DOCUMENT RESUME

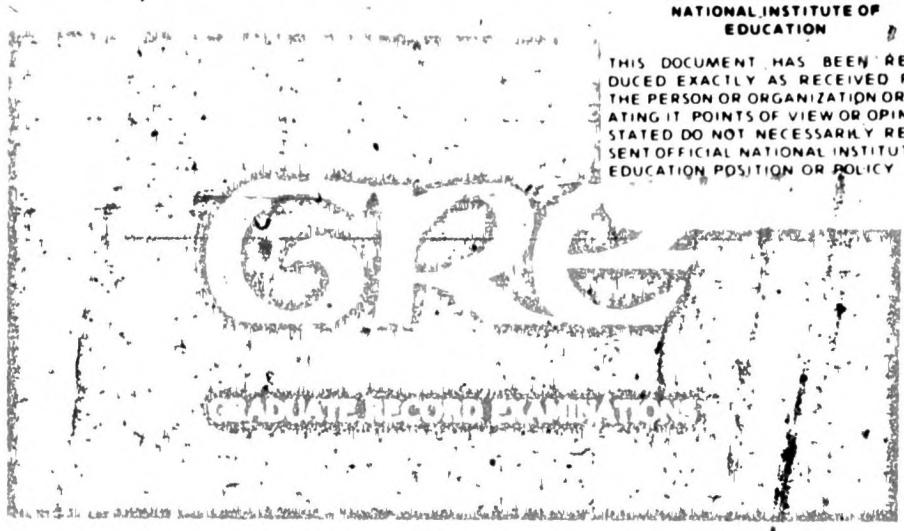
ED 181 026 TM 009 462

AUTHOR Wild, Cheryl L.
TITLE Summary of Research on Restructuring the Graduate Record Examinations Aptitude Test.
INSTITUTION Educational Testing Service, Princeton, N.J.
PUB DATE Feb 79
NOTE 14p.
AVAILABLE FROM Graduate Record Examinations, Educational Testing Service, Princeton, NJ 08541 (free while supplies last)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Abstract Reasoning; *Aptitude Tests; *College Entrance Examinations; College Mathematics; Comparative Testing; *Critical Thinking; Graduate Study; Higher Education; Technical Reports; *Test Construction; Thought Processes; Verbal Tests
IDENTIFIERS *Graduate Record Examinations; Quantitative Tests; Test Equivalence; *Test Format; Test Length

ABSTRACT

Three sections of the Graduate Record Examinations (GRE) Aptitude Test were reviewed before the introduction of the restructured test in October, 1977: research on (1) the GRE-Verbal section; (2) the GRE-Quantitative section; and (3) a planned third section, measuring analytical thinking skills. Research in all three areas focused on test reliability, validity, difficulty, speededness, and equivalence of restructured and former test sections. The restructured verbal measure was shortened from 75 to 50 minutes, and included a long as well as a short reading comprehension passage. Research on the quantitative ability test involved combinations of three item types: regular mathematics, quantitative comparison, and data interpretation. The restructured test was reduced from 75 to 50 minutes, and contained about thirty quantitative comparison items in place of regular mathematics and data interpretation items. Seven new item types were evaluated for inclusion in the abstract/analytical reasoning test, based upon their difficulty, reliability, speededness, validity, appropriateness for all college majors, efficiency, and independence from the other two tests. Three of the seven item types were accepted for use in the new GRE: analytical reasoning, logical diagrams, and analysis of explanations. (GDC)



SUMMARY OF RESEARCH ON RESTRUCTURING THE GRADUATE RECORD EXAMINATIONS APTITUDE TEST

Cheryl L. Wild

February 1979

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

EDUCATIONAL TESTING SERVICE PRINCETON NJ

SUMMARY OF RESEARCH ON RESTRUCTURING THE GRADUATE RECORD EXAMINATIONS APTITUDE TEST

Cheryl L. Wild

February 1979

SUMMARY OF RESEARCH ON RESTRUCTURING THE GRADUATE RECORD EXAMINATIONS APTITUDE TEST

This paper reviews the research on which the introduction (in October 1977) of the restructured Graduate Record Examinations (GRE) Aptitude Test was based. Consideration of a new test format began early in 1974, when the GRE Board and its Research Committee began a systemmatic review of the GRE Program offerings. In April 1975, a model for further research and development of the test format was proposed to these groups by staff and approved in principle. The goal of this research was to broaden the Aptitude Test and thus enable students to demonstrate a wider array of academic talents. The research can be divided into three areas: (1) research on the GRE verbal (GRE-V) section of the test; (2) research on the GRE quantitative (GRE-Q) section of the test; and (3) research to develop a third module, GRE analytical (GRE-A) that would allow for broadening skill measurement.

Research in all three areas focused on reliability, validity, difficulty, speededness, and comparability of the restructured to old format test sections. Technical definitions of each of these terms are presented in depth in the GRE Technical Manual (Conrad, Trismen, & Miller, 1977). Briefly, reliability is the extent to which a test is consistent in measuring whatever it measures. Validity is the extent to which a test measures what it purports to measure. Several types of validity exist: "face" validity, the extent to which the test questions appear to be related to the appropriate ability; "criterion" validity, the extent to which the test score is related to other measures taken at the same time (e.g., the relationshp of GRE scores with undergraduate grades) or at a future time (e.g., the relationship of GRE test scores with first-year graduate grade-point averages); and "construct" validity, the extent to which test scores relate to other measures (e.g., other ability measures) in a predictable manner. Validity and reliability are interdependent - a test can be very reliable but not valid, while an unreliable test cannot be valid. Difficulty of a test is measured by the proportion of examinees who answer each question correctly. The appropriate difficulty of a test depends on how the test is used and is related to reliability. Speededness is the extent to which test scores are related to the time limits of the test rather than the examinee's ability to answer the questions. Comparability of scores refers to whether scores on two forms of the same test have the same meaning.

In the process of investigating restructuring, it was found that both verbal and quantitative sections could be revised while maintaining comparability of scores and appropriate reliability, difficulty and speededness. Seven item formats were investigated for inclusion in the analytical measure, from which three were chosen for inclusion in the operational measure. Detailed summaries of this research are presented in the following sections.

GRE Verbal

Prior to October 1977, the verbal ability measure of the GRE Aptitude Test consisted of two sections - a 25-minute section of discrete verbal questions including analogies, antonyms, and sentence completions and a 50-minute section of questions based on six long reading passages. The verbal measure of the restructured GRE Aptitude Test consists of one 50-minute section including discrete verbal questions and reading comprehension questions based on both short and long passages.

This outcome was slightly different than that anticipated in the beginning of the research effort. Altman and Conrad (1979) investigated the following questions concerning restructuring of the verbal measure:

(1) Can the verbal ability measure be shortened while retaining appropriate reliability, difficulty, speededness, and validity?

(2) Can reading comprehension subscores be provided?

(3) Can reading subscores be based on different reading comprehension sections for students with different undergraduate majors?

Four 25-minute experimental tests of reading comprehension questions were included in the October 1975 GRE national test administration. They contained the following reading comprehension questions and content: (1) 25 questions based on humanities and social sciences, (2) 30 questions on humanities and social sciences; (3) 25 questions on natural and physical sciences; and (4) 30 questions on natural and physical sciences. Over 8,000 examinees took each of these experimental tests.

Using item analyses of the responses to the experimental tests, estimates of difficulty, speededness, and reliability of each of the four experimental tests were obtained for samples of humanities and social science majors and biological and physical science majors.

Results of the investigation indicated that reliability, item difficulty, and speededness of a 50-minute verbal section would be psychometrically appropriate. Correlations with self-reported undergraduate grades also suggested that shortening the reading comprehension section would have no substantial effect on validity. However, it was found that the humanities and social sciences group and natural and physical sciences group performed differently on the experimental tests depending on whether the reading material was from the "appropriate" discipline. For this reason, introducing optional reading sections would mean that comparable verbal scores could not be provided.

Finally, the feasibility of creating a reading subscore was considered. Although such a subscore was feasible, weaknesses were identified by the factor analysis study (Powers, Swinton, & Carlson, 1977). The factor analysis showed that a reading comprehension score should include not only reading comprehension but also sentence completion items. Although the sentence completion items would be included based on the statistics, sentence completion items would not have face validity as part of a reading comprehension subscore. Another problem was the high correlation of the proposed subscore with the total score. The idea of a reading subscore was therefore abandoned.

Additional input on restructuring the verbal section was obtained from student and institutional surveys. Approximately equal percentages (72%) of students and institutions favored offering optional reading comprehension sections if possible. Students also favored the inclusion of short as well as long reading comprehension passages.. Although optional reading comprehension sections were not feasible, as discussed above, both short and long reading comprehension passages are included in the restructured test.

GRE Quantitative

Prior to October 1977, the quantitative ability section of the GRE Aptitude Test was 75 minutes long, containing 55 questions on data interpretation, geometry, arithmetic, algebra, and miscellaneous item types. The restructured section lasts 50 minutes and contains the same number of items with about 30 quantitative comparison items in place of some of the regular math and data interpretation items.

Altman and Conrad (1979) investigated the feasibility of achieving a quantitative ability section that was only 50 minutes in length. The research questions were:

(1) Can the quantitative ability measure be shortened while retaining appropriate reliability, difficulty, speededness, and validity?

(2) Can a subscore for data interpretation be provided?

These questions were answered by analyzing results of four experimental tests given in the October 1975 test administration. Over 8,000 examinees took each of the following 25-minute tests:

(1) 30 regular math questions (those currently in the test, excluding data interpretation items),

(2) 40 quantitative comparison questions,

(3) 35 questions including quantitative comparisons plus mixed regular types (designed to be parallel to a module of the proposed test), and

(4) 20 data interpretation questions designed to comprise the second module for the proposed test.

Samples of all examinees taking each of the four experimental tests were selected and item difficulty, reliability, and speededness calculated for each sample. The second and third tests were also evaluated for a sample of humanities and social sciences majors and a sample of biological and physical sciences majors. Difficulty indices indicated that for each of the item types items could be written with appropriate difficulty for a final test form. Speededness information showed that the quantitative comparisons plus mixed regular items were approximately as speeded as the operational quantitative section, with the other three experimental test modules being more speeded. Reliability information suggested that both quantitative comparisons and the mix of quantitative comparisons and regular item types were about equally reliable.

Comparability of scores based on the existing quantitative section and the proposed quantitative section including quantitative comparisons was also investigated. Comparisons of performance of undergraduate majors in the humanities and social sciences and natural sciences majors on quantitative comparisons in relation to their performance on the operational quantitative section were made. Differences in performance were slightly magnified when the comparison was based on all quantitative comparison material; however, this was judged as not significant since quantitative comparisons would only be a portion of the restructured quantitative section.

A second measure of comparability was obtained by reviewing a factor analytic study based on the same experimental tests (Powers, Swinton, and Carlson, 1977). This indicated that slightly different patterns of abilities underlie performance on regular mathematics, data interpretation, and quantitative comparisons questions, although differences were not considered large.

A third method of checking comparability involved correlating each of the four experimental tests with the operational section and correcting for attenuation (i.e., for the differneces in raliability of the experimental tests). Correlations of regular mathematics, quantitative comparisons, quantitative comparisons plus regular mathematics, and data interpretation experimental modules with the quantitative score were .99, .96, .97 and .97 respectively.

Finally, Altman and Conrad (1979) evaluated the concurrent validity of the proposed item types by comparing correlation of self-reported grades with the operational quantitative score, the

quantitative comparisons module, and the quantitative comparisons plus regular math module. Correlations ranged between .24 and .29, with virtually no difference between the operational section and the experimental modules.

Conrad and Altman (1979) also surveyed institutions and students to elicit reactions to restructuring the quantitative section of the test. Although a majority of the 1,530 students surveyed favored shortening the quantitative section, differences were found by major field. Almost two-thirds of the humanities majors, approximately one-half of the social sciences majors, and less than one-third of the natural sciences majors were in favor of a shortened quantitative section. Ninety percent of the institutional representatives favored shortening both verbal and quantitative measures to provide for a new measure within current time allotments.

To summarize, information from the Altman and Conrad (1979) study indicated that by including quantitative comparison items in the quantitative section, reliability could be maintained while the amount of time necessary to complete the section could be decreased. Too high a proportion of data interpretation items detracted from reliability and increased speededness. Inclusion of quantitative comparisons would not decrease validity. Results suggest that, although the regular mathematics, data interpretation, and quantitative comparison questions are not parallel in the strict sense, the item types could be used together to obtain a quantitative score comparable to the existing score. Finally, results indicated that a data interpretation subscore based on a 25-minute module should not be provided, due to reliability and speededness considerations.

GRE Analytical

Throughout the discussions on restructuring the Aptitude Test, one idea remained constant - the addition of a third module to broaden skills measured by the Aptitude Test. In early discussions of the third module, many options were considered. It was decided to focus on a module requiring minimal research for the present restructuring, while continuing to do research on theoretical measures such as scientific thinking and documented accomplishments.

Altman and Conrad (1979) surveyed institutions and students. Of the possible new measures listed (abstract reasoning, scientific thinking, and "study style"), abstract reasoning was favored by both faculty and students. Based on this interest, interest expressed by the Board, and the availability of item types, it was decided to try to develop a new reasoning module.

Seven item types were identified as possible components of the new module. These were included as experimental sections of regular national administrations during the 1975-76 academic year. Each of the seven modules (including mixtures of question types) are described briefly below. For further information and sample items see the GRE Technical Manual (Conrad, Trismen, & Miller, 1977).

(1) Letter Sets - Each item consists of five groups of letters, only one group of which is unlike the others in alphabetic arrangement. The examinee's task is to identify that dissimilar group. This item type originated in the Kit of Factor-Referenced Cognitive Tests and was intended to measure inductive reasoning.

(2) Logical Reasoning and Letter Sets - Logical Reasoning items are based on brief arguments or statements presenting evidence of opinions. The questions require recognition of unstated presuppositions, logical flaws, methods of persuasion, and conclusions logically following from arguments. From earlier pretests, it was known that the item type correlated highly with the verbal score. However, it was hypothesized that a combination of Logical Reasoning and Letter Sets might be appropriate as a measure of reasoning. Experiments in the Law School Admission Test Program showed that Logical Reasoning questions had high criterion validity.

(3) Analytical Reasoning and Letter Sets - Analytical Reasoning questions are based on brief sets of statements expressing relationships among abstract symbols (letters) or sets of rules governing processes or procedures having few concrete referents. The examinee draws inferences from and/or critically assesses those sets of statements. It was hypothesized that the combination of Letter Sets and Analytical Reasoning items might be appropriate for the new module.

(4) Evaluation of Evidence - These questions are based on a brief narrative establishing a situation and a conclusion drawn from the facts presented. The items consist of bits of evidence that, in relation to the situation described, strengthen, weaken, confirm, disprove, or fail to affect the conclusion.

(5) Analysis of Explanations - These questions are based on brief narratives establishing a situation in which an action is taken in order to have a specific effect. A later result, which may or may not be directly related to the action, is described in a brief statement. Each question is a piece of information that must be evaluated in terms of facts and results.

(6) Logical Diagrams - This item type is derived from Venn diagrams and has been used in the Kit of Factor-Referenced Cognitive Tests. Each item consists of three nouns and the examinee is asked to select the circle diagram that best characterizes the relationship of the three.

(7) Deductive Reasoning - This item type consists of a relatively complex set of rules which the student is asked to apply in solving problems based on diagrams.

Research was designed to answer the following questions about the above item types:

1) Will the item types yield material of appropriate difficulty, reliability, and unspeededness?

2) Will the item types measure skills that are relatively independent of verbal and quantitative skills?

3) Will the item types have criterion validity?

4) What combination of item types appears to be best in terms of (a) efficiency, (b) face validity, (c) criterion validity, (d) independence of V and Q, and (e) appropriateness for both science and humanities-social science students.

Each of the experimental tests was taken by a substantial number of GRE test-takers. In all but one case, at least three samples were drawn: a representative sample of students, a sample of biological and physical sciences majors, and a sample of humanities and social science majors. In addition, separate analysis for logical (Venn) diagrams were based on samples of black males, black females, white males, and white females.

Statistical analyses were used to investigate the item types' efficiency, criterion validity, difficulty, reliability, speededness, independence of V and Q, and appropriateness for students with different academic backgrounds. To assess the face validity of each item type and the way in which different groups perceived its utility, surveys were administered to samples of students who had taken each of the experimental item types, and two student committees examined and reacted to samples of the item types. Presentations were made at a number of national and regional meetings of professional associations, and the item types were briefly discussed by some GRE Advanced Test committes of examiners. As a result of this work, it was possible to make a number of preliminary decisions about the appropriateness of each of the seven item types as a possible part of a new module.

Based on the results of both statistical analysis and questionnaire responses, the ratings for each of the item types are summarized in the following table, where (+) indicates a favorable

rating, (-) an unfavorable rating, and (0) a neutral rating. It should be noted that the ratings for a given item type are relative to the performance of the other item types as shown in the research.

Item Type	Difficulty	Efficiency	Face Validity	Criterion Validity	Independence of V and Q	
Letter Sets	-	+	0	0	+	+
Logical Reasoning	+	-	++	+	-	+
*Analytical Reasoning	+	-	0	+	+	+
Evaluation of Evidence	+	++	++	0	+	+
*Analysis of Explanations	+	++	0	++	+	+
*Logical Diagrams	+	++	++	-	+	+
Deductive Reasoning	+	--	+	-	+	+

*Chosen for inclusion in the analytical ability module.

Based on this research, three of the seven item types were accepted for use in the analytical ability module--analytical reasoning, analysis of explanations, and logical diagrams (asterisked item types in the table). These were chosen, not only for their individually good performance on the various criteria, but also for their combined balance on the criteria.

However, the evaluation of evidence item type was considered a viable alternative to the analysis of explanations item type and therefore warranted further study. Four hypotheses were studied by Thompson and Conrad (1979):

(1) That previous indications of student interest in evaluation of evidence were due primarily to the happenstance that

the subject matter content was more interesting, not to characteristics of the item type itself; that students will find analysis of explanations and evaluation of evidence equally interesting if topical content of the experimental tests is matched.

(2) That experimental tests combining evaluation of evidence and analysis of explanation items will be more speeded and will result in lower scores than tests containing only one item type.

(3) That previous indications of higher validity for analysis of explanations than for evaluation of evidence will be supported by this study, using the criteria of overall undergraduate grade-point average in the last two years and undergraduate grade-point average in major field. (This hypothesis was made despite the fact that the criterion data will differ in two ways from data collected in the previous study: a) the data will be on the registration form instead of the answer sheet; and b) the undergraduate grade-point average data do not match the overall four-year grade-point average data collected previously.)

Six new experimental modules were spiralled and administered at the April 1977 administration of the GRE. A sample of examinees were asked to comment on these modules. Examinees ranked evaluation of evidence items as more interesting than analysis of explanations. Most examinees would prefer having both types of items.

The evaluation of evidence test was more speeded than the analysis of explanations test. Tests composed of both item types were more difficult and speeded than tests containing only one type. Differences in validities of the two item types using the criteria of self-reported undergraduate grades in all four years, last two years, and major field were not marked.

Based on the Thompson and Conrad study, the decision was made to retain the original decision of an analytical measure consisting of analytical reasoning, logical diagrams, and analysis of explanations item types.

Conclusions

Since October 1977, the restructured GRE Aptitude Test has been administered. This test produces three scores -- verbal, quantitative and analytical ability. Further research on the analytical score is currently in progress, and institutions are advised to withhold the analytical score from use in the selection process until the relationship between the score and the performance

of graduate students is established. Further research on the first operational year of the restructured Aptitude Test will be published as it becomes available.

Cheryl L. Wild
Associate Program Director

References

Altman, R. & Conrad, L. Aptitude Test restructuring research: Proposed changes in verbal and quantitative measures. In R. Miller and C. L. Wild (Eds.), Restructuring the Graduate Record Examinations Aptitude Test, (GRE Board Technical Report). Princeton, N.J.: Educational Testing Service, 1979.

Altman, R. & Conrad, L. Aptitude Test restructuring research: Constituency surveys. In R. Miller and C. L. Wild (Eds.), Restructuring the Graduate Record Examinations Aptitude Test, (GRE Board Technical Report). Princeton, N.J.: Educational Testing Service, 1979.

Conrad, L. Aptitude Test restructuring research: Findings concerning projected development of an abstract reasoning measure. In R. Miller and C. L. Wild (Eds.), Restructuring the Graduate Record Examinations Aptitude Test, (GRE Board Technical Report). Princeton, N.J.: Educational Testing Service, 1979.

Conrad, L., Trismen, D. & Miller, R. (Eds.). GRE Technical Manual. Princeton, N.J.: Educational Testing Service, 1977.

Ekstrom, R. B., French, J. W., & Harmon, H. H. Kit of Factor-Referenced Cognitive Tests. Princeton, N.J.: Educational Testing Service, 1976.

Miller, R. & Wild, C. L. (Eds.). Restructuring the Graduate Record Examinations Aptitude Test. (GRE Board Technical Report). Princeton, N.J.: Educational Testing Service, 1979.

Powers, D. E., Swinton, S. S., & Carlson, A. B. A factor analytic study of the GRE Aptitude Test. (GRE Board Professional Report GREB No. 75-11P). Princeton, N.J.: Educational Testing Service, 1977.

Thompson, R. E., and Conrad, L. Aptitude Test restructuring research: Comparison of analysis of explanations and evaluation of evidence item types. In R. Miller and C. L. Wild (Eds.), Restructuring the Graduate Record Examinations Aptitude Test. (GRE Board Technical Report). Princeton, N.J.: Educational Testing Service, 1979.