Do Religion and Gender Matter on Low-Stakes Assessment Tests: A Field Experiment*

Yuval Ofek-Shanny

May 30, 2019

Abstract

Standardized assessment tests are increasingly used as an important policy and research tool. Thus, it is important that they are an accurate reflection of ability. We examine how 8^{th} graders perform, based upon gender and religion, on a real-life high-stakes test, presumably measuring true ability in comparison to a low-stakes test. Overall, Jews have a significantly smaller (15.4 points) grade difference between the high-stakes and the low-stakes test than Arabs (23.7). The smaller grade difference suggests that 60% of the Jewish-Arab performance gap in national assessment tests can be attributed to effort differences in the test itself rather than in student ability. This study shows that educational reforms and educational and economic policy must use caution when assessing the policy results according to low-stakes assessment tests.

*I would like to thank Todd Kaplan, Niza Sion, Fatena Marjie, Joel Rapp, Members of the Economics department in University of Haifa, Participants of the 2019 Asia-Pacific ESA Meeting, Abu Dhabi and many more, for helpful comments and suggestions.

1 Introduction

Improving education policies using standardized assessment tests is a methodology of rising use in recent years. The assessment tests are used to evaluate the quality of teachers, schools, regional and countrywide educational systems. The results of international assessment tests find that Israeli ninth grade students on average are behind students in most other developed countries. Furthermore, significant performance gaps were found between boys and girls, Hebrew and Arabic speakers, and religious and secular Jews. In the 2008-2017 Israeli Schools Growth and Effectiveness Measures (MEITZAV) math tests, secular Jews outperformed religious Jews by 9-15 points, Hebrew speakers outperformed Arabic speakers by 32-53 points and in 2017 Arab girls outperformed Arab boys by 14 points (all on a 335-675 scale) (RAMA, 2017).[1] In addition, on the math part of the 2015 Program for International Student Assessment (PISA) conducted by the OECD, Israel ninth grade students were ranked 30 of 35 OECD countries, 20 points under the OECD average. Measuring the gap between 5th and 95th percentiles, Israel had the largest of all OECD countries. Inside Israel, the gaps between Hebrew and Arabic speaking students was 104 points. Comparing gender results, Hebrew speaking secular boys outperformed secular girls by 12 points, religious boys outperformed religious girls by 39 points and Arabic speaking girls outperformed Arabic speaking boys by 12 points (all on a 200-800 points scale) (RAMA, 2016).

In response to the Pisa tests results, the Israeli Education Minister Naftali Bennett stated that the test results "emphasize the need for reducing gaps in Israel's education and indicate that significant improvement is needed" (Dattel, 2016).

The purported reasons for the large gaps between different sectors in the population vary. Most studies suggest natural gender differences, cultural differences and differential investment in education per student in the different sectors. (Rapp, 2015)

[1] Converting the 2017 GEM tests results to the 0-100 scale used in this study is done by using the formula $Grade_{0-100} = \frac{Grade_{335-675}-335.3}{3.397}$. Since we are only dealing with differences, only division by 3.397 is required

More recent, there is increasing interest in the possible effect of heterogeneous student motivation on the test itself. A difference in test taking motivation can create under or over estimation of performance gaps between the different sectors. This bias should be a concern especially in assessment tests with low stakes for the test taker. While it is clear that motivation is an important factor in determining performance, most of the assessment tests findings are presented implicitly assuming all students make the same effort in the tests. Previous studies found differences with respect to gender, cultural and race/ethnicity groups responses to Low-Stakes tests as well as to monetary and non-monetary incentives (Attali et al., 2011; Levitt et al., 2016; Gneezy et al., 2017).

In this paper we examine differences in the test behaviour between 8^{th} grade boys and girls from different religions - secular Jewish, religious Jewish, Muslim Arabs and Christian Arabs. To examine the difference in behaviour we conducted an experiment in nine Israeli junior high schools belonging to four different religion groups.[2] In order to explore the performance difference between high and low stakes tests, each 8^{th} grade student was given two similar math tests in a GEM format.[3] The first, with no stakes to the student, presented as a practice test. The second, a week later, was a final year's exam determining about 30% of the student final year's grade - high personal stake. If the religion and gender groups behave differently in the two tests, we expect heterogeneity over groups in the performance difference between the high and low stakes tests. This means that using only a low stakes assessment test might give a biased results when comparing the groups' abilities.

This study is unique in two of its characteristics: first, we observe student performance in a real-life high stakes situation. Most of the literature in the field is based on controlled experiments comparing performance in low-stakes

[2]Actually, five religion groups, but the experiment in the Druze school failed because the math coordinator told the students they will be graded for the low-stakes test making it a high-stakes test.

[3]The GEMS known as "MEIZAV" (Hebrew acronym for "School Growth and Efficiency Measures") is a national assessment system for 2^{nd}, 5^{th} and 8^{th} grade students. GEMS include student achievement exams as well as questionnaires designed to gather information about the school climate and pedagogical environment.

test to a second test with some artificial (usually monetary) incentives. Second, in the experiment we use a "within subject" design. This enables us to observe the same individual in a low and high stakes situations. Comparison of the same individual instead of two different groups, enables a richer estimation model including controlling for the student's high stakes grade as a proxy for ability.

Our results show that while overall, boys and girls exhibit similar differences ($HS_{grade} - LS_{grade}$) examining the gender difference across religion groups suggests Muslim girls achieve higher grades in low stakes assessment tests than similar ability Muslim boys and religious Jewish boys achieve higher grades in low stakes assessment tests than similar ability religious Jewish girls. The experiment results also demonstrate an example of how using low stakes assessment tests might lead to wrong conclusions. For example, while the Christian students in the experiment perform best of all religion groups in the high stakes test, Jewish students achieve the highest average grade in the low stakes test. By using the low stakes results we can conclude secular Jewish students have higher ability while the reality might be opposite. Further explanations and elaboration are found in the Results section.

Combining the experiments results with 2017 Israeli national 8^{th} grade math assessment tests (GEM) suggests that around 60% of the Jewish Arab performance gap can be attributed to effort exertion in the test itself rather than students' ability. Similar calibration to further GEM results is made in the conclusion section.

These results strengthens the need for a very careful examination and analysis of performance gaps across population groups in general and especially when based on low stakes assessment tests. On a broader scale, educational reforms and educational and economic policy must use precaution when assessing policy results according to low stakes assessment tests.

2 Literature Review

The impact of low motivation and effort on the validity of assessment tests is in dispute. Most researchers argue that low motivation and effort can create a significant bias in ability estimation. Wise and DeMars (2005) synthesize previous papers to show that in 24 out of 25 experiments, the more motivated group of examinees outperformed the less motivated group. The average score was 0.59 STD higher in the motivated group.

On the other hand, some researchers claim that the motivation differences are not meaningful in many assessments scenarios. Eklöf (2010) finds that 8th graders invest a high level of effort in the international assessment tests, but 12th graders do not. Butler and Adams (2007) argue that systematic cultural differences in effort do not pose a threat to the assessment validity.

In the attempt to understand students' behavior in low-stakes tests, many studies tried to evaluate the impact of monetary and non-monetary incentives on motivation, effort, and performance doing lab and field experiments. Recently, a few large-scale field experiments examined the effect of monetary incentives. Results vary across the different settings, but a few important observations arise. Incentives framed as losses improved the performance significantly in all experiments while framing incentives as gains improved the performance in most but not all experiments. Characteristics like age, test subject, and gender, had a significant effect on the incentive impact while race and ability did not (Levitt et al., 2016). Both fixed and conditional rewards had a significant effect on reported effort and score (Braun et al., 2011). Interestingly, all motivational power vanishes when rewards are distributed with a month delay (Levitt et al., 2016). In a different experimental design, List et al. (2016) compared the performance of elementary and middle school low-performing students in official standardized tests to similar tests with a financial incentive. The incentives had a substantial positive effect ranging between 0.31 to 0.46 standard deviations.

Although most scholars claim a significant effect for monetary incentives, some do not find such an impact (O'Neil et al., 1995; Baumert and Demmrich,

2001; Eisenkopf, 2011; O'Neil et al., 2005). In all of these studies, the reward was given in a delay so Levitt et al. (2016) findings presented above can explain the null results.

Many standardized assessment tests have low stakes to the students. For that reason we expect to see a difference between the effort and performance in high stakes tests and in these standardized tests. Several studies examine how the examinee's characteristics and test item characteristics affect the amount of effort and performance. Usually it is assumed that in high-stakes tests, students will invest full effort so what is examined is the **effort and performance decline** in low-stakes tests.

Many studies found that test items characteristics have a significant impact on the decline in effort and performance on low-stakes tests (Sundre, 1999; Wise et al., 2009; DeMars, 2000; Wolf and Smith, 1995). Few studies examine the effect of examinee's characteristics on the effort or performance decline. These studies show that race, gender, and ability may affect the decline in performance.

While Wise et al. (2009) did not find a significant gender effect on effort in low stakes tests, Attali et al. (2011) find that white male scores declined significantly more than the white females' scores. Comparing the performance decline among different race groups they find significant differences between Whites, Asians, Hispanics and African Americans. Previous studies that reported gender and racial differences in test taking attitudes might explain these findings (Chan et al., 1997; DeMars et al., 2013; Eklöf, 2007; OECD, 2015). Further study of the examinee's characteristics affecting the decline in effort and performance is required.

With respect to students in standardized assessment tests, Borgonovi and Biecek (2016) show that proxies for motivation such as fraction of questions left unanswered or ability to maintain effort during the test suggest females have higher motivation in standardized assessment tests. Rigbi et al. (2013) also compare the ability to maintain effort during the 2006 and 2009 PISA tests to show Hebrew speaking students have higher motivation to perform well in standardized assessment tests than Arabic speaking students. All the

above suggest similar ability students differing in gender and religion might obtain different grades in assessment tests thus posing a threat to tests results credibility.

3 Experimental Design

We conducted the experiment in the spring of 2017 in nine junior high schools in the northern part of Israel. Schools in Israel are usually segregated by religion or a branch of a religion. The experiment took part two secular Jewish schools, two religious Jewish schools, two Muslim schools, two Christian schools and one Druze school.[4] The schools were selected with the advice of Math Supervision Department in the Israeli Ministry of Education so that they represent an average ability schools for their religion. Ability of the schools is mainly measured using GEM. Every year, one-third of the junior high schools take part in the GEMS so that every junior high school participates once every three years. When participating, the tests are given to the students with no grade or other personal consequences making the test a low stakes test. All schools not participating in a specific year's GEMS are offered by the Ministry of Education to take them as an internal test. Many schools use it as their year's final test, usually determining approximately 30% of the year's final grade and thus making it a high stakes test. With the advice of Math Supervision Department in the Ministry of Education, we chose for the experiment schools that took the math achievements test as an internal high stakes test.

As a low-stakes test for differences comparison we used a math test written with the help of Math Supervision Department in a way that maintained the structure and difficulty level of GEM math test. The tests are composed of 21-23 questions about evenly divided between multiple choice questions, fill in the blank questions or a combination of the two.[5] Explanation was requested

[4]The 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.

[5]When writing the LS test, we didn't know how many questions the HS test will have so we wrote 23 questions based on previous year structure, the HS test written by RAMA

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.[6]

All 8^{th} 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 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 we verified that the students understood that the test was low stakes. Apart from one school, 95% answered correct to the verification questions. In the Druze school there was a problem in the instructions given to the children so we could not use the school's results in the analysis. in the after test questionnaire the students were requested to report their level of effort in the test and few demographic characteristics.

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

Being an experiment, the first test was not mandatory but thanks to the commitment of the schools' staff, almost all the students participated in the experiment.[7] The second test was part of the schools syllabus and thus mandatory. Despite that, not all students participated in the second test. One Muslim school took the second test during Ramadan fast, this had a big effect on the students behaviour so it also had to be excluded from the results analysis. Table 1 presents the population included in the experiment analysis by gender

was a 21 questions test.

[6]The translation from Hebrew to Arabic was done under the guidance of Arabic Math Supervision department in the Israeli Ministry of Education

[7]Although no the students did not receive any grade or feedback on the first test, they were encouraged to take the it as a preparation for the second (year's final, high-stakes) test.

and religion.

Raw GEM's grades range on a 0-100 scale.[8] The grading for the high stakes test was made by schools' teachers according to a very detailed grading sheet from RAMA. The grading of the low-stakes test was made by the experiment team using a similar detailed grading sheet.

4 Empirical Framework

To examine the change in student performance between the high and low stakes test, we estimate the following first difference equation[9]

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

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, Male denotes a dummy variable set to 1 for male students, Rel are dummy variables for Religion, Female and Secular Jew are omitted variables. Vector x_i are student characteristics that include the following covariates: dummies for - mother's and father's education, understanding the instructions,[10] the student's grade in the high stakes test (as a proxy for student's ability) and the student's grade squared (enabling the ability's effect on grade difference to change across the ability spectrum). The coefficients of interest are β's that denote the difference between high stakes grade and low stakes grade across religions for females and γ's that denote the difference between male's and female's grade difference across religions. Using the grades difference specification controls for an individual's fixed effect taking in consideration all factors that affect students grade in both tests.

The regression results, estimated by OLS are reported in Table 5. Robust standard errors are in parenthesis under the estimated coefficient. In column

[8]For comparison across years, RAMA publishes every year a calibration formula to a 200-800 scale. For this paper's needs the 0-100 scale is sufficient.

[9]Similar framework to Attali et al. (2011).

[10]After the low-stakes test, the students were asked will they receive a grade for the test and whether the grade has an effect on their final year's grade. Answering yes to both questions means that the student misunderstood the low stakes test for a high stakes test.

1, we report the regression without covariates. The first 3 rows report the difference from Secular Jewish girls' $HS_{grade} - LS_{grade}$. So Christian girls have a 7 points bigger $HS_{grade} - LS_{grade}$ difference than secular Jewish, etc. Rows 4-7 report the boys to girls difference in $HS_{grade} - LS_{grade}$ finding the difference for a Muslim boy is 4.6 points bigger than for a Muslim girl. Column 2 reports the estimation after controlling for covariates. The persistence in coefficients values with controls suggests that the results are not driven from students characteristics or ability (as reflected by high stakes grade).

5 Results

5.1 Gender Differences

Table 2 and Figures 1 and 2 present students' performance in low and high stakes tests by gender. While most studies find that boys outperform girls in math tests, we find a very small difference on both tests. On average, boys gained 1.8 points higher in the high stakes test and 2.4 points higher in the low stakes test. As expected, on average, students grades in the low stakes test are significantly lower with 18.5 points difference for boys and 19.2 for girls. As most studies, we assume the high stakes grade is a better proxy for ability then low stakes because most students exert full or close to full effort in these tests. So, assuming boys and girls exert similarly full effort in high stakes tests, the very small difference in performance decline suggests that, on average, boys and girls exert similar effort in the low stakes tests. Looking into the gender differences by religion allows us to see a more complex picture.

Table 4 presents students performance by gender and religion. In the last 3 rows we can see the differences between boys and girls for each religion. Jewish boys outperform girls on both high and low stakes tests (4.3 and 3.4 points respectively). Christian boys and girls achieve on average almost the same grade on both tests. While religious Jewish girls outperform religious Jewish boys by 5.6 points according to high stakes grades, they lag religious Jewish boys by 4 points in low stakes grades. This means that measuring the

religious Jewish students ability according to low stakes grades might lead to a possibly wrong conclusion that religious Jewish boys have higher abilities than girls (as indeed indicated by PISA 2015 math tests results). The last row of Table 4 reports the difference in boys grades minus the difference in girls grade $(Mean_{HS.boys} - Mean_{LS.boys}) - (Mean_{HS.girls} - Mean_{LS.girls})$ for each religion. As mentioned above, we can see that Jewish and Christian students are in line with no gender difference for this variable. For the religious Jewish students we see that the girls decline from high stakes to low stakes is 9.6 points bigger than the religious Jewish boys, suggesting the religious Jewish boys exert bigger effort than girls in the low stakes test. This result moderates to 6.8 points difference when controlling for covariates, but remains significant as shown in row:"Religious Jew:Male", Column 2 of Table 5. Similarly, but with opposite sign, looking again on the last 3 rows of Table 4 we see that Muslim boys achieve 2.5 points more than girls in the high stakes test but 2.1 points less on the low stakes test. Again, this means that measuring the Muslim students ability according to low stakes tests might lead to a possibly wrong conclusion that Muslim girls have higher abilities than Muslim Boys (as indeed suggested by Rapp, 2015). The $(Mean_{HS.boys} - Mean_{LS.boys}) - (Mean_{HS.girls} - Mean_{LS.girls})$ difference is positive 4.6 points suggesting Muslim girls exert higher effort in the low stakes test. As shown in Table 5 the result for Muslim students with and without controls is consistent but not significantly different from zero. For Muslim students, the gender difference is not consistent across ability scale (measured by high stakes grade). This is shown in Figure 3 that presents the grade difference $HS_{grade} - LS_{grade}$ on the y-axis and student's high stakes grade on the x-axis for each gender from each religion. Trend line is a local polynomial regression fitting. While lower ability Muslim girls have smaller $HS_{grade} - LS_{grade}$ difference than lower ability boys, above HS grade of 85 the trend flips and Muslim boys have a significantly lower $HS_{grade} - LS_{grade}$ difference.

5.2 Religion Differences

Table 3 and Figures 4 and 5 present students' performance in low and high stakes tests by religion. In the high stakes test, Christian students achieve the highest average grade of 68.9. Secular Jews and Muslims with almost the same average of 64.3 and 64.4 (but higher median for the Jews) and religious Jews with lowest 57.9 average.

All religion groups have significantly lower grades in the low stakes test with an average decline of 15.4-23.7 points. Because the decline differs across religions, looking only at low stakes results might lead to wrong conclusions. While Christian students perform best in the high stakes test, Jewish students achieve the highest average grade of 49 points in the low stakes test. 2.3 points **above** Christian students. The difference between Muslim students to religious Jewish students shrinks from 6.4 in high stakes grades to only 1.5 points in the low stakes grades.

Table 5 reports the raw and controlled difference in $HS_{grade} - LS_{grade}$ between secular Jewish girls and other religious and gender groups. Figure 5 and the last column of Table 3 reports $HS_{grade} - LS_{grade}$ for each religion. We can see that the difference for secular Jewish students is the smallest, suggesting they exert the highest effort in low stakes tests. Controlling covariates, the difference between secular Jewish students and other groups remains significant for all religion groups except religious Jewish boys ($ReligiousJew + ReligiousJew : Male$ coefficients in Table 5). The highest difference $HS_{grade} - LS_{grade}$ is estimated for Muslim and Christian students (23.7 and 22.2 respectively), suggesting these religion groups exert lowest effort in the low stakes test.

6 Conclusion

In this study we examined the differences in performance in high and low stakes GEM ("MEITZAV") tests of Israeli boys, girls, Secular and religious Jews, Muslim and Christian Arabs in 8^{th} grade. The main question this study

is focused on is "to what extent proper economic and educational policy and research can be based on low stakes assessment tests?" Differing from most studies in this field that base their study on monetary incentives, we examine students behaviour in a real life high stakes test determining the students final year's grade.

The results support the wide literature suggesting student perform better in high stakes test. This experiments' results suggest that heterogeneous grade decline across gender and religion groups might cause wrong and even opposite direction conclusions when based on low stakes assessment tests. For example, in this study's results, religious Jewish girls achieve higher grades than boys in high stakes tests but religious Jewish boys achieve higher grades in low stakes tests. Studies based on low stakes tests might falsely conclude that religious boys have higher ability than girls. This direction of effect suggests religious boys exert higher effort in low stakes tests than girls - different from previous literature that found girls exert higher effort in low stakes tests. Similarly, but on the opposite direction (and in line with previous literature), Muslim girls achieve lower grades in high stakes tests than Muslim boys but the girls achieve higher grades in the low stakes test. Again, conclusions based on the low stakes test might be in the wrong direction of performance difference. Interesting to note that while religious and secular Jews as well as Muslims and Christians (Arabs) belong to the same ethnicity, they differ in their behaviour in low and high stakes as well as in their gender differences.

It is important to note that this experiment population is not fully representative and it is not the aim of this study to estimate the ability difference between the population groups. This study's main aim is to examine the difference in behaviour in high and low stakes tests. Using this study's results, to examine assessment tests' findings, can offer interesting interpretation. As presented in the introduction to this paper, the 2017 math GEM test found an average of 10.8 points between grades of Jews and Arabs.[11] According to the experiment results, about 60% of this gap can be explained by Jews exerting

[11] 37 points on the 335-675 2017 scale are converted to 10.8 points on the 0-100 scale.

higher effort in the low stakes assessment test.[12] Similarly, GEM tests suggest secular Jews outperform religious Jews by 2.6 points while this experiments results suggest that controlling for ability, secular Jews achieve about 3.5 points higher than religious in low stakes assessment tests. Based on the GEM results and this study's results we can suggest that it is possible that religious Jewish students even have higher ability then secular Jewish students. Further interesting results arise from the gender comparison. While GEM results find that Muslim girls achieve 4.1 points higher grades than boys, combining it with the experiment‘s results presented in Table 5 brings to a possible conclusion of no ability difference. Similarly, GEM tests find no difference in Jewish religious boys and girls, but combining with this study's results, we suggest a possible conclusion of higher ability of girls. This study's findings support the GEM result of similar ability between secular Jewish boys and girls.

This study's results strengthens the need for a very careful examination and analysis of performance gaps across population groups. Specifically, it strengthens the need to specify if the assessment tests were a low stakes tests and whether the assumption that these results are a good proxy for ability is made. On a broader scale, educational reforms and educational and economic policy must use precaution when assessing the policy results according to low stakes assessment tests.

Further research can help us understand what causes these differences in behaviour across different population groups.

[12] As presented in Table 5, controling for ability, Christians and Muslims achieve 6 points less then Secular Jews on the low stakes test

References

Attali, Yigal, Zvika Neeman, and Analia Schlosser, "Rise to the challenge or not give a damn: differential performance in high vs. low stakes tests," *IZA discussion paper 5693*, 2011.

Baumert, Jürgen and Anke Demmrich, "Test motivation in the assessment of student skills: The effects of incentives on motivation and performance," *European Journal of Psychology of Education*, 2001, *16* (3), 441–462.

Borgonovi, Francesca and Przemyslaw Biecek, "An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test," *Learning and Individual Differences*, 2016, *49*, 128–137.

Braun, Henry, Irwin Kirsch, and Kentaro Yamamoto, "An Experimental Study of the Effects of Monetary Incentives on Performance on the 12th-Grade NAEP Reading Assessment.," *Teachers College Record*, 2011, *113* (11), 2309–2344.

Butler, Jayne and Raymond J. Adams, "The impact of differential investment of student effort on the outcomes of international studies," *Journal of applied measurement*, 2007, *8* (3), 279.

Chan, David, Neal Schmitt, Richard P. DeShon, Cathy S. Clause, and Kerry Delbridge, "Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation.," *Journal of Applied Psychology*, 1997, *82* (2), 300.

Dattel, Lior, "Israeli Pupils Underperform Yet Again on International Exams," *Haaretz*, Dec 2016.

DeMars, Christine E., "Test stakes and item format interactions," *Applied Measurement in Education*, 2000, *13* (1), 55–77.

_ , Bozhidar M. Bashkov, and Alan B. Socha, "The role of gender in test-taking motivation under low-stakes conditions," *Research & Practice in Assessment*, 2013, *8*.

Eisenkopf, Gerald, "Paying for better test scores," *Education Economics*, 2011, *19* (4), 329–339.

Eklöf, Hanna, "Gender differences in test-taking motivation on low-stakes tests," in "The annual meeting of the American Educational Research Association (AERA), Chicago, IL, April 2007" 2007.

_ , "Skill and will: test-taking motivation and assessment quality," *Assessment in Education: Principles, Policy & Practice*, 2010, *17* (4), 345–356.

Gneezy, Uri, John A List, Jeffrey A Livingston, Sally Sadoff, Xiangdong Qin, and Yang Xu, "Measuring success in education: the role of effort on the test itself," Technical Report, National Bureau of Economic Research 2017.

Jr, Harold F O'Neil, Brenda Sugrue, and Eva L Baker, "Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance," *Educational Assessment*, 1995, *3* (2), 135–157.

Levitt, Steven D, John A List, Susanne Neckermann, and Sally Sadoff, "The behavioralist goes to school: Leveraging behavioral economics to improve educational performance," *American Economic Journal: Economic Policy*, 2016, *8* (4), 183–219.

List, John A., Jeffrey A. Livingston, and Susanne Neckermann, "Do Students Show What They Know on Standardized Tests?," 2016.

OECD, "The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence, PISA," *OECD Publishing*, 2015.

O'Neil, Harold F., Jamal Abedi, Judy Miyoshi, and Ann Mastergeorge, "Monetary incentives for low-stakes tests," *Educational Assessment*, 2005, *10* (3), 185–208.

RAMA, "PISA 2015, PISA – Programme for International Student Assessment, Israeli Overview," 2016.

_ , "Meitzav 2017, Growth and Effectivness School Measures," 2017.

Rapp, Joel, "Gender Gaps in Mathematics and Language in Israel–What Can Be Learned From the Israeli Case?," Technical Report, Working paper, National Authority for Measurement and Evaluation in Education 2015.

Rigbi, Amihai, Joel Rapp, and Inbal Ron-Kaplan, "Student Motivation in Pisa Research: Comparison of 2006 and 2009 Findings," 2013.

Sundre, Donna L., "Does examinee motivation moderate the relationship between test consequences and test performance?," *James Madison University*, 1999.

Wise, Steven L. and Christine E. DeMars, "Low examinee effort in low-stakes assessment: Problems and potential solutions," *Educational assessment*, 2005, *10* (1), 1–17.

_ , Dena A. Pastor, and Xiaojing J. Kong, "Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice," *Applied Measurement in Education*, 2009, *22* (2), 185–205.

Wolf, Lisa F. and Jeffrey K. Smith, "The consequence of consequence: Motivation, anxiety, and test performance," *Applied Measurement in Education*, 1995, *8* (3), 227–242.

Table 1: Experiment Population

	full	jew	religious_jew	christian	muslim	other
N	599	242	92	141	80	44
Girls	311	127	54	72	51	7
Boys	267	114	38	67	29	19
NA	21	1		2		18

Table 2: Performance by Gender

	HS.mean	HS.std	HS.median	LS.mean	LS.std	LS.median	Diff.mean
female	63.7	25.5	67.5	44.5	21.7	44	19.2
male	65.5	26.3	73	46.9	24.7	47	18.5

Table 3: Performance by Religion

	HS.mean	HS.std	HS.median	LS.mean	LS.std	LS.median	Diff.mean
jew	64.4	27.2	73	49.0	25.1	50	15.4
christian	68.9	25.7	76	46.7	21.4	46	22.2
muslim	64.3	22.3	62	40.6	20.4	42	23.7
religious_jew	57.9	24.3	59.5	39.1	20.5	34.5	18.8

Table 4: Performance by Gender and Religion

	jew	christian	muslim	religious_jew
female.HS.mean	62.4	68.6	63.4	60.2
female.HS.std	26.8	27	22.7	22.3
female.HS.median	70.5	76	62	60
female.LS.mean	47.4	46.6	41.4	37.4
female.LS.std	23.6	20.7	18.1	20
female.LS.median	46	45	42	34
female.Diff.mean	15	22	22	22.8
male.HS.mean	66.7	69.3	65.9	54.6
male.HS.std	27.5	24.4	22	27
male.HS.median	75	75	68	59
male.LS.mean	50.8	46.7	39.3	41.5
male.LS.std	26.7	22.2	24.1	21.1
male.LS.median	54	48	40	36
male.Diff.mean	15.9	22.5	26.6	13.2
gen.HS.diff	4.3	0.7	2.5	-5.6
gen.LS.diff	3.4	0.1	-2.1	4
gen.diff.of.diff	0.9	0.6	4.6	-9.6

Table 5: Regression results, Grade Differecne in High and Low stakes tests on Gender and Religion

	High Stakes Grade - Low Stakes Grade	
	(1)	(2)
Christian	7.004***	5.981***
	(2.128)	(1.713)
Muslim	7.011***	6.107**
	(2.662)	(2.383)
Religious Jew	7.824***	6.963***
	(2.047)	(1.862)
Secular Jew:Male	0.917	0.995
	(1.937)	(1.768)
Christian:Male	0.550	−0.093
	(2.357)	(1.842)
Muslim:Male	4.607	3.935
	(3.390)	(3.163)
Religious Jew:Male	−9.603***	−6.797***
	(2.652)	(2.140)
Constant	14.968***	−16.793***
	(1.228)	(3.912)
Controls	*NO*	*Full*
Observations	546	532
R^2	0.074	0.333
Adjusted R^2	0.061	0.305
Notes:	***Significant at the 1 percent level.	
	**Significant at the 5 percent level.	
	*Significant at the 10 percent level.	
	Omited religion - Secular Jew	

Figure 1: Grades in High and Low Stakes Tests by Gender

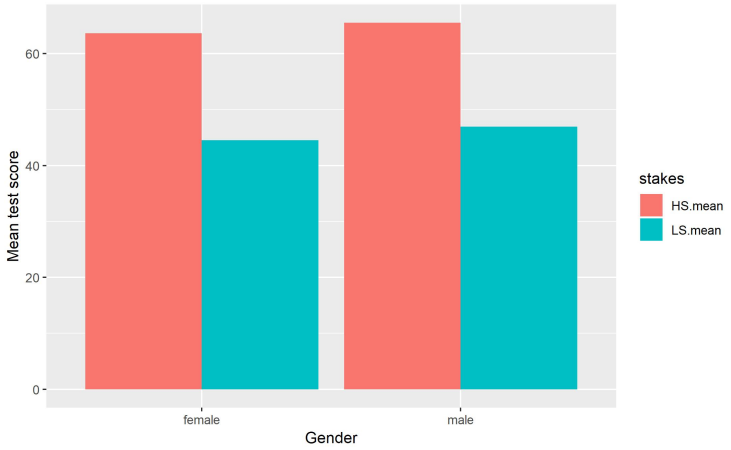

Figure 2: Grades Difference by Gender

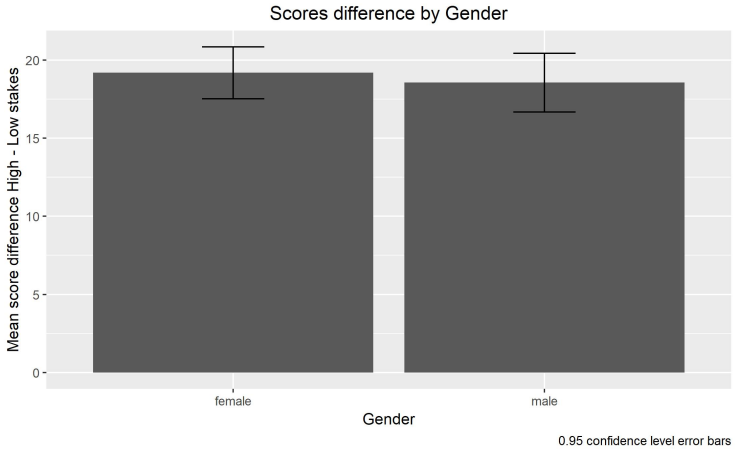

Figure 3: (HS grade - LS grade) by Gender and Religion

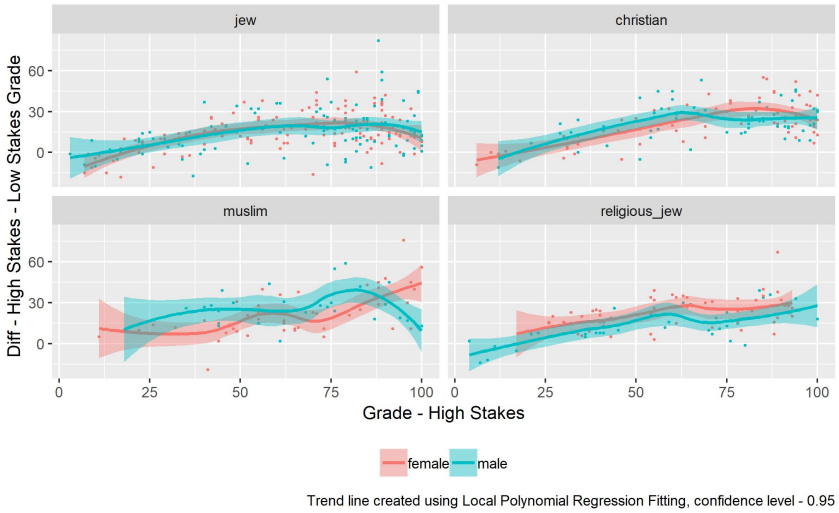

Figure 4: Grades in High and Low Stakes Tests by Religion

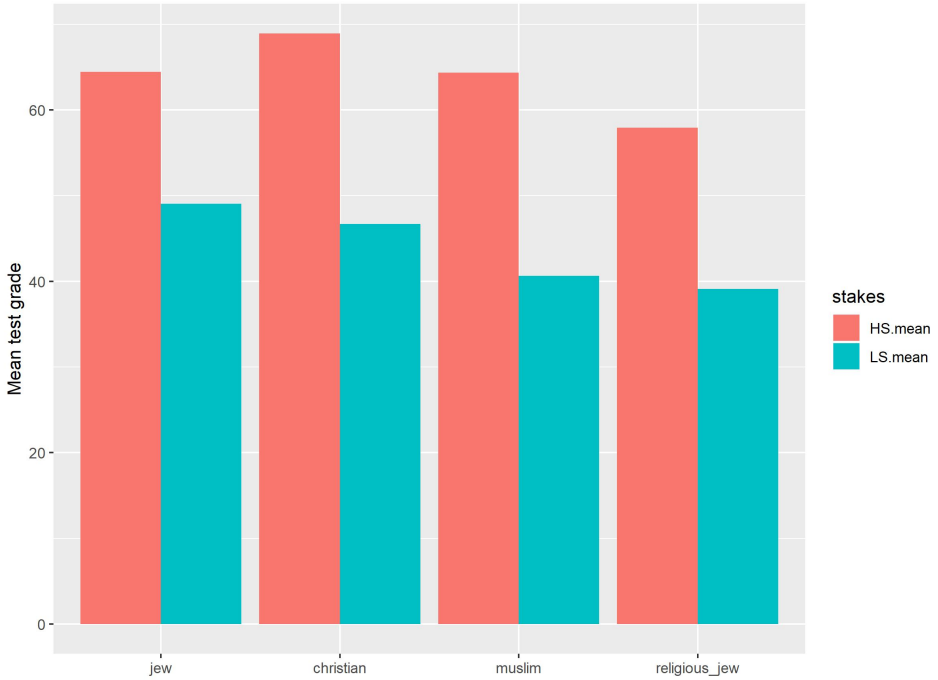

Figure 5: Grades difference HS-LS by Religion