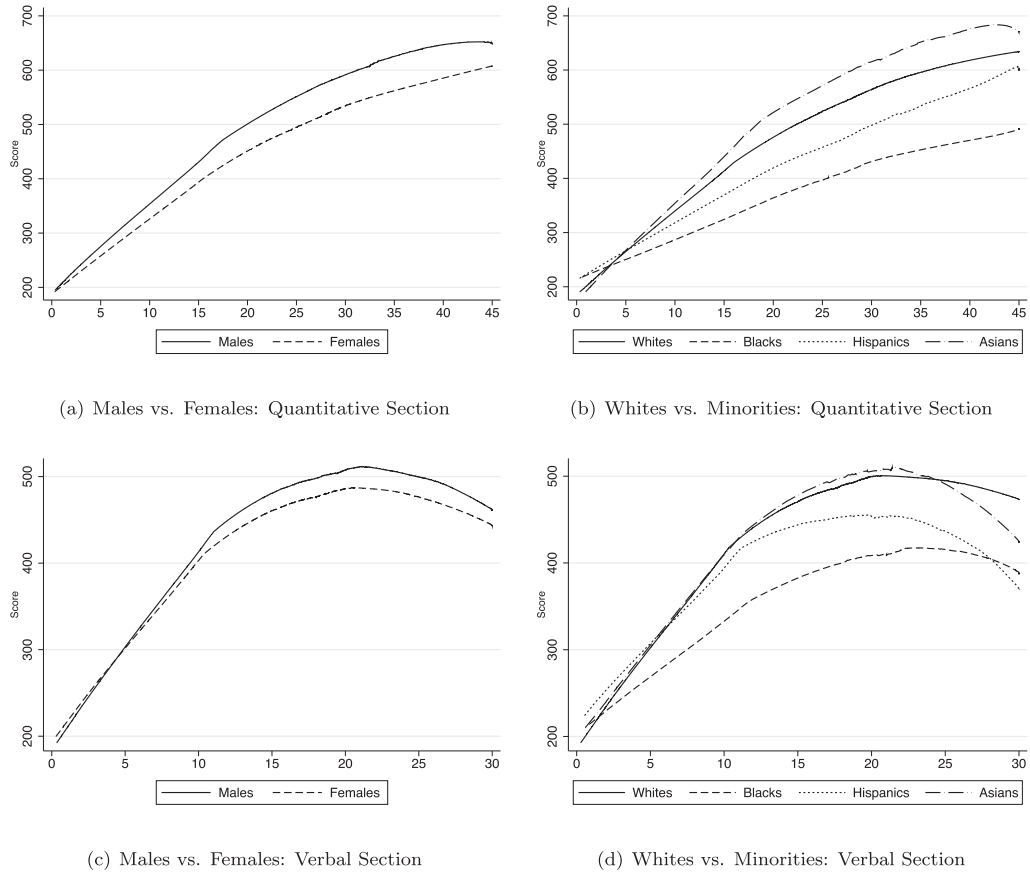


(a) Quantitative Section



(b) Verbal Section

Fig. 4. *Distribution of Time Invested in the Experimental Section.*
Notes: The histograms plot the distribution of time spent by examinees in the experimental Q- and V-sections by gender, race and ethnicity.



(a) Males vs. Females: Quantitative Section

(b) Whites vs. Minorities: Quantitative Section

(c) Males vs. Females: Verbal Section

(d) Whites vs. Minorities: Verbal Section

Fig. 5. *Relationship between Time Invested in the Experimental Section and Test Score Achieved in that Section.*

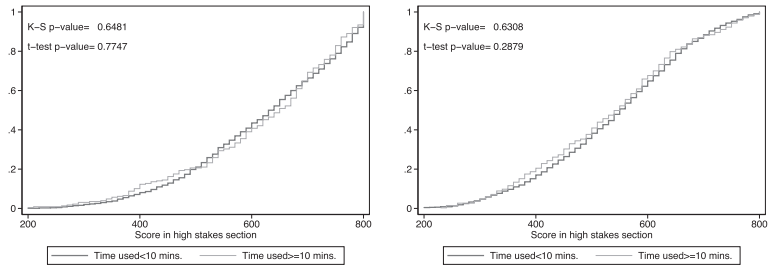
Notes: The figure exhibits the relationship between achievement in the experimental section and time invested in that section using a local weighted regression.

little time were obviously not exerting much effort. We define an indicator of low effort for individuals who invested less than 10 minutes in the experimental section. While the 10-minute cut-off is somewhat arbitrary, we chose a time threshold that clearly suggests low effort and cannot be confounded with the ability to solve a test quickly.[26]

We plot in Figure 6 the cumulative test score distribution in the high stakes section, stratifying individuals by time spent in the experimental section (below 10 minutes vs. at least 10 minutes). Each quadrant in the figure refers to a specific demographic group and section (quantitative or verbal). We also report *p*-values of Kolmogorov–Smirnov (K–S) tests of equality between the two distributions and *p*-values of t-tests of equality of means (assuming unequal variances).

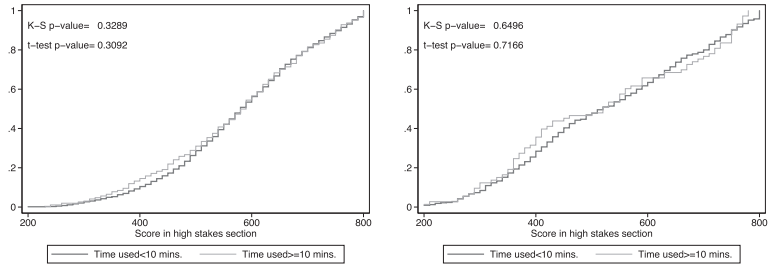
For the Q-section (panels a through d), we see no differences in the high stakes test score distribution between subjects who invested low effort in the experimental section and those who

[26] 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. A total of 94% of all those who spent less than 10 minutes in the V-section were also located below the 58th percentile.



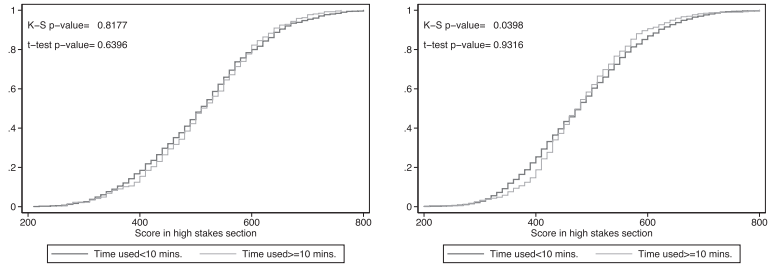
(a) Males: Quantitative Section

(b) Females: Quantitative Section



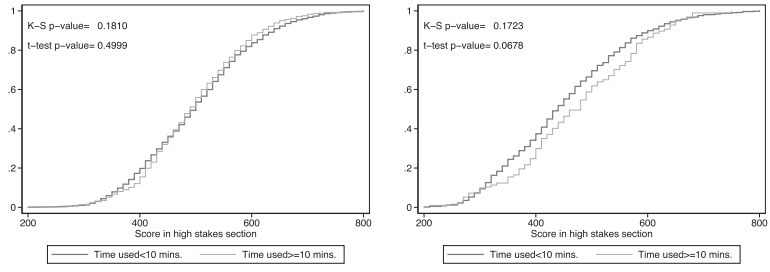
(c) Whites: Quantitative Section

(d) Minorities: Quantitative Section



(e) Males: Verbal Section

(f) Females: Verbal Section



(g) Whites: Verbal Section

(h) Minorities: Verbal Section

Fig. 6. *CDFs of Test Score in High Stakes Section by Effort Invested in Experimental Section.*
Notes: The figure plots the cumulative test score distribution in the high stakes section for individuals who spent less than 10 minutes versus at least 10 minutes in the experimental section.

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 shows that achievement in the high stakes section is unrelated to effort levels invested in the low stakes section, and implies that baseline differences in achievement in the high stakes section between demographic groups are unlikely to explain group differences in effort levels. Given that the chances of improving one's score are probably lower for individuals who obtained higher scores in the high stakes section, the result reported in Figure 6 suggests that individuals were not thinking about the chances of winning the prize when deciding about effort levels in the low stakes section.

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 others 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 find lower effort levels among those with lower scores in the high stakes section (although the difference in distributions is not statistically significant). These differences are the opposite of what we would expect if experiment participants were considering the monetary incentive when deciding about effort levels in the low stakes test. 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 6 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.

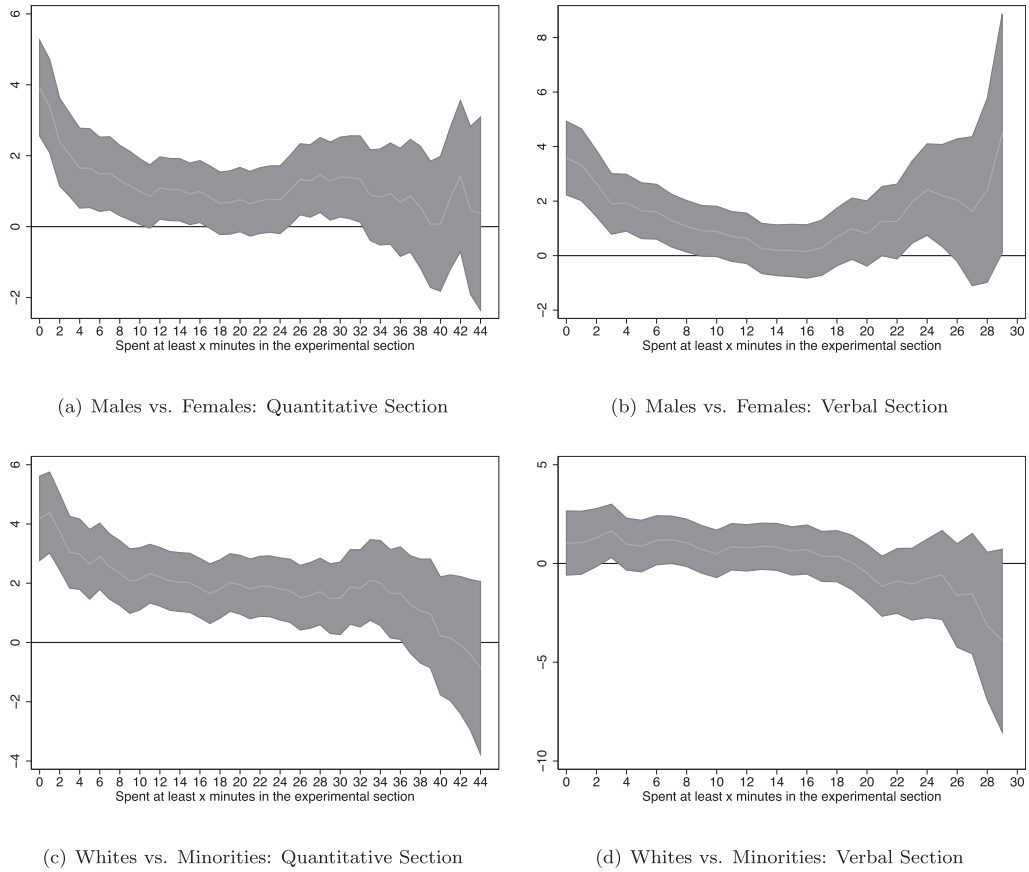
Table 6 reports the share of examinees who invested less than 10 minutes in the experimental Q- and V-sections, stratified by gender, race/ethnicity, academic achievement and parental education. We also report p-values that test for equality of proportions between groups. The results show that males appear to exert less effort in the experimental section compared with females: 17% of the males who participated in the Q-experiment spent less than 10 minutes in the experimental section, while the equivalent percentage among females is 13%. 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 smaller and not statistically significant.

The stratification of the sample by background characteristics and achievement shows 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' achievement, 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' academic performance, suggesting that other factors are likely to play a more important role in determining performance in low stakes situations. The lack of a relationship between students' academic performance and effort invested in the low

Table 6. *Share of Experiment Participants who Spent Less than 10 Minutes in the Experimental Section.*

Share who spent less than 10 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	0.0032	0.0002
Race/ethnicity		
Whites	0.152	0.154
Blacks	0.106	0.101
p-value of difference: Whites − Blacks	0.0405	0.0227
Hispanics	0.129	0.140
p-value of difference: Whites − Hispanics	0.3557	0.5714
Asians	0.071	0.161
p-value of difference: Whites − Asians	0.0010	0.7871
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	0.0860	0.2100
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	0.0580	0.1160
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	0.1300	0.0220
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	0.7220	0.0080
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; *p*-values for comparisons by gender and race are based on tests for equality of proportions; *p*-values for other categories are based on chi-squared tests.



(a) Males vs. Females: Quantitative Section

(b) Males vs. Females: Verbal Section

(c) Whites vs. Minorities: Quantitative Section

(d) Whites vs. Minorities: Verbal Section

Fig. 7. *Gap in Differential Performance for those Spending at Least X Minutes in the Experimental Section. Notes*: The figure plots estimates along with confidence intervals for differential performance by gender and race from a series of regressions that limit the sample to individuals who spent at least X minutes in the experimental section (for X = 0–45 in the Q-section and X = 0–30 in the V-section).

stakes section suggests also that our previous results on group differences in performance drop are unlikely to be explained by differences in academic achievement between groups.

We plot in Figure 7 estimates along with confidence intervals for differences in the change in performance from the high to the low stakes section between males and females or whites and minorities when we limit the sample to individuals who spent at least X minutes in the experimental section (for X = 0–45 in the Q-section and X = 0–30 in the V-section).[27] The figure shows that there is a larger gap in performance by gender or race among those who spent a short time in the experimental section. Nevertheless, we observe that the larger drop in performance among males and whites relative to females and minorities is evident along the whole distribution of time spent in the experimental section. Appendix Table A4 reports estimates for specific points of the figure (individuals who spent at least 10 minutes in the experimental

[27] The figure reports estimate and confidence intervals obtained from a series of regressions based on equation (1), where we limit the sample to individuals spending at least X minutes in the experimental section.

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 (1)	*Males − Females* (2)	*White − blacks* (3)	*Whites − Hispanics* (4)	*Whites − Asians* (5)	*Mean* (6)	*Males − Females* (7)	*Whites − blacks* (8)	*Whites − Hispanics* (9)	*Whites − Asians* (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		−0.004	−0.032	−0.016	−0.008
		(0.005)	(0.009)	(0.012)	(0.010)		(0.008)	(0.019)	(0.017)	(0.015)
Odds ratio		1.254	1.473	2.471	1.393		1.003	1.874	1.370	1.177

Notes: Columns 1 and 6 report the share of examinees who improved their score in the experimental Q- or V-section, 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 parentheses. The third row reports odds ratios relative to males/whites.

section and those who spent at least 3 minutes). The last row of the table reports estimates from a model that uses the full sample and controls for a fourth-order polynomial of time invested in the low stakes section.[28] 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 summarise, 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 males and whites can be partly explained by a lower level of effort exerted by these groups in the low stakes section.

4.2. *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 males 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 could be due to the volatility of, 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 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

[28] Results are very similar if we use a lower or higher degree of polynomial.

Table 8. *Differences in Gap in Performance Drop between Students Taking Test for Practice and Other Students.*

	Gaps by gender		Gaps by race/ethnicity					
	Female (1)	*Female × practice* (2)	*Black* (3)	*Black × practice* (4)	*Hispanic* (5)	*Hispanic × practice* (6)	*Asian* (7)	*Asian × practice* (8)
Quantitative section	− 3.595	− 4.948	− 4.314	0.536	− 5.002	− 2.262	− 2.480	− 10.524
	(0.848)	(3.436)	(1.014)	(6.174)	(1.470)	(4.469)	(1.781)	(4.053)
Verbal section	− 3.696	1.580	− 2.671	− 3.883	− 0.761	2.017	0.352	3.178
	(0.852)	(2.922)	(1.580)	(3.738)	(1.607)	(4.769)	(1.755)	(7.309)

Notes: The table reports estimates from a regression of test score change on indicators for the different demographic groups and the interaction between demographic groups and practice exam. The model also controls for an indicator of practice exam and student's background characteristics detailed in Table 2. Robust standard errors are reported in parentheses.

error of measurement of difference scores is greater than 1.65.[29] Roughly 1.5% of examinees have a significant score gain in the experimental Q-section, and 5.3% have one 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, the second row reports differences after controlling for students' background characteristics, and the third row reports odds ratios between females/minorities and males/whites. Overall, we find very small and insignificant differences in the likelihood of improving the score by gender. Odds ratios are close to one for both sections (i.e., small effect size), meaning that the odds of improving the score for males and females are similar. With the exception of Hispanics in the Q-section and of blacks in the V-section, all other differences between whites and minorities are small and insignificant, with odds ratios that are close to one.

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% marked 'practice' as one of the reasons for taking the exam.[30] If test anxiety hinders performance of females, blacks or Hispanics relative to males or whites in the high stakes section, we would expect to find smaller group differences in the performance drop between the high to the low stakes section among those taking the test for practice.[31] To examine this, we estimated our basic model of drop in performance (as in Table 3) while adding interactions between an indicator for taking the test for practice and the demographic groups. Estimates reported in Table 8 show that the gap between demographic groups among those taking the exam for practice is not smaller than the gap estimated among those who were taking the exam for admission to graduate school or fellowship application and were probably facing a more stressing situation.

[29] 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).

[30] The main reasons were 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%). Applicants were instructed to select all reasons that applied, so 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.

[31] 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.

Taken together, the evidence presented in Tables 7 and 8 suggests that test anxiety in the high stakes section is unlikely to be the reason for the smaller change in performance between the high and the low stakes tests observed among females and minorities.

4.3. *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 possible, 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 an amount distributed to such a small number of participants seems too low to have a significant differential effect in performance. Alternatively, it is arguably the case that differences in performance in the experimental section 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 in the state of residence of the examinee. We use two measures of earnings: median annual earnings of full-time workers and median annual earnings of college graduates computed separately by gender and state.[32] If the monetary prize or the opportunity cost of time had any impact on performance in the experimental section, we should expect a smaller reduction in performance in states with lower earnings levels. In Appendix Table A7 we report 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 in 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.

Another alternative explanation for differential changes in performance could be that performance of females and minorities is lower than expected in the high stakes section owing to stereotype threat (e.g., Steel and Aronson, 1995; Steele, 1997). 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 suggests that stereotype threat is unlikely to explain our main results, as the theory would predict that females would respond negatively only to the quantitative section. Moreover, stereotype threat theory implies that Asians should respond differently than blacks and Hispanics in the Q-section, but our findings are similar for the three groups.

We further assess the likelihood of stereotype threat explanation by examining the relationship between gender stereotypes in math and verbal achievement in the state of residence of the examinees, and the differential change in performance. To proxy for gender stereotypes in 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 favouring boys in math and science and favouring girls in reading and which was shown by the authors to be positively associated

[32] 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.

with other measures of gender stereotype attitudes at the state level.[33] Higher values in this index mean a stronger gender stereotype. To facilitate interpretation of the results, we transform this index into a z-score. 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 differential in the V-section in states with a higher stereotype index. In Appendix Table A6, we examine 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. Estimates for the interaction term between female and the stereotype index are all small and insignificant, meaning that there is no apparent relationship between state gender stereotypes and the gender gap in differential performance between the high and the low stakes section. Moreover, the estimates' 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 a differential drop in performance. However, as we note above, we observe the same pattern of gender and racial/ethnic differences across different subsamples and even in subsamples that exhibit similar performance in the high or the low stakes section.

It could also be the case that females and minorities become less fatigued by the GRE examination than males and whites, respectively, and therefore exhibit a smaller drop in performance in the experimental section. 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 compared with 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. Moreover, 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. Lastly, as shown in Appendix Table A4, the fact that we can replicate our results in the samples of students randomised into the extended time-limit sections provides strong evidence that mitigates this concern.[34]

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 familiarisation. 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 Educational Testing Service (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.

[33] 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% of the students.

[34] Cotton *et al*. (2013) find that the advantage of males over females erodes over a sequence of quizzes of equal importance. However, they note that this cannot be attributed to differential fatigue because the effect was also found when there was a two-week gap between the different quizzes. In fact, they interpret their result by claiming that males seem to become less excited about the quizzes over time, which lowers the stakes of these quizzes. This interpretationis thus consistent with our findings.

Results of these expanded models are reported in Appendix Table A7, together with results from our main specification. All estimates are highly similar to our main results, suggesting that learning or test familiarisation cannot explain our findings.

Finally, our results might also be explained by the fact that some groups might get 'bored' faster than others, and this might harm their performance. This would imply that there are group differences in the likelihood of getting bored when incentives are low. We believe that the implications of our results are still relevant under this alternative interpretation.

4.4. *Relation to Other Assessment Tests*

Our findings demonstrate 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 analysing test-score gaps. We show this in Appendix Table A8, where we compare the (low stakes) National Assesment of Educational Progress (NAEP) test scores in mathematics achieved by 12th-grade students in 2015 to the (high stakes) Scholastic Aptitude Test (SAT) scores in 2015. Consistent with our results, we see that test-score gaps between males and females, and whites and blacks, whites and Hispanics, and whites and Asians are larger for the SAT compared with the NAEP. While many explanations are offered for the differences in performance in SAT and NAEP exams, results from our study imply that males might just exert lower effort in NAEP exams. This is supported by Freund and Rock's (1992) finding of a higher prevalence of pattern-marking behaviour among males in the NAEP exam, which they interpret as an attempt to complete the exam as quickly and with as little effort as possible. Similar gender differences in response patterns were also found by Chiacchio *et al*. (2016), who analysed students' responses for the Italian Pisa Science exam of 2006. Our results are also consistent with the claim that standardised tests usually underpredict college and graduate school performance for women and overpredict performance for men (see, e.g., Willingham and Cole, 1997; Rothstein, 2004).[35]

4.5. *Nature or Nurture*

It is interesting to try to determine to what extent differences in performance between high and low stakes situations are socially constructed or innate.[36] 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. Women are generally perceived as more conscientious and/or altruistic than men.[37] Other possible explanations may be associated with experimenter demand effects and/or with higher salience of the reward for women and non-whites.

[35] Our findings also suggest the same pattern for whites compared with 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).

[36] See, e.g., Gneezy *et al*. (2009), Booth and Nolen (2011,2012).

[37] Interestingly, these stereotypes do not have firm empirical support. In a recent review of various meta-analyses, Hyde (2014) finds only small or no gender differences in the Big Five personality traits. The only exception is a higher score of women on 'tender-mindedness', which is part of the agreeableness factor (Cohen's d between 0.3 and 1.07 for U.S. and Japanese adults, but not for black South Africans).

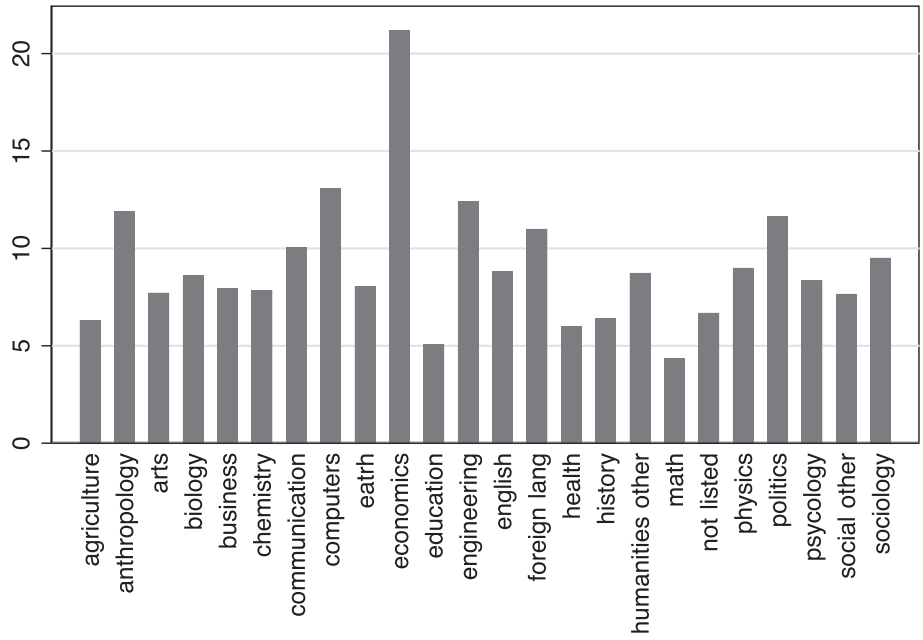


Fig. 8. *Performance Gap between High and Low Stakes Test by Undergraduate Major: Quantitative Section.*

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.

A curious finding that relates to this question is presented in Figure 8, 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 a result either of self-selection into economic majors or of 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 maximise profits relative to other undergraduate majors.

5. Conclusions

In this study, we examined the change in performance of females, males, whites and minorities between high and low stakes situations by comparing the performance of GRE examinees in the real and in an experimental section of the test. Our results show that males and whites have the highest change in performance relative to females, Asians, Hispanics and blacks. Males' drop in performance between the high and low stakes section is 0.16 SD larger than the drop of females; whites' drop in performance in the Q-section is 0.23 SD larger than the drop of blacks and the drop of Hispanics, and 0.19 SD larger than the drop of Asians. We showed that the larger change in performance observed among males and whites is partially due to the fact that these two groups invest less effort in the low stakes test. We ruled out alternative explanations for these findings, such as stereotype threat, differences in stress levels, learning or alternative cost of time.

Our findings suggest that men and whites who perform well in high stakes tests might not perform as well in ordinary assignments, and that women and minorities who do not perform so well in high stakes tests may do relatively better in low stakes tasks. Accordingly, our results may have 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 putting more weight on 'low stakes' measures of achievement and performance, such as high school performance and grades, extra-curricular activities, and recommendation letters (as opposed to focusing entirely on performance in standardised tests).

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. Oftentimes, job candidates are evaluated using high stakes tests. However, most jobs require excellence in tasks that are not directly attached to high-powered incentive schemes. This suggests that consideration of performance in high stakes tests should not come at the expense of consideration of other indicators of ability that reflect good performance in low stakes situations.

Tel Aviv University
Tel Aviv University
Educational Testing Service

Additional Supporting Information may be found in the online version of this article:

Online Appendix A.

References

Ackerman, P.L. and Knafer, R. (2009). 'Test length and cognitive fatigue: an empirical examination of effects of performance and Test-Taker reactions', *Journal of Experimental Psychology: Applied*, vol. 15(2), pp. 163–81.
Ariely, D., Gneezy, U. and Loewenstein, G. (2009). 'Large stakes and big mistakes', *Review of Economic Studies*, vol. 76(2), pp. 451–69.
Azmat, G. and Petrongolo, B. (2014). 'Gender and the labor market: What have we learned from field and lab experiments?', *Labour Economics*, vol. 30, pp. 32–40.
Azmat, G., Bagues, M., Cabrales, A. and Iriberri, N. (2019). 'What you don't know ... can't hurt you? A natural field experiment on relative performance feedback in higher education', *Management Science*, vol. 65:3449–3497.
Babcock, L., Recalde, M.P., Vesterlund, L. and Weingart, L. (2017). 'Gender differences in accepting and receiving requests for tasks with low promotability', *American Economic Review*, vol. 107(3), pp. 714–47.
Booth, A. and Nolen, P. (2011). 'Choosing to compete: How different are girls and boys?', *Journal of Economic Behaviour and Organization*, vol. 81 (2), pp. 542–555.
Booth, A. and Nolen, P. (2012). 'Gender differences in risk behaviour: Does nurture matter?', *Economic Journal*, 122 (558), pp. F56–F78.
Borghans L., , Meijers , H. and Weel, B.T. (2008). 'The role of noncognitive skills in explaining cognitive test scores', *Economic Inquiry*, vol. 46(1), pp. 2–12.
Bridgeman, B., Cline, F. and Hessinger, J. (2004). 'Effect of extra time on verbal and quantitative GRE scores', *Journal of Applied Measurement in Education*, vol. 17(1), pp. 25–37.
Cassaday, J.C. and Johnson, R.E. (2002). 'Cognitive test anxiety and academic performance', *Contemporary Educational Psychology*, vol. 27(2), pp. 270–95.
Chiacchio, C., De Stasio, S. and Fiorilli, C. (2016). 'Examining how motivation towards science contributes to omitting behaviors in the Italian PISA 2006 sample, *Learning and Individual Differences*, vol. 50, pp. 56–63.
Cole, J.S., Bergin, D.A. and Whittaker, T.A. (2008). 'Predicting student achievement for low stakes test with effort and task value', *Contemporary Educational Psychology*, vol. 33(4), pp. 609–24.
Cotton, C., McIntyre, F. and Price, J. (2013). 'Gender differences in repeated competition: evidence from school math contests', *Journal of Economic Behaviour and Organization*, vol. 86.

Cribbie, R.A. and Jamieson, J. (2000). 'Structural equation and the regression bias for measuring correlates of change', *Educational and Psychological Measurement*, vol. 60(6), pp. 893–907.
Croson, R. and Gneezy, U. (2009). 'Gender differences in preferences', *Journal of Economic Literature*, vol. 47(2), pp. 448–74.
Cunha, F. and Heckman, J.J. (2007). 'The technology of skill formation', *American Economic Review*, vol. 97(2), pp. 31–47.
Datta Gupta, N., Poulsen, A. and Villeval, M.C. (2005). 'Male and female competitive behaviour: experimental evidence', IZA Discussion Paper No. 1833, IZA institute of Labor Economics.
Dohmen, T.J. and Falk, A. (2011). 'Performance pay and multi-dimensional sorting: productivity, preferences, and gender', *American Economic Review*, vol. 101(2), pp. 493–525.
Duckworth, A.L. and Seligman, M.E.P. (2006). 'Self-discipline gives girls the edge: gender in self-discipline, grades, and achievement test scores', *Journal of Educational Psychology*, vol. 98(1), pp. 198–208.
Educational Testing Service.(2007). *Graduate Record Examinations, Guide to the Use of Scores 2007-2008*.
Flory, J.A., Leibbrandt, A. and List, J.A. (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.
Freund, D.S. and Rock, D.A. (1992). 'A preliminary investigation of pattern-making in 1990 NAEP data', paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20–24). Retrieved from ERIC Document Reproduction Service. (ED 347189)
Gneezy, U., Leonard, K.L. and List, J.A. (2009). 'Gender differences in competition: evidence from a matrilineal and a patriarchal society', *Econometrica,*vol. 77(5), pp. 1637–64.
Gneezy, U., Niederle, M. and Rustichini, A. (2003). 'Performance in competitive environments: gender differences', *Quarterly Journal of Economics*, vol. 108(3), pp. 1049–74.
Gneezy, U. and Rustichini, A. (2004). 'Gender and competition at a young age', *American Economic Review Papers and Proceedings*, vol. 94(2), pp. 377–81.
Heckman, J.J. and Rubinstein, Y. (2001). 'The importance of noncognitive skills: lessons from the GED testing program', *American Economic Review Papers and Proceedings*, vol. 91(2), pp. 145–9.
Hyde, J.S. (2014). 'Gender similarities and differences', *Annual Review of Psychology*, vol. 65, pp. 373–98.
Jurajda, Š. and Münich, D. (2011). 'Gender gap in admission performance under competitive pressure', *American Economic Review Papers and Proceedings*, 101 (3), pp. 514–518, May.
Lavy, V. (2008). 'Gender differences in competitiveness in a real workplace: evidence from performance-based pay tournaments among teachers', NBER Working Paper 14338.
Levitt, S.D., List, J. and Sadoff, S. (2016). 'The effect of performance-based incentives on educational achievement: evidence from a randomised experiment', NBER Working Paper 22107.
Liu, J., Allspach, J.R., Feigenbaum, M., Oh, H.J. and Burton, N. (2004). 'A study of fatigue effects from the new SAT', College Board Research Report No. 2004–2005, The College Board, New York.
Lord, F.M. (1967). 'A paradox in the interpretation of group comparisons', *Psychological Bulletin*, vol. 68(5), pp. 304–5.
Mattern, K.D., Patterson, B.F., Shaw, E.J., Kobrin, J.L. and Barbuti, S.M. (2008). 'Differential validity and prediction of the SAT', College Board Research Report No. 2008–4, The College Board, New York.
Niederle, M. (2016). 'Gender', in (J. Kagel and A.E. Roth, eds.), *Handbook of Experimental Economics*, 2nd edition, pp. 481–553, Princeton, NJ: Princeton University Press.
Niederle, M. and Vesterlund, L. (2007). 'Do women shy away from competition? Do men compete too much?', *Quarterly Journal of Economics*, vol. 122(3), pp. 1067–101.
O'Neil, H.F., Sugrue, B. and Baker, E.L. (1996). 'Effects of motivational interventions on the national assessment of educational progress mathematics performance', *Educational Assessment*, vol. 2(2), pp. 135–57.
Örs, E., Palomino, F. and Peyrache, E. (2008). 'Performance gender-gap: Does competition matter?', CEPR Discussion Paper No. 6891.
Paserman, D. (2010). 'Gender differences in performance in competitive environments: evidence from professional tennis players', Discussion Paper, Boston University.
Pope, D.G. and Sydnor, J.R. (2010). 'Geographic variation in the gender differences in test scores', *Journal of Economic Perspectives*, vol. 24(2), pp. 95–108.
Rothstein, J. (2004). 'College performance predictions and the SAT', *Journal of Econometrics*, vol. 121(1-2), pp. 123–44.
Rubinstein, A. (2006). 'A skeptic's comment on the study of economics', Economic Journal, vol. 116(510), pp. C1–C9.
Segal, C. (2010). 'Motivation, test scores, and economic success', Working Paper, Universitat Pompeu Fabra.
Spencer, S., Steele, C.M. and Quinn, D.M. (1999). 'Stereotype threat and women's math performance', *Journal of Experimental Social Psychology*, vol. 35(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(6), pp. 613–29.
Steele, C.M. and Aronson, J. (1995). 'Stereotype threat and the intellectual test performance of African Americans', *Journal of Personality and Social Psychology*, vol. 69(5), pp. 797–811.
Willingham, W.W. and Cole, N.S. (1997). *Gender and Fair assessment*, Mahwah, NJ: Erlbaum.
Yoon, T., Keller, M., Schinder De-Lap, B., Harkins, A., Lepers, R. and Hunter, S. (2009). 'Sex differences in response to cognitive stress during a fatiguing contraction', *Journal of Applied Physiology*, vol. 107(5), pp. 1486–96.