The Effect of Coaching on the Predictive Validity

# of Scholastic Aptitude Tests

Avi Allalouf

National Institute for Testing and Evaluation (NITE), Jerusalem

Paper presented at the annual meeting of the
American Educational Research Association
New York, NY, April 1996

Abstract

The present study was designed to examine whether coaching affect predictive validity and fairness of scholastic aptitude tests. Two randomly allocated groups, coached and uncoached, were compared, and the results revealed that although coaching enhanced scores of the Israeli Psychometric Entrance Test by about 25% of a standard deviation, it did not affect predictive validity and did not create a prediction bias. These results refutes claims that coaching reduces predictive validity and creates a bias against the uncoached examinees in predicting the criterion. The results are consistent with the idea that score improvement due to coaching does not result strictly from learning specific skills that are irrelevant to the criterion.

Early theories viewed intelligence as inborn and determined primarily by heredity (e.g., Galton, 1869; Pearson, 1904; Terman, 1916). According to these theories, environment has little effect on intelligence, which means that scores on intelligence tests cannot be `significantly improved by special preparation. Deviating from this view was a slow and gradual process. Eventually, with the accumulation of evidence, it became clear that intervention could improve performance on intelligence tests (see Caruzo, Taylor and Detterman, 1982, for a comprehensive review). The extent of this improvement depends on the intensity of the intervention. According to some reports, the mean improvement due to focused preparation is about one third of a standard deviation (Jensen, 1980; Bond, 1989). When very intensive interventions are carried out over a period of years, the mean improvement can be as much as one half or even two thirds of a standard deviation (Spitz, 1986; Brody, 1992). The prevailing attitude in the area of scholastic aptitude tests which developed in the wake of intelligence tests underwent a similar process. Until about twenty years ago, the commonly held view was that improvement due to focused preparation was very small. This view is clearly demonstrated by the following citation from an ETS publication: *"The magnitude of the gains resulting from coaching vary slightly but they are always small regardless of the coaching method used or the difference in the student coached"* (ETS, 1965).

Since the early seventies, many studies focusing on the effects of preparation on scholastic aptitude tests have been conducted. Recent meta-analyses of these studies (Messick, 1981; Powers, 1993) demonstrated that scores on scholastic aptitude tests can be improved by focused preparation. The expected fluctuations in an examinee's score following coaching are generally small and the mean gain on the SAT (Scholastic Assessment Test, which consists of a verbal and a mathematical section), is approximately one fifth of a standard deviation. This gain is greater than the gain that would be expected as a result of retesting. Similar results were obtained in a study based on examinee feedback questionnaires for the Israeli Inter-University Psychometric Entrance Test (PET), which, like the SAT, consists of a mathematical and a verbal section as well as an additional section which tests command of English as a foreign language (Oren, 1993). On both the PET and the SAT, coaching was more effective for the mathematical section than for the verbal section of the test.

Special preparation is particularly common for entrance exams to institutes of higher learning. There are several reasons for this phenomenon: (a) the institutes which construct and administer entrance exams generally provide information about their content; (b coaching materials and courses are readily available; (c) examinees believe that preparation will improve their scores dramatically. If we consider coaching course which is the most intensive form of focused preparation, (comparing to self preparation with preparation books), in the United States, according to an estimate made by Baydar, (1990), the number of examinees taking coaching courses for the SAT has increased moderately, from 5 percent of examinees in 1980 to 15 percent of examinees in 1990. In Israel, the number of examinees who take coaching courses for the PET has increased much more dramatically - from 1% of examinees in 1984 (the first administration of this test), through 42% in 1990, to 77% of examinees in 1996 (Allalouf; 1984, Arieli, 1996).

Coaching focuses on three interrelated elements: acquiring familiarity with the test, learning testwiseness (TW), and reviewing material which is relevant to test content. Familiarity with the test means being acquainted with the test instructions, item types, time limits and answer sheet format. Familiarity is achieved by answering questions which are similar to the test questions under conditions which are as similar as possible to those encountered during the actual administration of the test. The original and still accepted definition of TW is that of Millman, Bishop, and Ebel (1965, p. 707): *"..subject capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score."* They enumerated four strategies which are independent of test construction or test purpose: efficient use of the available time, error avoidance, guessing and deductive reasoning.

Much research has been conducted on coaching for tests of scholastic aptitude. Most of the studies have dealt with tests administered in the United States by the Educational Testing Service, and primarily with the SAT. Most of these studies have focused on the effect of coaching on test scores (e.g. Messick, 1981, Powers, 1993). Many researchers, among them Messick, 1981, and Anastasi, 1981, have raised the question of the possible detrimental effects of coaching on test validity. However, few research efforts have been devoted to studying the effects of coaching on the validity and fairness of scholastic aptitude tests. For example, Bond (1989, p. 440) wrote: "*A continuing concern on the part of testing specialists, admissions officers, and others is that coaching, if highly effective, could adversely affect predictive validity and could, in fact, call into question the very concept of aptitude."*

The earliest study dealing with the effect of coaching on predictive validity was conducted by Ortar (1960). The Triangle Test[1] was administered to a group of 397 children aged 6-14 who were unfamiliar with it. The test consisted of three parts: The first part served as a baseline, the second part was used for coaching, and the third part of the test was administered immediately after the coaching was completed. The scores computed for the first and the third parts were used as predictors, and the criterion was based on teacher evaluation of scholastic aptitude. The results indicated that the correlation with the criterion was significantly greater for the third part of the test than for the first part. Ortar's (1960) explanation for the improved predictive validity was that because coaching is a learning process the after-coaching scores reflect better learning ability.

Bashi (1976) conducted a study with a similar design to the one used by Ortar (1960). The Raven Progressive Matrices (RPM) test was administered to 4,559 Israeli Arab students aged 10-14. The scores on achievement tests in mathematics and Arabic, as well as the teachers' evaluations of the students' relative position in the class served as criteria. The test, which was not familiar to the students, was administered twice, with a very short coaching period of about one hour in between. The mean gain following coaching was high and statistically significant (between one half and three quarters of a standard deviation).

Results showed that the RMP had a small but significant improvement in predicting the above mentioned criteria as a result of coaching.

Marron (1965) studied the effects of a long-term coaching program for the SAT and for the College Board Achievement Tests on the validity of these tests for predicting freshman class standing at military academies and selective colleges. Score gains were very high (about three quarters of a standard deviation, on both the verbal and the mathematical parts of the test). Marron found that in some of the preparatory programs, in which the mean gain due to coaching was higher, coaching led to an overprediction of academic performance. It can be argued that thirty years ago almost all examinees had little familiarity with test content, and therefore the effect of coaching was large so the scores of those who underwent intensive coaching overpredicted their academic performance.

Powers (1985) examined the effects of variations in the number of hours of preparation on the predictive validity of the analytical section of the Graduate Record Examination (GRE). The self-reported grade averages of 5,107 undergraduates served as the "postdictive" criterion, and the preparation consisted solely of familiarization through self-testing. The data which served for the study was reported in Powers and Swinton (1984) and showed that the score improvement due to coaching was similar to the one reported in the meta-analytic study conducted by Messick and Jungeblut (1981). Powers' concluded that: "*preparation of the kind studied may enhance rather than impair test validity*" (p. 189).

Jones (1986) studied the effects of coaching on the predictive validity and bias of the Skilled Analysis section of the Medical College Admission Test (MCAT). The criterion used by Jones was whether or not a student experienced academic problems in medical school. He analyzed two groups of self-reported coached and uncoached students, each consisting of 2,127 subjects (Jones did not report whether coaching improved MCAT scores The findings indicated that coaching does not lead to an overprediction of students' subsequent medical school performance.

Baydar, (1990) using a simulated study, attempted to determine whether or not the decline in SAT validity (a decline of 8 percent in the years between 1976-1985) was related to changes in coaching density (i.e. the percent of coached examinees). Freshman Grade Point Average (FGPA) was used as the criterion and the simulation indicated that, at most, only ten

[1] This test was developed by Ortar for her study and it was based on the Arthur Stencil Design Test(1943) which measures nonverbal intelligence.

percent of the decline in predictive validity could be explained by the increase in coaching density.

Contrary to the concern raised by Bond (1989), most of the above studies indicated that coaching led to slight improvements in predictive validity of scholastic aptitude tests, while no consistent picture emerged regarding the question of whether these tests are biased against the uncoached examinees. However, these studies suffered from several methodological problems: (1) They contained insufficient information about whether or not examinees actually underwent coaching and about the intensity of the coaching. (2) In some of the studies, coaching consisted of a few hours and therefore it cannot be compared with today's commercial courses which offer much more intensive practice. (3) Most participants in the studies conducted 20 years ago were unfamiliar with the types of questions used in the test as well as with the test instructions, and therefore coaching had a relatively large impact on their scores. Today, most examinees who undergo coaching are already familiar with the test format prior to coaching. Clearly, there is a need for an up-to-date, well-designed study which will shed more light on the effect of coaching on predictive validity and fairness.

Theoretically, coaching can affect validity in various ways. For example, if it is assumed that coaching affects factors that are common to the test's scores and the criterion, then reducing the variance of these components through coaching might reduce predictive validity. On the other hand, if it is assumed that coaching affects factors which are unique to the test's scores, then reducing the variance of these factors through coaching should increase predictive validity. In addition to the question of the influence of coaching on predictive validity, there is also the question of bias which arises when examinees differ in the amount of coaching they have undergone. This matter was generally not dealt with in the studies mentioned above (with the exception of Jones, 1986). In the present study, test bias was defined on the basis of the regression model suggested by Cleary (1968). Cleary defined a test as biased against a certain group if it consistently underpredicts criterion scores for members of that group. On the basis of this definition a method which tests for bias between groups by comparing the different regression lines for predicting the criterion for each group was proposed by Lautenshlager & Mendosa (1986). This method was applied in the present study to examine whether the test is biased against the uncoached group.

The present study was designed to examine the effect of coaching on predictive validity and fairness of scholastic aptitude tests in two situations: 1) fixed coaching - a situation in which all examinees prepare for the test to the same extent, and 2) variable coaching - a more common situation in which there are different levels of test preparation in the population. The findings should provide an empirically-based answer to the oft-heard public criticism which is based on the belief that preparation improves scholastic aptitude tests scores significantly and therefore these tests cannot serve as valid predictive tools. Of course, if coaching does not impair predictive validity and fairness, it might be desirable. From a theoretical perspective, the results of this study can shed light on the question of whether or not coaching has an impact on factors which are common to the test and the criterion. From an applied perspective, institutes that use aptitude tests for admissions purposes would be able to take into account the impact of coaching on predictive validity, as well as, the test's bias against uncoached applicants.

METHOD

Participants

The study population consisted of students in eight pre-academic preparatory institutes throughout Israel during the academic year 1992-1993. These institutes offer programs which usually last one year and which provide their students with an opportunity to complete their high-school education and obtain matriculation certificates (see Beller and Ben-Shakhar, 1994, for further details). Students in these pre-academic preparatory institutes are generally highly motivated to obtain high PET scores in order to be accepted to universities. Both matriculation certificate and PET scores serve as criteria for admitting students into most higher learning institutions in Israel. All participants had some familiarity with PET because they took it before starting the pre-academic program.

Design

The students in each institute were randomly divided into a research group and a control group. The participants knew in advance that they would receive a coaching course, but did not know how many tests they would be given prior to the course. The Psychometric Entrance Test was administered to the research group (n=207), which then participated in a coaching course that lasted approximately one and a half months. Following the course they were retested with another version of the test (versions were fixed). The control group (n=67) also took two versions of the test, without attending a coaching course in between. The time interval between the tests administered to the control group was also about one and a half months, and the course was offered to them after they took the second test.

The test versions were comparable in content, structure and reliability. Each test contained three subtests: verbal, quantitative, and English. Each subtest was scored separately and standardized on a scale whose distribution in the base year (1984) was [100,20]. In addition to the scores on the different sections of the test, a total score *(VQ)*, based only on the quantitative and verbal sections was calculated for each examinee (there was no coaching for English). The VQ score was standardized on a scale whose original distribution was [500,100]. The experimental design is shown in **Table 1**.

TABLE 1 **Experimental Design**

Group	Test 1	Course	Test 2	Course	Test 3
Research	yes	yes	yes	--	--
Control	yes	no	yes	yes*	yes*

* Offered to participants in the control group, but not an integral part of the study.

A special coaching course was designed for the purposes of this study. Course instruction dealt only with the verbal and quantitative sections of the Psychometric Test. Items used in the course were provided by the National Institute for Testing and Evaluation, which constructs and administers the Psychometric Entrance Test. These items served as the basis for 24 study units, two of which were devoted to the subject of testwiseness. Ten experts in lesson planning and teaching in the relevant fields reviewed the study units during the various preparation stages. Answers to a feedback questionnaire which was administered to the participants upon completion of the course indicated that the students were reasonably satisfied with the coaching course.

Criterion

The weighted average of the study participants' scores on the matriculation exams was used as the criterion for validation. This criterion is similar to a concurrent validity criterion. The weighted average was computed by exactly the same method for all participants, based on their matriculation[1] scores in all subjects. These matriculation scores were obtained directly from the Ministry of Education. This criterion was chosen because the matriculation certificate is required for university registration, it is an accepted measure of success in studies, and it is measured on an identical scale for all of the study participants. Another advantage of this criterion is the fact that most of the matriculation subjects are mandatory for university acceptance and therefore are common to all study participants (further details on the method for calculating the matriculation score average can be found in The Hebrew University Information Booklet, 1994).

[1] Some of the scores were obtained from matriculation exams which were taken before entering the preparatory institute while other scores were obtained from matriculation exams taken after completing the preparatory institute.

RESULTS

Effect of Coaching on Test Scores

The coaching effect on test scores was defined as the difference between the mean gain (i.e. score on Test 2 minus score on Test 1) obtained by the research group and that obtained by the control group. **Table 2** presents the mean scores obtained by each group in each administration of the test (before and after coaching), the mean gain score obtained by each group and the coaching effect. Dependent samples t-tests were used to test whether the gains were statistically significant *(df = n-1 = 206 coached group, 66 uncoached group; p=.05*) , and t-tests for independent samples were conducted to compare the gains obtained in the experimental and the control groups (*df = n1+n2-2 = 207+67-2 = 272; p=.05).* Gains in the two subtests as well as the total score (VQ) were much larger and statistically significant in the coached group than in the control group. Gain scores were not significantly different from zero in the uncoached group. The difference between the gains were all statistically significant. These findings indicate that the mean gain in the total test score (VQ) of the coached group exceeded the mean gain of the control group by about 25% of a standard deviation. Gains on the quantitative test scores (Q) were larger than gains on the verbal test scores (V). The estimate of the coaching effect obtained in this study is similar in magnitude to estimates obtained in meta-analytic studies which focused on this topic (Messick, 1981; Powers, 1993). This similarity indicates that the coaching program applied in this study was as effective as commercial coaching programs, and thus reinforces the generalizability of the study findings. It is interesting to note that there were almost no gains in the control group. This finding can be explained by the short amount of time which elapsed between tests, and by examinee familiarity with the test entering the preparatory program.

TABLE 2 **Mean Scores and Gains, by Group**

(Standard deviations appear in brackets)

Score	**Research Group** Coached			**Control Group** Uncoached			**Difference Between Gains**
	Test 1 Before	**Test 2** After	**gain**	**Test 1** Before	**Test 2** After	**gain**	
Verbal	107.69 (13.9)	110.69 (14.6)	3.00* (10.3)	103.36 (15.9)	102.84 (15.6)	-0.52 (10.0)	3.52*
Quantitative	107.57 (15.2)	113.78 (15.0)	6.21* (11.2)	103.31 (16.4)	104.10 (16.1)	0.79 (9.6)	5.42*
VQ	541.30 (68.2)	565.76 (69.2)	24.46* (41.8)	518.51 (73.8)	519.21 (76.1)	0.70 (39.2)	23.76*

* Significant (p < .05)

Table 2 indicates that the before-coaching scores were higher for the research group than for the control group. Although the difference was not large (between one fifth and one quarter of a standard deviation), it deserves consideration and is discussed below (see the section entitled "**Sampling**").

It was expected that the significant effect of coaching on the scores of the coached group will reduce the similarity between the "before" and the "after" scores, in this group as compared with the uncoached group. Indeed, the "before-after" correlations, which are presented in **Table 3,** were lower for the research group than for the control group, but the correlation difference was statistically significant only for the quantitative section of the test.

TABLE 3 **Correlations Between "Before" and "After" Scores by Group and by Sub-tests**

Score	**Research group** Coached	**Control Group** Uncoached	**Difference Between Correlations**
Verbal	.738	.798	.060
Quantitative	.725	.826	.101*
VQ	.815	.864	.049

* Significant (p < .05)

Effect of Coaching on Predictive Validity

The main purpose of this study was to estimate the effect of coaching on predictive validity. This was done by examining the differences in predictive validity between the two test administrations (validity of the "after scores" minus that of the "before scores") , within each group, and by comparing the validity differences between the two groups.

Effect of Coaching on Predictive Validity **Within** Groups

Table 4 presents the validity data by groups and by sub-tests, the differences between correlations, and the percentage of these differences. The table indicates that in all cases the "after" correlations were higher than the "before" correlations. The correlation differences were analyzed by *t-tests* for correlation differences in matched samples (Weinberg and Goldberg, 1990). The differences in correlations between the "before" and "after" scores within groups were similar for the two groups, but they were higher for the verbal than for the quantitative sub-test. Only in a single case (the verbal sub-test in the research group) the "after" correlation was significantly larger than the "before" correlation..

TABLE 4 **Predictive Validity Data by Group: Correlations Between "Before" and "After" Scores and the Criterion; Differences[1] Between Correlations and Percentage[2] of Differences**

Score	**Research Group** Coached				**Control Group** Uncoached			
	Before	After	Dif	% Dif	Before	After	Dif	% Dif
V	.328	.404	.076*	23	.463	.557	.096	20
Q	.389	.421	.032	8	.475	.529	.054	11
VQ	.409	.469	.060	15	.548	.600	.052	9

* Significant (p < .10, two-tailed), t-test

1 - "after" correlation minus "before" correlation

2 - added percentage to the "before" correlation

An additional statistical method, based on confidence intervals obtained by *bootstrap* simulations (Efron, 1979, 1982), was used to examine the significance of the changes in validity within and between groups. The bootstrap method is a nonparametric resampling (with replacement) method which, like the *jackknife* nonparametric resampling method, can be used for estimating the sampling distribution of a statistic on the basis of the sample, without making any parametric assumptions about its distribution in the population. According to Robertson (1991), confidence intervals obtained by bootstrap simulations perform better than those obtained by jackknife simulations. Confidence intervals were computed by drawing k = 1000 samples sized = n group size (207 in the research group, 67 in the control group_, for each group. The correlations between the predictor and the criterion were computed in each sample for the "before" and "after" scores. The k differences between the "before" and "after" correlations served to define the 5th and the 95th percentiles. The results of the bootstrap simulation are presented in Table 5.

TABLE 5 **90% Confidence Intervals for the Gains in Predictive Validity* Computed by the Bootstrap Method (k=1000), Within Each Group**

Score	**Research Group** ***Interval***		**Control Group** ***Interval***	
	5th percentile	**95th percentile**	**5th percentile**	**95th percentile**
V	-.0112	.1712	-.0096	.2132
Q	-.0448	.1196	-.0040	.1196
VQ	-.0051	.1252	-.0116	.1252

* Gains in predictive validity were defined as the "after" correlation minus the "before" correlation

Although the mean gains in predictive validity were positive in both groups (see **Table 4**), the differences were not statistically significant according to the bootstrap confidence intervals. The findings obtained by the bootstrap method were almost identical to the findings obtained by the t-test, thus reinforcing the conclusion that predictive validity of PET was not affected by coaching.

Between Groups Differences in Predictive Validity Gains

An additional bootstrap analysis was conducted to determine whether the differences between the two groups in the observed gains in predictive validity were statistically significant. In each bootstrap sample, the correlation change in the control group was subtracted from the correlation change in the research group, and the 90% confidence interval for this difference was computed on the basis of 1000 bootstrap samples. The findings, which are presented in **Table 6,** indicate that no significant differences between the groups were obtained. Also, it should be noted that the 50th percentile is very close to zero. Thus, the conclusion that can be drawn from these results is that coaching does not have an effect on the gains in the predictive validity of PET.

TABLE 6 **90% Confidence Intervals and Median for the Differences in Predictive Validity Gains Between the Control and the Research Group Computed by the Bootstrap Method (k=1000)**

Score	***Interval***		
	5th percentile	**50th percentile**	**95th percentile**
Verbal	-.1439	-.0172	.1132
Quantitative	-.1411	-.0264	.1004
VQ	-.0842	.0107	.0995

Effect of Coaching on Predictive Validity for Varying Proportions of Coached and Uncoached Examinees

The within group results show that no significant gains in predictive validity were obtained within each group, and gains were similar in the coached and uncoached groups. It is therefore unlikely that a coaching effect will be found for a group composed of coached and uncoached examinees. Nevertheless, we examined this possibility on the basis of confidence intervals obtained by bootstrap simulations. Five varying proportions of coached and uncoached examinees were analyzed: 1:4 (coached : uncoached), 1:2, 1:1, 2:1, 4:1. We composed **k** artificial groups sized **n** = 60, which included participants from the two groups in the desired ratio (e.g., 12 coached and 48 uncoached examinees for the 1:4 ratio). No significant differences were found between the predictive power of any of the "before" and "after" scores in any of the simulated groupings of coached and uncoached examinees.

Effect of Coaching on Bias

Testing for bias was accomplished by analyzing the marginal increase in predictive validity resulting from the use of two regression lines, one for coached examinees and one for uncoached examinees, as compared with the use of a single (common) regression line for the two groups. The method used was Step-Down Hierarchical Multiple Analysis (Lautenshlager & Mendosa, 1986). According to this method, the following four models are defined:

Four Models for Regression Lines

Model 1 - One regression line	Y = B0 + B1T
Model 2 - Two regression lines differing in constant and slope	Y = B0 + B1T + B2D + B3DT
Model 3 - Two regression lines differing in slope	Y = B0 + B1T + B3DT
Model 4 - Two regression lines differing in constant	Y = B0 + B1T + B2D

Where: Y - criterion, T - predictor, D - dummy var: 1 research, 0 control, DT - interaction var.
B0 - constant, B1 - slope, B2 - difference between constants, B3 - difference between slopes.

The first comparison is made between the first two models; only if the marginal increase is significant are other comparisons made. The assumption was that before coaching, there would be no significant marginal increase between the proportion of variance explained by Model 2 relative to Model 1. The important question is: What happened after coaching? **Tables 7a & 7b**, show the percentage of explained variance in predicting the criterion through the use of PET scores for the four models, before and after coaching, respectively. In both tables, the critical value for F is 3.04 (df_1 = 2, df_2 = n-4 = 270; p=.05). As expected, the differences between Model 1 and Model 2 in the explained variance before coaching are far from significant.

The results displayed in **Table 7b** indicate that the marginal increase in predictive validity resulting from the use of two regression curves, one for coached examinees and one for uncoached examinees, was not significant for the "after" condition scores. Thus, it can be concluded that use of a single regression line for a combined population of coached and uncoached examinees does not create bias against the uncoached group.

TABLES 7a&7b **Percentage of Explained Variance in Predicting the Criterion Through the Use of PET Scores for the Four Models**

Results of Step-Down Hierarchical Multiple Regression Analysis

a. Test 1 - “Before” Condition

Score	**Percentage of Explained Variance by:**				**Significance of Difference Between Model 1 and Model 2 (F)**
	Model 1	**Model 2**	**Model 3**	**Model 4**	
V	.135232	.136636	.135621	.135815	F = .220
Q	.171725	.172473	.172125	.172236	F = .122
VQ	.200244	.201569	.200258	.200318	F = .224

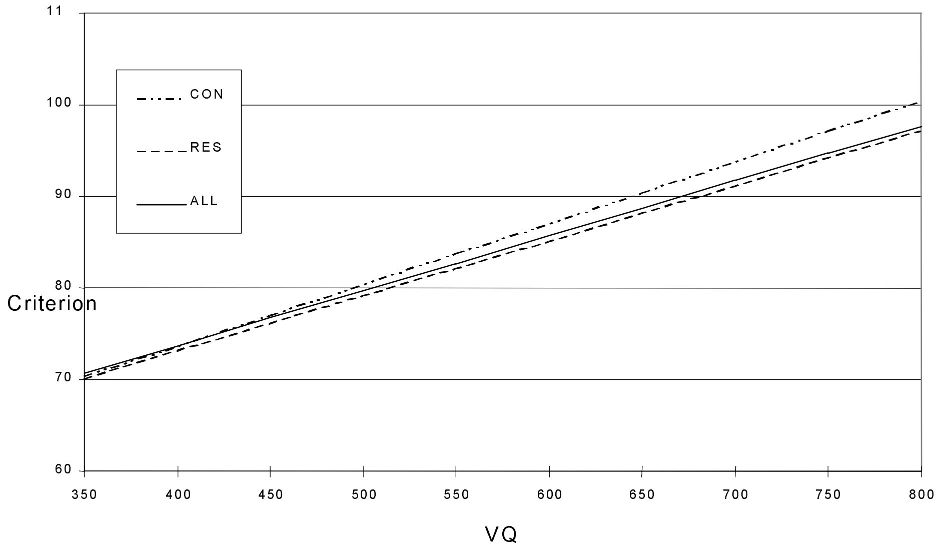
b. Test 2 - “After” Condition

Score	**Percentage of Explained Variance by:**				**Significance of Difference Between Model 1 and Model 2 (F)**
	Model 1	Model 2	Model 3	Model 4	
V	.198140	.200842	.199339	.198953	F = .456
Q	.201999	.205004	.204707	.204385	F = .510
VQ	.250604	.255824	.255592	.255192	F = .947

Figure 1 displays the three regression lines (a line for each group according to Model 2, and the common regression line computed across groups) for predicting the criterion scores from the VQ “after scores.” The fact that the line computed for the control group is located above the other two lines means that there is a tendency to underpredict the criterion scores of the uncoached examinees. As we have already seen, this tendency is small and not statistically significant.

Figure 1 - Regression Lines Computed Within Each Group and Across Groups for Predicting the Criterion by the VQ After Scores.

CON- Control Group, RES - Research Group, ALL - All Participants



Sampling

The results revealed that the initial predictive validity was higher in the control group than in the research group (see **Table 4**). This difference, along with the initial differences found for the mean scores (**Table 2**) indicates that although subjects were randomly allocated to the experimental and control conditions, these two groups were not equivalent. The bias findings show that, despite the differences between the two groups, the two regression lines for predicting the criterion before coaching were almost identical (see **Table 7a**). Nevertheless, to check whether our results and conclusions might have been affected by the initial differences between the two groups, an additional analysis was conducted using a subsample (n = 136) of the research group. The subsample was selected so that the distribution of its "before" scores, and the correlation of its "before" scores with the criterion, would resemble the control group as closely as possible. **Table 8** compares the findings for the three groups (research, research subsample and control). The first three rows display the resemblance between the subsample and the control group; the last two rows indicate that results were not affected by the initial differences between the two groups: The mean gain score in the research subsample is even greater than that obtained for the entire research group, and the gain in predictive validity is only slightly smaller (0.037 vs. 0.060). These two values are very similar to the predictive validity gain in the control group (0.052). Thus, the general conclusions drawn from the results do not seem to be affected by the initial differences between the groups.

TABLE 8 **Findings for the Research Subsample, Research, and Control Groups**

VQ		**Research Group**	**Research Subsample Group**	**Control Group**
Mean Score "Before"		541.3	521.1	518.5
Standard Deviation		68.2	70.8	73.8
Correlation with Criterion	"Before"	0.409	0.549	0.548
	"After"	0.469	0.586	0.600
	Difference	0.060	0.037	0.052
Mean Score Gain Following Coaching		24.5	29.2	0.7

CONCLUSIONS

The estimate of the coaching effect (about 1/5 s.d) obtained in this study is similar in magnitude to estimates obtained in meta-analytic studies. This similarity reinforces the generalizability of the study findings. The conclusion that can be drawn from the study results is that coaching does not have an effect on the predictive validity of PET or other similar scholastic aptitude tests. The theoretical conclusion is that coaching does not change the relative weights of the factors which are relevant and not relevant to the criterion. There is no bias when there is one regression curve for a combined population of coached and uncoached examinees.

REFERENCES

Allalouf, A. (1984). *Examinee feedback questionnaire, April 1984*. (In Hebrew). Report No. 7, National Institute for Testing and Evaluation, Jerusalem

Arieli, M. (1994). *Examinee feedback questionnaire, March 1994*. (In Hebrew). Technical Report No. 32, National Institute for Testing and Evaluation, Jerusalem

Bashi, Y. (1976). *Verbal and non-verbal abilities of 4th, 6th and 8th grade students in the Arab educational system in Israel*. (In Hebrew). Jerusalem: Hebrew University, School of Education.

Baydar, N. (1990). Effects of coaching on the validity of the SAT: Results of a simulation study. In W. W. Wilingham, C. Lewis, R. Morgan, and L. Ramist (Eds.), *Predicting college grades: An analysis of institutional trends over two decades.* Princeton, New Jersey: *ETS.*

Beller, M., Ben-Shakhar, G. (1994). *Pre-academic preparatory studies in Israel*. A paper presented at the 20th annual IAEA Conference, on "Bridging the Gap", October 17-21,1994, Wellington, New Zealand.

Bond, L., (1989). The effects of special preparation on measures of scholastic ability. In R. L. Linn (Ed.). *Educational measurement* (Third edition). New York: Macmillan.

Caruzo, D. R., Taylor, J., and Detterman, D. K. (1982). Intelligence research and intelligent policy. In D. K. Detterman and R. J. Sternberg (Eds.). *How and how much can intelligence be increased.* New Jersey: Ablex.

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

Efron, B. (1979). Introduces the bootstrap to the world. *Annual statistics, 7*, 1-26.

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. In *Regional Conference Series in Applied Mathematics, No. 38.* Philadelphia: SIAM.

ETS, (1965). *Effects of coaching on scholastic aptitude tests scores.* New York: College Entrance Examination Board.

Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences.* London: MacMillan.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Macmillan.

Jones, R. F. (1986). A comparison of the predictive validity of the MCAT for coached and uncoached students. *Journal of Medical Education, 61,* 325-338.

Lautenshlager, G. J., and Mendosa, J. L. (1986). A step down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. *Applied Psychological Measurement, 10,* 165-172..

Marron, J. E. (1965). Preparatory school test preparation: Special test preparation, its effect on College Board scores and the relationship of affected scores to subsequent college performance. West Point NY: United States Military Academy.

Messick, S. (1981). The controversy over coaching: Issues of effectiveness and equity. In B. F. Green (Ed.) *Issues in testing: Coaching, disclosure & ethnic bias.* San Francisco: Jossey Bass.

Millman, J., Bishop, C. H., and Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25,* 707-726.

Oren, C. (1993). *On the effect of various preparation modes on PET scores*. (In Hebrew). Report No. 170, National Institute for Testing and Evaluation, Jerusalem

Ortar, G. R. (1960). Improving test validity by coaching. *Educational Research, 2,* 137-142.

Pearson, K. (1904). On the laws of inheritance in man. *Biometrika, 3,* 90-131.

Powers, D. E. (1985) Effects of test preparation on the validity of Graduate Admission Test. *Applied Psychological Measurement, 9,* 179-190.

Powers, D. E. (1993). Coaching for the SAT: Summary of the summaries and an update. *Educational Measurement: Issues and Practice, 12,* 24-30.

Robertson, C. (1991). Computationally intensive statistics. In P. Lovie, and A. D. Lovie (Eds.) *New developments in statistics for psychology and the social sciences. Vol 2.* Exeter: BPCC Wheatons.

Terman, L. M. (1916). *The measurement of intelligence.* Boston: Houghton.

Weinberg, S. L., and Goldberg, K. P. (1990). *Statistics for the behavioral sciences*. Cambridge: Cambridge University Press.