Effects of Test Preparation on the Validity of a Graduate Admissions Test

Donald E. Powers
Educational Testing Service

Test score improvement has been the major concern in nearly all the extant studies of special preparation, or "coaching," for tests. Recently, however, logical analyses of the possible outcomes and implications of special test preparation (Anastasi, 1981; Cole, 1982; Messick, 1981) have suggested that the issue of test score effects is but one aspect of the controversy surrounding coaching; the impact of special preparation on test validity is an equally germane consideration. Although the assumption is sometimes made that coaching can serve only to dilute the construct validity and impair the predictive power of a test, some kinds of special preparation may, by reducing irrelevant sources of test difficulty, actually improve both construct validity and predictive validity. This study examined the relationships of both internal and external criteria to Graduate Record Examination (GRE) candidates' performance on several analytical ability item types, obtained under several test preparation conditions. The purpose was to assess the effects of these various preparations on test reliability and validity. The preparation conditions were those previously shown to be effective, in varying degrees, in improving examinee performance on two of three analytical item types (Powers & Swinton, 1982, 1984). The data for this study were those collected by Powers & Swinton (1982, 1984). The results suggest that GRE analytical ability scores may relate more strongly to academic performance after special test preparation than under more standard conditions and that they may relate less to measures of other cognitive abilities (verbal and quantitative scores). No consistent effects were detected on either the internal consistency or the convergent validity of the analytical measure.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 9, No. 2, June 1985, pp. 179–190

0146-6216/85/020179-12$1.85

The controversy over special preparation, or "coaching," for tests is not a single dispute but many. These include disagreements about the meaning of coaching, the adequacy of the evidence for the effectiveness of coaching, the implications of effective coaching for examinee performance and for test validity, and the ethical imperatives for testing practice (Messick, 1981).

The critics of standardized testing as well as the general public appear most concerned with the ethical issue of equal access to effective test preparation. To some extent, however, both the public and the test critics have expressed concerns related to test validity. For example, ". . . if coaching *does* work, then the tests themselves could be worthless because they . . . fail to provide a uniform indicator of academic potential" (Lieberman, 1979), and effective coaching ". . . would imply that perhaps the exams do not test what they are supposed to" (Levy, 1979).

Types of Coaching and Potential Outcomes

Both Messick (1981) and Anastasi (1981) have elucidated some of the issues surrounding the "coaching controversy." Anastasi (1981) specified three kinds of interventions that may differentially affect test performance. The first, test-taking orientation, is designed to ensure that all test candidates are familiar with the general procedures involved in taking a particular test. The second kind of intervention, which she termed "coaching," is

usually characterized by intensive, short-term practice on item types similar to those employed in the test. The third type is training in broadly applicable cognitive skills.

Both Messick (1981) and Anastasi (1981) discussed three possible outcomes of special test preparation. First, some special preparation may involve training in cognitive skills, thus resulting in improved test scores as well as improved criterion performance. Such increases should threaten neither the construct validity nor the predictive validity of the test (Messick, 1981).

A second possible outcome is that special preparation, particularly test orientation or familiarization, may enhance test-taking sophistication, thereby raising test scores that would ordinarily be inaccurately low because of construct-irrelevant sources of difficulty (such as lack of familiarity with item types or test directions, and test speededness). Overcoming these irrelevant sources of difficulty should improve both predictive validity and construct validity.

The third possible outcome, usually associated with the teaching of test-taking "tricks," is that test preparation may result in scores that are inaccurately *high* assessments of ability, insofar as it improves test performance without producing a corresponding improvement in criterion behavior. Thus, such improvement may serve not only to dilute the construct validity of the test, but to impair its predictive validity as well.

Cole (1982) has also specified three ways that coaching may affect test validity, primarily construct validity. First, coaching could raise test scores above the level of examinees' ability, thus precluding interpretations of scores as valid measures of individuals' ability. Second, coaching could enable examinees to demonstrate their abilities to the maximum; but, if such coaching were differentially available, not all examinees would achieve maximum performance, thus again affecting the interpretation of scores. Third, coaching that affects tests purported to measure relatively stable traits (like academic aptitude) would cast doubt on the interpretation of test scores as valid measures of these stable characteristics.

Research on Coaching and Test Validity

After reviewing the effects of special preparation on Scholastic Aptitude Test (SAT) scores, Messick (1981) found (1) that all of the studies concentrated primarily on the effects of special preparation on test scores, and (2) that no study to date has addressed the issue of improved abilities. The extent to which test validity may change with special preparation has also been virtually ignored. Marron (1965) did explore the extent to which SAT scores obtained after long-term coaching predicted freshman class standing at the U.S. Military Academy and at several selective colleges. Although the results were not entirely consistent, test scores appeared to overpredict academic performance at some of the schools. At the primary school level, Ortar (1960) found that nonverbal aptitude test scores obtained after coaching corresponded more closely with teachers' assessments of their pupils' abilities than did initial scores.

Investigations of the validity of initial and subsequent test scores are also relevant to the issue of coaching-induced changes in test validity. The first testing and any intervening test familiarization could reduce irrelevant difficulty, thereby rendering the second scores more accurate reflections of true ability than initial scores. Olsen and Schrader (1959) examined the validity of first and second SAT scores for a sample of test repeaters at nine colleges and found that initial and retest scores were about equally accurate predictors of college performance. In analyzing the multiple scores of Law School Admission Test repeaters, Linn (1977) found that no single method of treating multiple scores was clearly superior in terms of test validity, but he suggested that using the initial score only was clearly inferior to other procedures, such as using only the second score. Linn's (1977) conclusion is consistent with the notion that practice or increased test familiarity may enhance test validity.

In studying the effects of test disclosure, Stricker (1982) concluded that access to disclosed test materials had no appreciable effects on subsequent retest performance on the SAT—either in level, stability, or concurrent validity. Kendall (1964)

found that test validities became more homogeneous among regional groups of test takers as time limits were increased to a certain point. Kendall's work is germane here because test familiarization may increase test-taking facility, and in effect, increase time limits for efficient test takers. At the level of individual items, Kingston and Dorans (1982) found that item difficulty was affected by within-test practice on several Graduate Record Examination (GRE) General Test item types, but that there were no consistent effects of practice on item discriminations.

In summarizing the implications of coaching for student performance and test validity, Messick (1981) suggested the desirability of studies that relate both uncoached scores and scores obtained after various kinds and amounts of test preparation to multiple measures of cognitive abilities and processes. In this way, determinations could be made of the extent to which scores obtained under coached and uncoached conditions reflect the same ability factors and the extent to which they entail format-specific variance unrelated to other methods of measuring these abilities.

Opportunity for Further Research

The data assembled for a recently completed study of the effects of special preparation on GRE analytical scores[1] (Powers & Swinton, 1982, 1984) provided an excellent opportunity to study the effects of special preparation on test validity. Several features of the database made it especially attractive for this purpose:

1. Performance on two or four item types was affected substantially by the preparation that was given.
2. Because the study was a true experiment, the effect estimates were free from the equivocality that has plagued other quasi-experimental studies of coaching.
3. Several distinct kinds of special preparation were shown to be differentially effective (again under experimental conditions).
4. The various kinds of preparation used in the study were well documented.

With respect to the last point, Cole (1982) has argued that "the most disappointing aspect of the coaching literature is that the components of coaching have been so poorly identified" (p. 409). The Powers and Swinton (1982, 1984) study is an important exception: all of the self-preparation material was available in written form, the method of delivery (direct mailing to candidates) was controlled, and the amount of time devoted to preparation was documented.

Purpose

The overall goal of this study was to understand better the effects of special test preparation on important psychometric properties of a test. The major objective was to assess differences in the relationships of GRE analytical scores (and subscores based on analytical item types) to other criteria. These criteria included both internal ones (i.e., performance on other analytical item types) and external ones (e.g., performance on the verbal and quantitative sections of the test and previous academic achievement).

Procedure

The Database

As mentioned above, this study used data from an experimental study of the effects of several kinds of special preparation on GRE analytical ability

[1]The analytical section of the GRE Aptitude (General) Test was instituted in 1977. Although it has been administered with the operational verbal and quantitative sections of the test, the score derived from the analytical items has been given in a shaded area of the score report form in order to designate its experimental nature. The score report has also included a caveat, advising institutional score recipients that the analytical measure is not yet fully operational because additional evidence is needed to support the validity of the measure. In 1981, as a result of three research studies (Powers & Swinton, 1982, 1984; Swinton & Powers, 1982, 1983; Swinton, Wild, & Wallmark, 1983) that suggested the susceptibility of two item types to within-test practice and to preexamination special preparation, the analysis of explanations and the logical diagrams item types were deleted from the test.
In October 1985, the GRE Board will remove the experimental designation of the analytical measure, making it fully operational partner with the verbal and quantitative sections of the GRE General Test.

scores and item type subscores (Powers & Swinton, 1982, 1984). Random samples of GRE Aptitude (General) Test candidates (a total of 6,600 candidates) were provided with various combinations of test preparation materials to enable them, through self-study, to become more familiar with the analytical section of the GRE Aptitude (General) Test and with each of the item types it contains. These test candidates were selected randomly from the total number who had registered for the test at least seven weeks prior to its administration.

These prospective test takers were then randomly assigned to 10 treatment or control conditions. The materials included full-length sample analytical tests, explanations to sample analytical test questions, and hints or tips for approaching each of the analytical item types. The treatment conditions associated with various combinations of these materials were (1) extra test practice (sample tests), (2) feedback or knowledge of test results (explanations of sample test questions), and (3) strategies for test taking (tips for answering analytical questions).

Half the candidates in each treatment condition, including those in the control group, were given extra encouragement to use the materials that were sent to them. Encouragement took the form of a strongly worded persuasive letter designed to convey to examinees the importance and potential benefits of test preparation. All test takers had received the *GRE Information Bulletin* as part of the regular test registration process. Table 1 shows the various preparation conditions and the numbers of candidates for whom test scores were available.

The findings of the study were as follows. First, two of the analytical item types (analysis of explanations and logical diagrams), but neither subtype of a third kind (analytical reasoning), were susceptible to self-study of test preparation materials. The analysis of explanation item type involves a passage describing a situation and a result after which the examinee is asked to evaluate a number of statements in relation to the situation and result. The logical diagram type involves choosing the Venn-type diagram that best illustrates the relationship among several classes. Analytical reasoning is a puzzle-like item type requiring drawing inferences from a complete set of statements (for a complete description of each item type, see Powers & Swinton, 1981; Swinton & Powers, 1983). Second, some components of special preparation (e.g., extra test practice) were more effective than others; and third, the special preparation, which was designed for the analytical section, did not affect scores on either the verbal or

Table 1
Description of Study Sample

Treatment Code	Preparation Materials Available	Encouraged: Not Encouraged	Encouraged: Encouraged	Total
0	None	921	481	1402
1	Explanations to Sample Analytical Test 1	454	466	920
2	Sample Analytical Test 2	480	460	940
3	Explanations to Sample Analytical Test 1; tips for Answering Analytical Questions	460	482	942
4	Explanations to Sample Analytical Test 1; Sample Analytical Test 2; Explanations to Sample Analytical Test 2; Tips for Answering Analytical Questions	452	451	903
	Total	2767	2340	5107

the quantitative sections of the test. Across treatment groups the average amount of time devoted to preparation was strongly related to average analytical test performance and seemed to be more important than following any specific strategy or using any particular materials. Therefore, increased familiarity with the item types was thought to be the major source of improvement, and the differential effectiveness of some of the special preparation components was thought to result largely from the different amounts of time required for their use.

In addition to data on performance on analytical item types, a variety of other information was also available for these test takers, including the background information that they provided when registering for the test and scores from the verbal and quantitative sections of the test. The most pertinent background data were test takers' self-reports of their undergraduate grade averages, which were used as a "postdictive" criterion of academic success.

Hypotheses

The analyses by Anastasi (1981), Cole (1982), and Messick (1981) suggested the following hypotheses:

I. Because increased familiarity with the analytical item types should serve to decrease extraneous sources of test difficulty (e.g., misunderstanding of directions), the internal consistency reliability estimates should increase as the amount of time devoted to test preparation increases.

II. The correlations among the analytical item types should increase with increasing amount of test preparation, reflecting greater convergent validity, that is, the extent to which the alternative analytical item formats are measuring the same construct. This hypothesis was tested by relating the total analytical score to performance on each item type.

III. The relationship of undergraduate grade-point average (UGPA) to GRE analytical scores should increase with more test preparation, reflecting greater "postdictive" validity of the measure, that is, the extent to which it relates to previous academic achievement. Ideally, it would have been preferable to examine the effects of test preparation on the *prediction* of success in graduate school, as indexed, for example, by first-year graduate grade averages. However, because these data were not available for this sample, self-reported undergraduate averages were used as a surrogate for graduate performance. The rationale for using undergraduate grades as a proxy for graduate grades has been established by Wilson (1981), who observed patterns of postdictive validities that were generally similar to patterns of predictive validities for GRE verbal, quantitative, and analytical scores for a wide variety of graduate departments.

IV. Because extraneous variation was reduced only for the analytical section and not for the verbal or quantitative sections, the relationships of GRE-V and of GRE-Q with GRE-A should decrease with increasing amounts of test preparation. This situation would reflect the greater discriminant validity of the analytical measure, that is, the extent to which it measures something different from the verbal and quantitative sections of the test.

Results

Table 2 provides descriptive statistics for each of the 10 treatment conditions described in Table 1. As reported by Powers & Swinton (1982, 1984), there were no significant differences among verbal or quantitative mean scores across treatment conditions. GRE analytical mean scores did vary significantly across treatment conditions. The numbers of test takers reflect only those who received GRE scores, reported their UGPA, and completed the survey question on the amount of time they devoted to preparing for the analytical section of the test. The numbers of examinees for whom complete data were available were less than the numbers of test registrants initially assigned to treatment groups. There was no indication of any differences among groups with respect to missing data.

Table 2
Selected Descriptive Statistics by Treatment Group

Variable		Not Encouraged					Encouraged				
		0	1	2	3	4	0	1	2	3	4
UGPA in Major	M	5.24	5.29	5.42	5.32	5.37	5.32	5.35	5.31	5.41	5.25
	SD	1.11	1.17	1.12	1.02	1.07	1.10	1.14	1.07	1.07	1.09
UGPA Last 2 Years	M	5.37	5.40	5.49	5.38	5.42	5.38	5.36	5.45	5.49	5.26
	SD	0.99	1.03	1.05	1.03	1.05	0.99	1.05	1.00	1.07	1.06
GRE-Verbal	M	485.2	498.9	493.9	482.4	484.2	486.5	489.5	485.6	488.1	498.1
	SD	111.0	116.5	115.6	113.1	123.8	117.8	113.8	110.7	123.0	114.3
GRE-Quantitative	M	505.6	514.5	510.0	496.5	508.8	507.1	503.5	496.5	504.3	501.7
	SD	126.9	130.0	127.0	125.0	124.4	125.3	130.4	120.1	127.5	126.9
GRE-Analytical	M	509.4	536.9	534.2	528.3	545.3	541.9	543.6	546.0	552.0	566.6
	SD	120.0	121.3	124.3	120.6	120.6	119.1	119.1	118.6	115.7	114.1
Reported Hours of Analytical Preparation	M	2.52	2.61	2.89	3.12	3.42	2.90	3.30	3.58	3.41	3.96
	SD	2.55	2.38	2.43	2.34	2.36	2.34	2.38	2.27	2.37	2.28
	N	614	306	323	320	301	340	329	339	325	333

Note. 0, 1, 2, 3, 4 represent increasing amounts of test preparation material for the GRE analytical measure. UGPA = undergraduate grade point average expressed as follows: D- or lower = 1, C- = 2, C = 3, B- = 4, B = 5, A- = 6, A = 7.

Table 3 shows correlations between GRE analytical scores and selected variables by treatment group. Note that all correlations are significantly different from zero, except for the correlations of reported hours of preparation with GRE analytical scores within treatment conditions. Also note that the correlations between GRE analytical scores and scores on item types within the analytical section are overestimates in that they involve common items. The statistics in Tables 2 and 3 provide the data for the major analyses described below.

As stated earlier, Powers and Swinton (1982, 1984) discovered an extremely strong relationship over treatment groups between average GRE analytical scores and the average time within each group that students reported devoting to preparing for the analytical test section. For this study, this relationship was recomputed using the slightly smaller numbers of test takers who had reported their UGPAs. Again, a highly significant relationship was detected ($r = .84$, $p < .01$); on average, each additional hour of analytical test preparation was associated with an average analytical score increase of 28.3 points. This strong relationship is displayed in Figure 1. Because the time devoted to test preparation seemed to be more instrumental than the particular materials used, the average reported time was used as the major explanatory variable in the analyses reported here, rather than the particular features of each treatment. Further rationale for this data analytic strategy has been given by Powers and Swinton (1982, 1984).

Hypothesis I: Effects on Reliability

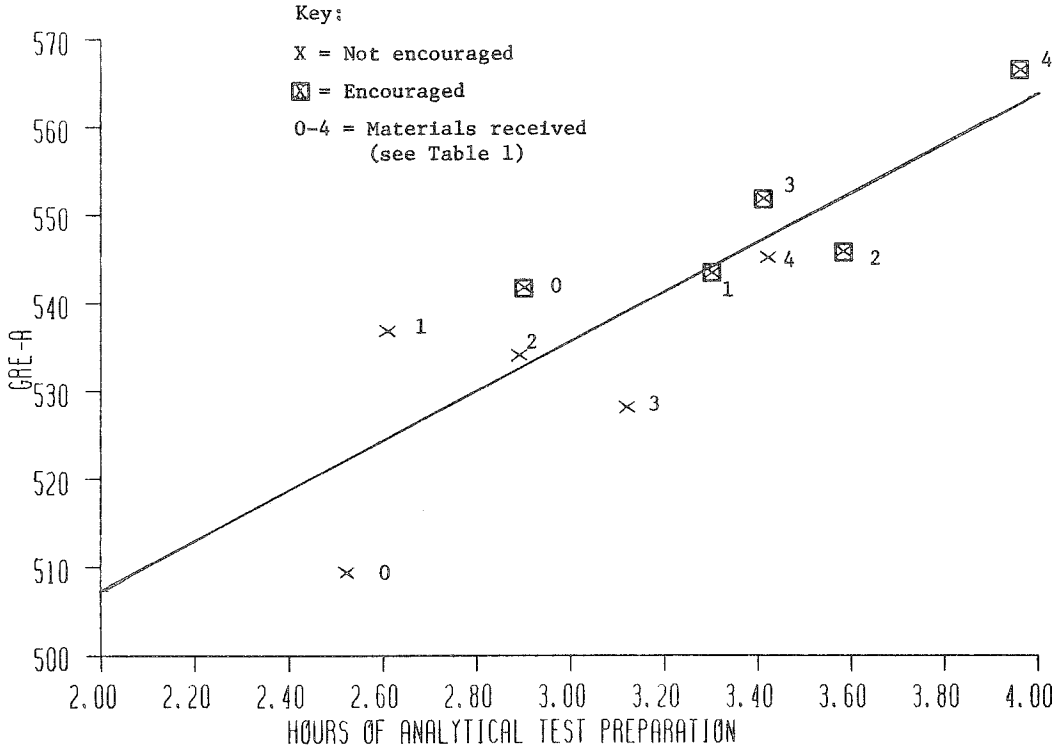
Coefficient alpha reliability estimates were computed for each treatment group for each of two test forms and for each of the two separately-timed analytical sections of the test. Reliability estimates exhibited relatively little variation across treatment groups. For Section III of the test, which was made up entirely of analysis of explanations items, reliability estimates ranged from .83 to .88 for one test form and from .87 to .91 for the other. For Section IV, which was comprised of three distinct

Table 3
Correlations Between GRE-Analytical Scores and Selected Variables by Treatment Group

Variable	Not Encouraged					Encouraged				
	0	1	2	3	4	0	1	2	3	4
UGPA in Major	26	28	19	29	31	24	31	35	35	27
UGPA Last 2 Years	22	24	22	28	30	19	27	32	30	27
GRE-Verbal	71	69	69	68	71	68	71	67	72	66
GRE-Quantitative	68	70	65	72	64	66	66	67	63	64
Analysis of Explanations	93	94	95	95	94	93	95	93	94	94
Logical Diagrams	83	80	84	77	82	81	79	81	77	80
Analytical Reasoning:										
Type I (Analytical Reasoning)	65	68	66	65	66	67	69	66	65	64
Type II (Logical Reasoning)	55	56	60	58	53	63	56	51	54	56
Reported Hours of Analytical Preparation	-01	-00	03	03	03	02	07	07	00	09

Note. Decimals have been omitted. With sample sizes of 300 or more, as is the case here, correlations of .11 are significant at the .05 level and correlations of .15 are significant at the .01 level.

Figure 1
Regression of GRE Analytical Mean Scores on Mean Hours of Analytical Test Preparation for 10 Treatment Groups



item types, over half of which were logical diagrams, reliability estimates ranged from .81 to .86 and from .80 to .85 for the two test forms. Reliability estimates for analytical total scores ranged from .90 to .93 and from .91 to .94 for each test form over the 10 treatment groups. Thus, there appeared to be no significant effect on the reliability of either of the two separately-timed analytical sections or of the total analytical section. When reliability estimates for each treatment group were regressed on the average times spent preparing for the test, no significant or consistent relationships were detected (Table 4).

Hypothesis II: Effects on Convergent Validity

As an indirect assessment of effects of test preparation on convergent validity, the correlations of the GRE total analytical score with each of the analytical item types were computed for each treatment group. These correlations were then related, through regression analysis, to the average hours of preparation for each group (Table 4). No significant relationships were detected. This pattern of results did not suggest that individual item types became interrelated measures of a common construct when examinees received more preparation for the test. That is, convergent validity was not affected.

Hypothesis III: Effects on "Postdictive" Validity

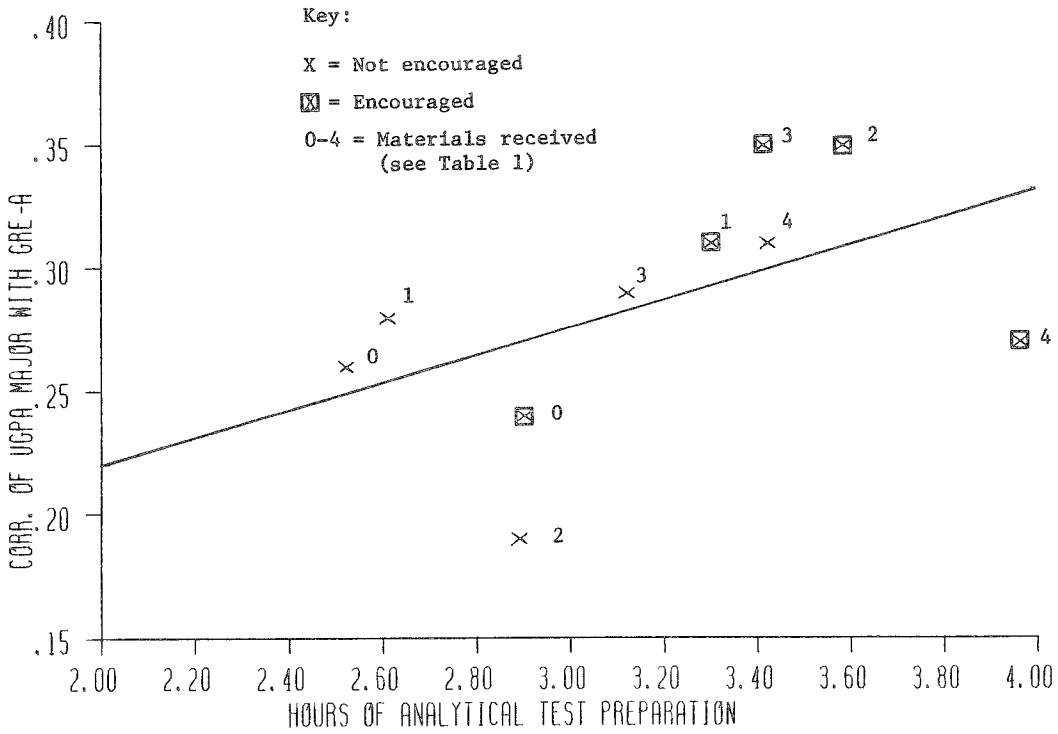
As can be seen in Table 4, the relationships of "postdictive" validities to the average time reportedly spent preparing for the analytical measure were larger than any others that were computed. The correlations of GRE analytical scores with both UGPA in major field and in the last two years of undergraduate school were positive, and significant ($p < .05$) for UGPA in the last two years. These results suggest that the criterion-related validity of the GRE analytical measure did not diminish and, as hypothesized, may have increased. The relationship between (1) the mean reported hours of preparation and (2) the correlation of GRE-A with UGPA in major field ($r = .51$) is displayed in Figure 2. The relationship between mean hours of preparation and the correlation of GRE-A with UGPA in the final two years was even stronger ($r = .69$). These results suggest the possibility that an

Table 4
Correlations of Validity and Reliability Coefficients on Average Reported Hours Of Analytical Test Preparation

Independent Variable	r
Reliability	
Form A	.162
Form B	-.532
Correlation of GRE-A with:	
Analysis of Explanations	.138
Logical Diagrams	-.322
Analytical Reasoning:	
Type I (Analytical Reasoning)	-.259
Type II (Logical Reasoning)	-.424
UGPA in Major	.512
UGPA in Last 2 Years	.698*
GRE-Verbal	-.352
GRE-Quantitative	-.506

Note. N = 10 for all analyses.
*p < .05

Figure 2
Regression of "Postdictive" Validities (Correlation of UGPA in Major with GRE-A) on Mean Hours of Analytical Test Preparation for 10 Treatment Groups



achievement or motivational component was introduced into test scores as a result of the presence of test item types that were susceptible to test familiarization effects.

Hypothesis IV: Effects on Discriminant Validity

Table 4 shows that the correlations between GRE analytical scores and both GRE verbal and quantitative scores decreased (though not significantly) with increasing amounts of reported preparation. The negative relationships suggest that the analytical ability measure did not become more like the verbal and quantitative measures when examinees were better prepared to take the test.

Speededness

One potential extraneous source of test difficulty is speededness. It was hypothesized that increasing amounts of test familiarization might result in more efficient test-taking behavior, and therefore enable examinees to consider and answer more questions in the allotted time. Generally, the differences among treatment groups in the percentages of examinees reaching the final test item were relatively small. For Section III of the test (analysis of explanations items), these percentages ranged from 74.0 to 79.2 and from 74.8 to 86.7 for the two forms; for Section IV (the other analytical item types), these percentages ranged from 55.9 to 63.6 for one form and from 24.4 to 29.7 for the other. The regressions of the mean percentages on mean time preparing for each treatment group did not reveal any significant relationships. The correlations were .54 and $-.44$ for Sections III and IV of one form, and $-.22$ and $-.19$ for the other. Thus, this pattern of results does not justify any conclusions about the role of test familiarization in reducing test speededness.

Discussion

Nearly all the available studies of test preparation, or coaching, for standardized admissions test have focused on improvements in test scores. Because so few investigations have employed true experimental designs, with examinees randomly assigned to treatments, conclusions regarding the effects of coaching on test scores have been somewhat equivocal. It is not surprising, therefore, that most of the studies have provided less than adequate opportunities to examine the effects of test preparation on test validity. Because a true experiment had been conducted for item types that were truly susceptible to special test preparation, data were available that enabled a reasonable empirical assessment of the effects of test preparation on test validity.

Relatively large improvements on the original GRE analytical measure resulted from relatively short periods of test preparation. It is, therefore, reasonable to question the extent to which the original measure actually reflected analytical abilities that are purported to develop over a long period of time. Though substantial score improvements provide good reason to question the validity of this particular interpretation of original analytical scores, the results reported here provide little if any evidence that effective test preparation may dilute certain other aspects of test validity. In fact, the data suggest that GRE analytical scores may relate more highly to academic performance when scores have been increased by special test preparation than when they are obtained under more standard conditions. Although no significant trends were found, data also suggest that, when examinees have had more preparation for the analytical section of the test, scores may be less strongly related to other predictors of academic performance (e.g., verbal and quantitative scores) that reflect different cognitive abilities.

The study did not reveal any consistent effects of special preparation on either the internal consistency or the convergent validity of the GRE analytical item types. No significant relationships were detected, though the correlations of three of the four item type subscores with total analytical scores decreased as the time reportedly devoted to test preparation increased. It is noted again that, when this study was conducted, the GRE analytical measure was considered to be experimental until further evidence could be accumulated on its validity. In 1981, the measure was revised substantially: Because of their susceptibility to special preparation, the analysis of explanations and the logical diagrams item types were deleted, and additional numbers of each of the two analytical reasoning subtypes (one is now called logical reasoning) were inserted.

The addition of greater numbers of analytical reasoning items, which occupied only a very small portion of the original measure, has made possible further research on the validity of the revised measure and of these two item types. Some of the most recent evidence (Wilson, 1985) has revealed that the two item types that make up the current analytical measure may be more reflective of either verbal ability or quantitative ability rather than of analytical ability, as measured by the original test. If this is indeed true, then the negative relationships between the reported amount of test preparation and the item type-total test correlations may merely reflect greater discriminant validity of these "analytical" item types rather than diminished convergent validity.

The conclusions of this study must be qualified on several grounds. First, only one particular kind of test preparation (self-test familiarization) was considered here. To the extent that they employ different methods or materials, other types of test preparation or coaching experiences might yield different patterns of results. Second, these results may not generalize completely to other kinds of test items. The item types considered here were relatively complex, and both coachable types (analysis of explanations and logical diagrams) employed fixed-response option formats. Third, the relationships of test scores to academic performance were studied in a "postdictive" sense. Although there is sufficient justification for this kind of analysis, results may not necessarily apply to the prediction of academic success. Finally, the analyses yielded few statistically significant results with the small number of units of analysis avail-

able, and may therefore deserve replication. Nonetheless, on balance, the modest evidence provided here seems more consistent with the notion that test preparation of the kind studied may enhance rather than impair test validity.

If such "coachable" item types were to be included in tests like the GRE General Test, then continued provision of appropriate test familiarization materials and encouragement of their use would seem advisable. This advice is probably equally sound for test item types that are less susceptible to coaching, practice, and other kinds of special preparation than those considered here. If, for these less "coachable" item types, test familiarization minimizes the number of test takers who receive inaccurately low scores, then the validity of test score interpretations should be maintained for tests composed of these item types as well.

References

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist, 36*(10), 1086–1093.

Cole, N. (1982). The implications of coaching for ability testing. In A. Wigdor & W. R. Garner (Eds.), *Ability Testing: Uses, consequences, and controversies, Part II: Documentation sections* (pp. 389–414). Washington DC: National Academy Press.

Kendall, L. M. (1964). The effects of varying time limits on test validity. *Educational and Psychological Measurement, 24*, 789–800.

Kingston, N. M., & Dorans, N. J. (1982). *The effect of the position of an item within a test on item responding behaviors: An analysis based on item response theory* (GRE Board Professional Report GREB 79–126P and ETS Research Report 82–22). Princeton NJ: Educational Testing Service.

Levy, S. (1979, March). ETS and the "coaching" cover-up. *New Jersey Monthly*, pp. 51–54, 82–89.

Lieberman, D. (1979, June 20). You can study for aptitude tests. *The Hartford Advocate*, pp. 6–7, 16.

Linn, R. L. (1977). On the treatment of multiple scores for Law School Admission Test repeaters (Report #LSAC–77–4). In Law School Admission Council, *Reports of LSAC Sponsored Research: Volume III, 1975–1977*. Princeton NJ: Law School Admission Council.

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, Research Division, Office of the Director of Admissions and Registrar.

Messick, S. (1981). The controversy over coaching: Issues of effectiveness and equity. In B. F. Green (Ed.), *Issues in testing: Coaching, disclosure, and ethnic bias* (pp. 21–53). San Francisco: Jossey-Bass.

Olsen, M., & Schrader, W. B. (1959). *The use of preliminary and final Scholastic Aptitude Test scores in predicting college grades* (College Entrance Examination Board Research and Development Reports, and Statistical Report #59–19). Princeton NJ: Educational Testing Service.

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

Powers, D. E., & Swinton, S. S. (1981). Extending the measurement of graduate admission abilities beyond the verbal and quantitative domains. *Applied Psychological Measurement, 5*, 141–158.

Powers, D. E., & Swinton, S. S. (1982). *The effects of self-study of test familiarization materials for the analytical section of the GRE Aptitude Test* (GREB Research Report GREB No. 79–9). Princeton NJ: Educational Testing Service.

Powers, D. E., & Swinton, S. S. (1984). Effects of self-study for coachable test item types. *Journal of Educational Psychology, 76*, 266–278.

Stricker, L. J. (1982). *Test disclosure and retest performance on the Scholastic Aptitude Test* (College Board Report No. 82–7 and ETS RR No. 82–48). Princeton NJ: Educational Testing Service.

Swinton, S. S., & Powers, D. E. (1983). A study of the effects of special preparation on GRE analytical scores and item types. *Journal of Educational Psychology, 75*, 104–115.

Swinton, S. S., Wild, C. L., & Wallmark, M. (1983). *Investigation of practice effects on item types in the Graduate Record Examinations Aptitude Test* (GREB Professional Report No. 80–1 CP and ETS RR 82–56). Princeton NJ: Educational Testing Service.

Wilson K. M. (1981). *A study of the validity of the restructured GRE Aptitude Test for predicting first-year performance in graduate study* (GREB Research Report GREB No. 78–6). Princeton NJ: Educational Testing Service.

Wilson, K. M. (1985). *The relationship of GRE General Test item-type part scores to undergraduate grades* (GREB Professional Report GREB No. 81–22P) Princeton NJ: Educational Testing Service.

Acknowledgments

The author thanks the GRE Board for its sponsorship of this research and the members of the GRE Research Committee for their continued support and advice. Thanks

also to Robert Altman, Gordon Hale, Spencer Swinton, and Cheryl Wild for thoughtful reviews of an earlier draft, to Richard Harrison for data analysis, and to Lorraine Simon for preparing the final manuscript and overseeing other project activities.

Author's Address

Send requests for reprints or further information to Donald E. Powers, Educational Testing Service, Princeton NJ 08541, U.S.A.