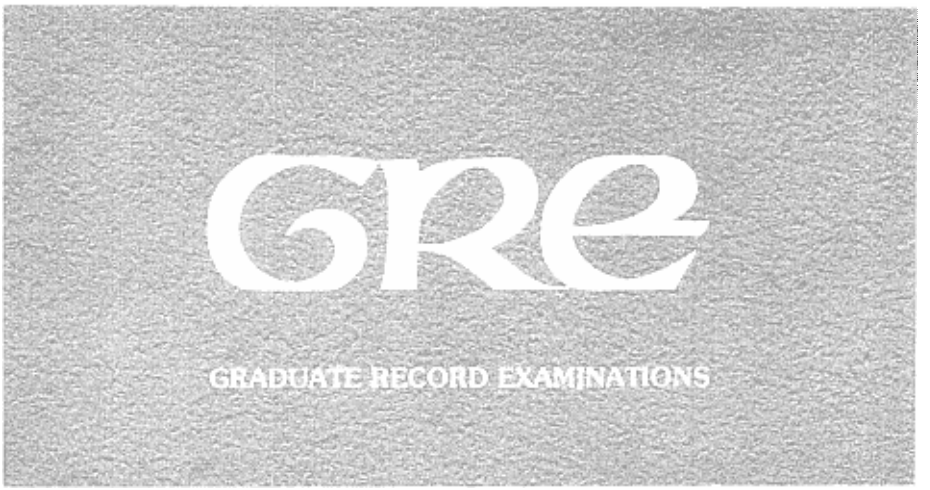

INVESTIGATION OF PRACTICE EFFECTS ON ITEM TYPES IN THE GRADUATE RECORD EXAMINATIONS APTITUDE TEST

Spencer S. Swinton
Cheryl L. Wild
Madeline M. Wallmark

GRE Board Professional Report GREB No. 80-1cP
ETS Research Report 82-56

May 1983

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

GRE BOARD RESEARCH REPORTS FOR GENERAL AUDIENCE

Altman, R. A. and Wallmark, M. M. A Summary of Data from the Graduate Programs and Admissions Manual. 74-1R, 1975.

Baird, L. L. An Examination of the Graduate Study Application and Enrollment Decisions of GRE Candidates. 79-11R, 1982.

Baird, L. L. An Inventory of Documented Accomplishments. 77-3R, 1979.

Baird, L. L. Cooperative Student Survey (The Graduates [$2.50 each], and Careers and Curricula). 70-4R, 1973.

Baird, L. L. The Relationship Between Ratings of Graduate Departments and Faculty Publication Rates. 77-2aR, 1980.

Baird, L. L. and Knapp, J. E. The Inventory of Documented Accomplishments for Graduate Admissions: Results of a Field Trial Study of Its Reliability, Short-Term Correlates, and Evaluation. 78-3R, 1981.

Burns, R. L. Graduate Admissions and Fellowship Selection Policies and Procedures (Part I and II). 69-5R, 1970.

Centra, J. A. How Universities Evaluate Faculty Performance: A Survey of Department Heads. 75-5bR, 1977. ($1.50 each)

Centra, J. A. Women, Men and the Doctorate. 71-10R, 1974. ($3.50 each)

Clark, M. J. The Assessment of Quality in Ph.D. Programs: A Preliminary Report on Judgments by Graduate Deans. 72-7aR, 1974.

Clark, M. J. Program Review Practices of University Departments. 75-5aR, 1977. ($1.00 each)

Clark, M. J and Centra, J. A. Conditions Influencing the Career Accomplishments of Ph.Ds. 76-2R, 1982.

Donlon, T. F. Annotated Bibliography of Test Speededness. 76-9R, 1979.

Flaugher, R. L. The New Definitions of Test Fairness In Selection: Developments and Implications. 72-4R, 1974.

Fortna, R. O. Annotated Bibliography of the Graduate Record Examinations. 1979.

Frederiksen, N. and Ward, W. C. Measures for the Study of Creativity in Scientific Problem-Solving. 1978.

Hartnett, R. T. Sex Differences in the Environments of Graduate Students and Faculty. 77-2bR, 1981.

Hartnett, R. T. The Information Needs of Prospective Graduate Students. 77-8R, 1979.

Hartnett, R. T. and Willingham, W. W. The Criterion Problem: What Measure of Success in Graduate Education? 77-4R, 1979.

Knapp, J. and Hamilton, I. B. The Effect of Nonstandard Undergraduate Assessment and Reporting Practices on the Graduate School Admissions Process. 76-14R, 1978.

Lannholm, G. V. and Parry, M. E. Programs for Disadvantaged Students in Graduate Schools. 69-1R, 1970.

Miller, R. and Wild, C. L. Restructuring the Graduate Record Examinations Aptitude Test. GRE Board Technical Report, 1979.

Powers, D. E. and Lehman, J. GRE Candidates' Perceptions of the Importance of Graduate Admission Factors. 81-2R, 1982.

Powers, D. E. and Swinton, S. S. Effects of Self-Study of Test Familiarization Materials for the Analytical Section of the GRE Aptitude Test. 79-9R, 1982.

Reilly, R. R. Critical Incidents of Graduate Student Performance. 70-5R, 1974.

Rock, D. and Werts, C. An Analysis of Time Related Score Increments and/or Decrements for GRE Repeaters across Ability and Sex Groups. 77-9R, 1979.

Rock, D. A. The Prediction of Doctorate Attainment in Psychology, Mathematics and Chemistry. 69-6aR, 1974.

Schrader, W. B. Admissions Test Scores as Predictors of Career Achievement in Psychology. 76-1aR, 1978.

Schrader, W. B. GRE Scores as Predictors of Career Achievement in History. 76-1bR, 1980.

Swinton, S. S. and Powers, D. E. A Study of the Effects of Special Preparation on GRE Analytical Scores and Item Types. 78-2R, 1982.

Wild, C. L., Swinton, S. S., and Wallmark, M. M. A Summary of the Research Leading to the Revision of the Format of the Graduate Record Examinations Aptitude Test in October 1981. 80-1aR, 1982.

Wild, C. L. Summary of Research on Restructuring the Graduate Record Examinations Aptitude Test. 1979.

Wild, C. L. and Durso, R. Effect of Increased Test-Taking Time on Test Scores by Ethnic Group, Age, and Sex. 76-6R, 1979.

Wilson, K. M. A Study of the Validity of the Restructured GRE Aptitude Test for Predicting First-Year Performance in Graduate Study. 78-6R, 1982.

Wilson, K. M. The GRE Cooperative Validity Studies Project. 75-8R, 1979.

Wiltsey, R. G. Doctoral Use of Foreign Languages: A Survey. 70-14R, 1972. (Highlights $1.00, Part I $2.00, Part II $1.50).

Witkin, H. A.; Moore, C. A.; Oltman, P. K.; Goodenough, D. R.; Friedman, F.; and Owen, D. R. A Longitudinal Study of the Role of Cognitive Styles in Academic Evolution During the College Years. 76-10R, 1977 ($5.00 each).

Investigation of Practice Effects on Item Types in the Graduate Record Examinations Aptitude Test

Spencer S. Swinton

Cheryl L. Wild

Madeline M. Wallmark

GRE Board Professional Report GREB No. 80-1cP

May 1983

Abstract

Three studies are reported that investigated within-test practice effects of item types that were, or might be, included in the Graduate Record Examinations Aptitude Test. Item types studied included reading comprehension, sentence completions, analogies, antonyms, discrete quantitative, data interpretation, quantitative comparison, supporting conclusions, analysis of explanations, logical diagrams, analytical reasoning, and logical reasoning.

Study 1 assessed the practice effect for analytical reasoning, logical diagrams, analysis of explanations, and logical reasoning item types, using three editions of the GRE Aptitude Test. Study 2 examined practice effects for supporting conclusions and logical reasoning item types in an experimental administration. Study 3 evaluated practice effects for all item types listed in the first paragraph except supporting conclusions, using two editions of the GRE Aptitude Test.

In studies 1 and 3, significant within-test practice effects were identified for analysis of explanations and logical diagrams item types. In study 2, a smaller practice effect was identified for the supporting conclusions item type. Within-test practice does not appear to affect the other item types investigated in these studies.

The analysis of explanations and logical diagrams items were removed from the analytical portion of the test in the fall of 1981.

Investigation of Practice Effects on Item Types in the Graduate Record Examinations Aptitude Test

Spencer S. Swinton
Cheryl L. Wild
Madeline Wallmark

The Graduate Record Examinations Aptitude Test results in three reported scores--verbal, quantitative and analytical ability.* Each of these scores is based on responses to three or four types of questions. The purpose of the series of studies reported here was to test whether scores on any of these types of questions (plus one new type of question) are improved with short term (within-test) practice. Results of the studies were available to inform the Graduate Record Examinations Board and its Research Committee in making a decision to reformat the test beginning in October 1981.

The impetus for this series of investigations was a problem (first noticed in October 1978) in predicting the difficulty of the analytical questions. In the edition of the Aptitude Test introduced in October 1978, the operational analytical measure proved to be more difficult than had been anticipated from the difficulties estimated when the questions had been pretested in Section V. No such problem exists with verbal or quantitative sections of the test. One hypothesis for the apparent increase in difficulty was a within-test practice effect, that is, that the analytical questions became easier for examinees exposed to questions of the same type earlier in the test.

Three studies are reported here. The first study, conducted in April 1979, investigated practice effects for only the analytical item types. The second study was done in a special administration after the February 1980 test administration and investigated practice effects for a possible new analytical item type and for a larger sample of logical reasoning items. The third study was conducted in June 1980 and looked into practice effects on all the item types then in the Aptitude Test. In all three studies, practice effect was defined as section score change as a function of order of administration and is thus net of practice over any fatigue effects or format differences that might offset positive score changes.

Study 1

Purpose

The purpose of this study was to test whether scores on analytical reasoning, logical reasoning, logical diagrams,

---

*The research reported in this study is based on the GRE Aptitude Test as it existed between October 1977 and October 1981. In October 1981, analysis of explanations and logical diagrams items were deleted from the analytical portion of the test.

and/or analysis of explanations item types are affected by within-test practice.

Procedures

In April 1979, three editions* of the Aptitude Test were administered. (See Figure I for an illustration of the test format.) An estimate of the magnitude of the practice effect was obtained by moving the analysis of explanations questions from section III in one test edition to section V in the other two editions. Similarly, the practice effect estimate for logical diagrams and analytical and logical reasoning was obtained by placing questions of both item types from section IV in one test edition into section V in the other editions. All three sections had the same 25-minute time limit. Thus, each test edition had one of four different experimental sections, section III from each of the two other editions and section IV from each of the two other editions. These twelve (three editions times four section Vs per edition) test versions were administered in spiralled order so that randomly equivalent groups could be expected to take each test.

Analysis of covariance, with group and form as main effects, was the principal mode of analysis. This design, with three forms, makes it possible to test for form-by-order interaction, a test not possible when only two forms are employed, as for example in a comparison of groups 1 and 2 only since, in this latter case, group is completely confounded with form-by-order interaction.

The design may be summarized as follows:

		First (Section III or IV)	Second (Section V)
Analysis of Explanations	Group 1	Form A	Form B
	Group 2	Form B	Form A
	Group 3	Form A	Form C
	Group 4	Form C	Form A
	Group 5	Form B	Form C
	Group 6	Form C	Form B
Logical Diagrams and Analytical and Logical Reasoning	Group 7	Form A	Form B
	Group 8	Form B	Form A
	Group 9	Form A	Form C
	Group 10	Form C	Form A
	Group 11	Form B	Form C
	Group 12	Form C	Form B

*Forms ZGR1, ZGR3, and 3BGR1.

Figure I

Contents and Format of the
Graduate Record Examinations Aptitude Test
(not necessarily in this order)

Section	Item Types	Number of Questions	Time
I.	Verbal	80 (75)*	50 minutes
	a) Reading Comprehension	25 (22)	
	b) Antonyms	20 (22)	
	c) Analogies	18 (18)	
	d) Sentence Completions	17 (13)	
II.	Quantitative	55	50 minutes
	a) Quantitative Comparisons	30	
	b) Regular Mathematics	15	
	c) Data Interpretation	10	
III.**	Analysis of Explanations	40	25 minutes
IV.**	Analytical	30	25 minutes
	a) Logical Diagrams	15	
	b) Analytical Reasoning Type I	10-11	
	Logical Reasoning Type II	4-5	
V.	Experimental Section	Variable	25 minutes

*The number of verbal questions has been decreased from 80 to 75 to decrease the speededness. Numbers on the left indicate the test content in October 1977, when the first restructured test was introduced. Numbers on the right indicate content of the more recent editions of the test prior to October 1981.

**The total analytical score was based on the items in Sections III and IV.

Study 2

Purpose

The purpose of this study was to test whether scores on supporting conclusions (Attachment I) or logical reasoning items (Attachment II) are improved by within-test practice on the item type. The information was directly relevant to two possible restructurings of the analytical measure.

The supporting conclusions item type had been identified as a potential replacement for analysis of explanations. Supporting conclusions items had been pretested in a Law School Admission Test (LSAT) experimental section at the same time an analysis of explanations experimental section was given. The correlations of these two experimental scores with LSAT operational sections suggested that the supporting conclusions item type may relate to logical reasoning and to quantitative comparison questions in the same manner as analysis of explanations items do. Corrected for attenuation, the correlations of logical reasoning items were .82 and .84 with analysis of explanations and supporting conclusions, respectively. The quantitative comparisons correlations with these two analytical item types were .63 and .61 respectively. Unlike the directions for analysis of explanations items, the directions for supporting conclusions are not complex (see Attachment I). No direct evidence of practice effect existed for this item type, and study 1 had not yielded sufficient evidence for the logical reasoning item type.

Another possibility, if the analytical measure was to be reformatted, was to expand the number of analytical and logical reasoning items. The first study reported here investigated practice for only 11 analytical reasoning questions and 4 logical reasoning items. Especially for the second type, investigation of practice effect within the operational constraints of the test as then constituted may not have provided sufficient practice to reveal an effect. For these reasons, study 2 was carried out.

Procedures

Examinees who registered to take the test in February 1980 at selected test centers were invited to participate in an experimental administration in the afternoon of the national administration (February 23, 1980). Examinees were paid $15 for two hours of experimental testing.

The experimental test consisted of four 30-minute sections--two consisting of supporting conclusions (25 items), and two of logical reasoning (23 items). Two versions of the test,

differing only in the order of the form A and B sections, were administered to about 425 examinees each (see Figure II).

Figure II

Experimental Tests Administered in February 1980

Experimental Test 1 (Separately timed 30-minute sections)		Experimental Test 2 (Separately timed 30-minute sections)	
I.	Supporting Conclusions--A	I.	Supporting Conclusions--B
II.	Logical Reasoning--A	II.	Logical Reasoning--B
III.	Supporting Conclusions--B	III.	Supporting Conclusions--A
IV.	Logical Reasoning--B	IV.	Logical Reasoning--A

This design resulted in some extra practice for the logical reasoning questions; that is, the examinees had encountered four of these items in the morning. One-seventh of the group also had been administered an experimental section with analytical reasoning questions. However, it was deemed unlikely that practice in four questions would substantially change item difficulty for this item type. This possible confounding was judged to be preferable to administering the experiment prior to the GRE Aptitude Test, thus giving subjects one hour of practice on an item type that could influence their reported analytical score. The one-seventh who received the analytical Section V in the morning testing were removed from the sample.

Repeated measures analyses of covariance, with group and form as main effects and verbal, quantitative, and analytical scores as covariates, were the principal mode of analysis. In these analyses, order, the effect of interest, is confounded with group-by-form interaction. The form difficulty main effect is confounded with any group-by-order interaction (Form A is the first test for group 1, but the second test for group 2), and group is confounded with form-by-order interaction. However, since group membership was randomly assigned and covariance-adjusted, this confounding was not considered a threat to interpretation. An equivalent analysis would call group and order the main effects, with group-by-form interaction still confounded with order. The existence of real form-by-order interactions is more likely than that of either of the group interactions since, in some cases, forms did differ notably in difficulty, and it is possible that an easy form followed by a harder form leads to a different practice effect than does the reverse situation. In contrast to the design of Study 1, form-by-order interactions cannot be examined in the present design, since they are completely confounded with group.

Study 3

Purpose

The purpose of this study was to investigate the practice effect for verbal and quantitative item types and to replicate the earlier findings on analytical item types.

Procedures

The study took place at the June 1980 administration of the GRE Aptitude Test. At that time two different editions of the operational test were administered by spiralling (ZGR1 and CGR1). Through this procedure, approximately random samples of examinees took each edition of the test. Material from one operational test was inserted in the fifth experimental section of the other operational test. Thus, there were six versions of ZGR1. The first four sections were identical, but the fifth section contained one of either the two verbal, two quantitative, or two analytical half-sections developed for the other test edition (CGR1). Similarly the six versions of CGR1 all had the same four operational sections, but differed in the contents of Section V.

The various combinations of test forms are illustrated in Figure III.

As in the earlier studies, analyses of covariance were the principal mode of analysis, with the design and its constraints essentially identical to that of Study 2. To obtain some idea of the impact of practice effect on validity, correlations of analytical item type scores in the operational and variable sections with the self-reported undergraduate grade-point averages were obtained. These correlations are reported in Wild, Swinton, and Wallmark (1982). Separate analyses were conducted to look at the implications of practice effect for each of the item types (reading comprehension [explicit and inferential], antonyms, analogies, sentence completions, quantitative comparison, discrete quantitative, data interpretation, analysis of explanations, logical diagrams, analytical reasoning, and logical reasoning).

One problem with this design is that examinees had to spend a proportionally larger amount of time reading directions in the 25-minute verbal and quantitative sections than in the

Figure III

Versions of Tests Administered in June 1980

Test Edition ZGR1	Test Edition CGR1
I. Verbal - 50 minutes	I. Verbal - 50 minutes
II. Quantitative - 50 minutes	II. Quantitative - 50 minutes
III. Analysis of Explanations - 25 minutes	III. Analysis of Explanations - 25 minutes
IV. Logical Diagrams & Analytical & Logical Reasoning - 25 minutes	IV. Logical Diagrams & Analytical & Logical Reasoning - 25 minutes
V. 25 minutes of one of the following:	V. 25 minutes of one of the following:
a) 1st parallel half of Section I, CGR1	a) 1st parallel half of Section I, ZGR1
b) 2nd parallel half of Section I, CGR1	b) 2nd parallel half of Section I, ZGR1
c) 1st parallel half of Section II, CGR1	c) 1st parallel half of Section II, ZGR1
d) 2nd parallel half of Section II, CGR1	d) 2nd parallel half of Section II, ZGR1
e) Section III, CGR1	e) Section III, ZGR1
f) Section IV, CGR1	f) Section IV, ZGR1
	g) Section III, ZGR1 with four questions deleted

comparable 50-minute sections. The result was that the experimental half-sections were slightly more speeded than the operational section. In order to investigate the effects of this factor, changes in item difficulty were examined as a function of item location. For the reading comprehension item type, the effect of being caught by the time limit before completing the final passage would affect a proportionately greater percentage of the items in a 25-minute section than in a 50-minute section. For this reason, a slight negative apparent practice effect was anticipated for this item type. The same reasoning does not apply to analysis of explanations (Section IV), which also contain series of passage-dependent items, because these items were initially in a 25-minute section and hence did not require breaking into parallel halves for the experimental section.

Results: Study 1

Adjusted cell means of the various analytical sections for the 12 groups are given in Table 1. Because nine analyses are summarized here, details of each analysis are not given. Each analysis is similar to the one reported in detail in Study 2.

Examination of the means in Table 1 shows that the analysis of explanations form appeared easier when administered in the second position. The Form A adjusted mean score increased by over three items when administered after the more difficult Form B and by almost two items when administered after Form C. Form C scores were over three items higher when administered after Form A and four items after Form B. Form B scores were one item higher after exposure to Form A and three-fourths of an item higher after encountering Form C. The F values, obtained from analyses of pairs of same-test groups, show these effects to be highly significant. F was calculated by the analysis of covariance program with repeated measures, using verbal and quantitative scores as covariates. The adjusted means for groups 1 and 2 are thus based on groups of covariance-adjusted ability, as are the means for groups 3 and 4, and other successive pairs. However, such pairs as groups 1 and 3, for example, have not been so adjusted for comparability.

The two groups that took Form B first happened to be the most able, as indicated by their verbal and quantitative mean scores. The apparently smaller practice effect for Form B may be partly accounted for by possible failure of the covariance adjustment to correct completely for between-group differences in ability, because of less than perfect subscore reliabilities.

Table 1

Adjusted Cell Means, and F* for Order Effect-Study 1

Item Type	Group	n	Adjusted Cell Means 1st Form		Adjusted Cell Means 2nd Form		F
Analysis of Explanations (40)	1	1440	A	18.29	B	17.96	372.87
	2	1455	B	16.92	A	21.66	
	3	1473	A	19.06	C	21.42	570.46
	4	1368	C	18.13	A	20.94	
	5	1381	B	17.62	C	21.64	552.04
	6	1400	C	17.64	B	18.36	
Logical Diagrams (15)	7	1378	A	9.53	B	8.31	213.42
	8	1389	B	7.69	A	10.48	
	9	1395	A	9.82	C	7.82	113.18
	10	1419	C	7.10	A	10.28	
	11	1411	B	7.97	C	8.01	139.04
	12	1442	C	6.93	B	8.24	
Analytical Reasoning (11) and Logical Reasoning (4)	7	1378	A	7.27	B	7.42	34.60
	8	1389	B	7.39	A	7.95	
	9	1395	A	7.45	C	5.22	10.54
	10	1419	C	4.86	A	7.46	
	11	1411	B	7.63	C	4.71	23.13
	12	1442	C	5.62	B	7.29	

*All F values significant beyond .001 level

It may also result from an asymmetry in practice effect, with a more difficult test inducing a larger practice effect on an easier test than vice versa. This hypothesis deserves further investigation because of its implications for the assumptions of most equating models.

In any case, the average practice effect for these three analysis of explanations forms, 2.4 items in a 40-item form, or .06 points per item, is not only statistically, but practically, significant and is unacceptably large.

The logical diagrams item type also showed positive practice effects for each form, ranging from .27 for Form B after Form C to 1.08 for Form C after Form B (each 15 items). Again, these effects are all statistically significant and, with an average value of .04 points per item, approach practical significance, in that practice on 25 items is estimated to increase scores by one raw score point.

The combined analytical reasoning and logical reasoning section showed mixed results, with less significant positive practice effects in two analyses, and a similarly less significant negative practice effect in the third analysis. The average estimated effect of less than .02 points per item was not considered to be practically significant in that it would require working through more than three sections of this length to realize a one raw-score point gain.

The results of this study suggest that the analysis of explanations item type is highly susceptible to within-test practice. Since the item type was new, the possibility remained that this susceptibility to practice would decrease with the passage of time and with increased examinee familiarity. However, the complexity of instructions coupled with the fact that instructions were presented within the time allotted for reading and responding to the items, argued against complacency concerning issues of equity.

The logical diagrams results raised a similar concern, although not as serious as that surrounding analysis of explanations because the effect was smaller and because logical diagrams comprised only about 21 percent of the analytical section, as compared to 57 percent for analysis of explanations.

The results for analytical and logical reasoning are more reassuring. However, the small number of logical reasoning items involved in the operational test made these results inconclusive and argued for further study of practice effects for this item type. The results were further complicated by the fact that these items appeared operationally after logical diagrams in the same timed section, and thus any benefit

from practice accruing to logical diagrams might be expected to carry over in more time available for analytical and logical reasoning.

Results: Study 2

Data arising from the study were analyzed using the repeated measures analysis of covariance program. The two forms of each test were analyzed as repeated trials within each group. The error term for the group effect was the residual after removing group effects and covariates. The error for form effect and for the form-by-group interaction (practice effect) was the pooled subject-by-trial interaction within groups and was only about half as large as the error term for the group effect. The analysis was thus highly efficient. It is necessary to keep in mind that any real form-by-group interaction was confounded with practice effect in this design. The likelihood that one form was relatively easier for one group than for the other for any reason other than the different order of exposure to the two forms was considered small, however. The results are described in detail for this study.

The logical reasoning test showed no practice effect. Adjusted cell means and the analysis of covariance results are given in Tables 2 and 3. (Numbers in parentheses are deviations from additivity, a measure of the magnitude of form-by-group interaction, or practice effect.)

Table 2

Means and Deviations from Additivity: Logical Reasoning

	n	Form A2	Form A4
Group 82 (A2 FIRST)	414	10.177(-.06)	10.2189(+.06)
Group 83 (A4 FIRST)	415	10.411(+.06)	10.283(-.06)

Table 3

Analysis of Covariance: Logical Reasoning

Source	df	Mean Square	F	p
Group (G)	1	5.360	.38	.537
V	1	3,438.533	244.36	.000
Q	1	176.379	12.53	.000
A	1	2,521.204	179.17	.000
Error	824	14.071		
Form (F)	1	.029	.00	.945
Practice Effect (GXF)	1	5.910	.97	.324
Error	827	6.066		

The practice effect is approximately .12 of an item for a 23-item test, or .005 per item, and the probability of an effect this large occurring by chance is .324. This estimate is about one fourth of that for analytical reasoning and logical reasoning combined in Study 1. To the extent that experience with four items of this type in the morning may have inflated scores, this may be a slight underestimate of the effect for logical reasoning.

It is interesting to note that the covariate that accounted for the most variance in logical reasoning was the GRE verbal score, with the analytical score accounting for somewhat less and the quantitative score contributing much less to prediction of the experimental test scores.

The analysis of the supporting conclusions item type led to the result that a small, but statistically significant, practice effect was present in the data. Adjusted cell means and the analysis of covariance results are given in Tables 4 and 5.

Table 4

Means and Deviations from Additivity:
Supporting Conclusions

Group 82 (S1 FIRST)	414	10.878(-.32)	11.262(+.32)
Group 83 (S3 FIRST)	415	11.308(+.32)	10.428(-.32)

Table 5

Analysis of Covariance: Supporting Conclusions

Source	df	Mean Square	F	p
Group (G)	1	16.826	1.01	.315
V	1	1,106.896	66.56	.000
Q	1	381.441	22.94	.000
A	1	2,529.746	152.13	.000
Error	824	16.629		
Form (F)	1	25.438	2.92	.088
Practice Effect (GXF)	1	165.450	19.00	.000
Error	827	8.709		

The practice effect is approximately .64 of an item for a 25-item test, or .026 per item, considerably less than that for logical diagrams in Study 1. The probability of an effect this large occurring by chance is less than .0001.

It is also interesting to note that the GRE analytical score accounted for about as much variance in the supporting conclusions scores as it did for logical reasoning, but that verbal ability was much less strongly related to supporting conclusions than was analytical ability. This contrasts with the relationship of verbal and analytical scores to logical reasoning.

The analyses suggest that the two logical reasoning forms were almost perfectly equated when the nonsignificant differences in group ability were taken into account, but that supporting conclusions form S1 was easier than form S2 by about .24 of an item, a degree approaching statistical significance ($p < .09$).

Statistical significance and practical significance are different. The practice effect for analysis of explanations is considered to be of practical significance, and a comparison of the .06 point per item analysis of explanations practice effect with the .026 points per item effect obtained for supporting conclusions is of help in evaluating the practical significance of this latter result. We believe that the practical significance of an effect of this size is of marginal concern.

Results: Study 3

Because 20 repeated measures analyses were required to assess practice effects for the entire set of operational item

types, the results will not be presented in detail, as were the results for Study 2. Rather, the values of the F tests for order effects (e.g., group-by-form interactions) and estimated effects per item are summarized for all three studies in Table 6. In study 3, there were two forms of each analytical section, yielding one estimate of practice effect, and four half-forms of each verbal and quantitative section, yielding two estimates of practice effect for each item type.

Examination of Table 6 reveals practically significant practice effects for analysis of explanations (.06 per item) and for logical diagrams (.04 per item). The estimate for supporting conclusions (.03 per item) is next highest among analytical item types, with analytical reasoning (.02 per item) and logical reasoning (.01 and -.01 per item) yielding the least evidence of practice effect.

It is not appropriate to compare verbal and quantitative practice effect estimates with the analytical section estimates, because the breaking of 50-minute sections into two 25-minute sections clearly added to the difficulty of the items in the shorter section. An analysis examining not total score but item difficulty, taking into account the ability of those attempting later items in each section, would be needed to make such comparisons. Such an analysis of these same data is reported by Kingston and Dorans (1982). However, within the verbal and quantitative sections, it appears that analogies show consistent relatively larger practice effects (whether these are absolutely positive cannot be determined from these data). The significant negative effect for one form of reading comprehension appears to be an artifact of placing a passage from the middle of the 50-minute section at the end of a 25-minute section. However, the other form yields essentially zero estimates (-.01 for explicit questions and .01 for inferential questions).

In mathematics, there is no evidence of a relatively more positive practice effect for quantitative comparisons (-.02) than for discrete quantitative questions (0 and -.03), while the evidence for the small number of data interpretation items is inconsistent. Analysis of the median position of items in the two forms in which they appeared suggests that, in some cases, later position could contribute to apparent negative practice effects, but does not explain all cases in which they appear. One possible hypothesis is that a certain proportion of examinees gambled that the half-length reading and math sections probably did not count in their scores and did not give them full effort. In any case, only relative comparisons in the verbal and quantitative sections are appropriate within the constraints of this design.

Table 6

Summary of Practice Effects Analyses, Studies 1, 2, and 3

Study	Item Type	n Items	n Individuals	Practice Effect	Practice Effect/Item	F
Study 1	Analysis of Explanations	40	2895	2.21	.06	372.87**
	Analysis of Explanations	40	2841	2.51	.06	570.46**
	Analysis of Explanations	40	2781	2.58	.06	552.04**
	Logical Diagrams	15	2767	.78	.05	213.42**
	Logical Diagrams	15	2814	.67	.04	113.18**
	Logical Diagrams	15	2853	.59	.04	139.04
	Analytical and Logical Reasoning	15	2867	.35	.02	34.60**
	Analytical and Logical Reasoning	15	2814	.29	.02	10.54**
	Analytical and Logical Reasoning	15	2853	.19	.01	23.13**
Study 2	Supporting Conclusions	25	829	.64	.03	19.00**
	Logical Reasoning	23	829	.12	.01	.97

Table 6 (Cont'd.)

Study	Item Type	n Items	n Individuals	Practice Effect	Practice Effect/Item	F
Study 3	Logical Reasoning	4	3958	- .02	-.01	.73
	Analysis of Explanations	38	4005	2.44	.06	852.40**
	Logical Diagrams	15	3958	.53	.04	107.82**
	Analytical Reasoning	11	3958	.19	.02	21.25
	Reading Comprehension (explicit)	6	3970	- .06	-.01	6.44*
	Reading Comprehension (explicit)	5	3976	- .25	-.05	96.35**
	Reading Comprehension (inferential)	7	3970	.07	.01	4.68*
	Reading Comprehension (inferential)	6	3976	- .37	-.06	122.39**
	Sentence Completion	7	3970	.00	.00	0.00
	Sentence Completion	8	3976	- .07	-.01	3.23
	Analogies	8	3970	.18	.02	31.87**
	Analogies	9	3976	.13	.01	16.61**
	Antonyms	10	3970	.20	.02	28.81**
	Antonyms	11	3976	.02	.00	43.99**
	Discrete Quantitative	8	3982	- .23	-.03	47.46**
	Discrete Quantitative	7	3976	- .01	-.00	0.04
	Quantitative Comparison	15	3982	- .26	-.02	33.60**
	Quantitative Comparison	15	3976	.26	-.02	30.25**
	Data Interpretation	4	3982	- .13	-.03	24.98**
	Data Interpretation	6	3976	.26	.04	61.11**

*$p < .05$

**$p < .01$

Conclusions

Results from these analyses show a large practice effect for analysis of explanations, moderate practice effects for logical diagrams and supporting conclusions, and an inconsistent effect for data interpretation. Little evidence of practice effect was found for analytical reasoning, logical reasoning, analogies, or antonyms. The reading comprehension and mathematics item types generally yielded negative order effect estimates. This may have been a result of a larger proportion of test-taking time required to read instructions when the items from a 50-minute section are divided into two 25-minute sections. In the case of reading comprehension, passage three out of four in the 50-minute section became passage two out of two in one of the 25-minute sections. Thus, there was a greater opportunity to be "caught by the bell" in this 25-minute section, contributing to the large estimated negative effect for that one of the two forms. It was not possible to obtain separate estimates of the effects of familiarity with instructions, as opposed to practice on individual items, in this study. It seems reasonable that both factors contribute to the observed total effects.

Although many statistically significant effects are reported, the practically significant results are those for analysis of explanations and logical diagrams, and, marginally, for supporting conclusions.

One hypothesis to explain practice effect has been that fixed-format questions (those whose response options are the same for all questions) may be more susceptible to practice (Vernon, 1954). In the current study, analysis of explanations, supporting conclusions, quantitative comparisons, and, to some extent, logical diagrams are fixed-format items. Results from this study tend to support this hypothesis in three cases out of four (logical diagrams, supporting conclusions, and analysis of explanations, but not in the case of quantitative comparisons).

Swinton and Powers (1983) obtained similar results in a small special preparation study of the experimental analytical section. As a result of that report, of the current study, and of other factors, the GRE Board decided to restructure the analytical measure to consist of only analytical reasoning and logical reasoning items.

We do not regard the existence of small practice effects as evidence against the validity of an item type. Indeed, an item type that is totally immune to experience would be hard to justify as a measure of developed ability. However, when within-test practice effects are encountered comparable to

those found here for analysis of explanations, not only equating and test development procedures, but also questions of equity demand serious consideration of the appropriateness of the item type. A problem that was initially encountered in the course of routine retesting, and that could have been "solved" by a rule-of-thumb adjustment of estimated item difficulties, has instead led to a thorough investigation of the range of practice effects across the entire range of GRE Aptitude Test item types and to fundamental corrective action. We believe that such joint attention to test quality by the GRE Board, Program, and Research staff is an important component of efforts to assure equity in graduate admissions procedures.

References

Kingston, N. M., & Dorans, N. J. The effect of the position of an item within a test on item responding behavior: An analysis based on item response theory (GRE Board Professional Report No. 79-12bP). Princeton, N.J.: Educational Testing Service, 1982.

Swinton, S. S., & Powers, D. E. A study of the effects of special preparation on GRE analytical scores and item types. Journal of Educational Psychology, 1983, 75(1), 104-115.

Vernon, P. E. Symposium on the effects of coaching and practice in intelligence tests. British Journal of Educational Psychology, 1954, 24, 57-63.

Wild, C. L., Swinton, S. S., & Wallmark, M. Research leading to the revision of the format of the Graduate Record Examinations Aptitude Test in October 1981 (GRE Board Professional Report 80-1bP). Princeton, N.J.: Educational Testing Service, 1982.

SECTION VII

Time—30 minutes

35 Questions

PLEASE NOTE THAT YOUR ANSWER SHEET HAS FIVE ANSWER POSITIONS MARKED A, B, C, D, E WHEREAS THE QUESTIONS IN THIS SECTION CONTAIN ONLY FOUR CHOICES. BE SURE NOT TO MAKE ANY MARKS IN COLUMN E.

Directions: Each passage in this section consists of numbered sentences. Following each passage are several statements. Under each statement, answer choices A, B, and C give the numbers of single sentences or combinations of sentences in the passage.

You are to decide whether the information in the statement is explicitly supplied by the passage. If it is, decide whether A, B, or C lists the fewest number of sentences from the passage necessary to supply all of the information in the statement. If one of them does, it is the correct answer, and you should blacken the corresponding space on your answer sheet. If the information in the statement is not explicitly supplied by the passage or if none of the answer choices A, B, or C lists the fewest number of sentences necessary to supply the information in the statement, blacken space D on your answer sheet.

SAMPLE SET

(1) A recent court ruling held that citizens defamed by public officials cannot sue in federal courts for deprivation of their civil rights. (2) This decision, made by the United States Supreme Court, did not affect the right to sue public officials in state courts for defamation. (3) However, such suits are difficult to bring in many states, because of statutes or legal rules favoring public officials in civil suits. (4) The decision appeared to extend a recent trend toward denying access to federal court by persons complaining of constitutional violations by public officials. (5) The Court divided five to three in this decision.

1. Statement: In a split decision the United States Supreme Court seemed to continue a recent trend toward limiting citizens' access to federal court for suits involving the violation of constitutional rights by public officials.

   (A) 1, 2, and 4
   (B) 2, 4, and 5
   (C) 1, 2, 3, 4, and 5
   (D) None of the above lists the fewest number of sentences necessary to supply the information in the statement.

   Ⓐ ● Ⓒ Ⓓ

2. Statement: The United States Supreme Court ruled that citizens defamed by public officials cannot sue for deprivation of their civil rights.

   (A) 1
   (B) 1 and 2
   (C) 1, 2, and 4
   (D) None of the above lists the fewest number of sentences necessary to supply the information in the statement.

   Ⓐ Ⓑ Ⓒ ●

ATTACHMENT II
Analytical Reasoning Item Types
Part B
(Suggested time—19 minutes)
15 Questions

Directions: Each question or group of questions is based on a passage or set of statements. In answering some of the questions it may be useful to draw a rough diagram. Choose the best answer for each question and blacken the corresponding space on your answer sheet.

Questions 16–17 **Type I**

In order to remodel her house, Joan has hired a plumber, a brickmason, an electrician, and a painter.
The plumber is available only on Monday morning, all day Tuesday, and on Wednesday afternoon.
The brickmason is available only on Monday afternoon, Wednesday morning, and all day Friday.
The electrician is available only all day Wednesday, Thursday, and Friday.
The painter is available only all day Tuesday and on Friday morning.

16. One of the workers asks to spend an entire day working alone in the house. Joan can grant this request, without losing any of another worker's available time, if the worker making the request is the

(A) plumber and the day is Monday
(B) painter and the day is Tuesday
(C) plumber and the day is Wednesday
(D) electrician and the day is Thursday
(E) brickmason and the day is Friday

17. The painter will need only half a day to do the job required, but cannot begin until all the other workers have finished. If the work begins on Monday, what is the earliest possible time the painter could begin?

(A) Tuesday afternoon
(B) Wednesday morning
(C) Thursday afternoon
(D) Friday morning
(E) Friday afternoon

Questions 18–20 **Type II**

Recent studies which prove that one may induce violent behavior in rats by crowding them together lend support to the view that the rising rate of violent crime in the cities is the result of crowding.

18. The argument above makes which of the following assumptions?

I. Rising rates of violent crime are a national catastrophe.
II. Conclusions about human behavior may be drawn from rat behavior.
III. It is not inhumane to do psychological experiments on rats.

(A) I only (B) II only (C) III only
(D) I and II only (E) I and III only

19. The argument would be strengthened by pointing out that

(A) rats are often a serious health problem in the city
(B) controversy exists over how to compute crime figures
(C) only one breed of rat was tested in the studies
(D) nonviolent crime is also on the rise
(E) a similar study of elephants produced a similar result

20. The argument would be weakened by pointing out that

(A) the urban crime rate has increased whereas crowding has decreased
(B) a blue-ribbon commission has been studying the causes of violence
(C) numerous independent studies confirmed the effects of crowding on rats
(D) many crimes are not reported to the police
(E) government rat-control measures have become increasingly effective

Questions 21–23 **Type I**

(1) Two men (George and Dave) and four women (Betsy, Ann, Ellen, and Carla) are seated around a circular table with ten seats.
(2) No two people of the same sex are sitting in adjacent seats.
(3) Carla sits next to George.
(4) Ellen sits between George and Dave and next to each of them.
(5) There is an empty seat next to Dave.
(6) There are fewer than three empty seats between Betsy and Ann.

21. Which of the six statements repeats information available elsewhere in the set of statements?

(A) (2) (B) (3) (C) (4) (D) (5) (E) (6)

22. If the number of empty seats between Ann and the next person on her right is added to the number of empty seats between Ann and the next person on her left, the sum must be either

(A) one or two (B) one or three (C) two or three
(D) two or four (E) three or four

23. Leonora takes a seat at the table, and she does not sit next to another woman. If no one has moved to accommodate her, it must be true that

(A) Leonora sits next to Dave
(B) Leonora sits next to George
(C) there is one empty seat between Leonora and Ann
(D) there is one empty seat between Leonora and Betsy
(E) there are two empty seats between Ann and Betsy

Questions 24–25 **Type II**

If people continue to reproduce at their present rate, the earth's population will double in 35 years. Therefore, it is not enough for us in the United States to keep our population from growing; we must decrease our birth rate.

24. The argument above is based on the assumption that

(A) the present world population is at an optimum size
(B) most Americans are anxious to decrease the birth rate
(C) government regulation of population growth is inevitable
(D) doubling the earth's population is undesirable
(E) the population explosion is unmanageable

25. The argument presented would be strongest if it were true that

(A) the birth rate in the United States has been steadily rising
(B) infant mortality in the United States has been steadily decreasing
(C) a drop in the birth rate of the United States would significantly affect world population
(D) the United States has already decreased its population more than has any other country in the world
(E) the population of the United States has doubled in the past 35 years

Questions 26–30 Type I

Professor Green is choosing a four-member research team from graduate students F, G, and H and undergraduate students W, X, Y, and Z.
There are to be at least two graduate students on the team.
Student F refuses to work with student Y.
Student G refuses to work with student W.
Student Y refuses to work with student Z.

26. If student Y is chosen, which of the following must be the other members of the research team?

(A) F, G, and X (B) G, H, and W (C) G, H, and X
(D) G, H, and Z (E) H, W, and X

27. If student Z is chosen and student F is rejected as a member of the research team, which of the following must be the members of the research team?

(A) G, H, W, and Z (B) G, H, X, and Z
(C) G, H, Y, and Z (D) G, X, Y, and Z
(E) H, W, X, and Z

28. If student G is chosen and student H is rejected as a member of the research team, which of the following statements must be true?

I. Student X is chosen.
II. Student Z is chosen.

(A) I only (B) II only (C) Either I or II but not both
(D) Both I and II (E) Neither I nor II

29. Which of the following must be true?

I. Students W and Y never work together.
II. Students X and Y always work together.
III. If student W works, student H works.

(A) I only (B) I and II only (C) I and III only
(D) II and III only (E) I, II, and III

30. Which of the following must be true?

I. If student F works, student Z works.
II. If student F does not work, student W does not work.
III. If student F does not work, student H works.

(A) I only (B) III only (C) I and II only
(D) I and III only (E) II and III only

STOP

IF YOU FINISH BEFORE TIME IS CALLED YOU MAY CHECK YOUR WORK ON PARTS A AND B OF THIS SECTION ONLY.
DO NOT WORK ON ANY OTHER SECTION IN THE TEST.

GRE BOARD RESEARCH REPORTS OF A TECHNICAL NATURE

Boldt, R. R. Comparison of a Bayesian and a Least Squares Method of Educational Prediction. 70-3P, 1975.

Campbell, J. T. and Belcher, L. H. Word Associations of Students at Predominantly White and Predominantly Black Colleges. 71-6P, 1975.

Campbell, J. T. and Donlon, T. F. Relationship of the Figure Location Test to Choice of Graduate Major. 75-7P, 1980.

Carlson, A. B.; Reilly, R. R.; Mahoney, M. H.; and Casserly, P. L. The Development and Pilot Testing of Criterion Rating Scales. 73-1P, 1976.

Carlson, A. B.; Evans, F.R.; and Kuykendall, N. M. The Feasibility of Common Criterion Validity Studies of the GRE. 71-1P, 1974.

Centra, J. A. Graduate Degree Aspirations of Ethnic Student Groups Among GRE Test-Takers. 77-7P, 1980.

Donlon, T. F. An Exploratory Study of the Implications of Test Speededness. 76-9P, 1980.

Donlon, T. F.; Reilly, R. R.; and McKee, J. D. Development of a Test of Global vs. Articulated Thinking: The Figure Location Test. 74-9P, 1978.

Echternacht, G. Alternate Methods of Equating GRE Advanced Tests. 69-2P, 1974.

Echternacht, G. A Comparison of Various Item Option Weighting Schemes/A Note on the Variances of Empirically Derived Option Scoring Weights. 71-17P, 1975.

Echternacht, G. A Quick Method for Determining Test Bias. 70-8P, 1974.

Evans, F. R. The GRE-Q Coaching/Instruction Study. 71-5aP, 1977.

Fredericksen, N. and Ward, W. C. Development of Measures for the Study of Creativity. 72-2P, 1975.

Kingston, N. and Dorans, N. Effect of the Position of an Item Within a Test on Item Responding Behavior: An Analysis Based on Item Response Theory. 79-12bP, 1982.

Kingston, N. M. and Dorans, N. J. The Feasibility of Using Item Response Theory as a Psychometric Model for the GRE Aptitude Test. 79-12P, 1982.

Levine, M. V. and Drasgow, F. Appropriateness Measurement with Aptitude Test Data and Esimated Parameters. 75-3P, 1980.

McPeek, M.; Altman, R. A.; Wallmark, M.; and Wingersky, B. C. An Investigation of the Feasibility of Obtaining Additional Subscores on the GRE Advanced Psychology Test. 74-4P, 1976.

Oltman, P. K. Content Representativeness of the Graduate Record Examinations Advanced Tests in Chemistry, Computer Science, and Education. 81-12P, 1982.

Pike, L. Implicit Guessing Strategies of GRE Aptitude Examinees Classified by Ethnic Group and Sex. 75-10P, 1980.

Powers, D. E.; Swinton, S.; Thayer, D.; and Yates, A. A Factor Analytic Investigation of Seven Experimental Analytical Item Types. 77-1P, 1978.

Powers, D. E.; Swinton, S. S.; and Carlson, A. B. A Factor Analytic Study of the GRE Aptitude Test. 75-11P, 1977.

Reilly, R. R. and Jackson, R. Effects of Empirical Option Weighting on Reliability and Validity of the GRE. 71-9P, 1974.

Reilly, R. R. Factors in Graduate Student Performance. 71-2P, 1974.

Rock, D. A. The Identification of Population Moderators and Their Effect on the Prediction of Doctorate Attainment. 69-6bP, 1975.

Rock, D. A. The "Test Chooser": A Different Approach to a Prediction Weighting Scheme. 70-2P, 1974.

Rock, D., Werts, C., and Grandy, J. Construct Validity of the GRE Aptitude Test Across Populations--An Empirical Confirmatory Study. 78-1P, 1982.

Sharon, A. T. Test of English as a Foreign Language as a Moderator of Graduate Record Examinations Scores in the Prediction of Foreign Students' Grades in Graduate School. 70-1P, 1974.

Stricker, L. J. A New Index of Differential Subgroup Performance: Application to the GRE Aptitude Test. 78-7P, 1981.

Swinton, S. S. and Powers, D. E. A Factor Analytic Study of the Restructured GRE Aptitude Test. 77-6P, 1980.

Ward, W. C. A Comparison of Free-Response and Multiple-Choice Forms of Verbal Aptitude Tests. 79-8P, 1982.

Ward, W. C.; Frederiksen, N.; and Carlson, S. B. Construct Validity of Free-Response and Machine-Scorable Versions of a Test of Scientific Thinking. 74-8P, 1978.

Ward, W. C. and Frederiksen, N. A Study of the Predictive Validity of the Tests of Scientific Thinking. 74-6P, 1977.

Wild, C. L., Swinton, S. S., and Wallmark, M. M. Research Leading to the Revision of the Format of the Graduate Record Examinations Aptitude Test in October 1981. 80-1bP, 1982.