

Assessment of Minorities' Skills Using Low-Stakes Tests - Evidence from PISA and a Field Experiment

Yuval Ofek-Shanny*

September 15, 2021

Abstract

The minority-majority educational performance gap is typically measured using low-stakes standardized assessment tests. However, these measures will be distorted if there is a difference in intrinsic motivation between students from different groups. I conduct a field experiment and use data from PISA 2015 to evaluate the differences in minority-majority motivation and performance-gaps when using high and low-stakes tests. I find that performance gaps substantially differ depending on the stakes of the test. Using high-stakes tests reduces the gap by 0.3 standard deviations. The experimental results suggest that 60% of the Arab-Jewish math performance gap measured on 8th grade Israeli National Assessment tests, could be due to differences in motivation rather than cognitive skills.

*Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

Acknowledgments

I would like to thank Todd Kaplan, Yoni Ben Bassat, Ed Baker, Przemyslaw Biecek, Chaim Fershtman, Uri Gneezy, Ro'ee Levy, Fatena Marjie, Markus Nagler, Joel Rapp, Sally Sadoff, Analia Schlosser, Niza Sion, Avner Strulov-Shlain, Noam Zussman, Members of the Economics department in University of Haifa, participants of the 2019 Asia-Pacific ESA Meeting in Abu Dhabi, the Tenth IWAE in Catanzaro, Italy 2019, and the Israel Economic Association Conference, Jerusalem 2019, for helpful comments and suggestions; Jonathan Rosenblatt for the computational resources (obtained with ISF grant 924/16); Gali Gevim and Deema Hamada for the research assistance.

Data Availability Statement

The data used in this article can be obtained from the PISA <https://www.oecd.org/pisa/data/2015database/>. The experimental data and additional replication materials will be made available online before publication.

Disclosure Statement

Funding Organizations: None.

Support: My work was supported by the University of Haifa and the Israeli Council for Higher Education, Planning and Budgeting Committee.

Positions: None. Relatives/Partners: None. Review of the Work: None.

Institutional Review Board (IRB) Approval: This research was conducted with approval from the Chief Scientist at the Israeli Ministry of Education, 9343 and the

Institutional Review Board at the Faculty of Social Sciences in the University of Haifa, 171/17.

1 Introduction

The educational underperformance of minority groups is a long-lasting issue of debate and concern.¹ One of the main tools used to assess educational performance is standardized assessment tests.² These tests have no consequences for the examinee (low-stakes tests) and therefore, the students' performance depends not only on their cognitive skills but also, to a higher degree than usual, on their intrinsic motivation. Heterogeneity in levels of intrinsic motivation across groups of interest can make cross-group performance comparisons sensitive to the level of test stakes. If minority students are less intrinsically motivated, low-stakes assessments will underestimate their cognitive skills compared to majority groups. Ignoring this characteristic of low-stakes assessments could lead policy-makers to the wrong conclusions and even distort good policy design.

Heterogeneous motivation levels in low-stakes assessments is a well known phenomenon, but most of the literature so far has focused on cross-country comparisons and used self reports or proxies for motivation such as performance endurance or rate of skipped items in the test or following questionnaire. These methods cannot

¹Examples include Hispanic and Black compared to White students in the United States (see, for example, Brey et al. 2019), immigrants to non-immigrants in OECD countries (Schleicher 2019), and Arabs compared to Jewish students in Israel (RAMA 2017).

²Such as the international PISA and national NAEP in the US and GEMS in Israel.

reliably measure to what extent heterogeneous motivation distorts between-group performance comparisons, because they are not able to identify the impact of motivation separately from the impact of skill. Two papers addressed this challenge using a field experiment.³ They use incentives to manipulate students' motivation and find that performance differences between groups of interest are meaningfully different with and without incentives. However, to the best of my knowledge, no paper has yet experimentally compared minority and majority school-age students. Measurements of the performance gap among school-age students are especially likely to be distorted as these measurements usually rely on low-stakes assessment tests.

This paper combines engagement measures based on observational data from PISA 2015 and direct evidence from a field experiment, to test the hypothesis that minority students have lower intrinsic motivation and to quantify the effect of this lower motivation on the change in performance-gap measurements using high and low-stakes tests.

Evidence from PISA indicates that minority students are less intrinsically motivated. I use two measures of test engagement in PISA as a proxy for intrinsic motivation - endurance of performance during the test, and the probability of skipping or guessing an answer. I find that in Israel and the United States, the average minority student shows lower engagement compared to the average majority stu-

³Gneezy et al. (2019) use between-subjects design to compare performance with and without financial incentives offered right before the test. Schlosser, Neeman, and Attali (2019) use a within-subject design to compare adult examinees' performance in the GRE test to a low-stakes non-incentivized part at the end of the test.

dent, indicating lower intrinsic motivation.⁴ To examine if this lower engagement leads to underestimation of the minorities' performance, I reassess the performance in PISA by filtering out unmotivated students and alternatively, answers that are not meaningful (skip or a rapid guess). Both methods find no significant change in minority-majority performance gap measurements before and after filtering. The method of motivation filtering requires two strong assumptions: students are either fully motivated or unmotivated, and there is no correlation between cognitive skills and motivation. Under these assumptions, filtering unmotivated students allow us to estimate low-stakes test results "as if" measured using a high-stakes test. However, it is likely that these assumptions do not hold as many students exert lower effort but do not skip or guess test items. To cleanly estimate the effect of motivation without relying on these assumptions, I conduct a field experiment.

I use a field experiment to learn if this between-group heterogeneity in intrinsic motivation manifests into meaningful differences in minority-majority performance gaps when measured using low or high-stakes assessments. 599 students in the 8th grade from seven Jewish and Arab schools in the north of Israel participated in the experiment. Using a within-subject design, each student took two similar math tests - one with no feedback or consequences (low-stakes) and the other, one week later, accounting for 30% of the final year's grade (high-stakes). I assume, that when stakes

⁴The Asian minority is an exception, in line with previous evidence (see, for example, Goldammer 2012; Nguyen et al. 2019)

are high enough, students exert close to full effort.⁵ In low-stakes assessments, the more the students are intrinsically motivated, the closer their grade is, to their high-stakes grade. So students with a very high level of intrinsic motivation, will do just as well on both tests and have no grade difference between the tests.

I find that Arab-minority students in Israel are significantly less motivated when taking a low-stakes test. Their average performance, measured using a low-stakes test was 1 standard deviation lower than their performance measured using a high-stakes test. For the Jewish students, this difference was only 0.7 standard deviations. Assuming these results are representative of the Israeli population, and that the high-stakes test fully captures cognitive skills, the Jewish-Arab cognitive-skill gap in the 8th-grade is 60% lower than implied by the official assessment test (0.15 instead of 0.38 standard deviations). Further implications are discussed in Section 4.4.

This paper improves on the previous experimental literature on low-stakes assessment tests by combining a within-subject design in a natural-incentives setting on a population of interest. Gneezy et al. (2019) and Schlosser, Neeman, and Attali (2019) are the two papers closest to this one, comparing cross-group heterogeneity in response to incentives on assessment tests. Gneezy et al. (2019) compare performance with and without financial incentives offered right before the test to high-school students in the US and Shanghai. They find that US students improve

⁵Of course this is not the case for all the students. Weak students for example, which fail to achieve a minimum grade, and so fail the test, might not care what exactly their grade is. I further discuss this when interpreting the results across the performance distribution in Section 4.2.

their performance substantially with incentives while Shanghai students do not. The difference in performance gaps between US and Shanghai students decreases by 0.33 standard deviations when students are financially incentivized. They interpret this result as evidence of high intrinsic motivation of Shanghai students that leads to full effort in the test without financial incentives. To the best of my knowledge, Schlosser, Neeman, and Attali (2019) is the only experimental study that compares the performance of minority and majority groups on low and naturally high-stakes tests. In contrast to my findings, they find that the white majority examinees are the least intrinsically motivated in the low-stakes part. However, they compare adult GRE examinees, thus, their results might not generalize well to the assessment of school students. This paper extends these two papers by combining the within-subject design in a natural-incentives setting and minority-majority comparison used in Schlosser, Neeman, and Attali (2019) with the test structure and age of interest as in Gneezy et al. (2019).

This paper also contributes to a developing strand of the literature that uses engagement on assessment tests as a proxy for intrinsic motivation and tries to verify the effect of heterogeneous intrinsic motivation on assessments of groups' performance and ranking. It highlights the strengths and limitations of these methods by comparing their findings to the experimental findings. Akyol, Krishna, and Wang (2021), Borghans and Schils (2018), Borgonovi and Biecek (2016), and Zamarro, Hitt, and Mendez (2019) find a high correlation between engagement and cross-country performance variation in PISA assessments. However, after controlling for this variation, only Borghans and Schils (2018) find a meaningful change in the ranking of the

countries. Few other papers have examined within-country comparisons. Anaya and Zamarro (2020) and Balart and Oosterveen (2019) find that girls are more engaged than boys in low-stakes tests. Wise, Soland, and Bo (2020) use the probability of rapid-guessing to identify lower test engagement and correct the ranking of 84 US schools but find minor changes in ranking. Soland (2018) uses a similar methodology to identify lower test engagement of Black students, but adjusting for differences in test engagement based on rapid-guessing, only changes the Black-White performance gap by 0-0.03 standard deviations.

These attempts, to correct student rankings using engagement measures, suffer from two main limitations. First, they assume no correlation between engagement and cognitive skills. If, for example, weaker students are more prone to lower engagement, filtering out these students will inflate the average score of the group/country/test. Rios et al. (2017) show this assumption does not always hold. Second, it assumes that the engagement measures capture the differences in test-taking effort between the students. Comparing the results of motivation filtering in Section 2 to the experimental results in Section 4, I show this assumption is too strong.

I conclude that methods based on engagement measures are good for indicating a possible heterogeneity in intrinsic motivation but less reliable in evaluating the magnitude of effect on performance measurements.⁶

⁶In line with these conclusions, Gneezy et al. (2019) experimentally find that accounting for low motivation of US students will increase their performance by 22-24 points on the PISA 2015 math test while Akyol, Krishna, and Wang (2021)

The paper proceeds as follows. Section 2 provides suggestive evidence that minority groups are less motivated in low-stakes PISA assessments. Section 3 presents the experimental design and empirical framework I use to compare performance in high and low-stakes tests. Section 4 presents and discusses the experimental results and their implications. Section 5 discusses possible mechanisms that might cause these differences between minority and majority groups and Section 6 concludes.

2 Heterogeneous engagement of minority and majority groups in PISA 2015

Studies using data from the different PISA waves, find that test engagement differs significantly across countries. Recent studies show this is also the case for gender groups within countries. Less attention has been given to engagement comparisons within countries across ethnicity groups. In this section I show that minority groups in the United States and Israel are less engaged with PISA 2015 tests, suggesting lower intrinsic motivation. Using motivation filtering methods does not change the performance gaps between minority and majority groups. This could suggest heterogeneous motivation is not a concern for the validity of minority-majority performance gaps measurements but stands in contrast to the experimental results presented in the following sections.

use engagement measures to filter unmotivated answers and find only a 3.5 points increase for the US students.

2.1 Background

The Programme for International Students Assessment (PISA) assesses the proficiency of 15-year-old students in the core school subjects of science, reading, and mathematics. PISA surveys take place every three years. I chose to study the 2015 results as they are the first to be almost fully computerized and were the newest available when I started the research. In this round, science was the major domain, so I focus my analysis on the science items as will be described below. Approximately 540,000 students participated in the 2015 round representing a sample from 29 million students in 72 countries and economies (OECD 2016).

In this analysis, I focus on Israel and the United States, which have significant minority populations who are identified using additional data supplied by the countries' research centers.⁷ The analysis of US data also allows comparisons to previous results in the literature.

⁷In Israel, schools are segregated by ethnicity. Data on the ethnicity of the schools was obtained from the Israeli Ministry of Education. I excluded from the analysis Jewish Ultra-Orthodox students because the institutions that participated are not representative of the population. In the United States, students are asked regarding their ethnicity in the PISA student's questionnaire.

The data is available on the website of the National Center for Education Statistics (NCES) - <https://nces.ed.gov/pubsearch/getpubcats.asp?sid=098>.

2.2 Endurance

Students' performance normally declines during a test (see, for example, Balart and Oosterveen 2019; Borghans and Schils 2018). The phenomenon is even more pronounced in low-stakes tests (Finn 2015; Gneezy et al. 2019). The magnitude of the decline in performance can suggest how motivated the students are.⁸ Gneezy et al. (2019) find that while unincentivized students in the United States declined from 47.5% correct answers in the first half of a math test to 35.1% correct in the second half (12.4 percentage points decrease), financially incentivized students decline from 50.5% to 43% (7.5 pp), a 4.9 percentage points smaller decrease. So, if two students start at the same level of performance and the performance of one of them declines more than that of the other, we interpret it as an indication of a lower level of motivation.⁹

I use the random assignment of test parts in PISA cognitive test to compare this relative performance decline across groups. The test is composed of four consecutive

⁸The more motivated the students are, the more they endure fatigue and maintain their performance during the whole test (Borgonovi and Biecek 2016).

⁹While Gneezy et al. (2019) show that more motivated students endure their performance better, endurance is not only a proxy for motivation, it is also, to some extent, a component of ability (see, Brown et al. 2019). In this section, I follow the literature in treating endurance as a motivational measure (Borgonovi and Biecek 2016; Gneezy et al. 2019). In the end of this section I expand on the limitations of this method.

30 minutes parts. On the 2015 wave, each part is either science, math, reading, or collaborative-problem solving. Each 30 minutes part is randomly assigned from a set of clusters (12 for science, 6 for math and reading, 3 for CPS).¹⁰

To compare the endurance of students across ethnic groups, I compare how their relative performance changes from one test cluster to the other.¹¹ I construct relative performance as their performance percentile compared to all other PISA examinees that took the same part under the same conditions. Since the students are randomly assigned to different test clusters, I calculate the performance percentile of each student relative to all other PISA participants that were assigned to the same cluster in the same position (first/second/third/fourth). If a student is ranked in the 60th percentile in the first test cluster and then in the 55th percentile in the second test cluster, this means her relative performance declined and is an indication of a lower level of motivation.¹²

Using the endurance in PISA as a proxy for intrinsic motivation, I find minority students are less intrinsically motivated. Figure 1 depicts the mean change in the relative position of different groups during the test. The percentiles of the different

¹⁰For example, a student can start a test with cluster S01 (science), then S07, R02 (reading), M04 (math).

¹¹Relative performance can also be termed - rank-order performance.

¹²Each cluster is calculated separately, so an observation is a student in a cluster. Since each student took two science clusters, we have two observations for each student. See Section B.1 in the appendix for a thorough description of the methodology.

groups vary in a meaningful sense during the test. In Israel, the percentile of all groups declines throughout the test, suggesting low endurance at the country level. Between groups, we see a stronger decline of the Arab minority, suggesting lower performance endurance during the whole test. In the United States, the White and Asian students improve their percentile during the test, and the mean percentile of the other minority groups declines.

This evidence suggests that different ethnicities maintain their performance differently during the test, with most minority groups declining in their percentile relative to the majority group in the country. The Asian students are an exception, in a manner that conforms with previous evidence on high intrinsic motivation levels of the Asian origin students (Goldammer 2012; Nguyen et al. 2019).

One limitation of the endurance method is that we can think of very low motivation students that will maintain their performance during the test perfectly - this is the case of students exerting very low levels of effort right from the first question. In this case, we will not see any performance decline. To partially account for this limitation, I add another measure of test engagement using item response times.

2.3 Guessing behavior

Starting 2015, PISA data include measurements of response time per item. That is, the time that took the examinee to answer each question. This allows us to identify items the student chose not to meaningfully engage with by - 1. skipping with no response, and 2. answering very quickly without devoting enough time to the solution of the problem - “*guessing-behavior*”. Based on a methodology developed

by Wise and Kong (2005) (See also a recent review in Wise (2017)), I identify the guessing-behavior of students if they answer a question in a very short time. See section B.2 in the Appendix for a detailed description of the method. Then, I use the level of skipped and guessed items as a proxy for the examinee’s level of intrinsic motivation. The more items the student chooses to skip or guess, the less motivated I conclude she is.

The results indicate that as in the endurance measure, minority groups in Israel and the United States are less intrinsically motivated. Figure 2 depicts the mean proportion of the items skipped or guessed in each group. We can see that in Israel, the Jewish students choose not to interact with a question less than the Arab students. In the United States, the most engaged groups are White and Asian students, and the least engaged are Black students.

2.4 Discussion

Both performance endurance and the proportion of items the student chose not to engage with indicate heterogeneous intrinsic motivation across ethnicity groups, with most minorities showing lower levels than the majority groups. Using two methods of motivation filtering I find that minority-majority performance gaps remain roughly the same after excluding unmotivated students or items that were not meaningfully answered. Lower motivation, leading to lower effort in the test, can cause a bias in the measurements or ranking of the different groups. If group A has weaker cognitive skills and is also less engaged in the test relative to group B, using the test results

would lead to overestimation of the performance gap.¹³

Biased results and comparisons could be misleading and lead to wrong conclusions and even distort good policy design. This is a major concern as many minority groups are the target of different educational policies that are later evaluated using this kind of low-stakes assessment tests. Wise and DeMars (2005) suggest identifying the students that do not make an effort and exclude them from the analysis. By comparing only engaged students, they argue that they improve the validity of their measurements. This method was used for example in Akyol, Krishna, and Wang (2021) and Anaya and Zamarro (2020). If engagement is not correlated with cognitive skills, and if we can identify students that are not engaged precisely, then we have a solution to the problem.

However, there are several reasons to think this is not the case. First, while Segal (2012) and Wise and DeMars (2005) do not find a correlation between engagement and performance as measured in another high-stakes incentivized test, Rios et al. (2017) do find that it is correlated, with weaker students also being less engaged with solving a low-stakes assessment test. They find that filtering the low-engagement students inflates the average score by 0.42 standard deviations.

Second, there are good reasons to believe we do not fully identify all the students with low levels of engagement using response time and endurance measurement. Students could be guessing the answers to the items, but doing it slowly, so their

¹³On the other hand, if group A is stronger but less engaged, this would lead to underestimation of the performance gap or even to a change in the ranking of the groups.

response will not be identified as a rapid guess. Also, the level of effort is not binary, while not guessing, partly motivated students tend to exert less effort in a low-stakes test (Gneezy et al. 2019). This adds a cause to measurement error that cannot be fixed when filtering guessing behavior.

Examining the effect of motivation filtering on the score and relative performance, I find that the rankings and performance differences between majority and minority groups are unchanged. This is similar to most papers that use these methods. Table 1 shows the effect of filtering unmotivated students or unengaged items on the score and performance of each group. *Percentile* represents the average relative percentile of the group members compared to all other PISA participants. *Score* represents the group's average rate of correct answers. In the filtering items method, the calculations are made after filtering items that were not answered or identified as a rapid guess. In the filtering students method, the calculations are made after filtering students identified as unmotivated. *N* in the last column represents the number of students removed in each group. Following Wise and DeMars (2010) I choose to exclude students that guessed or skipped more than 10% of the items.¹⁴ As expected, almost all groups improve their rate of correct answers using both methods. However, their percentile ranks do not meaningfully change as well as the percentile differences between the minority and majority groups. Based on the two filtering methods, we could have concluded that the low-stakes of PISA are not a major concern when comparing the performance differences between minority and majority groups. This

¹⁴The results do not meaningfully change when using thresholds of 20% or 30% of the items.

conclusion is similar to the findings of Soland (2018). However, as I will show in Section 4.1 this would be a mistaken conclusion.

The use of engagement measures can therefore provide some suggestive evidence regarding the effect of low-stakes testing on the validity of comparisons across groups, but it will fail to provide valid measurements of the magnitude and meaningfulness of the problem. To elucidate the answers to these questions, I conducted a field experiment in Israel.

3 Design

3.1 Experimental design

I use a field experiment to examine to what extent minority-majority performance gaps measurements change with the level of stakes. I use a within-subject design to compare the performance of 8th-grade Jewish majority and Arab minority students on high and low-stakes mathematics tests.¹⁵ The experiment was conducted in the spring of 2017 in the northern district of Israel.

The tests are written in the format of the National Israeli Standardized Assessment Test - GEMS.¹⁶ Each test is composed of 21-23 multiple-choice questions,

¹⁵See Figure A.1 in the appendix, for a graphical representation of the design.

¹⁶The GEMS known as “MEITZAV” (Hebrew acronym for “School Growth and Efficiency Measures”) is a national assessment system for 2nd, 5th and 8th-grade students. GEMS include student achievement exams as well as questionnaires designed

constructed response questions, and a combination of the two.¹⁷ An explanation was requested in about one-fourth of the questions. The questions in the tests are mixed in math subjects (e.g., algebra, geometry, etc.) and increase in difficulty level throughout the test. Hebrew speaking students took the tests in Hebrew and Arabic speaking students took the tests in Arabic.¹⁸

As the high-stakes test, I used the test written by RAMA for the GEMS of 2017.¹⁹ Every year, only one-third of the schools participate in the national assessment tests. The other two-thirds are offered to take the test as an internal test and many of them use it as a final year's test determining about 30% of the student's final year's grade. I chose schools that took this offer and thus had the same high-stakes test across all participating schools.

As a low-stakes test, I used a math test written with the help of the Mathematics Supervision Department in a way that maintained the structure and difficulty level

to gather information about the school climate and pedagogical environment.

¹⁷When writing the low-stakes test, I did not know how many questions the high-stakes test will have, so I wrote 23 questions based on the previous year's structure. Eventually, the high-stakes test was written by the Israeli National Authority for Measurement and Evaluation in Education (RAMA) with 21 questions.

¹⁸The translation from Hebrew to Arabic was done under the guidance of Arabic Math Supervision department in the Israeli Ministry of Education.

¹⁹RAMA - Israeli National Authority for Measurement and Evaluation in Education.

of the GEMS math test. Similar to the GEMS' grades, both tests' scores range on a 0-100 scale. The scoring for the high-stakes test was done by schools' teachers according to a very detailed scoring guide from RAMA. The scoring of the low-stakes test was done by the experiment team using a similar detailed scoring guide.

Due to IRB restrictions, I could not randomize the tests to high and low-stakes. Also, the motivation of the schools to participate was to allow their students a preparation test before the final high-stakes test. So the low-stakes test had to come before the high-stakes test. This makes the design less clean because it means the difference between the tests captures not only the difference in motivation but also a week's preparation. However, I do not consider this a big problem. First, because the students did not get any grade or feedback on the low-stakes exam. Second, because also in real-life, students usually prepare for a high-stakes test and do not prepare for a low-stakes test.²⁰

Schools in Israel are usually segregated by religion or a branch of a religion.²¹ In the analysis I use the results of four Jewish schools - two secular and two religious,

²⁰Even if there is some effect of the first low-stakes test on the performance in the second high-stakes test, since we are interested in the minority-majority comparison, it is only a concern if this effect is heterogeneous across the ethnicity groups and large enough to distort the results. I do not think this is the case.

²¹The segregation in the Israeli education system is enacted to allow teaching in Hebrew for the Jews and Arabic for Arabs as well as enabling every group to maintain its values and traditions.

and three Arab schools - one Muslim, two Christian.²²

While the sample is not statistically representative, I selected the schools with the advice of the Mathematics Supervision Department in the Israeli Ministry of Education so that they represent common ability schools for their ethnicity.²³

All 8th-grade students in the participating schools were assigned to take both tests. The first test was the low-stakes test. The students were notified a few days before the test that it would take place. They were informed several times that they will not receive a grade for the test and that the test results will have no implications for them. Still, they were requested to do their best, both for contribution to science and as preparation for the high-stakes test. In a questionnaire handed after the test, I verified that the students understood that the test was a low-stakes one.²⁴ The

²²Two additional Arab schools participated in the experiment but were excluded from the analysis. One due to misinterpretation of the instructions by the math coordinator, and the second, because it took the high-stakes test during the Ramadan fest. During Ramadan many students fest and sleep very little, so it makes their performance incomparable. See Section C of the Appendix for more details.

²³Based on the national assessment tests conducted in the years before the experiment - the Jewish schools are in deciles 8, 5, 4, 3 and the Arab schools are in deciles 9, 8, 4. Unfortunately, two Arab schools, ranked in deciles 3 and 4, dropped from the experiment, so the Arab schools are a bit above the average of the Israeli Arab schools.

²⁴Apart from one school, 95% answered correctly to the verification questions. See

second test was a high-stakes test. It was conducted a week later and all the students were aware of it a few weeks in advance. The students were well notified that this was a high-stakes test that will determine about 30% of their final year's grade.

The first test was not mandatory due to IRB requirements, but thanks to the commitment of the schools' staff, most of the students participated in the experiment. The second test was part of the schools' syllabus and thus mandatory. Of the seven schools that successfully participated, a total of 734 students took the high-stakes test and 599 took both tests. The only data I have on students that did not take the low-stakes test, is their high-stakes grade. As shown in Table 2, the students that took both tests had a slightly higher high-stakes grade. Table ?? presents the population included in the experiment analysis by gender and ethnicity after excluding outliers.²⁵

3.2 Empirical Framework

It is well known, that performance on high-stakes tests is significantly better than low-stakes tests. If this difference in performance is the same for all students, then it is not a big concern, because the ranking and standardized score differences will

appendix C.

²⁵I excluded student with a low-stakes grade of more than 50 points higher than their high-stakes grade. One student was excluded based on this criterion. 21 students did not report their gender and were excluded when gender was used as a control. This does not qualitatively change the results.

remain the same using low and high-stakes tests for comparing students and groups. If the performance difference between high and low-stakes tests is heterogeneous across the population groups, then stakes matter, and the ranking as well as grade differences might vary depending on the stakes of the test.

Heterogeneous performance difference comparing high and low-stakes tests across population groups questions the validity of performance measurement, and therefore, our dependent variable will be a student's grade difference - $Grade_{HS} - Grade_{LS}$. We are interested to examine, whether this difference is homogeneous across our population groups.

To examine the change in students' performance between the high and low-stakes test, I follow Schlosser, Neeman, and Attali (2019) and estimate a first difference equation:

$$Y_i^{HS} - Y_i^{LS} = \alpha + \beta Arab_i + \gamma Arab_i * Male_i + \delta x_i + u_i$$

where Y_i^{LS} denotes grade of student i in the LS test, Y_i^{HS} denotes grade of student i in the HS test, $Arab_i$ denotes a dummy variable set to 1 for Arab students (0 for Jewish), $Male_i$ denotes a dummy variable set to 1 for male students, Female and Jew are omitted variables. Vector x_i are student characteristics that includes dummy variables for - mother's and father's education, and for misunderstanding the low-stakes test to be a high-stakes one.²⁶ The coefficients of interest are β that

²⁶After the low-stakes test, we asked the students if they think that they will receive a grade for the test and whether the grade will affects their final year's grade. Answering yes to both questions means that the student misunderstood the

denote the difference between Jewish and Arab female in the difference between high-stakes grade and low-stakes grade and γ 's that denote the difference between male's and female's grade difference across ethnicities. For robustness, we use and present also performance measures in standard deviations and percentiles. Using the grades difference controls for an individual's fixed effect taking into consideration all factors that affect students' grades on both tests.

4 Results

Almost all the students performed better on the high-stakes test. The performance differences were largest for students above the median high-stakes performance and Arab boys specifically. The external validity of these findings is supported by the observational evidence from the PISA test and findings from previous experimental papers. A major implication of these results is that if assessment tests were high-stakes, performance differences between the Arab minority and Jewish majority would be 60% smaller.

4.1 Ethnicity

Keeping a student's high-stakes grade fixed, the Jewish students performed significantly better than the Arab students in the low-stakes test. Similar to most studies comparing high and low-stakes testing, the average grade in the high-stakes test is significantly higher than the low-stakes, 0.8 standard deviations (19.2 points) differ-

low-stakes test for a high-stakes test.

ence on average, in line with previous findings.²⁷ The average difference for the Arab students was 1 standard deviation (23.2 points) and the average difference for the Jewish students was only 0.7 standard deviations (16.8 points). While both groups are different in many observed and unobserved characteristics, the within-subject design allows us to control for all student characteristics that affect performance on both tests.

Table 4 presents the results of the OLS regression with the difference between a student's high and low-stakes grades as the dependent variable. Wild bootstrap standard errors clustered by school are in parenthesis under the estimated coefficient. Columns 1-2 report the regression results with $HS_{grade} - LS_{grade}$ as the dependent variable. Difference in percentiles is reported in column 3 and difference in Z-Score in column 4. The first row reports the difference from Jewish girls (all Jewish students in column 1).²⁸ The difference between the high and low-stakes grades for the Arab students is 6.4 points larger than the difference for the Jewish students. Adding student characteristics and controlling for gender does not significantly change the

²⁷Wise and DeMars (2005) review 12 empirical studies comparing high and low-stakes testing and find that performance differences vary in the range of 0-1.49 standard deviations with an average of 0.59.

²⁸The first row of column 2 shows that Arab girls have a 4.6 points bigger $HS_{grade} - LS_{grade}$ difference than Jewish girls, etc. Rows 2-3 report the boy to girl difference in $HS_{grade} - LS_{grade}$ finding the difference for an Arab boy is 2.8 points bigger than for an Arab girl.

results. The gender differences will be discussed in the next section.

This difference in engagement leads to significant differences in measurements and ranking, as shown in Table 5. If we use the low-stakes standardized assessment test, we find that the average grade for Jewish students is 45.5, slightly above the 44.4 of the Arab students. However, if we use the results from the high-stakes test, where all students were more engaged, we find that the Arab students significantly outperformed the Jewish students with average grades of 67.7 and 62.2 respectively (0.23 SD difference).

The level of heterogeneity in performance on low-stakes compared to high-stakes assessments across ethnicity groups is similar to other experimental studies and different from studies that used engagement measures to adjust low-stakes results. On average, the Arab students performed 1 standard deviation better in the high-stakes test while the Jewish students performed 0.7 standard deviations better. Gneezy et al. (2019) find that US students' performance change when moving from low to high-stakes test is 0.33 standard deviations larger than Chinese students, Schlosser, Neeman, and Attali (2019) find that the difference for White students is 0.24 standard deviations larger than Black students.²⁹ This is quite different from the results of papers that used methods based on identifying engagement on assessment tests. Soland (2018) finds significant Black-White differences in the probability of rapid-guessing

²⁹In Gneezy et al. (2019) I calculate using the results from Table 2 and an average standard deviation of the US and Chinese students. In Schlosser, Neeman, and Attali (2019) I calculate using the quantitative section results (Table 3) and the full sample standard deviation (Table 1).

behavior but using several methods to account for the differences only changes the performance gaps by 0-0.03 standard deviations. Most other studies that tried to assess cross-countries differences in performance referred to engagement in the test also found small effects (Akyol, Krishna, and Wang 2021; Borgonovi and Biecek 2016; Zamarro, Hitt, and Mendez 2019). These differences between experimental results and results based on engagement analysis suggest engagement could be useful in identifying the existence of differences in intrinsic motivation, but less useful in quantifying the effect of these differences.

4.2 Differences across the performance distribution

The within-subject design also allows us to see that the differences on $HS_{grade} - LS_{grade}$ between the Arab and Jewish students are meaningful and significant only for the above-median students. Each dot in Figure 3 represents the combination of one student's high-stakes (x-axis) and low-stakes (y-axis) grades. Almost all students perform better on the high-stakes test. The difference between the high-stakes and low-stakes grades is represented by the distance from the 45 degrees line. The trend lines indicate that for better-performing students, the ethnicity difference is more pronounced. Figure 4 statistically compares the differences in $HS_{grade} - LS_{grade}$ across ethnicity groups for each high-stakes performance decile.

For the worst 10% (according to their high-stakes grade) of the students in both populations, there is no performance difference on both tests. Similarly, Gneezy et al. (2019) and Schlosser, Neeman, and Attali (2019) do not find differences for the worst-performing students. A reasonable explanation for this is that they do not

know the answer to almost all questions, so they gain some points by guessing the multiple-choice questions and maybe some of the easiest questions. In this case, by chance, they can obtain a higher grade on the low-stakes test. Also, in our natural incentive design, it does not matter for the students if they get 20 or 30 in the high-stakes test because both mean that they fail.

For the students in the lowest 30%, both ethnicity groups perform worse on the low-stakes test, but the difference in performance between the two tests is similar for both populations. The difference between the groups is most pronounced above the median. The Arab students have an average performance difference $HS_{grade} - LS_{grade}$ of 29 points (1.3 SD) and the Jewish students, of 21 points (0.9 SD). Similar between-group heterogeneity in performance differences between high and low-stakes tests across the high-stakes performance distribution was also found in Gneezy et al. (2019) and Schlosser, Neeman, and Attali (2019).

4.3 Gender and Ethnicity

The difference in performance between the high and low-stakes tests on the experiment is not significantly different between boys and girls across the pooled experiment population but is significant within the ethnicity groups.

Comparing the overall average performance of boys and girls on both tests, Table 6 shows no significant difference. Boys slightly outperform girls on both tests with a 2.1 points difference (0.09 SD) in the high-stakes test and a 2.6 difference (0.11 SD) in the low-stakes test. So overall, the low-stakes tests give similar average results.

When we compare both gender and ethnicity, we find that Arab boys and girls

have almost the same average grade in the low-stakes tests but boys significantly outperform girls in the high-stakes test. Table 7, and the *Arab Boy* coefficient in Table 4 compare the performance differences between Arab boys and girls. They obtain similar low-stakes grades of 44.6 and 44.7 respectively, but in the high-stakes test, the Arab boys outperform the girls with 69.4 and 66.1 respectively (0.12 standard deviations diff). The difference in $HS_{grade} - LS_{grade}$ is 3.1 points (0.121 SD) and significant at the 1% level. I interpret this result as evidence of the Arab girls being more intrinsically motivated in the low-stakes test.

For the Jewish students, the boys outperform the girls on both the high and low-stakes tests. The gender performance difference is larger in the low-stakes tests but not significantly different from the high-stakes test. The average grades for boys and girls on the low-stakes test are 48.1 and 43.5 respectively (0.18 SD diff), and on the high-stakes test 63.6 and 60.9 respectively (0.09 SD diff). The coefficient *Jewish Boy* in Table 4 indicates that the difference in $HS_{grade} - LS_{grade}$ is smaller for boys but not statistically significant. This suggests Jewish boys might be slightly more intrinsically motivated in the low-stakes test.

These results are supported by the PISA data analysis. Comparing the results of the experiment to the PISA 2015 data, shown in Figure A.3 in the appendix, we see that using both engagement measures, we get a similar pattern, the Arab girls tend to skip fewer items and endure performance during the test better than the Arab boys, but for the Jewish students, the gender differences are negligible and point to the opposite direction if any, similar to the experimental results.

These experimental findings suggest gender differences in intrinsic motivation on

low-stakes are heterogeneous across population groups. Gender differences of the Arab students are similar to the results from PISA for the US students and to most evidence in the literature that find girls to be more motivated in low-stakes tests (see, for example, Azmat, Calsamiglia, and Iriberry 2016; Balart and Oosterveen 2019; Finn 2015).

4.4 Implications

The large and significant changes in grades and performance gaps between high and low-stakes tests across ethnicity groups contribute to an increasing body of evidence suggesting caution when using the results of low-stakes standardized assessment tests to assess cognitive skills.

To give a sense of the scale of changes in measurement, I calculated how low-stakes assessment results change if we “correct” them using the experiment’s results - “as if it was a high-stakes test”. This is possible under the assumption that the performance difference between the experimental low and high-stakes results is similar to the difference between the low-stakes national assessment tests results and hypothetical high-stakes national assessment test results. The Jewish-Arab math performance gap on the low-stakes national assessment tests is 10.89 on a 0-100 scale. $Grade_{HS} - Grade_{LS}$ in the experiment is 6.4 points larger for Arab students (Table 4). Assuming this is also the difference in the population, if the national assessment tests were high-stakes, the grades for both groups would increase, but the average grade of the Arab students would increase by 6.4 points more, and so, the performance gap would decrease from 10.89 (0.38 standard deviations) to 4.39 (0.15) points, a 60% decrease.

In the form of an equation -

$$New_Gap = Old_Gap - Diff_in_Gap_Experiment$$

Where *Old_Gap* is the average grade difference between the populations in the national assessment tests, $Diff_in_Gap_Experiment = (Grade_{HS} - Grade_{LS})_{Arab} - (Grade_{HS} - Grade_{LS})_{Jew}$, and *New_Gap* is the gap as if the assessment test was a high-stakes test. There are several other ways to do this transformation, but they provide similar results.³⁰ The same equation is used for the reassessment of the gender gaps within ethnicities.

The results, in Table 8, show that using this method, the performance gaps change significantly. The first column shows the gaps on a 0-100 scale. The second column shows what were the performance gaps if the assessment tests were high-stakes. The third and fourth columns show the gaps in standard deviations. The Jewish-Arab performance gap decreases from 10.9 to 4.4 points (0.38 to 0.15 standard deviations), the Arab gender gap in favor of girls almost disappears from 4.1 to 1.3 points (0.14 to 0.05 standard deviations), and the Jewish gender gap reverses from a small advantage to boys (0.59 points) to small advantage for girls (1.4 points. from 0.02 to 0.05 standard deviations). Figure 5 presents the implications of the same

³⁰For example, one can calculate the differences in standard deviations instead of grade units. Since both high and low-stakes experimental tests were in the GEM format and grading, I think both methods are valid and use the one that provides a more modest result.

calculations used on the performance gap during the years 2008-2017.³¹

An additional implication of this paper's results concerns attempts to adjust low-stakes results using motivation filtering or regression based on test engagement. Our results show that all students perform significantly better on the high-stakes test. Therefore, filtering very low motivation students cannot account for the lower motivation of all examinees. Also, heterogeneity in motivation across the performance distribution, suggests these measures could create an additional bias.

It is important to note that our population is not a representative sample of the Israeli students' population, so one should take these specific numbers with a grain of salt. However, the evidence from PISA supports the direction of the results and I believe they provide a scale and direction of the heterogeneous engagement effect on performance measurement.

5 Possible Mechanisms

The results from the experiment and the evidence from the PISA data suggest that most minority groups are less intrinsically motivated when taking low-stakes tests. Using a post-test questionnaire used in the field experiment, I find that student's performance difference between the high and low-stakes tests is correlated with reported

³¹The data for this analysis were retrieved from RAMA publication - Meitzav Hesegim Tasha"z and the conversion from 500 to 100 scale, is according to the 2017 conversion formula - $Y = 3.3970 * X + 335.3792$, see - https://cms.education.gov.il/EducationCMS/Units/Rama/AarachaBeitSifrit/Hamara_Sulam_Meitzav.htm

effort. However, the hypothesis that performance difference is negatively correlated with sense of belonging is not supported by the questionnaire results.

5.1 Effort

Using a post-test questionnaire, I find that on average, students that said they would make more effort on the high-stakes test, indeed made more effort. After the low-stakes test, the students were asked to rank their level of effort in the test and in a hypothetical high-stakes test on a 1 to 10 scale.³² To partly control for the subjectivity of the answer, I follow Butler and Adams (2007) and compare the difference in reported actual effort from hypothetical effort on a high-stakes test.³³ Figure 6 shows that according to their responses, the performance difference between the high and low-stakes tests is correlated with the difference in reported effort. The lower the average difference in reported effort between the tests, the lower the average performance difference between the tests. For the Jewish students, students reporting a very small or no difference in effort (zero or one) had on average a 0.58 standard deviations difference between performance in the high-stakes test and the low-stakes

³²Similar to the effort thermometer used in some of the PISA waves (see, Butler and Adams 2007)

³³I also tried to ask the students about their actual effort after the high-stakes test, but unfortunately, could not obtain the responses from all the participating schools. For the schools I did manage to get the responses, I find a very similar pattern to the hypothetical responses.

test compared to a 0.77 standard deviations difference for students reporting a large difference in effort. The Arab students show a similar decline change - 0.92 standard deviations for small effort difference and 1.14 for large effort difference. Students that did not answer the questionnaire at all had the largest performance difference.³⁴

5.2 Sense of belonging

Sense of belonging, as measured in the experiment's post-test questionnaire, is not correlated with low-stakes test engagement. Performance in high-stakes tests is considered less sensitive to a lack of intrinsic motivation than performance in low-stakes tests. So, if students from one group have higher intrinsic motivation, we expect to see them exerting higher levels of effort on the low-stakes tests and have a smaller performance difference between the high and low-stakes tests.

According to Self Determination Theory (SDT, Ryan and Deci 2000), relatedness (or belongingness, sense of belonging to a group, and interacting with its members) is one of the three basic psychological needs that foster intrinsic motivation. Previous

³⁴About 5% of the students did not report their level of effort at all. These students had the largest average performance difference (1.01 standard deviations for Jewish and 1.23 SD for Arab students). This evidence is in line with researches that use non-response to additional questionnaires as evidence of low motivation in a test (see, for example, Boe, May, and Boruch 2002; Butler and Adams 2007; Zamarro, Hitt, and Mendez 2019). Additional 5% of the students reported they will exert less effort on a hypothetical high-stakes test. No clear pattern is evident for these students.

studies find lower sense of belonging for minority students, suggesting a possible mechanism explaining lower intrinsic motivation of minority students.

In the post-test questionnaire we distributed after the low-stakes exam, we asked the students how much they care about their school and hometown.³⁵ Since a large majority answered either “agree” or “strongly agree”, I analyse the results comparing “strongly agree”, “agree”, and neutral or less. Figure A.4 in the appendix presents the average performance difference $HS_{grade} - LS_{grade}$ for each answer. I did not find meaningful differences in the performance difference across the students’ responses to the questions. Using this question as a proxy for “sense of belonging”, I do not find correlation between the performance difference and sense of belonging.

6 Conclusions

In this paper, I experimentally find that the lower intrinsic motivation of minority groups in low-stakes assessments can cause measures of performance gaps to be twice as high as when measured using high-stakes assessments that are less sensitive to this lack of intrinsic motivation. This difference in results of high and low-stakes tests can lead to biased or wrong conclusions when evaluating the skills of minority groups and effect of policy measures targeted at these groups. Researchers and policy makers must be aware of these differences and limitations and make sure they choose the

³⁵The students were asked to what extent they agree with the two statements - “I care about my school” and “I care about my town / village / community” on a five-level Likert scale from “Strongly Disagree” to “Strongly agree”.

proper method to evaluate students' abilities. If the main goal is to address cognitive skills, high-stakes testing, which are less sensitive to the heterogeneous effort, is a better choice. If the main goal is to predict future labor market outcomes, low-stakes assessments could be better, as they capture a combination of cognitive and non-cognitive skills, that are both of interest.

This paper has a few limitations. The participants of the experiment come from a small number of schools. Specifically, while in the general population the Jewish students outperform the Arab students on both high and low-stakes tests, in these experimental results, the Arab students slightly outperform the Jewish students in the high-stakes test. The results of the schools in the national assessment tests from the years before the experiment suggest this is because the Arab schools in the experiment are ranked higher than the average Arab schools.³⁶ The within-subject design partly addresses this limitation. We see that the differences between the groups are visible and significant across a large part of the high-stakes grade distribution, so while the average magnitude could differ due to the participants pool, the direction and scale likely reflect the general population. Since the experimental design is simple and relatively natural to the educational environment, a possible future avenue is to use a similar design on a larger scale and representative sample.

The work in this paper can be expanded in a few directions. Many schools and education systems (the Israeli Ministry of Education is one example) prohibit the use of monetary incentives when conducting experiments within the education

³⁶Based on information collected from the individual school reports on the Israeli Ministry of Education website. Unfortunately, the data is no longer available online.

system. The methodology developed in this paper uses end-of-the-year grade as a natural incentive for the high-stakes test. Implementing this methodology on a large scale and/or a representative sample of the students' population could allow for an exact measurement of the performance difference between high and low-stakes tests across different groups. A second development from this paper would be to combine the engagement measures used in the PISA data using the position of the question in the test and response time. This synthesis will allow additional useful insights on the PISA results and their use. An additional avenue for future research is the use of personality traits to predict and understand the personal performance difference between high and low-stakes tests³⁷ Combining engagement measures from the PISA test with personality traits and high and low-stakes results would make a significant contribution to the development of a general theory of high and low-stakes assessments.

To conclude, this paper shows that current estimates of educational performance gaps across groups are often significantly distorted since they rely on low-stakes assessment tests. Policymakers should be aware of these limitations and make sure they choose the proper evaluation method, in line with their policy goals.

³⁷Building on previous work in this direction as modeled by Borghans, Duckworth, et al. (2008) and implemented in Borghans, Meijers, and Ter Weel (2008), Borghans and Schils (2018), and Segal (2012).

References

- Akyol, Pelin, Kala Krishna, and Jinwen Wang (2021). “Taking PISA Seriously: How Accurate are Low-Stakes Exams?” In: *Journal of Labor Research*, pp. 1–60.
- Anaya, Lina and Gema Zamarro (2020). “The role of student effort on performance in PISA: Revisiting the gender gap in achievement”.
- Attali, Yigal and Maya Bar-Hillel (2003). “Guess where: The position of correct answers in multiple-choice test items as a psychometric variable”. In: *Journal of Educational Measurement* 40.2, pp. 109–128.
- Azmat, Ghazala, Caterina Calsamiglia, and Nagore Iriberry (2016). “Gender differences in response to big stakes”. In: *Journal of the European Economic Association* 14.6, pp. 1372–1400.
- Balart, Pau and Matthijs Oosterveen (2019). “Females show more sustained performance during test-taking than males”. In: *Nature communications* 10.1, pp. 1–11.
- Boe, Erling E, Henry May, and Robert F Boruch (2002). “Student Task Persistence in the Third International Mathematics and Science Study: A Major Source of Achievement Differences at the National, Classroom, and Student Levels.”
- Borghans, Lex, Angela Lee Duckworth, James J Heckman, and Bas Ter Weel (2008). “The economics and psychology of personality traits”. In: *Journal of Human Resources* 43.4, pp. 972–1059.
- Borghans, Lex, Huub Meijers, and Bas Ter Weel (2008). “The role of noncognitive skills in explaining cognitive test scores”. In: *Economic inquiry* 46.1, pp. 2–12.

- Borghans, Lex and Trudie Schils (2018). “Decomposing achievement test scores into measures of cognitive and noncognitive skills”. In: *Available at SSRN 3414156*.
- Borgonovi, Francesca and Przemyslaw Biecek (2016). “An international comparison of students’ ability to endure fatigue and maintain motivation during a low-stakes test”. In: *Learning and Individual Differences* 49, pp. 128–137.
- Brey, Cristobal de, Lauren Musu, Joel McFarland, Sidney Wilkinson-Flicker, Melissa Diliberti, Anlan Zhang, Claire Branstetter, and Xiaolei Wang (2019). *Status and Trends in the Education of Racial and Ethnic Groups 2018. NCES 2019-038*. Working Paper. National Center for Education Statistics.
- Brown, Christina, Supreet Kaur, Geeta Kingdon, and Heather Schofield (2019). “Cognitive Endurance as Human Capital”.
- Butler, Jayne and Raymond J Adams (2007). “The impact of differential investment of student effort on the outcomes of international studies”. In: *Journal of applied measurement* 8.3, p. 279.
- Finn, Bridgid (2015). “Measuring motivation in low-stakes assessments”. In: *ETS Research Report Series* 2015.2, pp. 1–17.
- Gneezy, Uri, John A List, Jeffrey A Livingston, Xiangdong Qin, Sally Sadoff, and Yang Xu (2019). “Measuring success in education: the role of effort on the test itself”. In: *American Economic Review: Insights* 1.3, pp. 291–308.
- Goldammer, Christian (2012). “Racial Gaps in Cognitive and Noncognitive Skills: The Asian Exception”.
- Kahneman, Daniel (2011). *Thinking, Fast and Slow*. Macmillan.

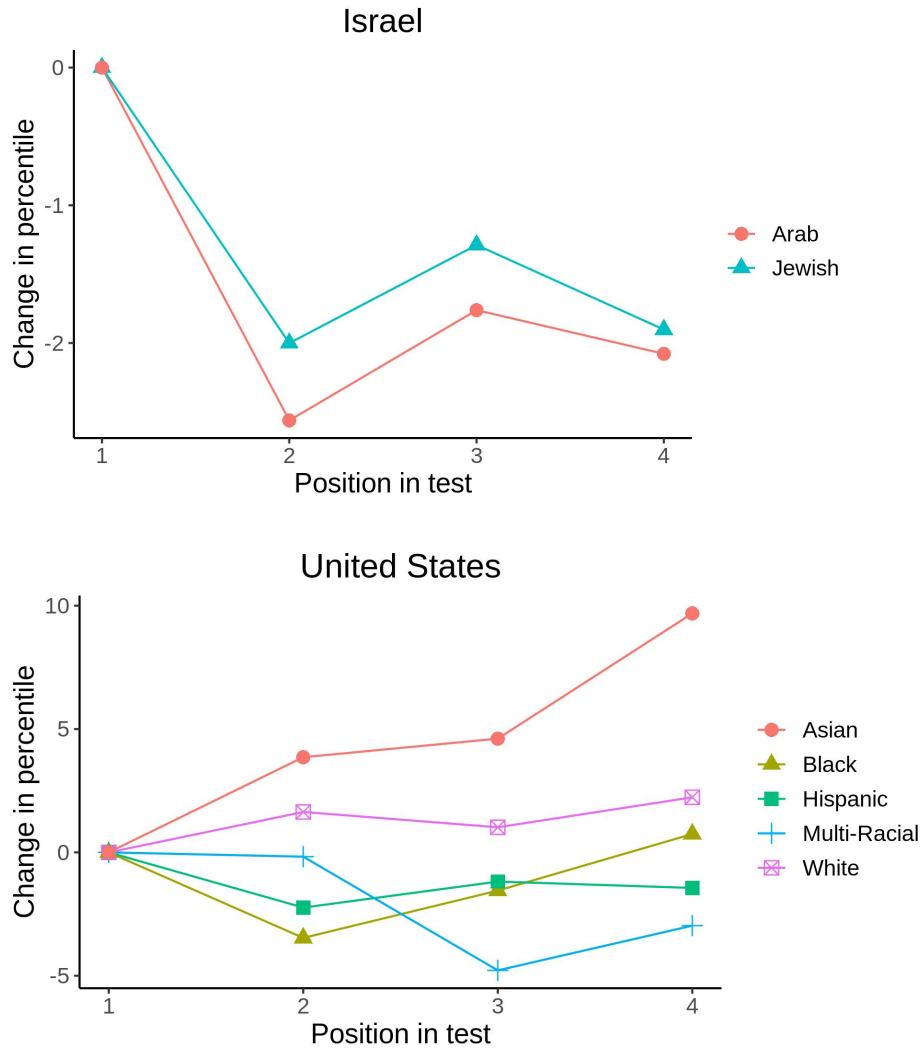
- Nguyen, Ha Trong, Luke B Connelly, Huong Thu Le, Francis Mitrou, Catherine Taylor, and Stephen R Zubrick (2019). “Sources of ethnicity differences in non-cognitive development in children and adolescents”.
- OECD (2016). *PISA 2015 technical report*. Working Paper. Paris: OECD.
- RAMA (2017). *Meitzav 2017, Growth and Effectiveness School Measures*. Working Paper. Ministry of Education, Culture, Sport, National Authority for Measurement, and Evaluation in Education, Jerusalem, Israel.
- Rios, Joseph A, Hongwen Guo, Liyang Mao, and Ou Lydia Liu (2017). “Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not?” In: *International Journal of Testing* 17.1, pp. 74–104.
- Ryan, Richard M. and Edward L. Deci (2000). “Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions”. In: *Contemporary Educational Psychology* 25.1, pp. 54–67. ISSN: 0361476X. DOI: 10.1006/ceps.1999.1020.
- Schleicher, Andreas (2019). *PISA 2018: Insights and Interpretations*. Tech. rep. OECD.
- Schlosser, Analia, Zvika Neeman, and Yigal Attali (2019). “Differential Performance in High vs. Low Stakes Tests: Evidence from the GRE Test”. In: *The Economic Journal*.
- Schnipke, Deborah L (1995). “Assessing Speededness in Computer-Based Tests Using Item Response Times”. In: *Annual Meeting of the National Council on Measurement in Education (NCME)*.
- Segal, Carmit (2012). “Working when no one is watching: Motivation, test scores, and economic success”. In: *Management Science* 58.8, pp. 1438–1457.

- Soland, James (2018). “The achievement gap or the engagement gap? Investigating the sensitivity of gaps estimates to test motivation”. In: *Applied Measurement in Education* 31.4, pp. 312–323.
- Wise, Steven L (2017). “Rapid-Guessing Behavior: Its Identification, Interpretation, and Implications”. In: *Educational Measurement: Issues and Practice* 36.4, pp. 52–61.
- Wise, Steven L and Christine E DeMars (2005). “Low examinee effort in low-stakes assessment: Problems and potential solutions”. In: *Educational assessment* 10.1, pp. 1–17.
- (2010). “Examinee noneffort and the validity of program assessment results”. In: *Educational Assessment* 15.1, pp. 27–41.
- Wise, Steven L and Xiaojing Kong (2005). “Response time effort: A new measure of examinee motivation in computer-based tests”. In: *Applied Measurement in Education* 18.2, pp. 163–183.
- Wise, Steven L and Megan R Kuhfeld (2020). “A cessation of measurement: Identifying test taker disengagement using response time”. In: *Integrating timing considerations to improve testing practices*. Routledge, pp. 150–164.
- Wise, Steven L and Lingling Ma (2012). “Setting response time thresholds for a CAT item pool: The normative threshold method”. In: *Annual meeting of the National Council on Measurement in Education, Vancouver, Canada*.
- Wise, Steven L, James Soland, and Yuanchao Bo (2020). “The (non) impact of differential test taker engagement on aggregated scores”. In: *International Journal of Testing* 20.1, pp. 57–77.

Zamarro, Gema, Collin Hitt, and Ildefonso Mendez (2019). “When students don’t care: Reexamining international differences in achievement and student effort”. In: *Journal of Human Capital* 13.4, pp. 519–552.

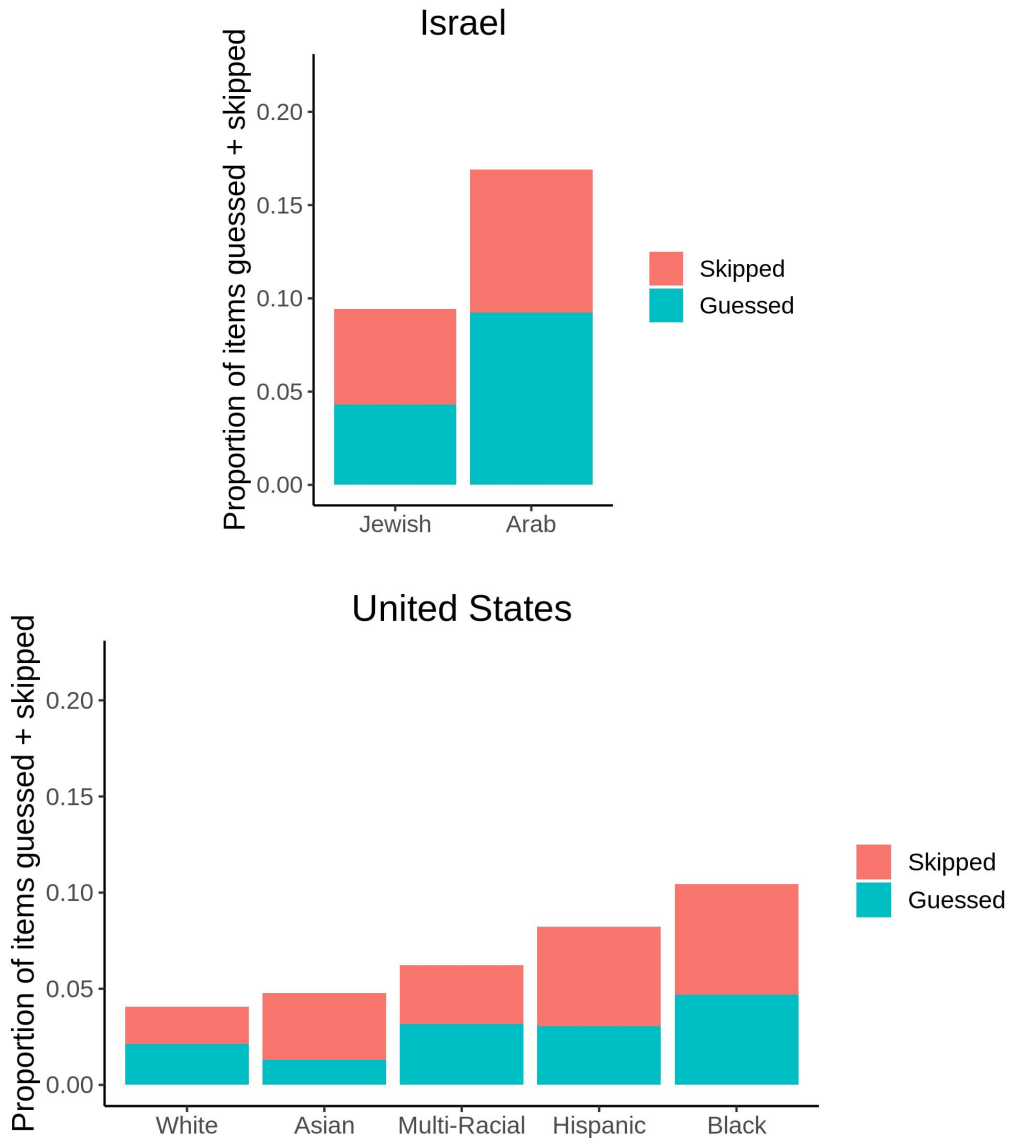
Figures

Figure 1: Performance percentile change of ethnicity groups during the test - science clusters



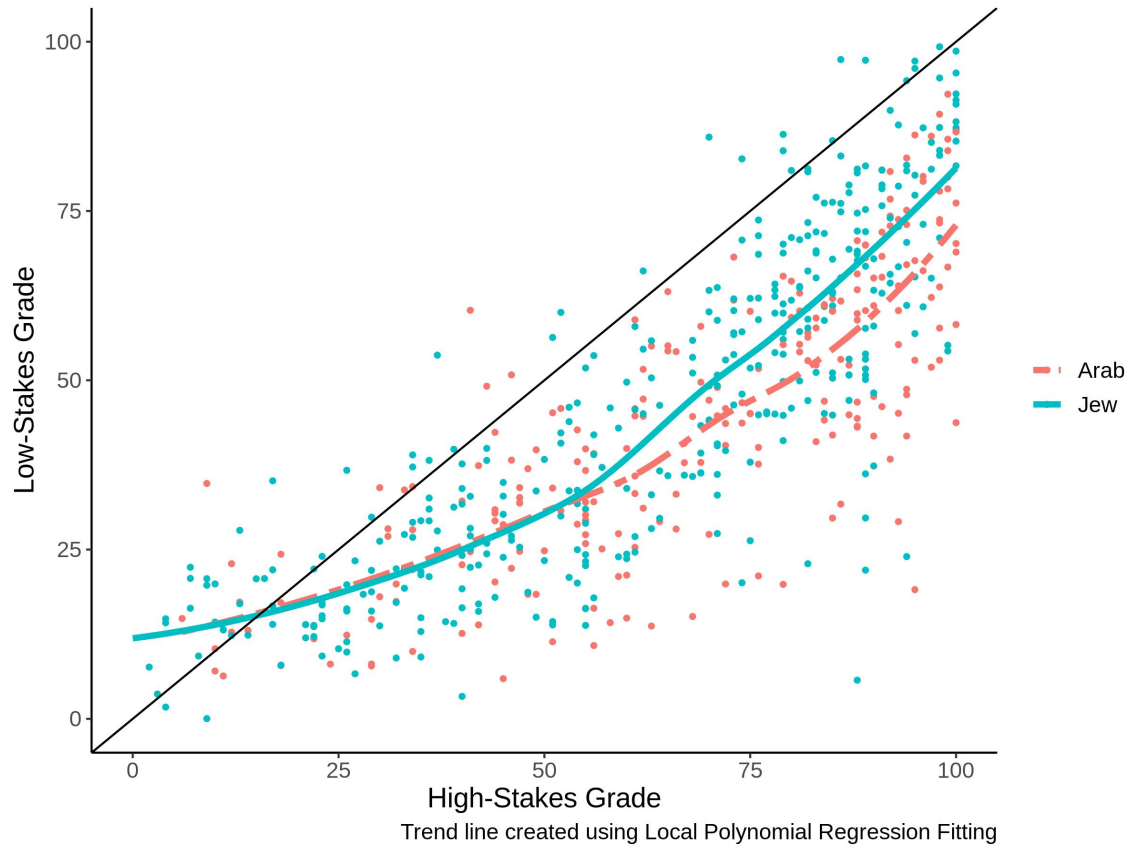
PISA test is composed of four 30-minutes parts. The graph depicts the change in performance percentile rank from the first part. The performance percentile of each student is calculated in comparison to all other PISA participants that took the same test cluster in the same position in the test (first, second, third or fourth).

Figure 2: likelihood of Guessing and skipping across ethnicities in Israel and the United States



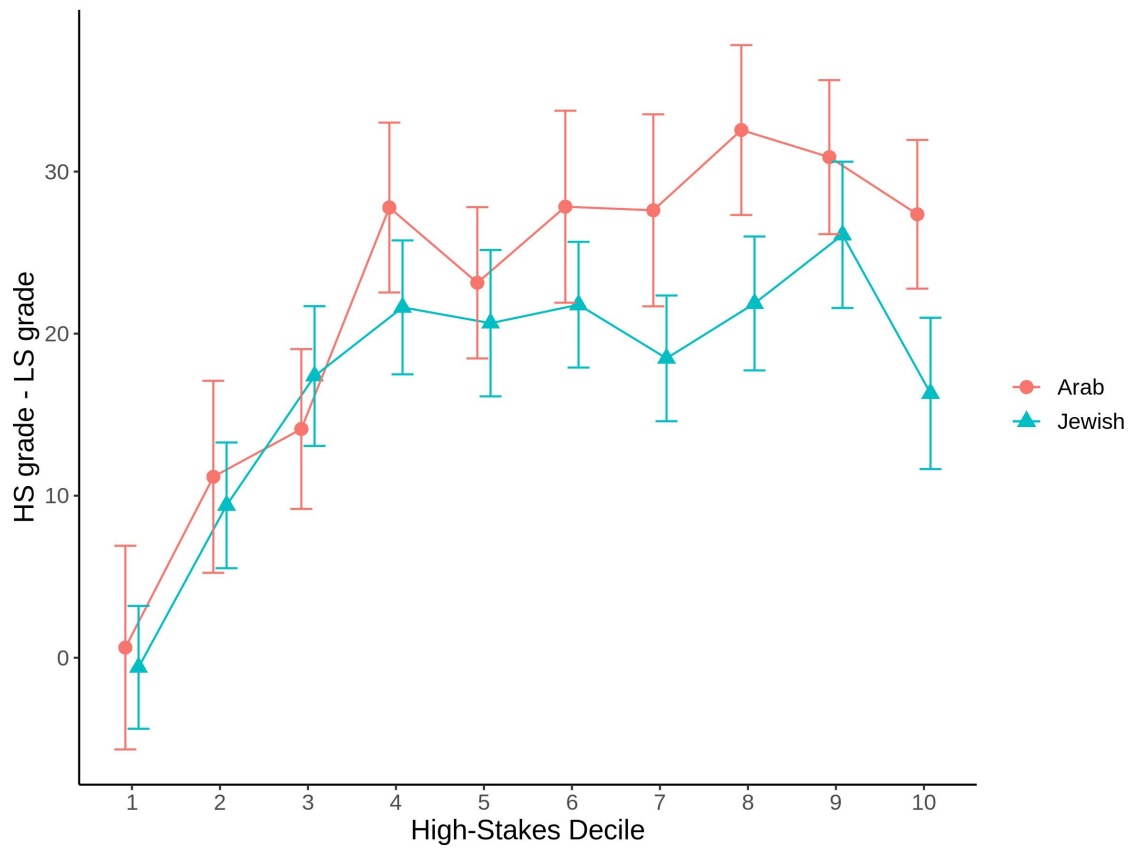
Proportion of items the students chose to skip or rapidly answer of all the PISA 2015 science items. Israel - excluding Ultra Orthodox.

Figure 3: High-Stakes and Low-Stakes Grades by Ethnicity



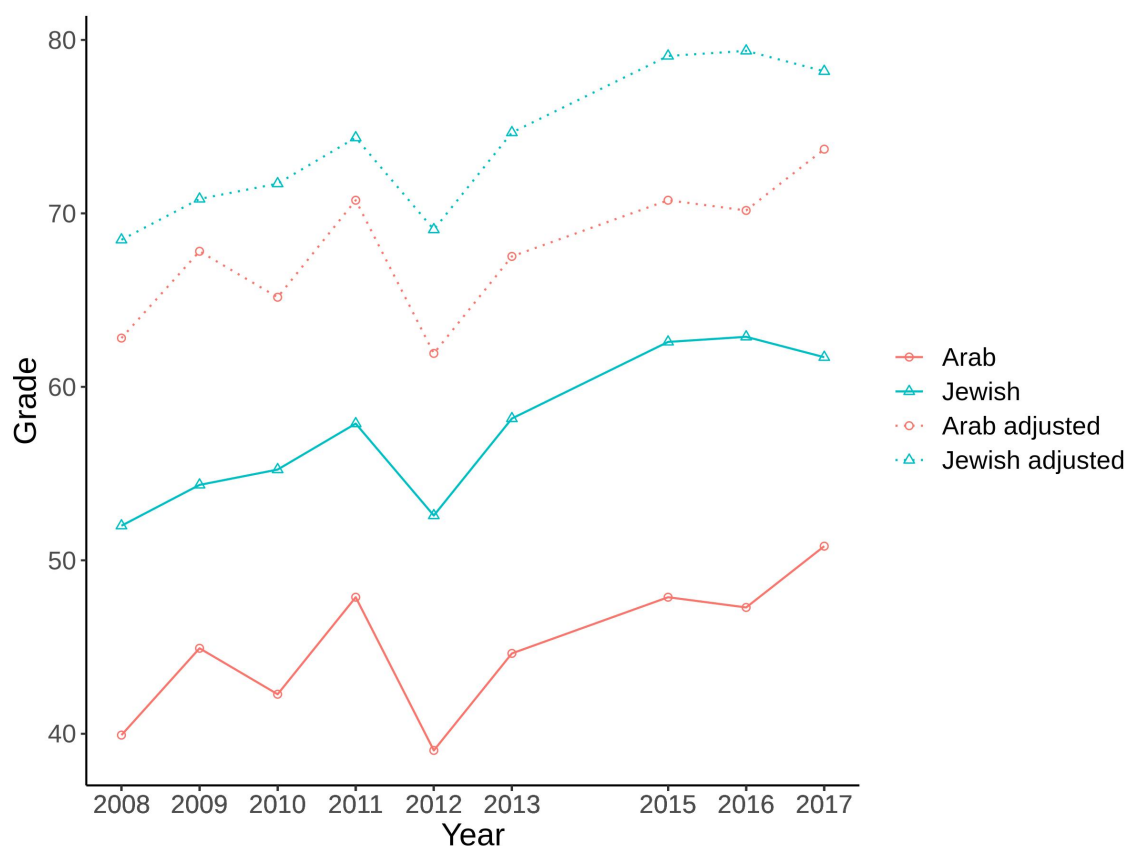
Each dot represents the combination of one student's grade in the high-stakes (x-axis) and low-stakes (y-axis) tests on the experiment. The difference between the high-stakes and low-stakes grades is represented by the vertical distance from the 45 degrees line.

Figure 4: High-Stakes and Low-Stakes Grades Difference by Ethnicity and High-Stakes Decile



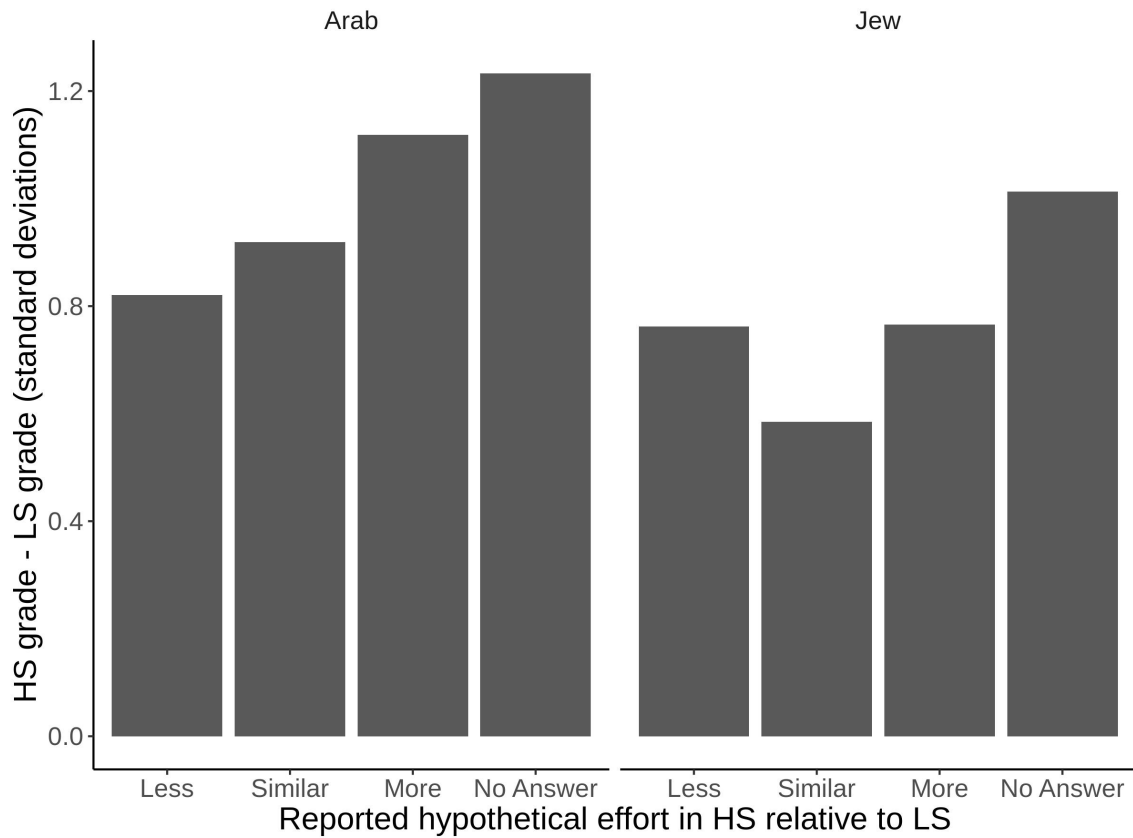
Comparison of performance difference between the high and low-stakes tests, across ethnicities and high-stakes performance deciles. Each dot represent the average performance difference for the decile's Arab or Jewish students. Error bars represent 95% confidence intervals.

Figure 5: Performance of Arab and Jewish Students on Israeli GEMS and Adjusted for Motivation



The experimental results suggest students perform better on high-stakes tests with Arabs improving significantly more. The adjusted results take the original results and adjust it “as if it was a high-stakes test” based on the performance differences between high and low-stakes test in the experiment. Section 4.4 further describes the procedure.

Figure 6: Grade Difference, High Stakes - Low Stakes vs. Reported effort



After the low-stakes test, the students were asked to rank their level of effort on the test and their level of effort had their grade accounted for the final year's grade, on a scale of 1-10. The X-axis categories describe the difference between the two answers. "More" - students that ranked their effort in the hypothetical high-stakes test 2 or more points above their reported effort in the low-stakes test, "Similar" - same or 1 point more than reported effort, "Less" - students that reported they will exert less effort in a high-stakes tests. "No Answer" are students that did not answer the effort questionnaire.

Tables

Table 1: Baseline performance compared to performance after filtering items and students

Country	Ethnicity	N	Baseline		Filtering Items		Filtering Students		
			Percentile	Score	Percentile	Score	Percentile	Score	N
Israel	Jewish	8,019	55.35	0.49	55.97	0.51	57.15	0.56	1,857
	Arab	3,075	31.17	0.30	31.84	0.32	32.27	0.37	1,391
United States	White, not Hispanic	4,810	65.58	0.55	63.55	0.56	62.95	0.59	554
	Hispanic/Latino	3,163	49.89	0.42	47.29	0.44	47.05	0.47	595
	Black/African American	1,405	40.19	0.35	37.58	0.36	36.44	0.38	352
	Multi-Racial	632	58.43	0.49	56.50	0.51	57.00	0.54	108
	Asian	383	65.21	0.56	63.17	0.56	62.42	0.58	35

This table shows the effect of filtering unmotivated students or unengaged items on the score and performance of each group. *Percentile* represents the average relative percentile of the group members compared to all other PISA participants. *Score* represents the group's average rate of correct answers. In the filtering items method, the calculations are made after filtering items that were not answered or identified as a rapid guess. In the filtering students method, the calculations are made after filtering students identified as unmotivated. *N* in the last column represents the number of students removed in each group.

Table 2: Selection to Experiment

	Arab		Jewish	
	All	Took both	All	Took both
N	270.00	230.00	464.00	369.00
High Stakes mean	65.37	67.43	60.32	62.23

All - students that took the obligatory high-stakes test.
 Took Both - students that took high and low-stakes tests and so participated in the experiment.

Table 3: Experiment Population

	Full	Arabic	Jewish
N	598	229	369
Girls	310	124	186
Boys	267	96	171
NA	21	9	12
Schools	7	3	4

Table 4: Grade difference: High - Low-stakes tests

	Grade		Percentile	ZScore
	(1)	(2)	(3)	(4)
Arab	6.402*** (1.500)	4.568** (1.797)	7.030*** (2.332)	0.175** (0.073)
Jewish Boy		-1.756 (1.379)	-1.244 (1.512)	-0.067 (0.053)
Arab Boy		3.158*** (1.196)	1.917* (1.131)	0.121*** (0.047)
Constant	16.485*** (1.278)	14.727*** (1.658)	14.260*** (1.670)	0.565*** (0.067)
Controls	<i>NO</i>	<i>Full</i>	<i>Full</i>	<i>Full</i>
Observations	577	577	577	577
R ²	0.042	0.062	0.068	0.062
Adjusted R ²	0.040	0.049	0.054	0.049

OLS regression with the difference between a student's high and low-stakes grades as the dependent variable. Wild bootstrap standard errors clustered by school are in parenthesis under the estimated coefficient. Columns 1-2 report the regression results with $HS_{grade} - LS_{grade}$ as the dependent variable. Difference in percentiles is reported in column 3 and difference in Z-Score in column 4.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Omitted Group - Jews (column 1), Jewish Girls (columns 2-4)

Table 5: Performance by Ethnicity

	Total	Arab	Jew	Diff
N	598	229	369	
HS.mean	64.3	67.7	62.2	5.4
HS.median	70	72	70	2
LS.mean	45.1	44.4	45.5	-1.0
LS.median	45	45	45	0
Diff.mean	19.2	23.2	16.8	6.5
Diff.in.std	0.8	1.0	0.7	0.3

HS - High Stakes, LS - Low Stakes
 Diff - difference between high and low-stakes grades.
 Diff.in.std - Difference in standard deviations units.

Table 6: Performance by Gender

	Boy	Girl	Diff
N	267	307	
HS.mean	65.7	63.6	2.1
LS.mean	46.9	44.3	2.6
Diff.mean	18.8	19.3	-0.5

Table 7: Performance by Gender and Ethnicity

	Arab	Jew
Girls.N	124	186
Girls.HS.mean	66.1	60.9
Girls.LS.mean	44.7	43.5
Girls.Diff.mean	21.4	17.4
Boys.N	96	171
Boys.HS.mean	69.4	63.6
Boys.LS.mean	44.6	48.1
Boys.Diff.mean	24.8	15.5
Gen.HS.diff	3.4	2.7
Gen.LS.diff	-0.1	4.6
Gen.diff.of.diff	3.5	-1.9

Table 8: Reassessing Gaps on Israeli National Assessment Tests Results

	Original	Corrected	Orig. (STD)	Corr. (STD)
Jewish - Arab	10.89	4.39	0.38	0.15
Arab Girls - Boys	4.12	1.30	0.14	0.05
Jewish Girls - Boys	-0.59	1.41	-0.02	0.05

Grades are on a 0-100 scale

To give a sense of the scale of changes in measurement, I calculated how low-stakes assessment results change if we “correct” them using the experiment’s results - “as if it was a high-stakes test”.

Appendices

A Data - PISA Analysis

The Programme for International Students Assessment (PISA) assesses the proficiency of 15-year-old students in the core school subjects of science, reading, and mathematics. PISA surveys take place every three years. I chose to study the 2015 results as they are the first to be almost fully computerized and were the newest available when I started the research. On this round, science was the major domain, so I focus the analysis on the science items as will be described below. Approximately 540,000 students participated in the 2015 wave representing a sample from 29 million students in 72 countries and economies (OECD 2016).

The main part of the PISA survey is a two-hour exam composed of four 30 minutes' parts (clusters). Students have 60 minutes for the first two clusters, followed by a few minutes break in some of the countries (depending on the local choice), and then 60 minutes for clusters 3 and 4. Since science was the major domain in 2015, each student was randomly assigned to 2 (out of different 12) science clusters and 2 other clusters from reading (out of 6), mathematics (out of 6), or Collaborative problem solving (out of 4) which was a special domain added to PISA 2015.

Of the 12 science clusters, 6 are composed of trend items, already used in previous PISA rounds and 6, of new items. Trend-items clusters accounted for one-third of the test clusters assigned and new-items clusters accounted for two-thirds of the clusters.

Publicly available data includes student-level data - students' characteristics and final scores calculated by the OECD and raw data for each response (answer to an

item) - raw answer, response time, number of actions - and score.³⁸ On this part I focus on the differences between ethnicities within Israel and the United States. Ethnicity is not part of the PISA student questionnaire, so I had to collect this information.³⁹

Table 9 presents the number of students in each country by ethnicity or religion group and gender. Of the 6598 Israeli students, 115 students under the educational trend “Jewish Other” and 610 Jewish Ultra-Orthodox were excluded from the analysis because they are not a representative sample.⁴⁰ Of the 5712 US students, 120 students did not mark their ethnicity or marked “Other”, so they were also excluded from the analysis.

³⁸Actions - the number of actions taken by the student while interacting with the item. Actions counted were clicks, double-clicks, key presses, and drag/drop events. Response time - the amount of time required to complete the item. Score - 0 for wrong, 1 for correct, and 0.5 for a partially correct answer on complex-multiple choice questions composed of several yes or no questions.

³⁹In the United States, the students were asked to state their ethnicity in a specific question added to the student questionnaire by the NCES. The US data is available at <https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2017120>. In Israel, I received data regarding the educational trend of each school. Since schools in Israel are segregated between Jews - Secular, Orthodox, and Ultra-Orthodox and Arabs, I could determine each student’s group according to the school.

⁴⁰See the ratio of female examinees as an evidence.

B Methodology - PISA Analysis

B.1 Endurance

Endurance, or the ability to maintain performance during the test, is a significant determinant of overall performance. Students with high cognitive skills, but low motivation, might be less willing to stay concentrated during the whole test and end up with low results interpreted as low level of cognitive skills. The decline in performance during the test is a well-known phenomenon in general and on the PISA test. Gneezy et al. (2019) and Akyol, Krishna, and Wang (2021) find that on PISA tests, the decline in performance occurs within and between clusters. Figure A.5 plots the probability of a correct science item across the 4 parts of the test. Between the first and second parts, there is a decline in average performance. However, after a break and a change in the test subject, there is a rise in performance in cluster 3 relative to cluster 2, followed by a further decline in cluster 4.⁴¹

To calculate individual endurance, I do the following: First, calculate each student's relative performance percentile for each cluster in each position. For example, I compare the score of student A who took science cluster number 10 in the first position (out of 4) to all other students who took science cluster number 10 in the first position. This controls for the possibility that some clusters are harder than

⁴¹All students take science either on the first and second parts of the test or on the third and fourth, so the third part is always a new test subject for the student.

others.⁴² Second, I compare the relative performance of the students across the survey clusters. If a student is ranked on the 75th percentile on both the first and the second cluster, this means she maintained her relative performance. If she declined from the 75th percentile on the first cluster to the 70th percentile on the second cluster, that means her performance decline from the first to the second cluster was greater than other students. She had a lower endurance.

If the student's first cluster was reading and the second was mathematics, relative performance will be affected by the students' endurance as well as the differences in her proficiency on the two domains. Therefore, we chose to focus on relative performance on science clusters only.

B.2 Effort according to response time

For an item to be answered in an effortful manner, the student must first read the question and then choose (in a multiple-choice item) or write (in an open response item) the answer. The amount of time this process requires could vary across students. Figure A.7 plots the response time (RT) distribution for item CS656Q01S (Birds Migration 1, see Figure A.8). The distribution has a spike at around 5 seconds followed by a decline and then the maximum at 70 seconds. This spike after a

⁴²Since students are randomly assigned to test clusters, the student's relative percentile in each group is similar. Figure A.6 shows that assignment to the different test clusters is balanced such that 1/3 of every group is assigned to clusters 1-6 and 2/3 to clusters 7-12.

few seconds is common to most PISA items and many other computer-based tests. It was identified by Schnipke (1995) and Wise and Kong (2005) as evidence of students guessing the answer without making a real effort to solve the question. They termed it rapid-guessing behavior. Wise (2017) uses the Dual Processing Theory to explain the spike phenomena. The students answering very fast are using Type 1 cognitive process - fast, non-effortful and autonomous thinking. The other students chose Type 2 thinking, which is characterized as slow, effortful, controlled, demanding of cognitive capacity, and involves analytical reasoning (Dual Processing Theory is thoroughly explained in Kahneman (2011)).

An answer to an item will be identified as a guess if response time is lower than some threshold. Wise (2017) surveys the commonly used methods for determining the threshold. The first and most common is the visual inspection of response time distributions. In this method, the researcher tries to identify the local minimum located to the right of the first spike, as it presumably indicates the transition from the majority of guessing students to the majority of not guessing. In Figure A.7 the minimum is at about 20 seconds. This method cannot be applied to all items because the two spikes distribution displayed in Figure A.7 is not always apparent. See Figure A.9 for example.

A second method uses the item characteristics such as the number of characters; the existence of a figure or a table, etc. The longer the item seems, a longer threshold is determined. This method cannot be applied to PISA items because their content is not publicly available. A third method uses a fixed threshold, usually of 3-5 seconds across all items. This method is straightforward to implement, but since it ignores

the features of the question, it might be too high for some questions identifying many solution answers as a guess, and vice versa. In a fourth method, the mean or the median RT of all responses to an item is calculated. The threshold is set as a proportion of this value. Wise and Ma (2012) compare values of 0.1, 0.15, and 0.2 and recommend setting an item threshold at 0.1 of the mean response time. The last method makes use of the item mean score as a function of response time. The threshold is determined when the mean score crosses a value that reflects the expected mean score in a random guess (for example, 0.25 for a four possible answers multiple-choice item). This method is sensitive in two aspects. First, under guessing behavior, some answers (B and C) are more frequently selected than others, so questions with B as the correct answer will have a higher than 25% correct under guessing behavior. Furthermore, tough questions, with a very low rate of correct answers, might even have a higher score under guessing behavior (Attali and Bar-Hillel 2003; Wise and Kuhfeld 2020).

In this paper, I computerize the visual inspection method to analyze PISA 2015 science items. I use R software to find the two maximum points of the RT distribution and then set the threshold at the minimum point between them. The vertical lines in Figure A.7 illustrate the three points, automatically identified. The thresholds are then manually verified. On 76 of the 184 items, the computerized visual inspection could not identify the threshold, usually because no first spike was apparent, as in Figure A.9. For these items, 0.1 of the mean was used as the threshold. The response to the item was classified as guessing behavior if it was faster than the determined threshold. The values of the thresholds varied between 2.7 seconds and 36.9 seconds

with a mean of 13.4 seconds (see Figure A.10 for the distribution of the thresholds). After identifying the guessing behavior thresholds, each response can be classified as a solution or a guess.

C Participating Schools - Field Experiment

Nine schools agreed to participate in the experiment. Due to technical problems, I could not use the results of two of the participating schools. This makes the population a bit less representative, but I believe it does not change the direction or scale of the results.

In the Arab-Druze school, the math coordinator that was responsible for conducting the tests, decided to improve the performance of the students in the low-stakes test by telling them their grades will count for the final year's grade. This made the low-stakes test a high-stakes for these students. We identified the problem according to their answers in the post-test questionnaire. Indeed their low-stakes grades were almost as high as their high-stakes grades.

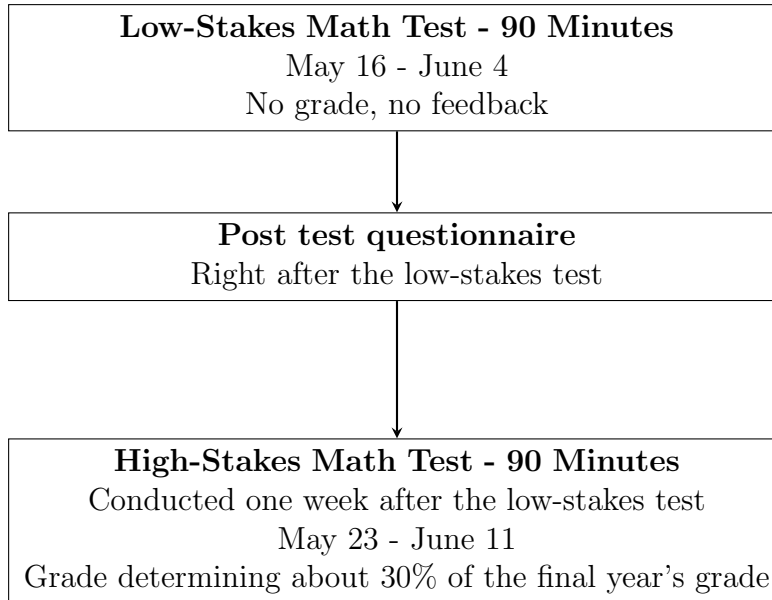
One of the Arab-Muslim schools conducted the high-stakes test during the Ramadan month. Some of the 8th fast during this month and all of them sleep very little because their families eat and party during the night. While this school's results in the low-stakes test were similar to the results of other Arab schools, the grades of the high-stakes test conducted during Ramadan were extremely low.

I believe the main findings are not affected by the exclusion of the two schools. The average grade of the Arab students might have changed to some extent, but

due to the within-subject design, the differences from the Jewish population are apparent across a large part of the performance distribution and I believe the results are therefore representative in their direction and scale.

Figures

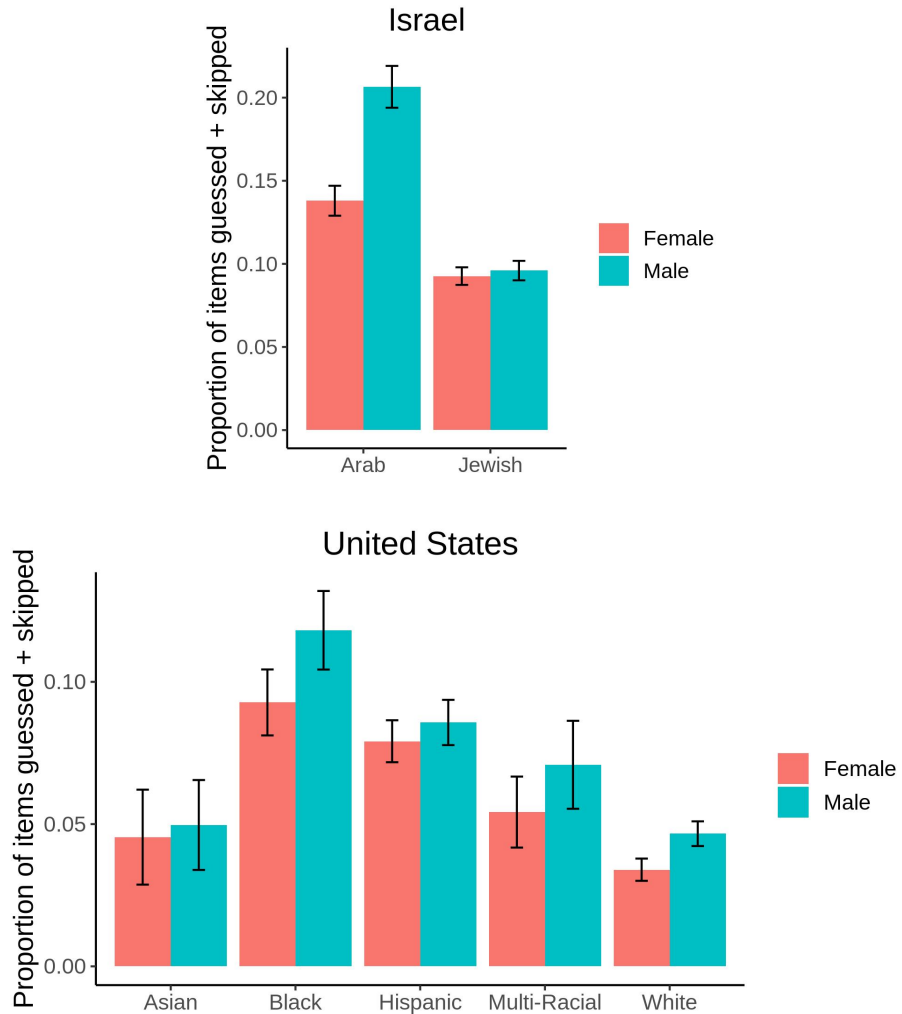
Figure A.1: Experimental Design



The low-stakes test was written in collaboration with the department of math instruction in the Israeli Ministry of Education.

The high-stakes test was written by the Israeli National Authority for Measurement and Evaluation in Education (RAMA)

Figure A.2: Gender Differences in Likelihood of Guessing and Skipping Across Ethnicities



The graphs show the average proportion of items identified as a skip or a rapid guess. A skipped item, is an item the student saw, but did not answer. Items the student did not reach are not included. Items identified as a rapid-guess are items the student answered very quickly, such that it is not likely she had the time to meaningfully answer the item. See Section B.2 of the appendix. Error bars represent 95% confidence intervals.

Figure A.3: Gender Differences in Endurance Across Ethnicities

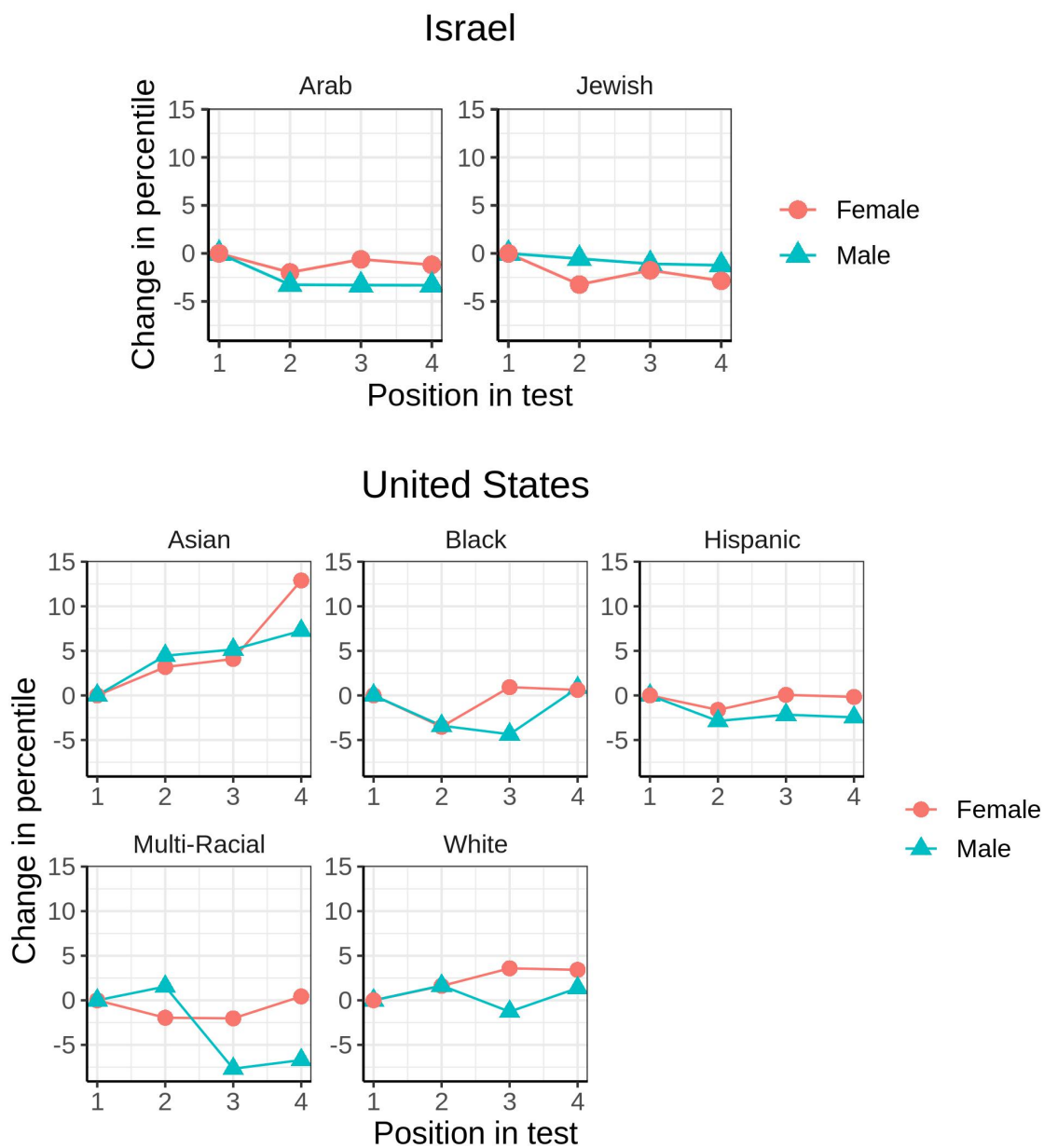
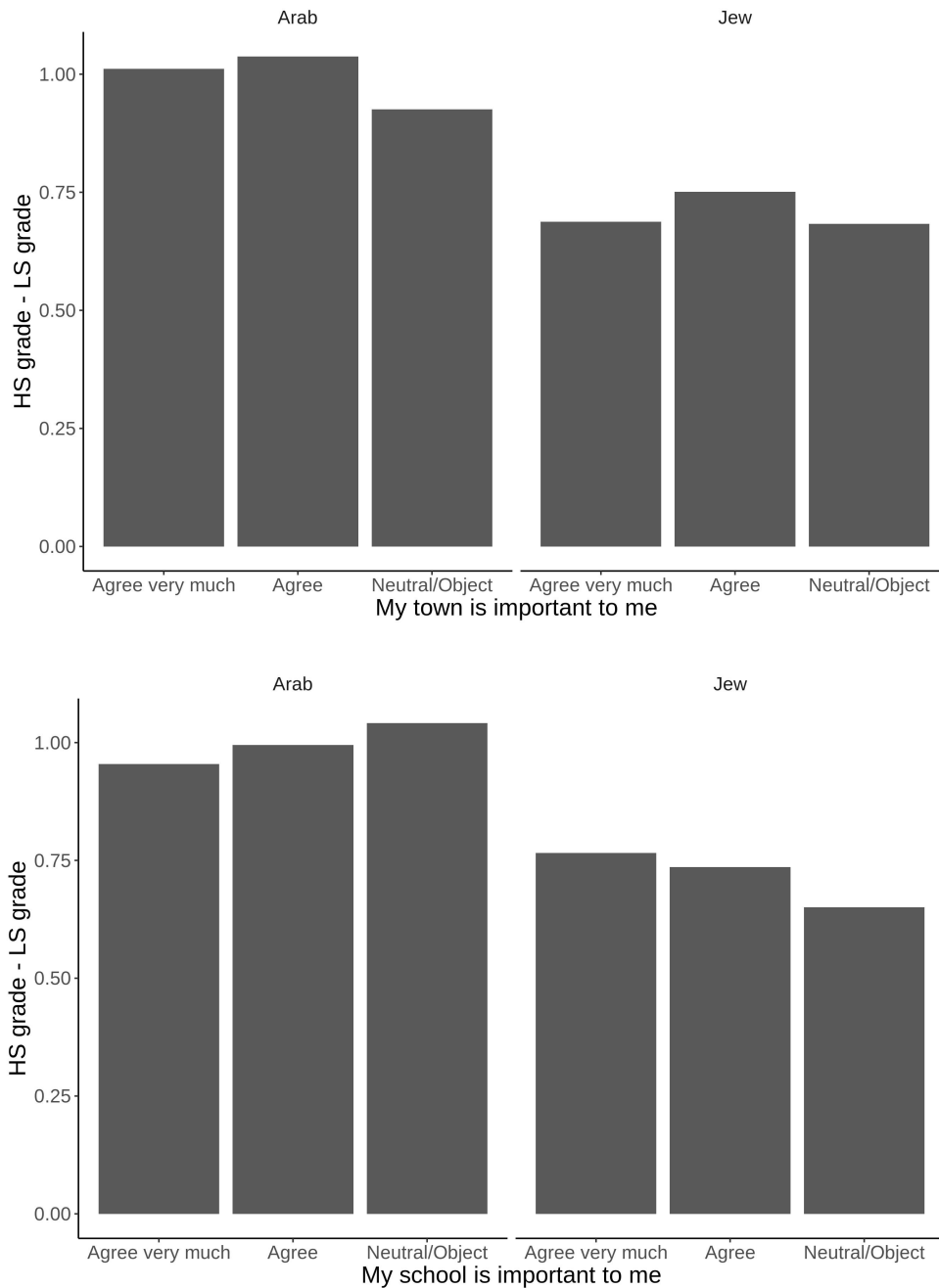


Figure A.4: Sense of Belonging and Performance Difference Between High and Low-Stakes Test



In the post-test questionnaire we distributed after the low-stakes exam, we asked the to what extent they agree with the two statements - “I care about my school” and “I care about my town / village / community” on a five-level Likert scale from “Strongly Disagree” to “Strongly agree”. Since a large majority answered either “agree” or “strongly agree”, I analyse the results comparing “strongly agree”, “agree”, and neutral or less.

Figure A.5: Mean Item Score According to Cluster Position

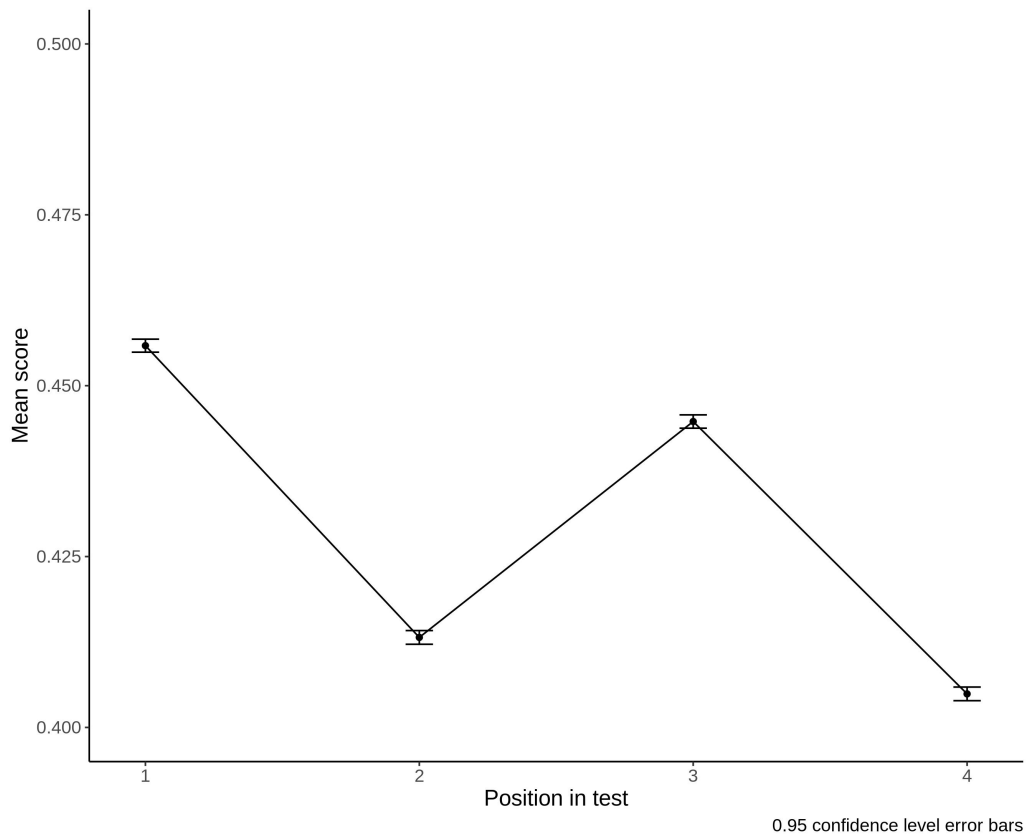


Figure A.6: Proportion of the Group Assigned to Each Science Test Cluster

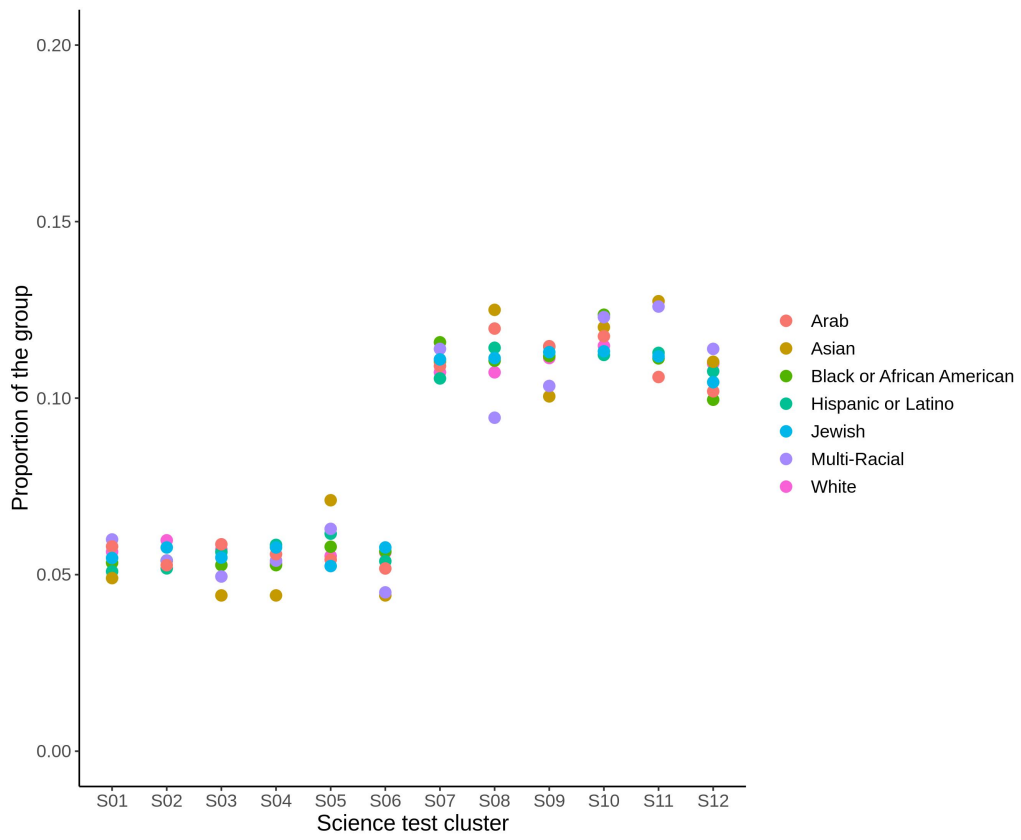


Figure A.7: Distribution of Students' Response Time on Item CS656Q01S

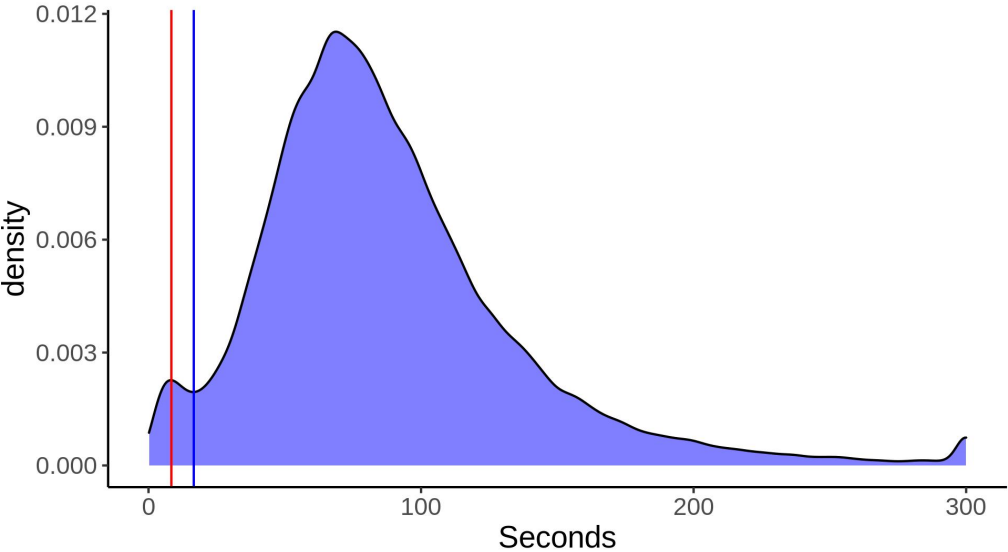


Figure A.8: Example of a PISA Question - Bird Migration 1

PISA 2015

Bird Migration
Question 1 / 3

Refer to "Bird Migration" on the right. Click on a choice to answer the question.

Most migratory birds gather in one area and then migrate in large groups rather than individually. This behaviour is a result of evolution. Which of the following is the best scientific explanation for the evolution of this behaviour in most migratory birds?

- Birds that migrated individually or in small groups were less likely to survive and have offspring.
- Birds that migrated individually or in small groups were more likely to find adequate food.
- Flying in large groups allowed other bird species to join the migration.
- Flying in large groups allowed each bird to have a better chance of finding a nesting site.

BIRD MIGRATION

Bird migration is a seasonal large-scale movement of birds to and from their breeding grounds. Every year volunteers count migrating birds at specific locations. Scientists capture some of the birds and tag their legs with a combination of coloured rings and flags. The scientists use sightings of tagged birds together with volunteers' counts to determine the migratory routes of birds.




Figure A.9: Distribution of Students' Response Time on Item CS521Q06S

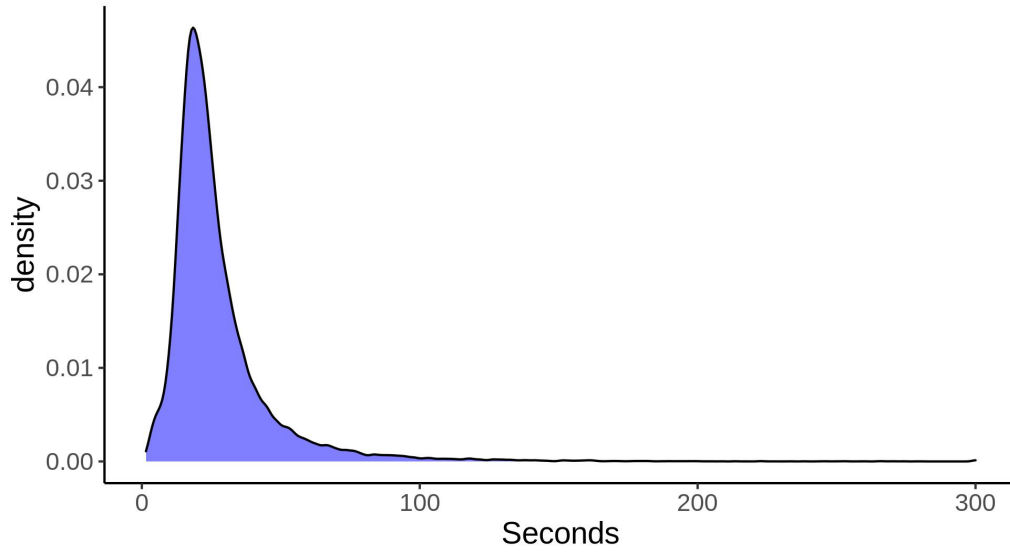
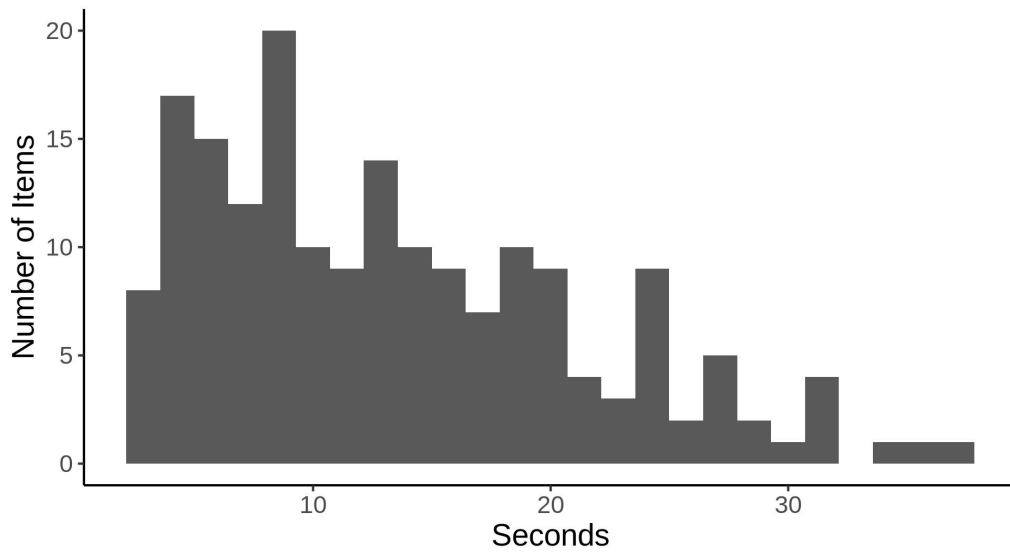


Figure A.10: Distribution of Rapid-Guessing Thresholds



Tables

Table 9: Research Population - PISA

	Country	Ethnicity	N	Female
1	Israel	Jewish Secular	3,309	0.50
2	Israel	Arab	1,622	0.55
3	Israel	Jewish Orthodox	936	0.58
4	Israel	Jewish Ultra-Orthodox	610	0.88
5	Israel	Jewish Others	115	0.31
6	United States	White, not Hispanic	2,481	0.48
7	United States	Hispanic/Latino	1,753	0.52
8	United States	Black/African American	782	0.53
9	United States	Multi-Racial	335	0.52
10	United States	Asian	207	0.46
11	United States	Other	62	0.39