A Cross-Cultural Perspective on Gender Differences in Higher Education: Admissions and Scholastic Achievement

Naomi Gafni, The National Institute for Testing and Evaluation, Israel
Michal Beller, The Open University of Israel
Shmuel Bronner, The National Institute for Testing and Evaluation, Israel

Naomi Gafni
National Institute for Testing & Evaluation
P.O.B. 26015
Jerusalem 91260
Israel
Tel: 972-2-6759506
Fax: 972-2-6759543
e-mail: naomi@nite.org.il

A Cross-Cultural Perspective on Gender Differences in Higher Education: Admissions and Scholastic Achievement

Abstract

This study investigated the extent to which results exhibiting gender differences on scholastic measures could be generalized across cultures. The analysis was conducted on twelve cohorts of Hebrew-, Arabic-, and Russian-speaking students in Israel's six universities. Gender differences in admission measures and in academic performance, as well as in the validity and fairness of the admissions process were examined. Across all ethnic groups, high school grades were higher for females, while males performed better on admissions tests. Different patterns of gender differences were found for each ethnic group. Males' advantage on admissions tests among the Hebrew-speaking examinees was similar to that found for the US and Sweden. As in the US, validity coefficients were slightly higher for females in all ethnic groups. It was concluded that gender differences reflect the social and cultural values of the society within which they occur.

Introduction

Comparing patterns of gender differences on tests and scholastic achievements within various ethnic groups provides a deeper understanding of gender as well as ethnic differences. As Willingham, Cole, Lewis, and Leung (1997) maintain, an examination of possible ethnic differences in the test performance of males and females is warranted because such variations can have considerable social and educational significance. To date, the research on gender differences has been conducted mainly across ethnic groups, while the question of whether gender differences may vary for different ethnic groups has been given less attention.

Some of the research in this area was conducted in the US on undergraduate admissions tests such as the SAT and ACT (e.g., Admissions Testing Program, 1992; Advanced Placement Program, 1992; American College Testing Program, 1988). These admissions data, in spite of being based on selective samples, represent the entire population of applicants to higher education. However, minority groups within these samples do not necessarily undergo the same self-selection processes as the majority group and therefore may not represent their ethnic origin and gender groups in the same manner.

The above studies used the ratio of the number of females to males (F/M) within each of the ethnic groups and the standard mean difference in the scores of females and males (D) (Cohen, 1998), with a positive D indicating a higher performance of females. In general, little variation in the pattern of gender differences was found in the above mentioned admissions tests between the different ethnic groups (American Indian, Asian-American, Black, Mexican-American, Other Hispanic, and White). The exception was Black examinees, who had a higher ratio of female to male examinees (F/M), and a standardized difference (D) that was more favorable to female examinees compared to Whites. Whereas the F/M ratio was 1.16 for White test takers, it was 1.40 for Black test takers. Moreover, the D values among

Blacks were 0.0 to 0.16 for the Verbal Reasoning domain and -0.20 to -0.06 for the Mathematical Reasoning domain. The corresponding D values for the Whites were -0.08 to 0.09 for the Verbal Reasoning domain and -0.39 to -0.26 for the Math domain. For Blacks, more women than men are taking admissions tests, and they are performing relatively better than men than is true in other ethnic groups (Willingham, Cole, Lewis, & Lueng, 1997). In addition, Willingham and his colleagues examined gender differences on nine AP examinations in different subject areas within the above groups. They found slight differences in average Ds from one ethnic group to another. The variations in the relative mean performance of females and males on the AP examinations were found to be associated with different types of tests, not with different ethnic groups. An exception was the Asian-American group, for whom the average D differed from that of White students. Relative to men in their group, Asian-American women performed slightly better than White women on each of the nine AP tests. As in the SAT and ACT, Black women were far more likely to take AP examinations than were Black men (by a factor of almost two to one) and at the same time they maintained the same level of performance relative to men.

In examining the GRE General data, the average Ds for Black and Asian-American examinees were some .15 to .20 more positive than those of the White group. In most ethnic groups (except for Blacks), women were not as heavily represented among GRE General test takers (F/M was around 1.00) as they were for the White majority students (F/M = 1.27).

As for the level of achievements in the first year of academic studies (FGPA), the standardized difference (D) was .11 and .18 for Black and White students, respectively (Dwyer and Johnson, 1997). These results were computed across all areas of study and were not controlled for variation in the courses chosen by males and females.

When dealing with the admissions process, it is not sufficient to examine differences on each variable separately. Rather, it is necessary to examine the relationships between the

predictors and the criterion for the various groups being studied. This would provide some evidence as to the predictive validity and fairness of the selection process. This is particularly important in cases where the different samples differ in the extent to which they represent their respective groups. When examining predictive validity for the various ethnic groups, there are two types of prediction errors of concern: First, the regression line of one group may be steeper, indicating that the test is a more effective predictor for that group than for the other group. Second, while the regression lines of the two groups may be parallel, they may differ in their intercept, indicating that the criterion (for example, GPA) tends to be over- (or under-) predicted for one group. These two types of error are respectively termed differential validity and differential prediction (Linn, 1982).

Ramist, Lewis, and McCamley-Jenkins (1994) conducted a detailed analysis of subgroup differences in predicting future college achievement. As a criterion, the study used grades of 46,379 students in 7,786 courses given at 45 institutions in the US. Using the SAT, high school record (HSR), and both measures combined, predictions were made for grades within individual courses, grades by types of courses, and freshman overall GPA. The predictions were analyzed by gender, ethnic and language group, level of academic composite at entry , and within more and less selective institutions. The average correlations with course grades across all colleges were .60 for SAT and .58 for HSR. The difference between the validities of the two predictors was larger for women (.64 and .59 for SAT and HSR, respectively) than for men. The above correlations were corrected for range restriction and unreliability of single course grades. Ramist et al. found that, in general, validity coefficients for all predictors tended to be higher for females than for males. In predicting freshman GPA from SAT and HSR combined, the correlations were .60 and .55 for women and men, respectively. Similar results were found for the ACT (ACT, 1973).

Ramist et al. (1994) found that for the 45 insitutions, the average under-prediction of women's freshman GPA based on the SAT alone was 0.09 (on the GPA scale). Adding HSR to the SAT and correcting for differences in grading stringency in the courses selected by women and men reduced the under-prediction to 0.03. Differential prediction by gender was particularly associated with less selective institutions. In the more selective colleges, women's course grades were under-predicted by .04 with SAT alone and over-predicted by .02 with HSR alone.

In Sweden, fair selection of male and female applicants to institutions of higher education, based on the SweSAT admissions exam, is perceived by the public as a crucial issue requiring research (e.g., Stage, 1994, 1997). It was found that, on average, females obtain higher course grades, whereas males obtain higher scores on standardized tests. Even when the course level was held constant, these differences were not eliminated. As for the SweSAT, for three out of the five sub-tests, the effect sizes were of medium magnitude (D ranging from 0.5 to 0.8) according to Cohen's criteria (1988). In two out of the three – the Data Sufficiency sub-test and the Diagrams, Tables & Maps sub-test (measuring mathematical reasoning) – the effect was in favor of males, while on the Swedish Reading Comprehension sub-test it was in favor of females. The effects for the other two sub-tests (Vocabulary and English Reading Comprehension) were small.

In the present study, gender differences in academic performance as well as performance on admissions measures were examined from a cross-cultural perspective. The admissions procedure to institutions of higher education in Israel can serve as an interesting case study because the applicant population is heterogeneous, consisting of applicants from several ethnic and language groups. Gender differences were analyzed for the following three groups of students: Hebrew-, Arabic-, and Russian-speaking. These groups differ not only in language but also in their cultural, educational and socio-economic background.

The main language spoken in Israel is Hebrew, and as such it is the lingua franca of higher education. Arabic is the second official language (spoken by 15% of the population). The Arabic-speaking population has its own K-12 educational system, where the language of instruction is Arabic with the exception of some subjects that are taught in Hebrew. In general, this educational system is less developed than the Hebrew one. Russian is spoken by the largest immigrant population group in Israel (about 10% of the population). The Russian-speaking group is currently in transition. A wave of immigrants arrived in the early 1970s, followed by a much larger group (one million) from the former Soviet Union, beginning in the 1990s. In spite of being a fairly heterogeneous group, the Russian-speaking immigrants share educational values that include gender equality and high standards of education.

In traditional societies, fewer women than men apply to higher education. Those who do apply come from more educated families and are better prepared relative to men from the same society. These features were expected to characterize the Arabic-speaking group in this study. On the other hand, it was hypothesized that Western modern societies, which share similar processes of social mobility, also share similar patterns of gender differences.

To explore these hypotheses, the pattern of gender differences within the three groups participating in this study – Hebrew-, Arabic- and Russian-speaking students - was examined and compared with those found in the US and Sweden.

Method

Population

The study population consisted of all students in Israel's six research universities who began their studies between the academic years 1985/6 and 1996/7. The analyses were performed separately for each of the three groups of students (Hebrew-, Arabic-, and Russian-speaking). The unit of analysis was a university department with at least five male and five female students in a given year in a university. The departments were aggregated into three clusters based on area of study: Verbal (humanities, sociology, political science, psychology, social work, and law); Quantitative (physics, math, engineering, statistics, computer science, economics, business, and accounting); and Life (biology, chemistry, and health sciences).

Table 1 presents the number of departments and students by gender, language group, and cluster (of areas of study) for students who completed their first year of studies. Results pertaining to the Russian- and Arabic- speaking students within Life and Quantitative clusters should be treated with caution due to the small number of participants.

Table 1

Number of departments, males and females by language and by academic cluster of fields of study for students who began to study during 1985-1996 and obtained FGPA

Russian			Arabic			Hebrew			
F	**M**	**Dep**	**F**	**M**	**Dep**	**F**	**M**	**Dep**	**Cluster**
1032	446	58	2527	1629	143	46823	23063	797	**Verbal**
1181	1400	92	622	994	56	14474	26750	524	**Quant.**
580	239	27	234	258	28	8060	4801	187	**Life**
2793	2085	177	3383	2881	227	69357	54614	1508	**All**

Table 2 presents the number of departments and students by gender, language group, and academic cluster for students who began their studies between the academic years 1985/6 and 1992/3 and completed at least two years of university studies. These data are somewhat incomplete, for technical reasons; therefore, attrition rates cannot be directly derived from a comparison of Tables 1 and 2. However, there is some indication that dropout rates are higher for Arabic- and Russian- speaking students, and that within the Hebrew- and Russian-speaking groups, attrition is slightly higher for females.

Table 2

Number of departments, males and females by language and by academic cluster of fields of study for students who began to study during 1985-1996 and obtained TGPA

Russian			**Arabic**			**Hebrew**			
F	**M**	**Dep**	**F**	**M**	**Dep**	**F**	**M**	**Dep**	**Cluster**
37	26	3	481	345	33	12576	6700	310	**Verb**
203	319	20	105	94	10	4784	8654	224	**Quant**
122	50	5	36	37	5	2845	1741	81	**Life**
362	395	28	622	476	48	20205	17095	615	**All**

Measures

Admissions measures

- Mean score on the high school record (HSR) – based on both external national exams and teacher evaluation. The range of the HSR scale in Israel is 40-130. Most Russian-speaking applicants did not graduate from Israeli high schools, and this measure is therefore not available for them.

- Four scale scores on the Psychometric Entrance Test (PET) – a scholastic assessment test - a total score (ranging from 200 to 800), and three sub-test scores (each ranging from 50 to 150): Verbal Reasoning (V), Quantitative Reasoning (Q), and English as a Foreign Language (E) (for more details on PET see Beller, 1994). PET is translated and adapted for each of the language groups (see Beller, Gafni, and Hanani, in press).
- Admissions score (Adm) - a composite score based on equal weights of PET and HSR. Adm was computed within applicants to each university and then standardized to a scale with a mean of 50 and a standard deviation of 10 (within each university). This measure was not computed for Russian-speaking examinees because the HSR measure is unavailable for them.

Criteria – academic performance

- First-year grade point average (FGPA), ranging from 0 to 100.
- Grade point average for at least two years of study (TGPA), ranging from 0 to 100.

Analyses

The following measures were computed within each unit of analysis:

- F/M - Ratio of the number of females to males
- D - Standardized difference in performance between females and males. A positive value of D indicates a higher level of performance for females.
- Validity - Correlation coefficients between the admissions variables and academic performance, corrected for range restriction and computed separately for each gender group[1].
- Test Bias - To detect bias, as defined by differential prediction, methods based on the definitions given in Darlington (1971) and the discussion by Linn (1984) were used. Linn

(1984) demonstrates that one can represent the unbiased model as one in which group membership (C) may influence an individual's latent (unobserved) qualifications (Q), and these latent qualifications then influence both the test score (X) and the university grade (Y). However, for an unbiased model to hold true, group membership should not influence X or Y directly. Such a model involves some constraints on coefficients involving X, Y, and C. Violation of these constraints implies that the no-bias condition is untenable. Specifically, 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. $\beta_{XC \cdot Y} < 0$.

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. In addition, bias was also examined according to Darlington's (1971) first and third definitions of bias. Darlington's first definition coincides with Cleary's (1968) definition of bias relating to the regression of the criterion on predictor for each group, while the third definition coincides with Cole's (1973) definition which relates to the reverse regression. Linn's criteria is more conservative since it flags only those cases where bias exists against a certain group according to all above definitions.

Results

Ratio of Females to Males (F/M)

Table 3 presents F/M for applicants and students for the academic years 1991/2 and 1992/3 by language group and academic cluster (data regarding applicants was available only for these two years).

Table 3

F/M for applicants and students by academic cluster and by language group for the academic years 1991/2 and 1992/3

Russian		Arabic		Hebrew		
students	**applicants**	**students**	**applicants**	**students**	**applicants**	**Cluster**
1.3	2.0	1.5	1.0	2.1	2.6	**Verbal**
0.9	0.5	0.8	0.3	0.5	0.6	**Quant**
1.8	1.3	0.8	0.6	1.6	2.1	**Life**
1.3	0.9	1.2	0.6	1.3	1.4	**All**

Different patterns of the female to male ratio (F/M) were found for the three groups. Across academic clusters, F/M in the Hebrew-speaking group was greater than one, for both the applicant and student populations. For Arabic-speaking applicants, F/M was smaller than one, but it was greater than one for the student group. Within the Russian-speaking group, F/M was close to 1.00 for applicants but larger for the student group. There were generally more females than males in all three groups for the verbal academic cluster, while the reverse pattern was true for the quantitative academic cluster. For the life sciences, F/M was greater than one for both the Hebrew- and Russian-speaking groups, while for the Arabic-speaking group it was smaller than one.

Standardized Difference (D)

For each variable Table 4 presents D for students within each language group and academic cluster.

Table 4
Standardized Difference (D) between males and females by academic cluster and language

Measures	Verbal	Quant	Life	All
Hebrew				
FGPA	0.10	-0.06	0.00	0.04
Adm	0.03	0.12	0.13	0.07
HSR	0.37	0.40	0.47	0.39
PET	-0.36	-0.27	-0.29	-0.32
V	-0.17	-0.10	-0.09	-0.14
Q	-0.43	-0.32	-0.30	-0.38
E	-0.15	-0.17	-0.23	-0.16
Arabic				
FGPA	-0.11	-0.12	-0.16	-0.11
Adm	0.23	0.21	-0.00	0.21
HSR	0.43	0.41	0.33	0.42
PET	-0.13	-0.11	-0.31	-0.14
V	0.09	0.07	-0.08	0.07
Q	-0.39	-0.39	-0.42	-0.39
E	0.20	0.21	0.02	0.19
Russian				
FGPA	0.39	0.05	0.34	0.20
PET	-0.19	-0.26	-0.19	-0.23
V	-0.13	-0.11	-0.02	-0.10
Q	-0.24	-0.29	-0.21	-0.26
E	-0.02	-0.14	-0.17	-0.11

In the Hebrew-speaking group, there were no meaningful differences between females and males in the admissions score or in achievement at the end of the first year of studies. A different pattern was found for the Arabic-speaking students, where the admissions score was higher for females, while academic performance was somewhat higher for males. In each of these language groups, the pattern within each academic cluster more or less resembled the pattern found across academic clusters. Russian-speaking females performed better academically than Russian-speaking males, with the exception of the quantitative academic cluster, where no gender difference in academic performance was found (note that Adm and HSR were unavailable for the Russian speakers).

In each language group, males scored lower than females on HSR (where available) and higher on PET. D for HSR was somewhat larger for the Arabic-speaking students than for the Hebrew-speaking students. The Hebrew-speaking group had the largest negative D for PET, and the Arabic-speaking group, the smallest. In each language group, the largest negative difference was found for Q. Much smaller differences were found for V and E, which were negative for the Hebrew- and Russian-speaking groups and positive for the Arabic-speaking group.

When comparing D's for the two components of Adm (HSR and PET), it can be seen that for the Hebrew-speaking students, D was approximately one third of a standard deviation for both components (positive for HSR and negative for PET). Among the Arabic-speaking students, the advantage of females on HSR was much higher than the advantage of males on PET, resulting in a positive Adm.

As mentioned above, no gender differences were found in academic achievement for Hebrew-speaking students and fairly small difference was found for the other groups. There was a slight tendency for males to perform better in the quantitative academic cluster and for females to perform better in the verbal academic cluster. Arabic-speaking males tended to

perform slightly better than females in all areas of study, while Russian-speaking females performed better than males mainly in the verbal and life academic clusters.

When examining academic achievement in more advanced years, differences between females and males were similar to those found for first year students. The only exception was for E, where the differences were more extreme for the advanced Arabic- and Russian-speaking students (for results, see Appendix A).

Validity

Table 5 presents the validity coefficients for males and females by language group and academic cluster.

Table 5
Correlations with FGPA (corrected for range restriction)

E	Q	V	PET	HSR	Adm	Gender	L	Cluster
								Verb
0.20	0.26	0.29	0.33	0.35	0.40	M	H	
0.24	0.32	0.34	0.39	0.42	0.46	F	H	
0.16	0.19	0.23	0.28	0.34	0.36	M	A	
0.21	0.27	0.24	0.33	0.42	0.44	F	A	
0.36	0.23	0.21	0.38	.	.	M	R	
0.30	0.32	0.30	0.37	.	.	F	R	
								Quant
0.20	0.38	0.29	0.38	0.43	0.47	M	H	
0.24	0.41	0.32	0.42	0.47	0.51	F	H	
0.11	0.30	0.14	0.25	0.32	0.33	M	A	
0.23	0.37	0.19	0.37	0.54	0.49	F	A	
0.30	0.27	0.22	0.32	.	.	M	R	
0.27	0.33	0.24	0.33	.	.	F	R	
								Life
0.16	0.34	0.25	0.35	0.39	0.44	M	H	
0.21	0.44	0.32	0.43	0.50	0.53	F	H	
-0.04	0.14	0.03	0.04	0.27	0.22	M	A	

0.28	0.14	0.29	0.37	0.48	0.47	F	A	
0.26	0.30	0.31	0.38	.	.	M	R	
0.34	0.28	0.24	0.34	.	.	F	R	
								All
0.20	0.31	0.29	0.35	0.38	0.43	M	H	
0.24	0.37	0.33	0.41	0.45	0.48	F	H	
0.13	0.21	0.19	0.25	0.33	0.34	M	A	
0.22	0.29	0.23	0.35	0.45	0.46	F	A	
0.31	0.27	0.24	0.35	.	.	M	R	
0.29	0.32	0.26	0.34	.	.	F	R	

In general, the validities of PET and its components, as well as those of HSR and Adm, were higher for females than for males in the Hebrew- and Arabic-speaking groups across all fields of study as well as within each academic cluster (this was more evident in the Arabic-speaking group). No clear pattern of differences in validity between females and males was found for the Russian-speaking group.

The above differences in validities could not be accounted for by differences in variances in the variables, because variances in this selected group of women were actually somewhat smaller than for males. To investigate the hypothesis that the somewhat higher validities among females are related to higher reliabilities of the criterion, the correlations between FGPA and second year GPA were calculated separately for males and females within each unit of analysis. These correlations were used as rough estimates of the reliability of the criterion. The results are presented in Table 6.

Table 6

Correlations of FGPA with second year GPA by academic cluster and gender

Gender	Verbal	Quant	Life	All
Hebrew				
M	.63	.62	.71	.63
F	.61	.58	.72	.61
N Departments	300	223	78	601
N Students	18,011	12,961	4,355	35,327
Arabic				
M	.58	.55	.40	.57
F	.50	.49	.67	.51
N Departments	27	9	4	40
N Students	671	173	62	906
Russian				
M	.84	.55	.65	.60
F	.71	.55	.56	.57
N Students	3	20	4	27
N Total	61	504	157	722

The correlations did not support the hypothesis that the higher validities obtained for females were related to a higher reliability of the criterion within the female group. On the contrary, these correlations were even slightly higher for males.

The validity coefficients of the various predictors were also calculated against TGPA as the criterion. As in FGPA, the correlations of all predictors with TGPA were generally larger for females than for males. When measured for the same group, the magnitude of the validities was the same for FGPA and TGPA (and therefore, the data are not presented).

Prediction-Bias

The focus of this analysis was the detection of bias against or in favor of females for each language group. Table 7a presents the proportion of university departments for which bias was detected against or in favor of females across areas of study. Tables 7b – 7d present this information for each academic cluster. The first column of the table presents results based on Darlington's first definition (Cleary's), and the second column presents results based on Darlington's third definition (Cole's). The third and fourth columns in the table relate to Linn's two boundary conditions for bias.

Table 7a
Percent of departments for which bias was detected ($\alpha = 0.05$) against or in favor of females across academic clusters according to Darlington's 1st and 3rd definitions, and Linn's criteria (FGPA)

Against Linn	In favor Linn	Against 3rd Equ	In favor 1st Equ	Pred
Hebrew (N Dep =1,508)				
4	6	23	32	**Adm**
1	16	12	54	**HSR**
12	1	49	12	**PET**
6	1	35	15	**V**
14	0	52	9	**Q**
0	0	19	8	**E**
Arabic (N Dep =227)				
0	10	16	47	**Adm**
1	14	11	55	**HSR**
1	3	26	31	**PET**
2	3	20	37	**V**
4	0	36	18	**Q**
0	0	14	20	**E**
Russian (N Dep = 177)				
11	1	51	9	**PET**

7	1	43	12	**V**
11	0	47	10	**Q**
4	0	37	10	**E**

Table 7b
Percent of departments for which bias was detected ($\alpha = 0.05$) against or in favor of females for the Verbal Academic cluster according to Darlington's 1st and 3rd definitions, and Linn's criteria (FGPA)

Against Linn	In favor Linn	Against 3rd Equ	In favor 1st Equ	Pred
Hebrew (N Dep = 797)				
6	3	24	21	**Adm**
2	12	13	37	**HSR**
17	0	41	7	**PET**
9	0	33	9	**V**
18	0	43	6	**Q**
0	0	23	6	**E**
Arabic (N Dep =143)				
1	11	16	39	**Adm**
1	17	11	40	**HSR**
2	3	22	31	**PET**
3	4	17	36	**V**
6	0	29	20	**Q**
0	0	15	22	**E**
Russian (N Dep = 58)				
16	0	38	7	**PET**
10	0	36	10	**V**
8	0	40	10	**Q**
2	0	33	7	**E**

Table 7c

Percent of departments for which bias was detected ($\alpha = 0.05$) against or in favor of females for the Quantitative academic cluster according to Darlington's 1st and 3rd definitions, and Linn's criteria (FGPA)

Against Linn	In favor Linn	Against 3rd Equ	In favor 1st Equ	Pred
Hebrew (N Dep =524)				
1	10	13	30	**Adm**
0	20	7	41	**HSR**
6	2	31	15	**PET**
3	2	23	19	**V**
10	1	32	12	**Q**
0	0	13	11	**E**
Arabic (N Dep = 56)				
0	7	11	39	**Adm**
0	11	4	50	**HSR**
0	2	21	30	**PET**
0	2	16	36	**V**
4	0	30	13	**Q**
0	0	16	14	**E**
Russian (N Dep = 92)				
8	1	37	9	**PET**
7	0	35	12	**V**
8	0	33	10	**Q**
2	0	32	13	**E**

Table 7d

Percent of departments for which bias was detected ($\alpha = 0.05$) against or in favor of females for the Life academic cluster according to Darlington's 1st and 3rd definitions, and Linn's criteria (FGPA)

Against Linn	In favor Linn	Against 3rd Equ	In favor 1st Equ	Pred
Hebrew (N Dep =187)				
2	11	14	35	**Adm**
0	21	10	37	**HSR**
7	2	35	17	**PET**
3	3	29	20	**V**
9	1	35	11	**Q**
0	0	21	8	**E**
Arabic (N Dep = 28)				
0	0	7	21	**Adm**
0	4	14	32	**HSR**
0	4	43	11	**PET**
0	0	21	25	**V**
0	0	50	14	**Q**
0	0	7	21	**E**
Russian (N Dep = 27)				
7	0	56	7	**PET**
4	4	41	11	**V**
15	0	41	7	**Q**
11	0	41	7	**E**

Across all academic clusters, bias was detected for Adm in only some 10% of the departments, based on Linn's criteria. For the Hebrew-speaking students, bias against or in favor of females was found in about equal number of cases (departments). In all cases where bias was found for Arabic-speaking students, it was in favor of females. No data regarding Adm were available for the Russian-speaking group.

Inspection of the various predictors separately revealed that for approximately 15% of the departments there was a tendency for HSR to be biased in favor of females in the Hebrew- and Arabic-speaking groups (no data were available for the Russian-speaking group). In the Hebrew- and Russian-speaking groups, PET was found to be biased against females in 12% of the cases. Hardly any gender bias was found for PET in the Arabic-speaking group. In all language groups, the Q sub-test was the major contributing factor to the bias in PET, and this was most evident in the Hebrew-speaking group. No bias was detected for the E sub-test for the Hebrew and Arabic-speaking students, and it was biased (against females) in only 4% of the cases in the Russian-speaking group. Examining bias along Darlington's first and third definitions reveals the same bias pattern and direction, but in a greater number of cases. This is not surprising, given Linn's more conservative criteria for defining bias.

Inspection of detected bias within each academic cluster revealed that there was some tendency of the admissions measures to be biased against females for the Verbal academic cluster, whereas for the Quantitative academic cluster it was found to be biased in favor of them.

A similar pattern of bias was found when TGPA was used as the criterion (and therefore, it is not presented here).

Discussion

The purpose of the study was to investigate the extent to which results exhibiting gender differences on scholastic measures could be generalized across cultures. The results regarding gender differences in performance (D) for three language groups (Hebrew-, Arabic- and Russian-speaking students) resembled those summarized by Willingham and Cole (1997) for the US. High school grades were higher for females across all ethnic groups, while males performed better on the standardized tests. The difference of one third of a standard deviation

in favor of males among the Hebrew-speaking examinees on the total PET score was similar to that found for the SAT in the US (Willingham and Cole, 1977), as well as to that found for the SweSAT, which is used for admissions decisions for higher education in Sweden (Stage, 1997). In all three countries, the largest difference was on the quantitative section of the scholastic admissions test. D of about one third of a standard deviation was also found for HSR. The magnitude of D was similar to that found both for HSR in the US. Arabic-speaking female applicants performed relatively better than males, on HSR, PET and Adm as compared with Hebrew-speaking female applicants relative to Hebrew-speaking males. This phenomenon was also found in the US for the Black examinees.

The difference between females and males on FGPA was relatively small for all three groups (none for the Hebrew-speaking examinees), with some tendency for females to perform better within the Russian-speaking group and for males to perform slightly better within the Arabic-speaking group. The slightly lower performance on academic studies of Arabic-speaking females relative to males remains unexplainable in the light of their higher admissions performance.

The ratio of females to males (F/M) among the applicant and student populations in the various language groups seems to be related to social and economic factors. These factors differentially affect the characteristics of the particular females and males comprising each group. The composition of each group, in terms of the social characteristics of its males and females, was related to the level of performance of the two gender groups on each of the variables studied. This was especially apparent when comparing F/M for the Arabic- and the Hebrew- speaking applicants and students. In contrast to the Hebrew-speaking group, in the Arabic-speaking group the number of female applicants is about half that of male applicants. On the other hand, among Arabic-speaking students, the ratio of females to males is higher than one and is similar to that of Hebrew-speaking students.

As in the US, validity coefficients were slightly but consistently higher for females in all ethnic groups. Statistical artifacts, such as differences in variances, could not explain the higher validities for females. It was hypothesized that the differences in validity could be attributed to higher reliability of the criterion for females, but the results did not support this hypothesis. A closer examination of the criterion reliability for each language group revealed that they were somewhat lower for Arabic- and Russian-speaking students as compared with Hebrew-speaking students, for both males and females. This may be related to difficulties faced by non-Hebrew-speaking students in a university in which instruction is carried out in Hebrew.

Bias issues have always been discussed in various contexts (e.g., social, political, educational), which are frequently inseparable from social values. Notwithstanding the importance of social values within this context, the aim of this study was to provide some empirical evidence as to the psychometrics of the bias issue. In general, bias was found in only a small number of cases. For all language groups, using either high school grades alone or standardized test scores alone in the admissions process would, in some cases, have resulted in bias in favor of or against females, respectively. These opposing effects tend to offset each other, and it is therefore not surprising that the actual admissions score, which consists of both predictors, is generally unbiased. Hardly any bias was found for the Hebrew-speaking group; within the Arabic-speaking group, the admissions score was biased in favor of females in 10% of the cases. Similar results, in both magnitude and trend, were reported by Willingham and Cole (1997) regarding classroom grades and standardized test scores, thus implying that test scores or grades should not be used alone in the prediction of academic success. Moreover, it is hard to expect that tests can be constructed such that all relationships between admission measures and criteria are simultaneously unbiased for all

subgroups. Inspection of detected bias within each academic cluster revealed that there was some tendency of the admissions measures to be biased against females for the Verbal academic cluster, whereas for the Quantitative cluster it was found to be biased in favor of them. It can be of interest to explore how these results are related to factors that affect the differential self-selection processes that occur for the two genders when choosing an area of study.

The over-all consistency between the results found for students in the USA, Sweden, and the Hebrew-speaking students in Israel is striking. This was true for both the direction as well as the magnitude of the differences for all the variables measured in this study. In light of the fact that these are three separate cultures, which differ in language, high-school system, university entrance exams and higher educational system, this is not a trivial phenomenon. It is likely that the influence of social dynamics on who will study in institutions of higher education and the pattern of preferences of study areas exhibited by the two gender groups are similar in Western countries.

The different pattern found for Arabic-speaking male and female students in Israel is probably the result of different social and cultural norms. Arab society in Israel is more traditional than the society as a whole; therefore, Arab women and men who apply to institutions of higher education do not represent the same socio-economic strata of the population as their respective Hebrew-speaking counterparts. Fewer Arabic-speaking women apply to higher education relative to men, but more are admitted, probably indicating that Arabic-speaking women candidates come from more educated and well-established families. A similar pattern of self selection was found for Black females when compared with other ethnic groups in the US, perhaps deriving from the fact that both Arab and Black female applicants are a more select group than their male counterparts (Willingham and Cole, 1997).

The Russian-speaking population investigated in this study is an immigrant group still in the process of constant change. It still reflects the social norms that were prevalent in the former Soviet Union. This is reflected to some extent in the relatively higher representation of women in the quantitative and life academic clusters of areas of study and in their relatively high achievements in these areas. It is likely that in the future, as part of the socialization process, this population will bear greater resemblance to the majority population of Hebrew-speaking women and men.

Cognitive gender differences are relative and reflect the social and cultural values and norms of the society and period of time within which they occur. Therefore, gender differences should always be discussed within the relevant cultural context. The cross-cultural consistency observed in the above patterns of gender difference calls for a closer examination of the various variables (cognitive, motivational, social, and cultural) that might explain these patterns.

References

American College Testing Program. (1973). *Assessing students on the way to college: Technical report for the ACT assessment program* (1). Iowa City, IA: Author.

American College Testing Program. (1988). *ACT assessment program technical manual*. Iowa City, IA: Author.

Admissions Testing Program, The College Board. (1992). *Profile of SAT and Achievement Test takers.* New York: College Entrance Examination Board.

Advanced Placement Program, The College Board. (1992). *National summary reports*. New York: College Entrance Examination Board.

Beller, M. (1994). Psychometric and social issues in admissions to Israeli universities. *Educational Measurement*: *Issues and Practice*, *13(2)*, 12-20.

Beller, M., Gafni, N. and Hanani, P. (in press). Constructing, adapting, and validating admissions tests in multiple languages: The Israeli case. In R. K. Hambleton, P. Merenda, & C. Spielberger. (Eds.), *Adapting educational and psychological tests for cross-cultural assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, 5(2), 115-124.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cole, N.S. (1973). Bias in selection. *Journal of Educational Measurement, 10 (4)*, 237-255.

Darlington, R. B. (1971). Another look at "cultural fairness." *Journal of Educational Measurement*, *8(2)*, 71-82.

Dwyer, C. A., & Johnson L. M. (1997). Grades, accomplishments, and correlates. In W. W. Willingham, & N. S. Cole, (Eds.) *Gender and fair assessment*. NJ: Lawrence Erlbaum Associates.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley & Sons. (Reprinted in 1987. Hillsdale, NJ: Lawrence Erlbaum).

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

Linn, R. L. (1983). Pearson selection formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement*, *20*, 1-15.

Linn, R. L. (1982). Ability testing: Individual differences, prediction, and differential prediction. In A. Wigdor & W. Garner (Eds.), *Ability testing: Uses, consequences, and controversies, Part II: Report of the National Academy of Sciences committee on Ability Testing* (pp. 335-388). Washington, DC: National Academy Press.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). Student group differences in predicting college grades: Sex, language, and ethnic group (CB Rep. No. 93-1; ETS RR-94-27). New York: College Entrance Examination Board.

Stage, C. (1994). *Use of assessment outcomes in selecting candidates for secondary and tertiary education: a comparison.* Paper presented at the 20th Annual Conference of the IAEA, Wellington, New Zealand.

Stage, C. (1997). *Do males and females with identical test scores solve test items in the same way?* Paper presented at the 23rd Annual Conference of the IAEA, Durban, South Africa.

Willingham, W. W., Cole, N. S., Lewis, C., & Lueng, S. W. (1997). Gender differences within ethnic groups. In W. W. Willingham., & N. S. Cole., *Gender and fair assessment*. NJ: Lawrence Erlbaum Associates.

Willingham, W. W., & Cole, N. S. (1997). Gender and fair assessment. NJ: Lawrence Erlbaum Associates.

Acknowledgments

The authors extend their appreciation to Gershon Ben-Shackahr, Tamar Kennet-Cohen, Ruth Fortus, and Cathrael Kazin for their helpful comments.

Appendix A

Standardized Difference (D) between males and females by academic cluster and language for students who studied at least two years

Variable	Verbal	Quant	Life	All
Hebrew				
TGPA	0.14	-0.01	0.06	0.07
FGPA	0.09	-0.09	-0.07	0.00
Adm	0.08	0.13	0.13	0.10
HSR	0.41	0.40	0.48	0.42
PET	-0.32	-0.24	-0.29	-0.29
V	-0.15	-0.10	-0.12	-0.13
Q	-0.40	-0.30	-0.31	-0.35
E	-0.07	-0.12	-0.17	-0.10
Arabic				
TGPA	-0.12	-0.02	-0.50	-0.13
FGPA	-0.13	-0.13	-0.74	-0.17
Adm	0.16	0.41	-0.30	0.17
HSR	0.37	0.54	0.08	0.38
PET	-0.15	0.07	-0.43	-0.12
V	-0.03	0.02	0.02	-0.02
Q	-0.37	-0.10	-0.68	-0.34
E	0.41	0.58	-0.02	0.41
Russian				
TGPA	-0.16	-0.09	0.79	0.10
FGPA	0.08	-0.16	0.47	0.00
PET	0.05	-0.38	0.07	-0.24
V	0.18	-0.20	0.25	-0.07
Q	0.05	-0.34	-0.17	-0.27
E	-0.40	-0.32	0.09	-0.23

[1] The approach adopted by the present study for correcting the sample statistics (the correlation coefficients and regression coefficients in the multiple regression analyses) is based on the assumption that the selection is carried out on the basis of Adm. It should be noted that Adm is composed of HSR and PET, and PET is composed of V, Q, and E. Therefore, the explicit selection is based on Adm, resulting in an incidental selection of all other predictors. The appropriate formula for univariate selection in a three-variable case such as the above, as formulated by Gulliksen (1950), is used to correct the validity coefficients of each of the predictors (followed by a corresponding adjustment of the multiple regression coefficients). The estimates of population variance are based on the mean standard deviation of candidates, computed across all departments for each gender and language group.