`

The Effect of Coaching on the Predictive Validity of Scholastic Aptitude Tests

Avi Allalouf

National Institute for Testing and Evaluation, Jerusalem

Gershon Ben Shakhar

The Hebrew University of Jerusalem

Author Notes

Avi Allalouf, National Institute for Testing and Evaluation, Jerusalem. Gershon Ben Shakhar, The Hebrew University of Jerusalem.

The research presented in this manuscript was supported by the National Institute for Testing and Evaluation, The Society For Advancement of Education in Israel, and the Israeli Ministry of Education. We would like to thank Michal Beller, Yoav Cohen, Naomi Gafni, Ronald K. Hambleton, as well as two anonymous reviewers for their helpful comments on an earlier version of this paper, and Chava Cassel for her editorial help. Special thanks are due to Hagit Efrati Bar-Gai for her assistance throughout the study.

Correspondence concerning this article should be addressed to Avi Allalouf, National Institute for Testing and Evaluation, P.O.B 26015, Jerusalem 91260, Israel, E-mail: avi@nite.org.il

Suggested Running Head: COACHING AND PREDICTIVE VALIDITY

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. The conclusions 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.

The Effect of Coaching on the Predictive Validity of Scholastic Aptitude Tests

The question of whether intelligence and scholastic aptitude test scores could be affected by interventions has been extensively discussed (e.g., Bond, 1989; Brody, 1992; Caruzo, Taylor, & Detterman, 1982; Spitz, 1986). Until about twenty years ago, the commonly held view was that improvement due to coaching (In this paper the term "coaching" is used to refer to all types of test 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 differences in the student coached" (ETS, 1965, p. 4). 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 and Jungeblut, 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 several weeks of coaching, are generally small and the mean gain on the SAT (Scholastic Assessment Tests, which consists of a verbal and a mathematical section), according to these meta-analyses is approximately one fifth of a standard deviation (beyond the gain that would be expected as a result of retesting only, which is, according to Donlon, 1984, about one seventh of a standard deviation). 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 (about one fourth of a standard deviation) than for the verbal section (about one sixth of a standard

deviation). According to Messick and Jungeblut (1981) the improvement which resulted from the first 20 hours of coaching is about 20% of a standard deviation in the mathematical subtest, and about 12.5% of a standard deviation in the verbal subtest. The number of hours needed to double these gains is estimated as 120 in the mathematical subtest and 250 hours in the verbal subtest.

Special preparation is particularly common for scholastic aptitude entrance exams to institutes of higher learning. For example, in the United States, according to Powers (1988), 11% of the SAT examinees in 1986-87 took coaching courses, and 41% used preparation books. In Israel, the number of examinees taking coaching courses for the PET has dramatically increased from 1% in 1984 (the first administration of PET), to 42% in 1990 and to 77% in 1996. In 1996, 90% used preparation books. (Allalouf, 1984; Stein, 1990; Arieli, 1996).

Coaching involves three interrelated elements: (1) Acquiring familiarity with the test (i.e., getting acquainted with the test instructions, item types, time limits, and answer sheet format), which can be 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, (2) Reviewing material which is relevant to the test's contents, for example, learning mathematics when the test contain mathematical reasoning, and (3) Learning testwiseness (TW), which can be defined as: "subject capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score" (Millman, Bishop, & Abel, 1965, p. 707). Four TW strategies, independent of test content or purpose, have been identified by Millman et al. (1965): efficient use of the available time, error avoidance, guessing and deductive reasoning.

Many studies on coaching for scholastic aptitude tests have dealt with the SAT. These studies focused primarily on the effects of both commercial and noncommercial coaching on test scores. Many researchers, among them Messick and Jungeblut (1981), Anastasi (1981) and Bond (1989), have raised the question of the possible detrimental effects of coaching on test validity. 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." Surprisingly, however, few research efforts have been devoted to studying the effects of coaching on the validity and fairness of scholastic aptitude or intelligence tests.

The earliest study dealing with the effect of coaching on predictive validity was conducted by Ortar (1960). The Triangle[1] Test 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 of 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 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 since coaching is a learning process, the after-coaching scores better reflect 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 also showed small but statistically 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. Mean Score gains were very high (about three quarters of a standard deviation). Marron found that in some of the preparatory programs, in which the mean gain due to coaching was higher than in others, coaching led to an overprediction of academic performance.

Powers (1985) examined the effects of variations in the number of preparation hours 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. 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 (it was not reported 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 simulation 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 the percentage of coached examinees. Freshman Grade Point Average (FGPA) was used as the criterion and the simulation indicated that, at most, only ten percent of the decline in predictive validity could be explained by the increase in coaching density.

In contrast to the concern raised by Bond (1989) that coaching could adversely affect predictive validity, most of the above mentioned 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, the empirical studies suffer three problems: (1) Insufficient information in most of the studies about whether or not examinees actually underwent coaching and the intensity of the coaching; (2) The coaching in some of the studies consisted of only a few hours, and therefore cannot be compared with commercial courses which offer much more intensive practice; (3) In some of these studies, examinees were not randomly selected into the coaching programs, and no control group were used. This may explain some of the differences in the findings of these studies. In addition, most participants in the studies conducted 30 years ago were unfamiliar with the types of questions 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. It should be noted also that some of the studies focused on intelligence tests rather than scholastic aptitude tests. 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 of scholastic aptitude tests.

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. With the exception of Marron (1965) and Jones (1986), this matter was generally not dealt with in the studies mentioned above.

This study was designed to examine the effect of coaching on predictive validity and fairness (or bias) of scholastic aptitude tests. Two main forms of test bias have been discussed in the literature (see Millsap, 1995). Measurement bias which refers to the relationship between the test and latent variables measured by it, and bias in prediction which refers to the relationship between the test and a relevant criterion. The present study is focused only on the second type of test bias, and to examine whether coaching creates bias in prediction, we adopted the definition of bias proposed by Cleary (1968), known as the regression model. In other words, we intend to examine whether the criterion scores of the uncoached group are systematically underpredicted by their test scores, relative to the coached group. The method proposed by Lautenshlager and Mendosa (1986), on the basis of the regression model, was applied in the present study to examine whether the test is biased against the uncoached group

The findings should provide an empirically-based answer to the oft-heard public criticism of these tests, 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 actually be desirable. From the 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 (if such bias is demonstrated).

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, lasting usually one year and providing 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 Psychometric Entrance Test (PET) scores in order to be accepted to universities. Both matriculation certificate and PET scores serve as criteria for admitting students into most institutes of higher learning in Israel. All participants had some familiarity with PET because they took this test before starting the pre-academic program. Our initial sample consisted of 366 students from the eight pre-academic preparatory institutes. Because the major question examined in this study was whether and to what extent predictive validity might be affected by coaching, we decided to allocate most of the participants (about 75%) to the experimental (coaching) condition, and to use a relatively small control (uncoached) group. Thus, the participants in each institute were randomly allocated, with a 3 to 1 ratio, to the experimental and control conditions, respectively, and the total number of participants was 271 in the experimental condition, and 95 in the control condition. All 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.

Design

PET was administered to the research group which then participated in a coaching course that lasted approximately one and a half months. Following the course they were

retested with another, parallel, form of the test. The control group was also given two forms of the test, but without attending a coaching course in between. The time interval between the two 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 reasoning (V), quantitative reasoning (Q), and English proficiency (E). Each subtest was scored separately and standardized on a scale which was determined, on the basis of the population of examinees who took PET in 1984, to have a mean of 100 and a standard deviation of 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 (no coaching was provided for English proficiency). The VQ score was standardized on a scale whose original distribution was [500,100]. The experimental design is shown in Table 1

Insert Table 1 about here

The Coaching Program

A special coaching course was designed for the purposes of this study. We considered to cooperate with commercial companies, but decided no to do so because we wanted to have full control. In addition, we anticipated that the commercial companies would not agree to participate in a study that might show the real distribution of the gain following coaching, which is considerably less than the gain they claim for. Course instruction dealt only

with the verbal and quantitative subtests of PET (and not with the English subtest which is mainly achievement based) . Most of the items which where used in the course were operational items and 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. In each study unit, a specific item type was explained to the students. Then they solved some representative items, followed by additional and more advanced explanation. This process took place two or three times and finally a short test was given on the specific item type. After each meeting, relevant assignments were given to the students, to be handed in by the next meeting .For example, the item type "verbal analogies" was explained by, first, introducing several examples of verbal analogies which were solved by the instructor who explained the solutions. Then, verbal analogies were classified into about ten subtypes, such that each subtype is based on a specific relation between the two words in the stem (e.g., a contained in b, a was once b, a needs b in order to do something, a and b are two equal members in the same family). This classification is very helpful to the students because it makes them search for the specific relation between the two words that comprise the stem of the analogy they have to solve. This explanation was followed by several word-analogy items the students had to solve finally a timed test was given on verbal analogies. The course duration was designed to last about 40 hours: 27 hours in class, and about 15 hours at home; Verbal and quantitative about 20 hours each. Ten experts in lesson planning and teaching in the relevant fields reviewed the study units during the various preparation stages. Because the purpose was to design a course which is similar in nature to commercial courses in Israel and in the US, some commercial coaching books and teachers served as an “inspiration source”. 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[2] average of the study participants' scores on the matriculation exams was used as a criterion for validation. It was computed by the same method, and measured on an identical scale for all participants. Some of the scores were obtained from matriculation exams taken before entering the preparatory institute, while other scores were obtained from matriculation exams taken after completing the preparatory institute and before entering the university. Most of the matriculation subjects are mandatory for university acceptance (e.g., Hebrew language, Mathematics, English, Biblical studies) and therefore are common to all study participants. Thus, this criterion, which differs from GPA obtained in college, may be regarded as a concurrent criterion. It should be noted that for the purpose of this study, the matriculation criterion has several advantages over the more commonly used freshman GPA. First, unlike freshman GPA, which may strongly depend on the specific university, and the specific program chosen by each student, the matriculation average is relatively standard for all participants. Second, the matriculation examinations are achievement tests based primarily on open ended, rather than on multiple choice questions, which enhance the generalizability of the study results to performance-based criteria. Unfortunately, we were unable to retrieve the criterion measure for all 366 participants and therefore the initial sample was reduced by about 25%. We ended up with 207 participants in the experimental condition (76% of the initial 271 participants), and 67 participants in the control condition (71%).

Results

Effect of Coaching on Test Scores

Table 2 presents the mean scores obtained by each group in each administration of the test (before and after coaching), the mean gain scores (i.e. scores on Test 2 minus scores on Test 1) of each group and the differences between the gain scores of the two groups. As indicated in Table 2, 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 presents a question regarding the equivalence of the two groups, that deserves consideration. This issue will be discussed and elaborated on (see the section entitled "Sampling").

Dependent samples t-tests were used to test whether the gains were statistically significant. These comparisons revealed that gain scores exceeded 0 (at a statistically significant level) in the coached group [for the main predictor VQ: $t(206)= 8.42$], but not in the control group [for the main predictor VQ: $t(66)= .15$]. The effect of coaching on test scores was defined as the difference between the mean gain obtained by the research group and that obtained by the control group, and t-tests for independent samples (comparing the gains obtained in the experimental and the control groups) were conducted to test whether statistically significant effects of coaching on test scores were obtained. Gains in the two subtests, as well as the total score (VQ), were much larger in the coached group than in the control group, and in all three cases the coaching effects were statistically significant [$t(272) =$ 2.43, 3.56, and 4.08, for V, Q and VQ, respectively]. In the English subtest, for which no coaching was provided, the difference was not statistically significant [$t(272) = 0.78$].

Insert Table 2 about here

These findings indicate that coaching has an effect on test scores. 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 of this topic, which included both commercial and noncommercial coaching programs (i.e. Powers, 1993). This similarity indicates that the coaching program in this study was as effective as other 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 time interval between the two tests, and by examinee familiarity with the test before entering the preparatory program.

It was expected that the effect of coaching on the scores of the coached group would reduce the similarity between the "before" and the "after" scores in this group as compared with the uncoached group (assuming that the coaching effect is not constant across all examinees). Indeed, the "before-after" correlations, which are presented in Table 3, were lower for the research group than for the control group (where the correlations express test-retest reliabilities), showing that the coaching effect was not uniform. The correlation difference was statistically significant only for the quantitative section of the test. In English, where there was no coaching, the correlation difference was near zero.

Insert Table 3 about here

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 of the two groups.

The mean and standard deviation of the criterion measure were 83.29 and 8.89, respectively, for the research group, and 81.84 and 8.53 for the control group. Table 4 presents the validity data by groups and by sub-tests, the differences between correlations, and the percentages 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). In all cases and in both groups (except English for which no coaching was provided) the "after" correlations exceeded the "before" correlations, but none of the correlation differences was statistically significant. The correlation differences in both groups were somewhat larger for the verbal as compared with the quantitative sub-test. The differences in correlations between the "before" and "after" scores within groups were similar for the two groups.

Insert Table 4 about here

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 between groups. The findings indicate that no statistically significant differences between the groups were obtained: the 90% confidence interval for the differences in predictive validity of VQ was [-.0842 - .0995] with a median of 0.0107 which is very close to zero. The findings obtained by the bootstrap method and the t-test lead to the conclusion that predictive validity of PET was not affected by coaching.

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, and for the criterion measure (Table 2), indicates that although subjects were randomly allocated to the experimental and control conditions, these groups were not equivalent. The analysis of test bias in the following section shows that, despite the differences between the two groups, the two regression lines for predicting the criterion before coaching were almost identical, but 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. We used a subsample (n=136) of the research group which 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. The results for the three groups (research, research subsample and control) are presented in Table 5. 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.

Insert Table 5 about here

Effect of Coaching on Bias

Test bias was examined 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. Before applying this analysis, the error variances of the two groups were compared by the method recommended by DeShon & Alexander (1996). The variances of the two groups were relatively equal on all measures (V, Q, E, and VQ), and the variance ratio has never exceeded the 1.5 limit proposed by DeShon & Alexander (1996). Thus, it is not essential to apply alternative (to the F-test) procedures for testing regression slopes homogeneity. Therefore the method of Step-Down Hierarchical Multiple Analysis (Lautenshlager & Mendosa, 1986) was applied and the following four models are defined:

Four Models for Regression Lines

Model 1 - One regression line	$Y = B_0 + B_1T$
Model 2 - Two regression lines differing in constant and slope	$Y = B_0 + B_1T + B_2D + B_3DT$
Model 3 - Two regression lines differing in slope	$Y = B_0 + B_1T + B_3DT$
Model 4 - Two regression lines differing in constant	$Y = B_0 + B_1T + B_2D$

Where: Y - criterion, T - predictor, D - dummy var: 1 research, 0 control, DT - interaction var.

B_0 - constant, B_1 - slope, B_2 - difference between constants, B_3 - difference between slopes.

The first comparison is made between the first two models; only if the marginal increase is statistically significant are other comparisons made. The assumption was that there

would be no statistically significant marginal increase between the proportion of variance explained by Model 2 relative to Model 1 for the "before coaching" scores. But the critical question addressed by this analysis relates to the "after coaching scores. Tables 6a & 6b, 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. As expected, the differences between Model 1 and Model 2 in the explained variance before coaching are not statistically significant.

More importantly, the results displayed in Table 6b 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 statistically significant for the "after" condition scores as well. Thus, it can be concluded that use of a single regression line for a combined population of coached and uncoached examinees does not create any bias against the uncoached group.

Insert Tables 6a & 6b about here

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 control group regression line is located above the other two lines means that there is a slight tendency to underpredict the criterion scores of the uncoached examinees. As we have already seen, this tendency is not statistically significant.

Insert Figure 1 about here

The regression slopes of the two groups, on all measures, were compared also by Alexander's Normalized-t statistic (DeShon & Alexander, 1996). The results of this analysis are in complete agreement with those obtained by the F test, and indicate that in all cases the difference between the regression lines of the two groups was not statistically significant.

Discussion

Most of the research on coaching for scholastic aptitude tests has focused on the effect of coaching on test scores, while few attempts have been made to study the effects of coaching on the validity and fairness of these tests. The present study was designed to examine whether coaching has an adverse effect on predictive validity and fairness of scholastic aptitude tests. Two randomly allocated groups, coached and uncoached, were compared, and . The results showed that in both groups the predictive validities of the "after scores" were higher (though not at a statistically significant level), than the predictive validities of the "before scores" and that the gains in validity were similar for the two groups. These findings indicate that coaching has no effect on predictive validity. Moreover, the results demonstrated that the linear function for predicting the criterion from scholastic aptitude test scores is not affected by coaching (i.e. there is no bias when a single regression line is used for the combined population of coached and uncoached examinees). The main conclusion that can be drawn from the present results is that coaching does not seem to affect predictive validity and fairness of the Israeli Psychometric Entrance Test (PET). This conclusion refutes critics' claims that coaching

reduces predictive validity and creates bias against the uncoached examinees in predicting the criterion.

Our principal conclusion concerns the educational institutions which use scholastic aptitude tests for selecting students. Our results show that the selection process can include both coached and uncoached examinees, in differing proportions, without affecting predictive validity or test fairness. Because the within group results show that no statistically significant gains in predictive validity were obtained within each group, and gains were similar in the coached and uncoached groups, it is unlikely that validity would be effected by coaching in groups composed of both coached and uncoached examinees. Because a slight improvement in predictive validity (though not statistically significant) was observed in the coached group, it is even possible that if all examinees were to undergo a special test preparation (to an equal extent), predictive validity might increase. This increase may be explained by an improvement, resulting from coaching, of the inaccurately low scores due to poor test taking skills.

Three possible methodological and statistical criticisms of the current study may be raised: (1) Predictive validity was not affected by coaching because the manipulation (the learning program) was ineffective. However, our estimate of the coaching effect on test scores, which was between 1/5 and 1/4 standard deviations is similar in magnitude to estimates obtained in the meta-analytic studies of Messick and Jungeblut (1981) and Powers (1993). This similarity supports and strengthens the internal validity of the present study by confirming that the manipulation was effective. (2) Results were affected by the fact that the two groups (coached and uncoached examinees) were not equivalent. Indeed, the results revealed that the initial predictive validity was higher in the control group than in the research group. This difference, along with the initial differences found in the mean scores, indicates that although

subjects were randomly assigned to the experimental and control groups, the two groups were not equivalent. Two answers can be provided for this sampling problem: (a) the analysis of test bias demonstrates that, despite the differences between the two groups, the two regression lines for predicting the criterion before coaching were almost identical in the two groups, and (b) an additional analysis conducted on a subsample of the research group, which was selected to resemble the control group as closely as possible, revealed that results were not affected by the initial differences between the two groups. (3) the statistical test for the differences between correlations did not have sufficient statistical power. The VQ validity results of the coached group served for a statistical power analysis regarding the t-test for the differences between correlation coefficients (after Cohen, 1977; see Appendix A). Statistical power is considered sufficient when it is at least 0.8, and therefore our power analysis demonstrates that the t-test for correlation differences had low statistical power for detecting small validity changes (0.10), and sufficient or high power for detecting medium or large validity changes (0.15 and above). Moreover, it should be noted that the observed differences in predictive validity were in the opposite direction to the hypothesis that coaching hurts predictive validity, and therefore this hypothesis seems unlikely in light of the present findings

The results of this study are generalizable to other types of test preparation and to other types of scholastic aptitude tests, for five reasons: (1) The type of coaching used in the study - familiarization with the types of questions, providing opportunity for practice, and self-examination using tests of a similar format - was very similar to what examinees do when they participate in commercial coaching courses, or use preparation books. (2) The PET which was administered in this study is similar in content to other scholastic aptitude/assessment tests, such as the SAT and the GRE. For example, Oren (1984), found a correlation of 0.77 between

the VQ score of an earlier PET version and the SAT total score. The current PET, which is more similar to the SAT (see Beller, 1994, for a description), is likely to correlate even higher with the SAT. (3) The score distribution in the study population was similar to that of PET examinees[3], which does not differ greatly from that of other populations which take scholastic aptitude tests. (4) The mean gain due to coaching was similar to the one reported by Messick & Jungeblut (1981) and Powers (1993) meta-analytic studies, and (5) The criterion used was mostly performance based and not multiple choice based, which make it similar to the criterion of interest. Nonetheless, in applying these results to other situations, the following three qualifications and reservations should be kept in mind: (a) The conclusions are applicable to situations in which uncoached examinees have some familiarity with the test content and structure (b) The conclusions are applicable to conventional scholastic aptitude tests which include verbal and quantitative sections, and may not apply to tests which differ greatly from the Psychometric Entrance Test in their content and structure. (c) The conclusions are applicable when the validity criterion is based on scores. However, when different criteria for validating aptitude tests are used, such as dropping out from school, or success at the work place, our conclusions might not be applicable.

Critics of scholastic aptitude tests typically claim that preparation improves scores on these tests by teaching examinees special techniques for solving multiple choice items. Consequently, according to these critics, the validity of the tests is adversely affected by such preparation, and the tests are biased against examinees who cannot afford expensive coaching programs. The results refutes these claims and they 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.

.

.

References

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

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. American Psychologist, 36, 1086-1093.

Arieli, M. (1995). Pre-academic preparatory studies . (In Hebrew). Technical Report No. 42, National Institute for Testing and Evaluation, Jerusalem

Arieli, M. (1996). Examinee feedback questionnaire, March 1996. (In Hebrew). Technical Report No. 51, 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. (1994). Psychometric and social issues in admissions to Israeli universities. Educational Measurement: Issues and Practice, 13, 12-21.

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.

Brody, N. (1992). Intelligence. Sun Diego: Academic Press.

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.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. Hillsdale. N.J.: Lawerence Erlbaum Associates.

DeShon, R. P., Alexander, R. A. (1996). Alternative procedures for testing regression slope homogeneity when group error variances are unequal. Psychological Methods, 1, 261-277

Donlon, T. F. (Ed.) (1984). The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests. New York: College Entrance Examination Board

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.

The Hebrew University (1994). Information booklet. (In Hebrew). Jerusalem: The Hebrew University.

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., and Jungeblut, A. (1981). Time and method in coaching for the SAT. Psychological Bulletin, *89*, 191-216.

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

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. Multivariate Behavioral Research, 30, 577-605

Oren, C. (1984). The ccorrelation between SAT and PET scores. (In Hebrew). Report No. 17, National Institute for Testing and Evaluation, Jerusalem

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.

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

Powers, D. E. (1988). Preparing for the SAT: A survey of programs and resources. (College Board Rep. No. 88-7). New York: College Entrance Examination Board.

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

Spitz, H. H. (1986). The raising of intelligence. New Jersey: Lawrence Erlbaum Associates

Stein, M. (1990). Examinee feedback questionnaire, April 1990. (In Hebrew). Report No. 130, National Institute for Testing and Evaluation, Jerusalem

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

Appendix A

Statistical Power Analysis for the Differences between Correlation Coefficients

The results of this study reveal that no statistically significant changes in predictive validity occurred between the two administrations of the test, in both the coached and the uncoached groups. This raises the issue of statistical power, because no statistically significant differences may be the result of weak statistical power. The VQ validity results of the coached and uncoached groups served for a statistical power analysis of the statistical test for the differences between correlation coefficients (after Cohen, 1977). According to the validity results, the "before correlation" between VQ and the criterion was 0.409 in the coached group, and 0.548 in the uncoached group (see table 4). According to the null hypothesis no change in this correlation was expected. Four possible correlation differences ("after correlation" minus "before correlation") were considered in the power analysis (0.10, 0.15, 0.20 and 0.25), and the significance level was 0.05. The results of the power analysis which are displayed in Table 9, indicate that our t-test for correlation differences within groups had low statistical power for detecting small validity changes (0.10), but had sufficient statistical power for detecting medium changes in validity (0.15) in the coached condition. The statistical power was high for detecting large changes in validity (0.20 and above) in both conditions. There is no conventional way for estimating statistical power for the bootstrap method because no parametric assumptions about the correlation distribution in the population were made.

Insert Table B1 about here

TABLE 1

Experimental Design

Group	Test 1	Course	Test 2	Course
Research	yes	yes	yes	--
Control	yes	no	yes	yes [a]

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

TABLE 2

Mean Scores and Gains, by Groups

(Standard deviations appear in brackets)

Score	Research Group Coached			Control Group Uncoached			Coaching Effect [b]
	Test 1 Before	Test 2 After	Gain [a]	Test 1 Before	Test 2 After	Gain [a]	
Verbal	107.69	110.69	3.00*	103.36	102.84	-0.52	3.52*
	(13.9)	(14.6)	15.0%	(15.9)	(15.6)	-2.6%	17.6%
Quantitative	107.57	113.78	6.21*	103.31	104.10	0.79	5.42*
	(15.2)	(15.0)	31.0%	(16.4)	(16.1)	3.9%	27.1%
English	107.24	109.37	2.13*	100.87	101.93	1.06	1.07
	(15.2)	(17.2)	10.6%	(16.2)	(15.7)	5.3%	5.3%
VQ	541.30	565.76	24.46*	518.51	519.21	0.70	23.76*
	(68.2)	(69.2)	24.5%	(73.8)	(76.1)	0.7%	23.8%

a Expressed as the score difference, Test 2 minus Test 1, in the first line, and as percentages of the population standard deviation (20 for V, Q and E, 100 for VQ), in the second line.

b Expressed as the gain difference, Research Group Gain minus Control Group Gain, in the first line, and as percentages of the population standard deviation (20 for V, Q and E, 100 for VQ), in the second line.

* significant ($p < .05$)

TABLE 3

Correlations Between "Before" and "After" Scores by Groups and Subtests

Score	Research group Coached	Control Group Uncoached	Difference Between Correlations
Verbal	.738	.798	.060
Quantitative	.725	.826	.101*
English	.829	.838	.009
VQ	.815	.864	.049

* significant ($p < .05$)

TABLE 4

Predictive Validity Data by Groups: Correlations Between "Before" and "After" Scores and the Criterion; Differences Between Correlations, and Percentage of these Differences

Score	Research Group Coached				Control Group Uncoached			
	Before	After	Dif [a]	% Dif [b]	Before	After	Dif	% Dif
V	.328	.404	.076	23.2	.463	.557	.094	20.3
Q	.389	.421	.032	8.2	.475	.529	.054	11.4
E	.444	.471	.027	6.1	.388	.350	-.038	-9.8
VQ	.409	.469	.060	14.7	.548	.600	.052	9.5

[a] "after" correlation minus "before" correlation

[b] added percentage to the "before" correlation

TABLE 5

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

TABLES 6a & 6b

Percentage of Explained Variance in Predicting the Criterion

by PET Scores, for the Four Models

a. Test 1 - "Before" Condition

Score	Percentage of Explained Variance by:				F Statistic for the Difference Between Models 1 and 2
	Model 1	Model 2	Model 3	Model 4	
V	.135232	.136636	.135621	.135815	F(2,270) = .220
Q	.171725	.172473	.172125	.172236	F(2,270)= .122
E	.188975	.190883	.188977	.189004	F(2,270) = .318
VQ	.200244	.201569	.200258	.200318	F(2,270) = .224

b. Test 2 - "After" Condition

Score	Percentage of Explained Variance by:				F Statistic for the Difference Between Models 1 and 2
	Model 1	Model 2	Model 3	Model 4	
V	.198140	.200842	.199339	.198953	F(2,270) = .456
Q	.201999	.205004	.204707	.204385	F(2,270) = .510
E	.201319	.203277	.201360	.21491	F(2,270) = .332
VQ	.250604	.255824	.255592	.255192	F(2,270) = .947

TABLE B1

Power Analysis for the Differences between Correlation Coefficients

Correlation Difference	Power	
	Coached	Uncoached
0.10	0.57	0.31
0.15	0.89	0.69
0.20	1.00	0.88
0.25	1.00	0.98

Figure Captions

FIGURE 1

Regression Lines Computed Within each Group and Across Groups for Predicting the Criterion by the VQ “After” Scores

Footnotes

[1] This test was developed by Ortar (1960) for her study and it was based on the Arthur Stencil Design Test which measures nonverbal intelligence.

[2] To complete the matriculation requirements, each student must take matriculation examinations in several subject matters. Each subject can be studied at a certain level (usually ranging from 2 to 5 units). The level determines both the scope of studies and the level of the matriculation exam. The average matriculation score is computed as a weighted average of the individual matriculation scores, weighted by their respective levels (further details on the computation of the Israeli matriculation scores can be found in The Hebrew University information booklet, 1994).

[3] The standard deviation of the study participants' score was about 70, which is less than the usual 100 in the entire test population since the participants' range was between 400 and 700 and not between 200 and 800 as in the entire population: The examinees with scores below 400 are a special group for whom PET is very difficult and who usually benefit less than expected from coaching (see Arieli, 1995); The examinees with scores above 700 do not really need coaching courses.