Examination of Gender Bias in Admission to Universities in Israel

Razia Azen[*]

University of Illinois at Urbana-Champaign, Champaign, IL, USA

Shmuel Bronner and Naomi Gafni

National Institute for Testing and Evaluation, Jerusalem, Israel (NITE)

June 1999

The authors wish to thank David V. Budescu, Yoav Cohen, and Tamar Kennet-Cohen, for their insightful comments.

[*] This research was performed in the summer of 1998, while Razia Azen was on a fellowship at the National Institute for Testing and Evaluation in Jerusalem, Israel.

Abstract

This study examined whether the measures used in the admission of students to Universities in Israel are gender biased. The criterion used to measure bias was performance in the first year of University study, and the predictors consisted of an admission score, a high school matriculation score, and a standardized test score as well as its component subtest scores. Statistically, bias was defined according to the boundary conditions given in Linn (1984). No gender bias was detected when using the admission score (which is used for selection) as a predictor of first year performance in University. Bias in favor of women was found predominantly using school grades as predictor whereas bias against women was found predominantly in using the standardized test scores. It was concluded that the admission score is a valid and unbiased predictor of first year University performance for the two genders.

Introduction

Males and females differ in their performance on school work and standardized tests in a very complex manner (Willingham & Cole, 1997). Since admission to University is almost always determined by a combination of some high school achievement score and some standardized test score, it is important to distinguish between what may be called real gender differences, which reflect differences in aptitude, talent or ability, and differences due to biased testing, which result from using tests that consistently misrepresent the aptitude, talent or ability of a specific group of examinees.

Gender differences

There are certain gender differences in scholastic aptitude which are consistent and prevalent to such an extent that they are probably due to true gender differences rather than to faulty measurement instruments. In their comprehensive examination of gender differences, Willingham and Cole (1997) found that "females and males show broad similarity as well as distinctive patterns of difference." (p. 349). For example, Willingham and Cole state that

> "When test scores and school grades are compared for the same samples of students, the tendency for women to make somewhat better grades and men to make somewhat better scores was a consistent finding across different testing programs and in different subject areas." (p. 353).

The fact that men obtain better test scores is explained at least in part by the finding that there is greater variability in the male population (Feingold, 1992; Hedges & Nowell, 1995). Willingham and Cole (1997) explain that

> "differential variability is important because it means that there will be more males at extreme score levels – more low-performing males at the bottom and more high-performing males at the top – even if there is no mean gender difference. Differential variability also means that, in a selected group of outstanding students, high-scoring males will tend to outnumber high-scoring females." (p. 356).

Another general finding in the area of test performance is that there seems to be a small mean score difference in favor of males in self-selected samples of high school seniors, and no mean gender difference in tests administered to representative samples of all seniors (Stanley, Benbow, Brody, Dauber & Lupkowski, 1992). Willingham and Cole (1997) explain this phenomenon by noting that

> "This difference did not appear to result from the types of tests employed, but from the statistical effects of restricting the samples to higher scoring students where males tend to outnumber females. A detailed analysis of this phenomenon indicated that three factors are at work in sample restriction: the tendency to greater variability in male scores, restriction in the range of scores that comes from testing mostly higher scoring students, and the relative number of females and males in the selected group." (p. 349).

In studying the differences between the genders over time, the pattern of results appears to depend on the type of ability measured (e.g., verbal or mathematical). In mathematical ability, a meta-analysis conducted by Hyde, Fennema, and Lamon (1990) showed that the selectivity of the sample as well as age and the cognitive level of the test played a role. Specifically, a small mean difference favoring females was found in the elementary and middle school years, but a larger difference favoring males was found in the high school and college years. In verbal ability, the conclusions from specific studies tend to disagree somewhat on the nature of the differences – for example, Maccoby and Jacklin (1974) report an advantage for girls that becomes more pronounced after age 10 or 11; Hyde and Linn (1988) found no meaningful gender difference overall and no change over time; and Halpern (1992) and Cleary (1992) report an advantage for females starting in preschool or early grades with no change over time.

Therefore, it would probably be reasonable to expect that where gender differences exist they manifest in males doing better in math and science related fields and tests, and females performing better in predominantly verbal fields and tests. It also appears that if a test is administered to a self-selected sample, one may expect that the test results will favor males simply as the statistical artifact of self-selection and higher variability in the male population. Whether this expectation is true or not, it implies that test bias is a complicated question, the answer to which probably depends more on the use or misuse of test results than on the actual test scores.

Test Bias

Males and females differ in their chosen interests and activities (for a review of this literature, see Willingham and Cole, 1997), and these differences are likely to have an influence on school activities, grades and test scores; therefore, measurement instruments need not be modified simply because they find gender differences. However, when measuring some aptitude it is important to ensure that the aptitude being measured is truly relevant in terms of the eventual use of the test, and is indeed being measured in a fair and accurate manner for various segments of the population. If it is not, the use of the measurement instrument may be questioned. As Willingham and Cole (1997) put it,

> "If examinees have varied interests and experience, their ability to score well will certainly vary. Comparable opportunity to demonstrate skills is not the same as comparable opportunity to acquire skills. Test fairness can only address the former. Although our educational goal may properly be a comparable opportunity to acquire skills, the assessment goal is not equality of group scores. Indeed, a valid and fair test must show group score differences that appropriately reflect differences in interests and experience." (p. 359)

Bias is an issue combining measurement and value-dependent considerations. Hunter and Schmidt (1976) distinguish between three ethical positions which affect one's interpretation of what constitutes discrimination and test bias. Those who endorse "unqualified individualism" advocate striving to "make a scientifically valid prediction of each individual's performance and always select those with the highest predicted performance"; therefore, the unqualified individualist would interpret discrimination as "treating unfairly". Alternatively, those who fall under "qualified individualism" reject the use of certain group memberships (such as race, religion, sex) as predictors even if it is scientifically valid to do so; the qualified individualist would thus interpret discrimination as "treating differentially". Finally, those who endorse "quotas" advocate using selection procedures that result in selecting individuals from different groups with proportions equal to their representation in the population of interest; the quotas supporters thus interpret discrimination to mean selecting a "higher proportion of persons from one group than from the other group". Hunter and Schmidt (1976) conclude their discussion by stating that "any purely statistical approach to the problem of test bias is doomed to rather immediate failure", and that the issue of test bias can not be resolved objectively but instead depends on the individual and his or her ethical position for the problem at hand.

Our position on test bias here is probably closest to that of "unqualified individualism". We believe that it is important to distinguish between calling a measurement instrument "biased" because it finds group differences, and calling an instrument "biased" because its use leads to some misrepresentation of one group relative to another. Bias here will thus refer to group misrepresentation, which depends on how a measurement instrument is ultimately used but not necessarily on the measurement process itself. If one group consistently performs more poorly on a test relative to other groups, this does not necessarily constitute test bias per se. It is perfectly reasonable to expect that one group may truly be inferior on the construct measured by the test. However, if a group of examinees always performs more poorly on a test, and this group does not subsequently exhibit inferior achievement, then the test misrepresents this group and is said to be biased.

The ultimate goals of University admission tests are selection and prediction. In this paper we will focus on the fairness of the selection process by examining two aspects of this process: differential validity and differential prediction. According to Linn (1982), differential validity refers to the degree of similarity of the validity coefficients obtained within the two groups. Differential prediction refers to "the question of whether test scores have the same predictive meaning for members of different groups. This question is generally approached by comparing within-group regression equations" (Linn, 1982, p. 367). It is to be expected that when the regression lines for the two genders are compared, the gender with the higher mean on both the predictor and the criterion will produce a relatively higher regression line. However, since this is the expected pattern of results, although it gives rise to underprediction of the criterion scores for the "better" (or higher-scoring) group, it is not indicative of test bias against this group. In other words, even if the validity of the test was zero, we would expect a higher regression line for the group with the higher predictor and criterion means, and this does not constitute test bias.

To detect bias, as defined by differential prediction, we use the method discussed by Linn (1984), which is related to the definitions given by Darlington (1971). The notation will refer to the score on the predictor admission test (or a linear composite of multiple tests) as X, the criterion university grade as Y, and the group category (male or female in our case) as C.

Darlington (1971) presented these three possible definitions of an unbiased test, based on the assumptions that the test has non-zero validity (i.e., $r_{XY} > 0$), that X and Y are bivariate normal within each group category C, that r_{XY} as well as the standard deviations s_X and s_Y are constant across groups, and given that r_{CY} is positive:

1. $r_{CX} = r_{CY}/r_{XY}$
2. $r_{CX} = r_{CY}$
3. $r_{CX} = r_{CY}r_{XY}$

Note that all the definitions are given in terms of the correlation between group membership and admission test score (r_{CX}), where r_{XY} is the test's validity coefficient and r_{CY} is the correlation between group membership and the criterion, university grades. Figure 1 shows the values of r_{CX} considered to constitute a fair test by each of these three definitions for a fixed value of $r_{CY} = .2$ and various values of test validity r_{XY}.

The first definition states that a test is fair (or unbiased) if, given a certain test score, all examinees with that score have the same probability of achieving a certain grade in university regardless of group membership. Stated another way, $P(Y|X) = P(Y|X,C)$, or $r_{CY \cdot X} = 0$, which implies that $r_{CY} - r_{CX}r_{XY} = 0$ and, therefore, $r_{CX} = r_{CY}/r_{XY}$ (definition 1). This definition has been criticized, however, because it implies that for a fixed value of r_{CY}, a test with lower validity (lower r_{XY}) is allowed to correlate more highly with group membership and still be considered fair.

The second definition implies that a test is fair if the proportion of examinees selected using the test (X) is the same as the proportion of examinees that would be selected if one had available the university grades (Y).

The third definition states that given university grades (Y), group membership has no effect on test score for a test to be considered fair. This follows from the same argument as in definition 1, but with $P(X|Y) = P(X|Y,C)$, or $r_{CX \cdot Y} = 0$, implying the reverse causal relationship. In other words, given a certain criterion score, all examinees with that score have the same probability of obtaining a certain test score regardless of group membership.

Linn (1984) showed that one can represent the unbiased model as one in which group membership (C) influences the individual's latent qualifications (Q) directly, and these latent qualifications then influence both the test score (X) and the university grade (Y). However, for the unbiased model to hold, group membership should not influence X or Y directly. Linn goes on to show that if group membership is coded such that there is a positive correlation between group and qualifications, then group membership should be unrelated to both X and Y for a given qualification level. That is, $\beta_{XC \cdot Q} = 0$, and $\beta_{YC \cdot Q} = 0$ for an unbiased test. Since the qualifications are latent (unobserved), it can also be shown that these conditions imply that $\beta_{YC \cdot X} = 0$; however, this will only hold for unrealistic special cases. Linn (1984) then explains that if one group is more highly qualified, its members will usually have a higher regression line than members of the other group when regressing Y on X, and thus the predicted scores for members of the more highly qualified group will be underpredicted by the overall regression line; however, this finding does not imply bias against this more qualified group. In order to detect bias, Linn presents two boundary conditions on the regression coefficients that imply clear bias:

1. $\beta_{YC \cdot X} < 0$, or
2. $\beta_{XC \cdot Y} < 0$.

Furthermore, these conditions can be translated into boundary conditions on the correlation between group membership and criterion score (Y), such that for an unbiased test $\rho_{CX}\rho_{XY} < \rho_{CY} < \rho_{CX}/\rho_{XY}$. Note that this is equivalent to $\rho_{CY}\rho_{XY} < \rho_{CX} < \rho_{CY}/\rho_{XY}$, which can be directly related to Darlington's first and third definitions; that is, when the r_{CX} value falls between the two lines representing definitions 1 and 3 in Figure 1 (i.e., between the top and bottom lines) then the test may be considered unbiased, otherwise there is clear bias. We will use Linn's boundary conditions to define test bias. The first boundary condition is used to detect bias against the lower ability group and the second condition is used to detect bias in favor of the lower ability group.

Had we a measure with a perfectly valid and reliable scores, all definitions of bias would overlap. The problem is that we do not have perfect measurements. Furthermore, if only mean performance level of first year studies was important, Darlington's (1971) first definition of bias alone could be used. However, we believe that the shortcomings of this definition, which arise from imperfect measurement, force the use of another aspect of bias detection – in our case, the third definition. The integration of the first and third definitions, given by Linn's (1984) boundary conditions, seems more ethical to both examiners and the general public.

The Study

The goal of this study was to determine whether the measures used in the admission of students to Universities in Israel are gender biased, and, if so, whether the bias was more often against males or females. The criterion used to measure bias was performance in the first year of University study, and bias was defined according to the boundary conditions given in Linn (1984) and described above. A detailed description of the method follows.

Method

Predictors

Six predictors, briefly described here, were used (for a more detailed description see Gafni & Bronner, 1998):

1. High school graduate certificate (Bagrut) score. In Israel, most high school graduates receive a matriculation certificate called “Bagrut”, which is based on a combination of high school grades and scores on national tests in various general high school subjects. The score on the Bagrut constitutes one’s high school graduation “score”.
2. Psychometric Entrance Test (PET) total score. The PET is designed to measure various cognitive and scholastic abilities with the purpose of providing a good estimate of success in future studies. The PET is similar to the SAT in the United States, and measures aspects of developed ability. It includes three multiple-choice subtests, which are discussed below.
3. Admission score. This is generally a weighted average of the Psychometric Entrance Test (PET) score and the Bagrut score.
4. Verbal reasoning subtest of PET. This section of the PET includes 60 items focusing on the verbal skills and abilities needed for academic studies: the ability to analyze and understand complex written material, the ability to think systematically and logically, and the ability to perceive fine distinctions in meaning among words and concepts. The verbal sections include items such as synonyms and antonyms, analogies, sentence completions, logic, and reading comprehension.

5. Quantitative reasoning subtest of PET. The quantitative section of the PET includes 50 items focusing on the ability to use numbers and mathematical concepts (algebraic and geometrical), to solve quantitative problems, and the ability to analyze information presented in the form of graphs, tables, and charts. Solving problems in this area requires only basic knowledge of mathematics – the math level acquired in the ninth or tenth grades in most high schools in Israel. Formulae and explanations of mathematical terms that may be needed in the course of the exam appear in the test booklet.
6. English as a foreign language subtest of PET. This section of the PET includes 54 items designed to test command of the English language (reading and understanding texts) at an academic level. The English subtest contains three types of items: sentence completions, restatements, and reading comprehension. This subtest serves a dual purpose: it is a component of the PET total score, and it is also used for placement of students in remedial English classes.

Criterion

The criterion investigated here is the Grade Point Average in the first year of University studies (FGPA). This score was measured on a scale of 0 to 100.

Sample

The sample consisted of 61885 Hebrew-speakers who had scores on the PET and the Bagrut as well as a reported first year GPA from their respective universities, and who began their studies between 1991 and 1995. A total of six Israeli universities were included, and within each university there were several general areas of study, each consisting of several departments. The analyses were performed only for departments that contained at least 5 men and 5 women (within each year cohort and university). The departments were clustered into areas of study based on content or administrative considerations; for example, the Social Sciences – verbal area of study consists of departments such as Sociology, Political Sciences, International Relations, Psychology and Education, while the Social Sciences – quantitative area of study consists of departments such as Economics and Business Administration (Gafni & Bronner, 1998). The numbers of men and women in the sample, by area of study, are given in Table 1. University was not considered as a variable in this study because it has been found that the universities in Israel do not differ much in their selectivity (Kennet-Cohen, Bronner, and Oren, 1995).

Procedure

Unit of Analysis

A total of 685 departments were used in the analyses (see Table 1). Analyses were carried out for each department separately, although the results are reported by area of study.

Descriptive statistics

Means and standard deviations on the criterion and predictors were computed. Effect size (d) was computed, according to Cohen (1988), as the difference between the means for the two genders divided by a pooled estimate of the standard deviation. In addition, the Pearson product-moment correlation between the criterion and each predictor (i.e., test validity) was computed for the two gender groups separately. Because the validity coefficients are computed on the selected group of examinees only, it is standard practice to correct the validity coefficients for range restriction to estimate the validity in the population (e.g., Allen & Yen, 1979, pp. 196-200). However, since we do not in this case have information on the unselected examinees (who did not attend university), the validity coefficients are uncorrected and, therefore, probably underestimate the true validity coefficients. All statistics reported at the area of study level are the weighted averages of the statistics computed for each department, weighted by the total number of students in each department.

Differential prediction detection

Each of the predictors, in turn, as well as a gender variable were regressed on the criterion, first year GPA. Bias was considered to occur only when the boundary conditions given by Linn (1984) were satisfied. The two conditions were examined for each unit of analysis. For each unit the gender group with the higher predictor mean value was coded 1 and the other 0. Thus for the cases in which the mean score on X was higher for males, C=1 for males and C=0 for females, and the following four regression models were fit to the data set (Lautenschlager & Mendoza, 1986):

1. $Y = b_{10} + b_{11}X + e$

2. $Y = b_{20} + b_{21}X + b_{22}C + e$

3. $Y = b_{30} + b_{31}X + b_{32}XC + e$

4. $Y = b_{40} + b_{41}X + b_{42}C + b_{43}XC + e$

The null hypothesis $\underline{H_{01}}$: $b_{42} = b_{43} = 0$ was tested by using the usual F test to compare the relative decrease in R^2 between Model 4 and Model 1. If $\underline{H_{01}}$ was not rejected, then the null hypothesis, which implies no bias, would be retained. However, if $\underline{H_{01}}$ was rejected, then some predictive test bias is to be suspected. If this was the case, Model 2 was compared to Model 4 (a test of $\underline{H_{02}}$: $b_{43} = 0$). Rejection of $\underline{H_{02}}$ indicates that there are slope differences between the genders. If the slopes were found to be significantly different, then Model 3 was compared to Model 4 in order to test the equality of the intercepts ($\underline{H_{03}}$: $b_{42} = 0$); otherwise, if the slopes were not found to be different, Model 2 was compared to Model 1 in order to test the equality of the intercepts ($\underline{H_{04}}$: $b_{22} = 0$). Therefore, bias was indicated if differences in intercept, slope or both were found between the two genders. Figure 2 depicts these model comparisons schematically. The predictor was male mean-centered so that comparison of the intercepts reflected the difference between the regression equations at the center of the distribution. The significance level used for each test was .05, so if bias was detected then the type I error rates were quite liberal (i.e., higher than .05). Given that X was higher for males in this discussion, if the regression line for the females was found to be higher than that for the males then this was taken to indicate bias against females according to Linn's first boundary condition. The same procedure was followed in regressing X on Y, such that if bias was detected and the regression line for males was found to be higher then this was indicative of bias against females according to Linn's second boundary condition. The analysis in each unit was described in terms of bias in favor of females, bias against females, or undecided.

Results

A summary of the sample sizes of men, women, and the number of departments within each area of study are presented in Table 1. The means and standard deviations for males and females on each variable are presented in Table 2. The effect size (d) was computed as the difference between the male mean and the female mean, standardized by a weighted average of the standard deviations, where the weights are the male and female sample sizes (see Table 3). The value of the effect size was computed for each department, and then averaged over all departments (weighted by total number of students). Note that in our case positive values of d indicate a higher mean for males, while negative values of d indicate that females score higher on average. Absolute values of d that fall between .20 and .49 are considered to be "small differences", values between .50 and .79 are considered "medium differences", and values of .80 or larger are considered to reflect "large differences" (Cohen, 1988). Very few mean differences here reach the "medium" level, and many do not even reach the "small" mean difference level. However, overall it does appear that women score higher on the Bagrut while males score higher on the PET, and in particular on the quantitative subtest of the PET. Note also that the effect sizes for both the Admission score and first year GPA are very close to zero, indicating no mean difference between the two genders.

Table 4 shows the test validity for each variable as a predictor or first year GPA. All validity values are reasonable (especially in the absence of a correction for restriction of range), with the possible exception of the English subtest of the PET, which has relatively low validity. In general the validity coefficients for the two gender group are highly similar.

Table 5 presents the counts of significant cases of bias detected within each area of study. The pattern of results is similar to the results for effect size – there are more cases of bias in favor of women when using the Bagrut as a predictor of first year GPA; on the other hand, there are more cases of bias against women when using the PET, particularly the quantitative subtest of the PET, as a predictor of first year GPA. When combined into the Admission score, the percentage of cases of bias either against (0.7%) or in favor of (1.3%) women is almost negligible.

Discussion

The purpose of this study was to examine gender differences in scholastic performance and standardized test scores, as well as to investigate whether the gender differences in these measures might translate into gender bias when used to predict first year university performance in Israel. The specific measures investigated consisted of first year GPA as the criterion, as well as six predictors: an Admission score (combining standardized test and high school scores), Bagrut (the high school matriculation score), the PET (standardized test) score, and three PET subtest scores (Verbal, Quantitative and English).

In terms of average performance differences, as measured by effect size, the findings here support the general pattern reported by Willingham and Cole (1997) of higher average performance by males on standardized tests and higher average performance by females on school grades. The findings also show that within the standardized test scores the male advantage is much more prominent on the quantitative than on the verbal sections of the test. In general, there is greater variability in the test scores for the male population relative to the female population, and this has been argued to statistically result in higher mean test scores for males in self-selected samples (Willingham and Cole, 1997).

This study examined the possibility that higher PET scores for males constitute test bias against females as reflected in first year university GPA in Israel, as well as any further consequences of this phenomenon on the Admission score (which consists in part of the PET score) used for selection.

No difference in validity coefficients was found between the two groups to indicate differential validity. In terms of differential prediction, it is important to first note that little bias was detected in general, and even the highest percentage of significantly biased departments (found using Bagrut as a predictor in the Social sciences – quantitative area of study) was about 17% (10 of 59 departments) and was in favor of women. It is also interesting to note that in study areas which would traditionally be thought of as more “male oriented”, when bias was detected it was more often in favor of women than against women (e.g., using Bagrut and, to a lesser extent, the Admission score in Social sciences – quantitative, Math, Statistics and Computer Science, Biological and Physical Sciences, Engineering and Architecture). On the other hand, in study areas which are traditionally considered to be women’s strengths (such as

Arts and Humanities and Social sciences – verbal), when bias was detected it was more often against women than in their favor (using the PET or its subtest scores). This is most likely the result of "self-selection", in the sense that, for example, women that select a traditionally "male oriented" field as their major may be more likely to perform better on the verbal predictors, and not necessarily excel on the more relevant (to the criterion) quantitative part of the predictor scores. Similarly, it is more likely that males who are admitted to the verbal faculties are relatively high on the quantitative part of the predictor. This ability is less relevant for achievements in this field. Therefore, it looks like their performance is overpredicted.

It is interesting to note that when bias was detected in favor of women it was predominantly for Bagrut as the predictor and never for the PET or any of its subtests. On the other hand, bias against women was predominantly detected using the PET and especially its quantitative subtest as predictors. It should be mentioned here that the most important variable for examination of fairness is the Admission score, because this is the score that is actually used for selection, and that the examination of separate components of this score (i.e., PET and Bagrut) is not entirely appropriate; we do so to better understand the ways by which each component affects the selection process but with an awareness of its limitation. In addition, the problem of different power across programs exists, as some departments have larger numbers of students than others. Therefore, the results for departments in which there are very few students should be interpreted with caution, while results for departments with large numbers of students may be regarded with more confidence. It might also be worth noting, however, that with a .05 significance level one would expect that in conducting one test for each department (.05)(685)=34.25 of the tests would be significant by chance. With one exception (a total of 40 departments show bias in favor of women using Bagrut as a predictor) this threshold was not reached. Therefore, the results of this study confirm that the Admission score is arguably the best (most valid) and least biased predictor to use in order to predict first year university GPA.

Any predictor of future academic performance is open to objections based on charges of bias, and any measure of bias is imperfect and open to criticism. As Linn (1984) wrote,

> "perfectly reliable and valid measurement of qualifications is not feasible. Hence it is impossible to select only those who rank highest in terms of the unobserved qualifications and always treat equally those with equal qualifications. This is so

> because equally qualified persons will not always have equal observed scores on the test or other indicator variables, and the less qualified will sometimes have higher scores than the better qualified…It should be emphasized, however, that the procedures may produce highly efficient results that are less unfair than practical alternatives." (p. 38).

Although Linn's definition of bias may be on the conservative side in that it allows only extreme cases of bias to be recognized as such, it nonetheless provides a good indication of those cases which are clearly biased. The trend of biased cases here was found to be in agreement with previous knowledge regarding the general differences in test performance and scholastic aptitude between men and women.

In conclusion, it would appear that in general there is little gender bias in the Admission score that is ultimately used to admit students to universities in Israel. Moreover, when bias was found it was found both in favor of women and against them, but there was no consistent or disturbing trend when using the Admission score as the predictor and first year GPA as the criterion. Though the issue of test bias often involves considerations to both the intended use of the test and its possible consequences (Cole & Moss, 1989), the bias analyses presented in this paper focus on the meaning, use and immediate purpose of the test scores (i.e., university admission and success), and do not address the construct validity of the test or the consequences of its use (for example, the subsequent differences in gender representation across areas of study).

Finally, it should be noted that the culture of Hebrew-speaking examinees in Israel is similar to that of American students (this is not necessarily the case for Arabic- and Russian-speaking students in Israel, but we included only Hebrew-speaking examinees in our analyses). There has been very little research on the direct effect of culture on gender differences in academic performance (cf. Beller & Gafni, 1996; Feingold, 1994), but what is known about gender differences in the United States appears to hold in Israel also; this points to the possibility that Western culture gives rise to similar gender differences. Therefore, these test bias results could very well generalize to the United States as well as other Western cultures, and this possibility could be examined by future studies investigating similar variables in other Western cultures.

References

Allen, M. J., & Yen, W. M. (1979). Introduction to Measurement Theory. Monterey, California: Brooks/Cole.

Beller, M., & Gafni, N. (1996). The 1991 International Assessment of Educational Progress in mathematics and sciences: The gender differences perspective. Journal of Educational Psychology, 88, 365-377.

Cleary, T. A. (1992). Gender differences in aptitude and achievement test scores. In Sex equity in educational opportunity, achievement, and testing: Proceedings of the 1991 ETS Invitational Conference (pp. 51-90). Princeton, NJ: Educational Testing Service.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cole, N. S. & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 201-219). New York: American Council on Education & Macmillan.

Darlington, R. B. (1971). Another look at "cultural fairness". Journal of Educational Measurement, 8, 71-82.

Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. Review of Educational Research, 62, 61-84.

Feingold, A. (1994). Gender differences in variability in intellectual abilities: A cross-cultural perspective. Sex Roles, 30, 81-92.

Gafni, N., & Bronner, S. (1998). An examination of criterion-related bias for Hebrew- and Russian-speaking examinees in Israel (Report No. 244). Jerusalem, Israel: National Institute for Testing and Evaluation.

Halpern, D. F. (1992). Sex differences in cognitive abilities (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. Science, 269, 41-45.

Hunter, J. E., & Schmidt, F. L. (1976). Critical analysis of the statistical and ethical implications of various definitions of test bias. Psychological Bulletin, 83, 1053-1071.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. Psychological Bulletin, 107, 139-155.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. Psychological Bulletin, 104, 53-69.

Kennet-Cohen, T., Bronner, S., & Oren, C. (1995). A meta-analysis of the predictive validity of the selection process to Universities in Israel (Report No. 202). Jerusalem, Israel: National Institute for Testing and Evaluation.

Lautenschlager, G. J., & Mendoza, J. L. (1986). A step-down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. Applied Psychological Measurement, 10, 133-139.

Linn, R. L. (1984). Selection bias: Multiple meanings. Journal of Educational Measurement, 21, 33-47.

Linn, R. L. (1982). Ability testing: Individual differences, prediction, and differential prediction. In A. K. Wigdor and W. R. Garner (Eds.), Ability testing: Uses, consequences, and controversies. Washington, D.C.: National Academy Press.

Maccoby, E. E., & Jacklin, C. N. (1974). The psychology of sex differences. Stanford, CA: Stanford University.

Stanley, J. C., Benbow, C. P., Brody, L. E., Dauber, S. & Lupkowski, A. E. (1992). Gender differences on eighty-six nationally standardized aptitude and achievement tests. In N. Colangelo, S. G. Assouline, & D. L. Ambroson (Eds.), Talent development: Proceedings from the 1991 Henry B. and Jocelyn Wallace National Research Symposium on Talent Development (pp. 41-48). Unionville, NY: Trillium.

Willingham, W. W., & Cole, N. S. (1997). Gender and fair assessment. Mahwah, New Jersey: Lawrenece Erlbaum Associates.

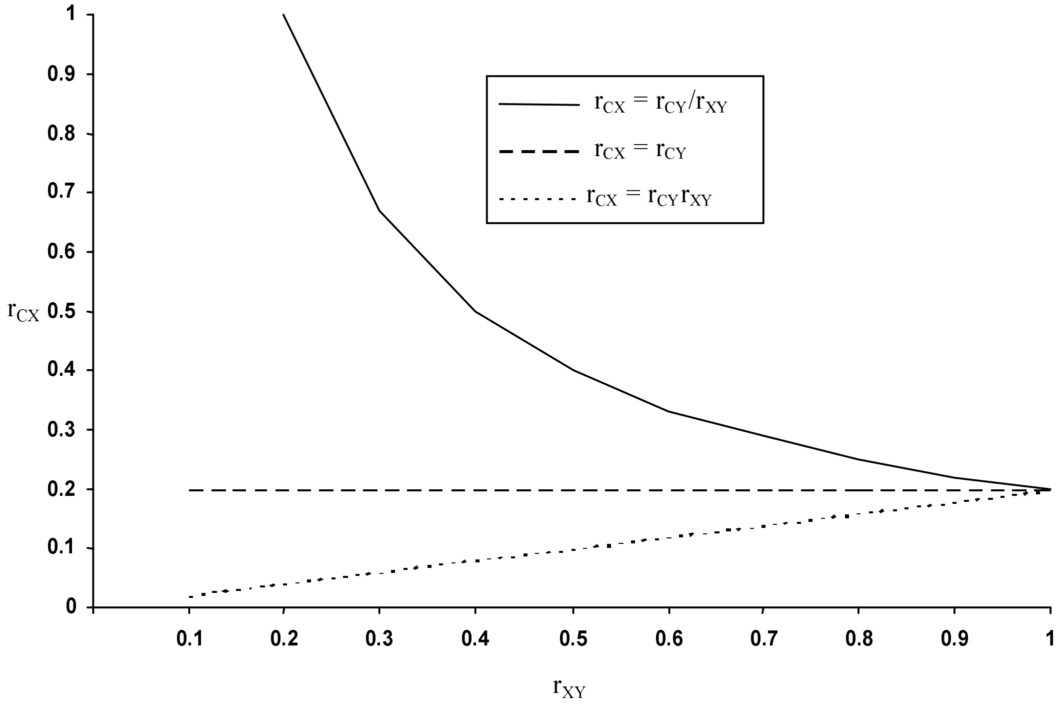
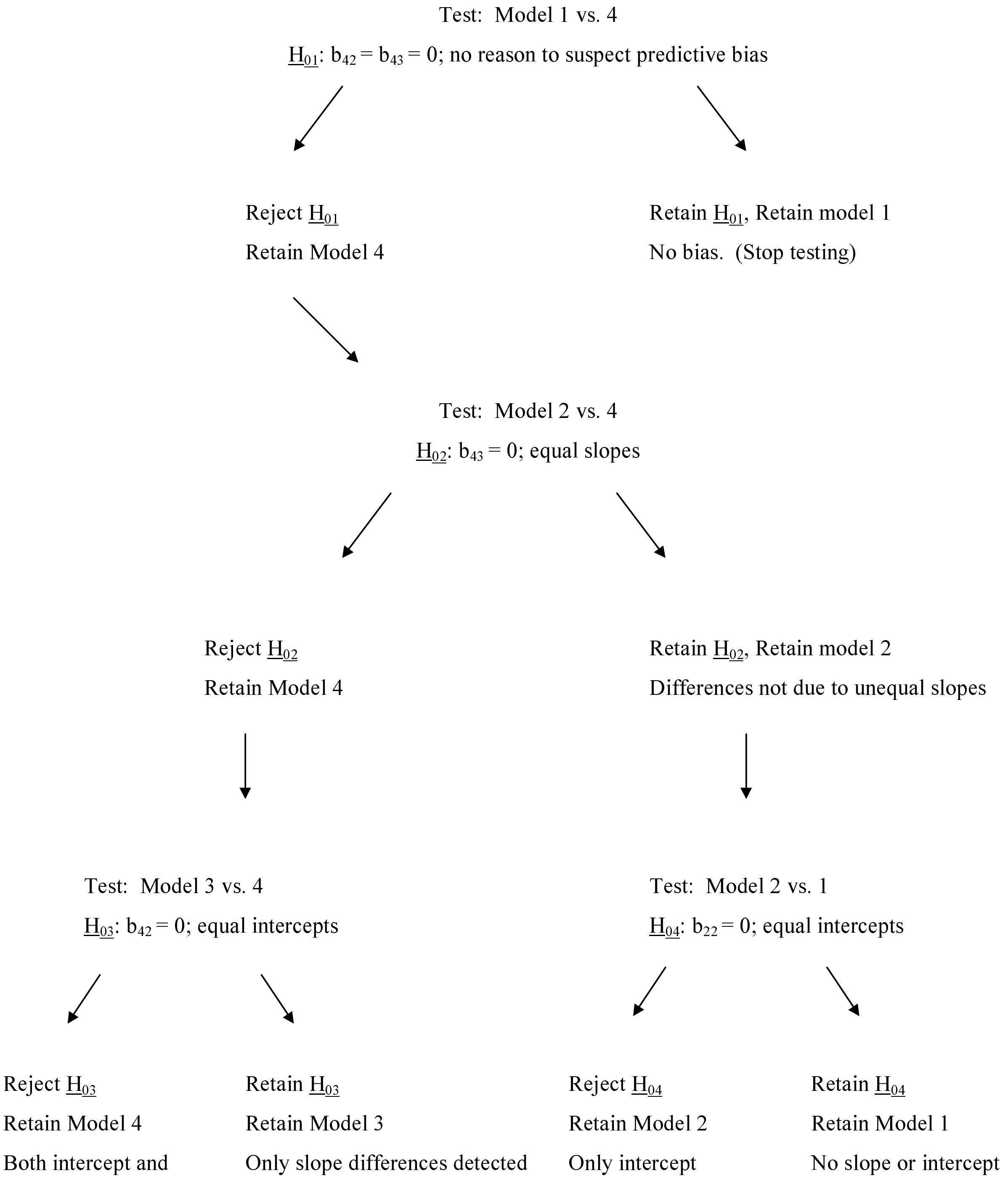


Figure 1: Values of r_{CX} considered to constitute a fair (or unbiased) test by each of Darlington's (1971) three definitions for a fixed value of $r_{CY} = .2$ and various values of test validity r_{XY}. Note: Adapted from "Another look at "cultural fairness"" by R. B. Darlington, 1971, Journal of Educational Measurement, 8, p. 74. Copyright 1971, Journal of Educational Measurement, National Council on Measurement in Education. Adapted with permission.

Models (Y is the criterion, X is the predictor, and C is group membership):

1. $Y = b_{10} + b_{11}X + e$
2. $Y = b_{20} + b_{21}X + b_{22}C + e$
3. $Y = b_{30} + b_{31}X + b_{32}XC + e$
4. $Y = b_{40} + b_{41}X + b_{42}C + b_{43}XC + e$



Test: Model 1 vs. 4

$\underline{H_{01}}$: $b_{42} = b_{43} = 0$; no reason to suspect predictive bias

Reject $\underline{H_{01}}$

Retain Model 4

Retain $\underline{H_{01}}$, Retain model 1

No bias. (Stop testing)

Test: Model 2 vs. 4

$\underline{H_{02}}$: $b_{43} = 0$; equal slopes

Reject $\underline{H_{02}}$

Retain Model 4

Retain $\underline{H_{02}}$, Retain model 2

Differences not due to unequal slopes

Test: Model 3 vs. 4

$\underline{H_{03}}$: $b_{42} = 0$; equal intercepts

Test: Model 2 vs. 1

$\underline{H_{04}}$: $b_{22} = 0$; equal intercepts

Reject $\underline{H_{03}}$

Retain Model 4

Both intercept and

Retain $\underline{H_{03}}$

Retain Model 3

Only slope differences detected

Reject $\underline{H_{04}}$

Retain Model 2

Only intercept

Retain $\underline{H_{04}}$

Retain Model 1

No slope or intercept

slope differences detected.	differences detected.	differences detected: unlikely outcome.

Figure 2: Schematic representation of the regression model comparison procedure for detecting predictive bias (Lautenschlager & Mendoza, 1986). Note: Adapted from "A Step-Down Hierarchical Multiple Regression Analysis for Examining Hypotheses About Test Bias in Prediction" by G. J. Lautenschlager and J. L. Mendoza, 1986, Applied Psychological Measurement, 10, p. 136. Copyright 1986, Applied Psychological Measurement Inc. Adapted with permission.

Table 1

Frequencies (percentages) of men, women and departments, by area of study, in the sample of Hebrew-speakers.

Area of study	Men		Women		Departments	Total students
Arts and Humanities	3867	(31.72%)	8324	(68.28%)	175	12191
Social Sciences -- Verbal	5528	(28.01%)	14206	(71.99%)	158	19734
Law	2265	(51.70%)	2116	(48.30%)	20	4381
Social Sciences -- Quantitative	4811	(58.10%)	3469	(41.90%)	59	8280
Biological sciences	1210	(35.89%)	2161	(64.11%)	46	3371
Physical sciences	864	(54.93%)	709	(45.07%)	45	1573
Math, Statistics and Computer science	2952	(66.01%)	1520	(33.99%)	57	4472
Engineering and Architecture	3852	(71.75%)	1517	(28.25%)	80	5369
Medicine	803	(46.93%)	908	(53.07%)	32	1711
Nursing	126	(15.69%)	677	(84.31%)	13	803
Total:	26278	(42.46%)	35607	(57.54%)	685	61885

Note. Percentages represent the proportion of men and women out of the total number of students.

Table 1a

Frequencies as a percentage of the total over faculties.

Area of study	**Men**	**Women**	**Departments**
Arts and Humanities	14.72%	23.38%	25.55%
Social Sciences -- Verbal	21.04%	39.90%	23.07%
Law	8.62%	5.94%	2.92%
Social Sciences -- Quantitative	18.31%	9.74%	8.61%
Biological sciences	4.60%	6.07%	6.72%
Physical sciences	3.29%	1.99%	6.57%
Math, Statistics and Computer science	11.23%	4.27%	8.32%
Engineering and Architecture	14.66%	4.26%	11.68%
Medicine	3.06%	2.55%	4.67%
Nursing	0.48%	1.90%	1.90%
Total:	100.00%	100.00%	100.00%

Table 2

Male and female means (and standard deviations) for each variable, by area of study.

Area of study	**Sex**	**FGPA**	**Admission**	**Bagrut**	**PET**	**Verbal**	**Quant**	**English**
			score		**score**			
Arts and Humanities	Males	81.19	-0.42	86.97	584.0	116.1	111.4	118.50
		(8.48)	(1.77)	(8.38)	(73.28)	(14.78)	(15.92)	(17.19)
	Females	80.56	-0.71	88.32	551.8	111.9	103.9	113.5
		(7.95)	(1.60)	(7.54)	(66.39)	(14.03)	(14.59)	(16.16)
Social Sciences - Verbal	Males	81.78	-0.12	88.23	594.90	117.94	115.7	116.10
		(7.01)	(1.32)	(7.09)	(58.16)	(12.50)	(13.37)	(16.66)
	Females	82.94	-0.00	91.09	578.8	116.21	110.81	114.8
		(6.42)	(1.19)	(6.15)	(52.68)	(11.85)	(12.47)	(14.84)
Law	Males	79.55	2.78	100.9	695.2	133.7	135.4	133.5
		(5.83)	(0.96)	(5.70)	(40.28)	(9.07)	(9.32)	(11.89)
	Females	80.04	2.84	103.01	681.0	132.0	131.24	132.91
		(5.01)	(0.83)	(4.78)	(37.78)	(9.10)	(9.51)	(10.33)
Social Sciences - Quantitative	Males	78.60	1.68	95.00	666.24	126.94	132.74	126.9
		(9.84)	(0.98)	(6.30)	(40.00)	(10.10)	(9.05)	(13.06)
	Females	78.03	1.78	97.70	650.00	125.16	128.80	123.90
		(9.48)	(0.83)	(5.04)	(37.82)	(9.94)	(8.97)	(12.33)
Biological Sciences	Males	79.09	0.61	90.25	629.6	122.3	122.5	124.3
		(11.28)	(1.11)	(6.10)	(49.35)	(11.76)	(11.51)	(14.45)
	Females	78.51	0.68	93.19	609.65	120.0	118.50	119.12
		(9.91)	(1.14)	(5.70)	(50.62)	(11.95)	(11.08)	(14.96)
Physical Sciences	Males	76.78	1.22	93.65	644.9	123.10	127.8	125.3
		(12.80)	(1.22)	(7.36)	(56.05)	(13.30)	(10.72)	(15.51)
	Females	76.58	1.31	96.29	628.9	118.65	122.63	117.50
		(13.68)	(1.40)	(7.13)	(54.74)	(13.20)	(10.68)	(14.99)
Math, Statistics and Computer Science	Males	74.48	1.54	95.14	655.1	123.7	131.7	125.5
		(14.23)	(1.21)	(6.66)	(50.30)	(12.17)	(9.59)	(14.69)
	Females	72.78	1.49	96.98	635.8	121.4	128.3	119.9
		(14.00)	(1.18)	(6.24)	(50.77)	(12.49)	(9.46)	(15.06)
Engineering and Architecture	Males	78.12	1.75	96.61	657.4	123.70	132.5	125.8
		(8.57)	(1.06)	(6.21)	(43.80)	(11.26)	(8.83)	(13.57)
	Females	77.44	1.90	98.34	653.2	124.2	130.9	124.5

		(7.94)	(0.97)	(5.43)	(43.59)	(11.32)	(8.63)	(13.57)
Medicine	Males	84.54	2.90	101.3	700.0	132.9	136.8	136.5
		(6.68)	(0.73)	(5.67)	(27.57)	(7.96)	(7.54)	(8.93)
	Females	84.55	3.10	103.6	694.9	133.5	134.7	134.9
		(6.13)	(0.60)	(4.22)	(28.10)	(8.09)	(7.65)	(9.18)
Nursing	Males	80.89	-0.23	88.3	586.1	113.6	114.4	119.4
		(5.19)	(0.94)	(6.45)	(44.35)	(10.64)	(13.13)	(13.27)
	Females	82.31	-0.06	90.40	580.7	114.7	112.9	115.6
		(4.90)	(0.98)	(5.56)	(43.27)	(11.03)	(10.69)	(13.49)
Overall	**Males**	**80.03**	**0.71**	**91.62**	**625.1**	**121.5**	**122.4**	**122.0**
		(8.88)	**(1.31)**	**(6.99)**	**(55.43)**	**(12.22)**	**(12.30)**	**(15.22)**
	Females	**80.03**	**0.73**	**93.91**	**607.0**	**119.5**	**117.8**	**119.2**
		(8.39)	**(1.19)**	**(6.13)**	**(51.56)**	**(11.86)**	**(11.58)**	**(14.30)**

Table 3

d-values for the difference between male and female means for each variable, by area of study.

Area of study	**FGPA**	**Admission score**	**Bagrut**	**PET score**	**Verbal**	**Quant**	**English**
Arts and Humanities	0.09	0.16	-0.18	0.47	0.29	0.51	0.31
Social Sciences - Verbal	-0.19	-0.09	-0.46	0.32	0.14	0.39	0.10
Law	-0.09	-0.05	-0.42	0.43	0.20	0.49	0.09
Social Sciences - Quantitative	0.06	-0.11	-0.47	0.43	0.18	0.44	0.24
Biological Sciences	0.06	-0.07	-0.51	0.40	0.19	0.35	0.36
Physical Sciences	-0.00	-0.06	-0.39	0.31	0.15	0.26	0.35
Math, Statistics and Computer Science	0.13	0.03	-0.30	0.39	0.18	0.37	0.40
Engineering and Architecture	0.09	-0.14	-0.30	0.10	-0.04	0.19	0.11
Medicine	0.03	-0.26	-0.43	0.19	-0.08	0.27	0.18
Nursing	-0.29	-0.23	-0.46	0.15	-0.07	0.13	0.30
Overall	**-0.02**	**-0.04**	**-0.38**	**0.36**	**0.16**	**0.40**	**0.21**

Table 4

Validities: Pearson product-moment correlations with First year GPA (FGPA).

Area of study	Sex	Admission score	Bagrut	PET score	Verbal	Quant	English
Arts and Humanities	Males	0.44	0.39	0.38	0.37	0.28	0.30
	Females	0.46	0.42	0.39	0.34	0.31	0.27
Social Sciences --Verbal	Males	0.28	0.21	0.23	0.19	0.19	0.08
	Females	0.32	0.27	0.24	0.20	0.20	0.10
Law	Males	0.37	0.31	0.22	0.20	0.15	0.08
	Females	0.31	0.26	0.16	0.14	0.11	0.05
Social Sciences -- Quantitative	Males	0.31	0.27	0.17	0.08	0.24	0.01
	Females	0.30	0.26	0.17	0.07	0.27	-0.04
Biological sciences	Males	0.33	0.29	0.21	0.10	0.28	0.04
	Females	0.39	0.37	0.27	0.16	0.34	0.07
Physical sciences	Males	0.50	0.45	0.37	0.29	0.37	0.18
	Females	0.45	0.43	0.34	0.25	0.33	0.18
Math, Statistics and Computer science	Males	0.39	0.33	0.30	0.21	0.31	0.14
	Females	0.41	0.37	0.29	0.19	0.31	0.16
Engineering and Architecture	Males	0.38	0.35	0.22	0.11	0.28	0.08
	Females	0.42	0.40	0.24	0.14	0.28	0.11

Medicine	Males	0.30	0.29	0.02	-0.06	0.11	-0.03
	Females	0.28	0.29	0.06	0.01	0.11	-0.04
Nursing	Males	0.19	0.08	0.23	0.28	0.03	0.09
	Females	0.40	0.35	0.26	0.18	0.23	0.08
Overall	**Males**	**0.35**	**0.29**	**0.25**	**0.20**	**0.23**	**0.12**
	Females	**0.37**	**0.33**	**0.26**	**0.19**	**0.25**	**0.11**

Table 5

Frequency of significant cases of test bias.

Area of study	Predictor	Bias Against women	Bias in Favor of women	Number of Departments	Number of subjects
Arts and Humanities	Admission score	1	2	175	12191
	Bagrut	0	5	175	12191
	PET	4	0	175	12191
	Verbal	2	0	175	12191
	Quantitative	3	0	175	12191
	English	2	0	175	12191
Social Sciences -- Verbal	Admission score	4	0	158	19734
	Bagrut	1	6	158	19734
	PET	14	0	158	19734
	Verbal	4	0	158	19734
	Quantitative	15	0	158	19734
	English	2	0	158	19734
Law	Admission score	0	0	20	4381
	Bagrut	0	3	20	4381
	PET	3	0	20	4381
	Verbal	2	0	20	4381
	Quantitative	3	0	20	4381
	English	0	0	20	4381
Social Sciences -- Quantitative	Admission score	0	3	59	8280
	Bagrut	0	10	59	8280
	PET	1	0	59	8280
	Verbal	0	0	59	8280
	Quantitative	2	0	59	8280
	English	0	0	59	8280
Biological Sciences	Admission score	0	1	46	3371
	Bagrut	0	6	46	3371
	PET	0	0	46	3371
	Verbal	0	0	46	3371
	Quantitative	0	0	46	3371
	English	0	0	46	3371
Physical Sciences	Admission score	0	0	45	1573
	Bagrut	0	1	45	1573

	PET	0	0	45	1573
	Verbal	0	0	45	1573
	Quantitative	0	0	45	1573
	English	0	0	45	1573
Math, Statistics and Computer science	Admission score	0	2	57	4472
	Bagrut	0	4	57	4472
	PET	1	0	57	4472
	Verbal	0	0	57	4472
	Quantitative	1	0	57	4472
	English	0	0	57	4472
Engineering and Architecture	Admission score	0	0	80	5369
	Bagrut	0	4	80	5369
	PET	0	0	80	5369
	Verbal	0	0	80	5369
	Quantitative	0	0	80	5369
	English	0	0	80	5369
Medicine	Admission score	0	1	32	1711
	Bagrut	0	1	32	1711
	PET	0	0	32	1711
	Verbal	0	0	32	1711
	Quantitative	0	0	32	1711
	English	0	0	32	1711
Nursing	Admission score	0	0	13	803
	Bagrut	0	0	13	803
	PET	0	0	13	803
	Verbal	0	0	13	803
	Quantitative	0	0	13	803
	English	0	0	13	803
Overall	**Admission score**	**5**	**9**	**685**	**61885**
	Bagrut	**1**	**40**	**685**	**61885**
	PET	**23**	**0**	**685**	**61885**
	Verbal	**8**	**0**	**685**	**61885**
	Quantitative	**24**	**0**	**685**	**61885**
	English	**4**	**0**	**685**	**61885**