DOCUMENT RESUME

ED 395 020 TM 025 037

AUTHOR	Wild, Cheryl L.; And Others
TITLE	Concurrent Validity of Verbal Item Types for Ethnic and Gender Subgroups. GRE Board Professional Report No. 84-10P.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RR-89-23
PUB DATE	Dec 89
NOTE	48p.
PUB TYPE	Reports - Research/Technical (143) -- Information Analyses (070)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Comparative Analysis; *Concurrent Validity; *Ethnic Groups; Grade Point Average; Higher Education; Majors (Students); Reading Comprehension; Sample Size; *Sex Differences; Test Construction; *Test Items; Test Validity; *Undergraduate Students; *Verbal Tests
IDENTIFIERS	Analogies; *Graduate Record Examinations

ABSTRACT

The verbal item types of the Graduate Record Examinations (GRE) were studied to explore possible reasons for any differences found in subgroup performance and validity. Statistical differences among item types in active forms of the GRE verbal measure were documented; experimental subtests of matching statistical characteristics for the item types were developed; and correlations of the matched and operational item type scores with self-reported grade point averages were compared. Comparisons were made by gender and ethnic group within undergraduate major field categories. Sample sizes ranged from 1,815 to 1,930 for the 8 new test editions between October 1985 and April 1987. Results suggest that all verbal item types studied exhibited concurrent validity, with only small differences among the item types. All the item types were valid, and they were very highly correlated. Because of this overlapping variance, little concurrent validity is lost by deleting any one item type. Reading comprehension items may, overall, be slightly more valid than the other item types; and the analogy item type may contribute slightly less to the concurrent validity of the verbal measure. However, these small differences do not suggest any specific revisions to the verbal measure of the GRE General Test. An appendix discusses equating the item-type subscores. (Contains 16 tables, 2 figures, and 24 references.) (Author/SLD)

GRE®
RESEARCH

Concurrent Validity of Verbal Item Types for Ethnic and Gender Subgroups

Cheryl L. Wild, W. Miles McPeek, and Stephen L. Koffler
with Henry I. Braun and William Cowell

December 1989

GRE Board Professional Report No. 84-10P
ETS Research Report 89-23

Educational Testing Service, Princeton, New Jersey

Concurrent Validity of Verbal Item Types for Ethnic and Gender Subgroups

Cheryl L. Wild, W. Miles McPeek, and Stephen L. Koffler
with Henry I. Braun and William Cowell

GRE Board Report No. 84-10P

December 1989

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, N.J. 08541

Concurrent Validity of Verbal Item Types

For Ethnic and Gender Subgroups

ABSTRACT

The validity of the Graduate Record Examinations (GRE) has been a high-priority research topic. Research to date concerning the GRE verbal measure suggests that for the GRE, as for the Scholastic Aptitude Test, the reading comprehension and sentence completion item types appear to carry the weight of the predictive validity of the verbal measure. However, this finding may have been a result of differences in difficulty and discrimination (as suggested by Schrader [1984a]), rather than a result of inherent differences in the item types.

The purpose of the present study was to examine the verbal item types for the GRE to explore possible reasons for any differences found in subgroup performance and validity. Statistical differences among item types in active forms of the GRE verbal measure were documented, experimental subtests of matching statistical characteristics for the item types were developed, and correlations of the matched and operational item type scores with self-reported grade point averages were compared. Comparisons were made by gender and ethnic group within undergraduate major field categories.

The results of this study suggest that all the verbal item types studied exhibit concurrent validity. Differences among the item types are small. All the item types are valid, and they are very highly correlated.

Because of this overlapping variance, little concurrent validity is lost by deleting any one item type. Results suggest that, of the four item types, reading comprehension may, overall, be slightly more valid than the other item types. The analogy item type may contribute slightly less than the other item types to the concurrent validity of the verbal measure. However, these differences are small and do not suggest any specific revisions to the verbal measure of the GRE General Test.

CONCURRENT VALIDITY OF VERBAL ITEM TYPES

FOR ETHNIC AND GENDER SUBGROUPS

Cheryl L. Wild, W. Miles McPeek, and Stephen L. Koffler
with Henry I. Braun and William Cowell

The purpose of the study described in this paper was to investigate the concurrent validity of the item types in the Graduate Record Examinations (GRE) General Test verbal measure within various ethnic and gender subgroups. The present study provides information about one of the relevant factors necessary to help assess whether the item types in the verbal measure are appropriate for use in the examination. In deciding whether an item type is appropriate for continued use in an examination, it is important to consider a number of factors in addition to concurrent validity. Among these additional factors are reliability, time available for testing, predictive validity, construct validity, appropriateness of the distribution of item difficulties, face validity, and impact on subgroups. Thus, results from this study will provide only a partial indication of the continuing appropriateness of the existing test specifications for the verbal section of the GRE General Test.

LITERATURE REVIEW

The literature review is divided into three major parts. The first presents an overview of all the GRE Tests and discusses the purpose of the verbal measure. Because the current interest in item type validity first surfaced in research on the SAT verbal measure, the SAT literature is reviewed next. The GRE verbal measure consists of the same item types as the SAT verbal measure, thus, findings about the SAT verbal measure could have important implications for the GRE General Test. Finally, the research on the verbal measure of the GRE General Test is discussed.

Background of the GRE Tests

The Graduate Record Examinations are administered to applicants to graduate and professional schools, and scores are typically used by graduate admissions committees and fellowship sponsors as one of several pieces of information in making admissions decisions.

The program offers a General Test and 17 Subject Tests intended for students who have majored as undergraduates in the subjects the tests measure. The GRE General Test measures verbal, quantitative, and analytical abilities and consists of seven 30-minute sections. Six sections of the General Test contribute to the examinees' test scores; one unidentified, separately timed section has trial questions that are not included in the examinees' test scores.

The verbal component of the General Test is the central focus of this paper and is described in the 1988-1989 GRE Information Bulletin:

The verbal ability measure is designed to test one's ability to reason with words in solving problems. Reasoning effectively in a verbal medium depends primarily upon the ability to discern, comprehend, and analyze relationships among words or groups of words and within larger units of discourse such as sentences and written passages....

The verbal measure consists of four question types: analogies, antonyms, sentence completions, and reading comprehension sets....

Analogy questions test the ability to recognize relationships among words and the concepts they represent and to recognize when these relationships are parallel....

Although antonym questions test knowledge of vocabulary more directly than do any of the other verbal question types, the purpose of the antonym questions is to measure not merely the strength of one's vocabulary but also the ability to reason from a given concept to its opposite....

The purpose of the sentence completion questions is to measure the ability to recognize words or phases that both logically and stylistically complete the meaning of a sentence....

The purpose of the reading comprehension questions is to measure the ability to read with understanding, insight, and discrimination. This type of question explores the examinee's ability to analyze a written passage from several perspectives, including the ability to recognize both explicitly stated elements in the passage and assumptions underlying statements or arguments in the passage as well as the implications of those statements or arguments... (p. 28-31).

Research about the SAT Verbal Measure

The major impetus for focusing on verbal item type research for the GRE originates from findings from the College Board Validity Study Service. Ramist (1981), in a routine compilation of validity results for the College Board, found that the reading subscore of the SAT (reading comprehension and sentence completion item types) correlated higher with freshman grade point average than did the vocabulary subscore (antonym and analogy item types). This indicated that the reading subscore might have greater predictive validity than the vocabulary subscore.

Ramist reported the results from 96 colleges that conducted validity studies in which the subscores were predictors. For those colleges, the reading subscore was almost identical to the full SAT verbal score with respect to average correlations with the freshman grade point average. In fact, in 44% of the schools, the reading subscore correlation with grade point average was actually higher than the total verbal score correlation. Similar mean correlations with the verbal subscore as predictors were considerably lower. The average zero order correlations

were as follows:

SAT Verbal Score	.374
Reading Subscore	.370
Vocabulary Subscore	.320

In a related study, Schrader (1984a) found that, for the SAT, vocabulary item types provided a substantially greater number of difficult items than did the reading item types. Schrader's results suggest that antonym and analogy items are likely to have lower biserial correlations than reading comprehension and sentence completion items. However, this finding is not conclusive because these differences could have resulted in part from the manner in which items were selected for the test. Specifically, antonym and analogy items are typically used to obtain the difficult items required by the test specifications. Since, for all item types, more difficult items tend to have lower biserials, the finding may be an artifact of the differences in the spread of difficulties among the item types. Further, interpretation of the results is complicated by the fact that subscores are not based on separately timed sections.

Evidence from another SAT study conducted by Schrader (1984b) suggested that the relative validities of vocabulary and reading subscores vary with different criteria. In that study, Schrader reported higher correlations for vocabulary subscores than for reading subscores with essay, multiple-choice, and total composition scores on the English Composition Test and with the Test of Standard Written English. However, he found higher correlations for reading subscores than for vocabulary subscores with other Achievement Test scores, high school rank and self-reported course grades.

In a third related study, Schrader (1984c) examined the validity of the SAT verbal item types for predicting first-year grades. He obtained simple and multiple correlations for the four item types when scores were based on all items of each type and when scores were based on subsets of eight items of each type matched for difficulty. Although the results from this study must be interpreted with caution because only one edition of the SAT verbal measure was studied, the results indicated that the four item types are about equal in predictive validity when differences in difficulty are controlled.

In summary, the reading comprehension and sentence completion item types (the reading subscore) tend to carry the weight of the predictive validity of the verbal measure of the SAT. However, this finding may be a result of differences in difficulty and discrimination (as suggested by Schrader's studies), rather than a result of inherent differences in the item types.

GRE General Test Research

There has been less research conducted about the differential validity of the GRE vocabulary and reading item types than there has been for the SAT. Differences such as those observed for the SAT would not appear in

routine GRE validity or summary data reports because subscores are not reported for the GRE General Test. Further, it would be difficult to obtain and interpret predictive validity results from existing data because sample sizes for GRE validity data are small and subscores would have to be obtained from a test not constructed to report them. As a way of addressing the sample size difficulties, GRE studies on verbal item types have generally used a concurrent validity approach (i.e., examining the relationship between a predictor and a criterion measure obtained at the same time as the predictor).

Wilson (1984) examined the relative contribution of a vocabulary subscore (composed of antonym and analogy items) and a reading subscore (composed of sentence completion and reading comprehension items) based on the GRE verbal measure to the prediction of self-reported undergraduate grade point average (SR-UGPA) for students in 12 undergraduate major fields. Wilson's results for the GRE were comparable to Ramist's findings on the SAT. Reading subscore correlations with SR-UGPA were about .04 higher than the vocabulary subscore correlations (.30 vs. .26). Combining the two item types increased the correlations with SR-UGPA by less than .01. However, scores were derived from items in existing test editions, and, therefore, there was no control for item difficulty or item-test correlation. And, as Schrader (1984c) showed, the differences may be confounded by differences in difficulty and item-test correlation.

Wilson (1986) extended his previous study and investigated the possibility that item type validity might vary by sex or ethnic group, but he did not control for differences in difficulty or item-test correlation of the various item types. Rock, Werts, and Grandy (1982) examined the psychometric properties of the GRE item types for Black female, Black male, White female, and White male social science majors. In both studies there was no evidence of psychometric bias in any of the item types.

Thus, the research for the GRE suggests that, as for SAT, the reading comprehension and sentence completion item types appear to carry the burden of the predictive validity of the verbal measure. However, similar to the SAT findings, this result may be a consequence of differences in difficulty and discrimination instead of variations in the item types.

Procedures for the current study were developed based on the literature reviewed above and are described in detail in the next section.

PROCEDURES

Purpose and Overview

The purpose of this study was to investigate the validity of the item types in the GRE verbal measure within various ethnic and gender subgroups. Previous research about the item type validity issue (Wilson, 1984) was based on item type scores constructed from operational test forms. Although the study's results were informative, subscores obtained from operational test forms confound differences in item type validity

with the differences in difficulty and discrimination of the items chosen for inclusion in a particular test. Operational tests are assembled to have prespecified average difficulties, spreads of difficulty, and average discrimination indices. Since the reading comprehension items are in sets, there is usually less flexibility in adjusting the difficulty in these items. Thus, the discrete items are usually used to adjust the statistical characteristics of a test form in order to meet the statistical specifications.

To avoid this confounding, experimental subtests of different item types with matching statistical characteristics were developed for this study and administered with an operational test. To better understand the implications of the results of the experimental data for operational tests, comparisons between existing GRE test editions and experimental editions are made. Thus, the data analysis can be divided into two major parts: (1) an analysis of the statistical characteristics of item types in existing and experimental GRE test editions and (2) an analysis of multiple correlations of item type subscores with self-reported undergraduate grade point averages.

The analysis of the statistical characteristics will address the following questions:

1. What are the mean item difficulties and mean item-test correlations (correlations with the total 76-item verbal score) for the reading comprehension, sentence completion, analogy, and antonym item types in existing editions of the GRE verbal measure? Do these vary from edition to edition?

2. What are the mean item difficulties and mean item-test correlations (correlations with the 64-item experimental score) for the reading comprehension, sentence completion, analogy, and antonym item types in the experimental editions? How do these compare to each other and to those of operational editions?

The multiple correlational analyses will address the following questions:

3. When matched on statistical characteristics, do reading comprehension, sentence completion, analogy, and antonym item types contribute equally to the concurrent validity of the GRE verbal measure?

4. Are there differences in the concurrent validity of reading comprehension, sentence completion, analogy, and antonym item types for Black female, Black male, Hispanic female, Hispanic male, White female, and White male groups?

Samples

Statistical characteristics of operational verbal item types were obtained from test analysis reports for eight GRE General Test final forms (Cowell & Tessema, 1986a, b, c, d; Cowell, Tessema, & Black, 1987a, b, c, d). These reports are based on spaced samples of all examinees at

the first administrations of a test. The sample sizes ranged from 1,815 to 1,930 for the eight new test editions given between October 1985 and April 1987.

Data on the experimental subtests were collected from three GRE administrations to obtain a sufficiently large sample size for minority groups. Analyses were based on native English-speaking examinees tested in June 1985, December 1985, and February 1986. No attempt was made to select only recent college graduates. Ethnic group, gender, self-reported undergraduate grade point average (SR-UGPA), undergraduate major field, and undergraduate institution were determined from responses to the questions asked of examinees when they register for the test. Black female, Black male, Hispanic female, Hispanic male, White female, and White male groups had sufficiently large sample sizes for inclusion in the correlational analyses.

Regression analyses were computed for each undergraduate institution separately to avoid the problems of comparability created by pooling grades across institutions. Of the 91,562 total examinees, 83,843 indicated the undergraduate institutions they attended. They attended 2,193 different institutions. For the purposes of this study, it was determined that the 300 undergraduate institutions with the greatest number of examinees would be used. Sixty-six percent of the total examinees attended these institutions. From the pool of 300 institutions, those attended by at least 10 examinees of the specific ethnic-by-gender combination (e.g., Black females) were included in that ethnic group-by-gender analysis.

Table 1 summarizes the mean GRE General Test scores for the examinees actually included in the data analysis. For example, among the 300 institutions, 292 had at least 10 White female native English-speaking examinees for a total of 19,646 female examinees. The Black males constituted the smallest sample, 131 examinees from 8 undergraduate institutions.

Comparison of the GRE General Test scores for examinees in this study with those from summaries of test takers across a three-year period (Educational Testing Service, 1986) suggests that this sample is comparable in performance to a national sample of test takers. In the national sample, women performed best on the analytical score and men performed best on the quantitative score. This pattern is also present in the data collected for this study. Mean scores by ethnic group are also generally comparable to those of the national sample.

Experimental Subtests

Sixteen-item subtests of each of the four verbal item types (reading comprehension, sentence completion, analogy and antonym) were constructed for this study. The subtests had the same statistical and subject matter specifications--i.e., the same mean and standard deviation of difficulty, the same mean biserial correlation, and the same balance of science and nonscience content. It was desirable to obtain scores on all four item types for each examinee. However, only one 30-minute experimental

TABLE 1

Mean GRE General Test Scores for Examinees Included in the Correlational Analyses

Group	Number of Undergraduate Institutions	Number of Examinees		GRE Verbal	GRE Quantitative	GRE Analytical
Black females	28	499	Mean	387	380	408
			S D	102	119	101
Black males	8	131	Mean	364	407	385
			S D	95	139	118
Hispanic females	21	470	Mean	404	424	427
			S D	99	115	112
Hispanic males	19	373	Mean	433	530	461
			S D	108	127	116
White females	292	19,646	Mean	517	520	554
			S D	106	116	113
White males	286	16,170	Mean	534	603	579
			S D	107	120	117

section was available for each examinee, and 30 minutes was not enough time to give all four item types. Thus, each 16-item reading comprehension subtest was selected from the operational edition of the test. The decision to use the reading comprehension items from the operational sections was made because this item type would have required an inordinate amount of the time available in the 30-minute experimental section. The selection was possible because there are 22 reading comprehension items in the operational tests and these items are similar in their statistical characteristics from one test edition to another. Further, any unique effect of a single edition of the reading comprehension items was controlled since three editions of the test were involved in the study.

Each 30-minute experimental section contained 48 items--16 analogy items, 16 sentence completion items and 16 antonym items (see Note 1). Two editions of the 48-item experimental tests were developed to reduce dependence of the final inferences on idiosyncratic aspects of a single edition. The order of the item types in the experimental sections was counterbalanced to ensure that response rate was not a major factor in interpretation. Thus, four experimental sections were developed and administered at each of the three administrations, as described in Figure 1. Random samples of about 25% of the overall sample took each of these subtests.

Data Analysis

Statistical Characteristics of Item Type Subscores. Item difficulty, item discrimination, and reliability data were obtained for both operational and experimental item type subtests. The item difficulties are reported on a delta scale (mean of 13 and standard deviation of 4), with larger numbers indicating more difficult items. The delta values have been "equated" or put on a common scale so that average difficulties can be compared from form to form. The biserial correlation is used as the item discrimination statistic for this study. For the operational verbal items, the 76-item verbal score is used as the criterion measure. For the experimental tests, the 64-item score based on the sum of the four 16-item experimental sets (three 16-item sets from the 30-minute experimental sections and one 16-item reading comprehension set chosen from the operational sections) is used as the criterion. Thus, biserial correlations for the experimental and operational items are not strictly comparable because of the difference in the length of the criterion. Reliabilities of each item type subscore are computed by the Kuder-Richardson formula 20.

Equating Scores. Because total verbal scores are equated across GRE test editions, those scores are comparable across test editions. However, the item type subscores on both the operational and experimental forms had to be equated in order to pool data across test forms and administrations. Scores on counterbalanced sections were compared to assure that administering the sections in different orders did not affect the scores. If differences occurred, they were adjusted by equating (see Appendix A for details about the equating process). Types of test scores, editions, and equating information are summarized in Table 2.

FIGURE 1

Experimental Section Design

	Questions	Edition 1 (H61)*	Edition 2 (H62)
Order 1	1-16	Antonym I	Antonym II
	17-32	Sentence Completion I	Sentence Completion II
	33-48	Analogy I	Analogy II

	Questions	Edition 1 (H63)	Edition 2 (H64)
Order 2**	1-16	Analogy I	Analogy II
	17-32	Sentence Completion I	Sentence Completion II
	33-48	Antonym I	Antonym II

*H61, H62, H63, and H64 refer to the particular experimental form designations.

**Note that Order 2 is a counterbalanced form of Order 1.

TABLE 2

Summary of Type of Test Scores, Editions, and Equatings

Score	Number of Items	Subset of Which Score	Number of Editions	Counterbalanced Orders	Possible Equatings
1. Verbal Total	76	1	4	0	4*
2. Reading Comprehension Operational (RC-O)	22	1	4	0	4
3. Sentence Completion Operational (SC-O)	14	1	4	0	4
4. Analogies -- Operational (ANA-O)	18	1	4	0	4
5. Antonyms -- Operational (ANT-O)	22	1	4	0	4
6. Reading Comprehension Matched (RC-M, 16-item subset of RC-O)	16	1,2	4	0	4
7. Sentence Completion Matched (SC-M, 16 items from the experimental section)	16	---	2	2	4
8. Analogies Matched (ANA-M, 16 items from the experimental section)	16	---	2	2	4
9. Antonyms Matched (ANT-M, 16 items from the experimental section)	16	---	2	2	4

*Operational forms are already equated.

Regression Analyses. To obtain multiple correlation coefficients, regression analyses were computed separately within each race-by-sex group for each undergraduate institution. SR-UGPA for the two preceding years served as the criterion.

The minimum sample size for the regression analyses was set at 10 because the focus of the study was on groups with traditionally low representation in the GRE test-taking population (combinations of ethnic groups and gender). The unit of analysis was determined to be the institution, rather than the department, to maximize the number of examinees available for each analysis. This provides generalizability of the results across departments but does not allow us to draw conclusions about validity in a specific field. Bayesian procedures were also employed as a way of compensating for small sample sizes. Multiple correlations determined by empirical Bayes techniques are more stable than those determined by least squares approaches for small sample sizes, such as those in this study, because they reduce the effects of sampling fluctuations (Braun, 1988).

Since grading standards are known to vary across academic departments within institutions (Goldman & Widawski, 1976; Klitgaard, 1985; Strenta & Elliott, 1987; Willingham, 1985), estimates of validity can be depressed by pooling samples across departments. This can be a particular problem when using the undergraduate grade point average based on the last two years of course work as opposed to the first year grade point average. Willingham (1985) has shown that there is increasing variation in grading practices in fields across the four years of undergraduate education. To partially control for major field differences, variables were created for the regression analyses to represent the four primary major field areas (social sciences, biological sciences, physical sciences, and humanities; see Figure 2). This procedure allows differences in grading practices among the four major field areas to be part of the regression analyses.

However, the observed differences in grades across major field areas may be due to two sources: (1) differences in grading standards and (2) real differences in performance. The use of variables in the regression equation is appropriate if the primary source of differences is differences in grading standards, but not appropriate if the primary source of differences is real differences in performance among the groups. If the primary source of differences in grades is performance, then the variance due to field should be part of the item type analysis rather than be controlled by variables representing field. There is no real solution to this problem, since there may be some interaction of the reasons for differences in grades across fields. However, by being aware that the possible confounding exists, it is possible to consider the possible implications. In order to address this question, we performed the basic analyses two ways--both with and without major field area variables in the regression equation. Based on the results of these analyses (which are described later) the major field area variables were used in the majority of analyses as described below.

A "step-up" regression approach was used to determine the contribution of each item type to the prediction of undergraduate grade point average.

FIGURE 2

Fields Classified as Humanities, Social Sciences, Biological Sciences, and Physical Sciences

HUMANITIES
Archaeology
Architecture
Art History
Classical Languages
Comparative Literature
Dramatic Arts
English
Far Eastern Languages and Literature
Fine Arts, Art, Design
French
German
Linguistics
Music
Near Eastern Languages and Literature
Philosophy
Religious Studies or Religion
Russian/Slavic Studies
Spanish
Speech
Other Foreign Languages
Other Humanities

SOCIAL SCIENCES
American Studies
Anthropology
Business and Commerce
Communications
Economics

SOCIAL SCIENCES, continued
Education (including M.A. in Teaching)
Educational Administration
Geography
Government
Guidance and Counseling
History
Industrial Relations and Personnel
International Relations
Journalism
Law
Library Science
Physical Education
Planning (City, Community, Urban, Regional)
Political Science
Psychology, Clinical
Psychology, Educational
Psychology, Experimental/Developmental
Psychology, Other
Psychology, Social
Public Administration
Social Work
Sociology
Other Social Sciences

BIOLOGICAL SCIENCES
Agriculture
Anatomy
Audiology
Bacteriology
Biochemistry
Biology
Biomedical Sciences
Biophysics
Botany
Dentistry
Entomology
Environmental Science/Ecology
Forestry
Genetics
Home Economics
Hospital and Health Services Administration
Medicine
Microbiology
Molecular And Cellular Biology
Nursing
Nutrition
Occupational Therapy
Pathology
Pharmacology
Pharmacy
Physical Therapy
Physiology

BIOLOGICAL SCIENCES, continued
Public Health
Speech-Language Pathology
Veterinary Medicine
Zoology
Other Biological Sciences

PHYSICAL SCIENCES
Applied Mathematics
Astronomy
Chemistry
Computer Sciences
Engineering, Aeronautical
Engineering, Chemical
Engineering, Civil
Engineering, Electrical
Engineering, Industrial
Engineering, Mechanical
Engineering, Other
Geology
Mathematics
Metallurgy
Oceanography
Physics
Statistics
Other Physical Sciences

From 1986-87 GRE Information bulletin

It consisted of:

- o determining the R^2 (squared multiple correlation) for a regression equation with only the variables for the major field areas as independent variables
- o computing the R^2 when a single independent variable (corresponding to the score on a particular item type) was added to the variables representing major field areas
- o determining the magnitude of the increment in the R^2's due to the addition of the item score.

The specific procedures used were as follows (using the White male analysis as an example):

1. For each of the undergraduate institutions (N = 286) for which there were at least 10 White male examinees, five least squares regression equations were computed, with the dependent variable being SR-UGPA. The five regression equations (E_o, E_1, ..., E_4) are described below in terms of R_o^2, R_1^2, ..., R_4^2 (R^2 is the squared multiple correlation, the proportion of variation of the dependent variable explained by the combination of independent [or predictor] variables):

   $R_o^2 = R^2$ for equation E_o with only the variables for the major fields as independent variables

   $R_1^2 = R^2$ for equation E_1 with the variables for the major field and the reading comprehension score as independent variables

   $R_2^2 = R^2$ for equation E_2 with the variables for the major field and the sentence completion score as independent variables

   $R_3^2 = R^2$ for equation E_3 with the variables for the major field and the analogy score as independent variables

   $R_4^2 = R^2$ for equation E_4 with the variables for the major field and the antonym score as independent variables

2. For each of the five regression equations (E_j), the multiple correlation (R_j) was determined ($j = 0, \ldots, 4$).

3. The R_j's were transformed to Z_j's via the Fisher Z transformation:

$$Z_j = \frac{1}{2} \log \left[\frac{1 + r}{1 - r}\right].$$

4. A sum-of-cross-products matrix was created for each of the sets of coefficients Z_0, ... Z_4. These matrices were the input for the empirical Bayes analysis.

5. Five empirical Bayes analyses were conducted, producing a 286 x 5 matrix of empirical Bayes coefficients (Z_e).

6. These Z_e's were transformed back to R_e's via the inverse Fisher transformation process. Thus, for each of the 286 institutions, there were five empirical Bayes-determined R_e's for White males.

7. $R_{eik}^2 = R_{ei}^2$ for the kth attending institution (i = 0,...,4; k = 1,...286).

8. $Increase_{ik} = R_{eik}^2 - R_{eok}^2$ (i = 1, ... 4; k = 1, ..., 286) represents for each institution the difference in R_e^2 attributed to the addition of a score for a particular item type compared to the R_o^2 for the equation with only variables for major field area included.

A "step-down" regression approach was also used to determine how much each item type contributed to the prediction of undergraduate grade point average beyond that predicted by undergraduate major field and the other three verbal item types. It consisted of:

- o determining the R^2 for a regression equation with the total experimental verbal score and the variables for major field as independent variables
- o for each item type, determining the R^2 for a regression equation with a 48-item verbal score (calculated by taking the total verbal score minus the score for the item type) and the variables for major field as independent variables
- o determining the magnitude of the differences between the R^2's

The specific procedures are similar to those described for the "step-up" procedure.

RESULTS

Statistical Characteristics of Item Types

The statistical characteristics of verbal item types in eight existing GRE General Test editions are summarized in Tables 3-7. Tables 3 and 4 present the means and standard deviations of the item difficulties expressed in delta units for the total verbal score. Generally, the verbal score is about middle difficulty. The sentence completion items are easier on the average than the other verbal items, and this is consistent for all eight forms reviewed. Reading comprehension items are slightly harder than middle difficulty on the average, and antonyms on the average are the hardest items. However, within a given test edition, the relative difficulty of these item types varies. The difficulty of the analogy items is the most variable from form to form. These results confirm the need to develop special forms of the subtests matched for

TABLE 3

Means of Equated Deltas and Difference of the Mean of the Equated Deltas of Each Item-Type Subset from the Mean of the Equated Deltas of Total Verbal by Test Form

	Mean Equated Delta					Differences from Mean Delta of Total Verbal			
Form	SC	Ana	RC	Ant	Total Verbal	SC	Ana	RC	Ant
3HGR3	11.4	11.9	12.3	12.3	12.1	-0.7	-0.2	0.2	0.2
3HGR4	11.6	12.4	12.2	12.4	12.2	-0.6	0.2	0.0	0.2
3IGR1	11.1	12.1	12.0	12.2	11.9	-0.8	0.2	0.1	0.3
3IGR2	11.5	12.0	12.2	12.2	12.0	-0.5	0.0	0.2	0.2
3IGR3	11.4	12.6	12.2	12.5	12.2	-0.8	0.4	0.0	0.3
3IGR4	11.4	12.1	12.5	12.2	12.1	-0.7	0.0	0.4	0.1
3JGR1	11.5	11.6	12.5	12.6	12.1	-0.6	-0.5	0.4	0.5
3JGR2	11.6	12.3	12.6	12.7	12.4	-0.8	-0.1	0.2	0.3
Mean	11.4	12.1	12.3	12.4	12.1	-0.7	0.0	0.2	0.3
Maximum	11.6	12.6	12.6	12.7	12.4	-0.5	0.4	0.4	0.5
Minimum	11.1	11.6	12.0	12.2	11.9	-0.8	-0.5	0.0	0.1
Range	0.5	1.0	0.6	0.5	0.5	0.3	0.9	0.4	0.4
Number of items	14	18	22	22	76	14	18	22	22

Delta is an index of item difficulty related to the proportion correct, p. Delta = 13 + 4z, where z is the standard normal deviate corresponding to the area under the normal curve of 1-p. Values of delta range from 6 for very easy items to 20 for very difficult items. Middle difficulty for an item is defined as the level at which half of the group would know the answer and the remaining half would guess at random. For the verbal sections, composed of 5-choice items, middle difficulty reference delta is 12.0.

The equated delta for an item is the estimated difficulty level of the item for the GRE reference group; a spaced sample of those who took Forms 3DGR1, 3DGR2, or 3DGR3 of the GRE General Test at the October 1981 administration.

TABLE 4

Standard Deviations of Equated Deltas and Difference of the Standard Deviation of the Equated Deltas of Each Item-Type Subset from the Standard Deviation of the Equated Deltas of Total Verbal by Test Form

	S.D. of Equated Delta					Differences from S.D. of Delta of Total Verbal			
Form	SC	Ana	RC	Ant	Total Verbal	SC	Ana	RC	Ant
3HGR3	1.6	2.5	1.9	2.8	2.3	-0.7	0.2	-0.4	0.5
3HGR4	2.0	1.8	2.3	3.1	2.4	-0.4	-0.6	-0.1	0.7
3IGR1	2.7	2.6	2.0	2.6	2.5	0.2	0.1	-0.5	0.1
3IGR2	2.1	3.5	1.5	2.7	2.5	-0.4	1.0	-1.0	0.2
3IGR3	2.8	2.1	1.5	2.9	2.4	0.4	-0.3	-0.9	0.5
3IGR4	1.8	2.6	1.8	2.9	2.4	-0.6	0.2	-0.6	0.5
3JGR1	2.1	2.7	1.6	2.6	2.4	-0.3	0.3	-0.8	0.2
3JGR2	2.8	2.6	1.4	2.6	2.4	0.4	0.2	-1.0	0.2
Mean	2.2	2.6	1.8	2.8	2.4	-0.2	0.1	-0.7	0.4
Maximum	2.8	3.5	2.3	3.1	2.5	0.4	1.0	-0.1	0.7
Minimum	1.6	1.8	1.4	2.6	2.3	-0.7	-0.6	-1.0	0.1
Range	1.2	1.7	0.9	0.5	0.2	1.1	1.6	0.9	0.6
Number of items	14	18	22	22	76	14	18	22	22

TABLE 5

Mean Biserial Correlations and Differences of the Mean Biserial Correlation of Each Item-Type Subset from the Mean of the Biserial Correlation of Total Verbal

	Mean Biserial Correlations					Differences from Mean R-Bis of Total Verbal			
Form	SC	Ana	RC	Ant	Total Verbal	SC	Ana	RC	Ant
3HGR3	.57	.49	.49	.51	.51	.06	-.02	-.02	.00
3HGR4	.53	.50	.53	.52	.52	.01	-.02	.01	.00
3IGR1	.59	.47	.52	.51	.52	.07	-.05	.00	-.01
3IGR2	.55	.51	.49	.57	.53	.02	-.02	-.04	.04
3IGR3	.57	.48	.54	.52	.53	.04	-.05	.01	-.01
3IGR4	.54	.52	.49	.53	.52	.02	.00	-.03	.01
3JGR1	.56	.46	.49	.51	.50	.06	-.04	-.01	.01
3JGR2	.57	.46	.48	.48	.49	.08	-.03	-.01	-.01
Mean	.56	.48	.50	.52	.52	.05	-.03	-.01	.00
Maximum	.59	.52	.54	.57	.53	.08	.00	.01	.04
Minimum	.53	.46	.48	.48	.49	.01	-.05	-.04	-.01
Range	.06	.06	.06	.09	.04	.07	.05	.05	.05
Number of items	14	18	22	22	76	14	18	22	22

TABLE 6

Standard Deviations of Biserial Correlations and Differences of the Standard Deviation of the Biserial Correlation of Each Item-Type Subset from the Standard Deviation of the Biserial Correlation of Total Verbal by Test Form

	S.D. Biserial Correlation					Differences from S.D. R-Bis of Total Verbal			
Form	SC	Ana	RC	Ant	Total Verbal	SC	Ana	RC	Ant
3HGR3	.07	.08	.10	.10	.10	-.03	-.02	.00	.00
3HGR4	.09	.09	.10	.11	.10	-.01	-.01	.00	.01
3IGR1	.07	.09	.09	.13	.11	-.04	-.02	-.02	.02
3IGR2	.12	.13	.09	.09	.11	.01	.02	-.02	-.02
3IGR3	.07	.10	.07	.12	.10	-.03	.00	-.03	.02
3IGR4	.07	.13	.07	.15	.11	-.04	.02	-.04	.04
3JGR1	.08	.12	.07	.13	.11	-.03	.01	-.04	.02
3JGR2	.08	.11	.10	.13	.12	-.04	01	-.02	.01
Mean	.08	.11	.09	.12	.11	-.03	.00	-.02	.01
Maximum	.12	.13	.10	.15	.12	.01	.02	.00	.04
Minimum	.07	.08	.07	.09	.10	-.04	-.02	-.04	-.02
Range	.05	.05	.03	.06	.02	.05	.04	.04	.06
Number of items	14	18	22	22	76	14	18	22	22

TABLE 7

Reliabilities and Difference of the Reliability of Each Item-Type Subset from the Reliability of Total Verbal by Test Form

	Reliability					Differences from Reliability of Total Verbal			
Form	SC	Ana	RC	Ant	Total Verbal	SC	Ana	RC	Ant
3HGR3	.76	.72	.80	.81	.93	-.17	-.21	-.13	-.12
3HGR4	.71	.74	.81	.78	.92	-.21	-.18	-.11	-.14
3IGR1	.75	.69	.81	.80	.93	-.18	-.24	-.12	-.13
3IGR2	.74	.71	.80	.84	.93	-.19	-.22	-.13	-.09
3IGR3	.74	.72	.84	.79	.93	-.19	-.21	-.09	-.14
3IGR4	.73	.73	.80	.80	.92	-.19	-.19	-.12	-.12
3JGR1	.75	.67	.80	.79	.92	-.17	-.25	-.12	-.13
3JGR2	.73	.69	.80	.78	.92	-.19	-.23	-.12	-.14
Mean	.74	.71	.81	.80	.93	-.19	-.22	-.12	-.13
Expected mean	.71	.76	.79	.79	.93	-.22	-.17	-.14	-.14
Maximum	.76	.74	.84	.84	.93	-.17	-.18	-.09	-.09
Minimum	.71	.67	.80	.78	.92	-.21	-.25	-.13	-.14
Range	.05	.07	.04	.06	.01	.04	.07	.04	.05
Number of items	14	18	22	22	76	14	18	22	22

* The reliability of a test is related to the number of items in the test. Thus, the reliability coefficients for the subsets cannot be expected to be as high as those of total verbal. For that reason, a row has been included below the row of mean differences to show the values that would be expected because of the smaller number of items in the subsets. For example, if a test with a reliability coefficient of .93 were reduced from 76 items to 14 items, the reliability would be expected to decrease to about .71, a decrease of .22.

Reliability estimates for the item-type subscores within each section were computed by Kuder-Richardson formula 20 (Kuder & Richardson, 1937).

difficulty and discrimination rather than use the subtests available in operational editions as Wilson (1986) did.

The standard deviations (Table 4) of item difficulties gives an indication of the spread of the difficulties of items by item type. The reading comprehension items tend to have the least spread in item difficulties of the verbal item types, while antonyms tend to have the greatest spread.

Tables 5 and 6 present the means and standard deviations of biserial correlations for the verbal measure. Sentence completion items tend to have the highest average biserial correlations while analogies have the lowest. However, these relationships vary considerably from form to form.

Table 7 shows that the reliability of the total verbal score is consistently .92 or .93. The reliability of a score is closely related to the number of items contributing to the score and the discrimination power of the items. As would be expected, scores based on the two item types (reading comprehension and antonyms) with the most items (22) are more reliable than scores based on the other two item types. The 14-item sentence completion item type is slightly more reliable than the 18-item analogy item type. This difference in reliability is consistent with the substantially higher average biserial correlations of the sentence completion items found in Table 5.

The statistical characteristics of verbal item types in the experimental sections used for this study are summarized in Tables 8-10. The mean equated deltas of the experimental sections range from 11.7 to 12.6. The average biserial correlations range from .47 to .54, with reading comprehension slightly higher and analogies slightly lower on the average. Although the mean biserials are not the same for all the experimental sections, they are less variable than within individual operational forms.

Table 11 provides data about the self-reported undergraduate grade point average for the examinees according to their major field areas. This information helps in the interpretation of the regression analyses, which included major field as a variable. The mean deviations provide a way of comparing UGPA for the four major field areas while controlling for the differences in grades across the institutions. Since the unit of analysis is the institution and since grading scales vary across institutions, the mean deviations in grades by field were calculated in the following way:

1. The mean UGPA for all examinees within an institution was determined.

2. Within the institution, the mean UGPAs for all examinees according to their major field areas were determined.

3. The mean residual UGPA for each major field area was determined by subtracting the mean determined in #1 above from each of the four means determined in #2 above.

TABLE 8

Means and Standard Deviations of Equated Deltas
For the Sentence Completion, Analogy, and Antonym Subscores
From the Experimental Sections

		Subscore		
		Sentence Completion	Analogy	Antonym
Test Form				
H 61	mean	12.2	12.7	12.0
(N=2,980)	S D	1.8	1.7	1.9
H 62	mean	12.3	12.6	12.5
(N=2,985)	S D	1.6	2.0	1.7
H 63	mean	12.2	12.5	12.1
(N=2,980)	S D	1.9	1.6	1.9
H 64	mean	12.2	12.4	12.6
(N=2,980)	S D	1.7	1.8	1.8

TABLE 9

Means and Standard Deviations of R-Biserial Correlations For the Sentence Completion, Analogy, and Antonym Subscores From the Experimental Sections

		Subscore		
		Sentence Completion	Analogy	Antonym
Test Form				
H 61	mean	.51	.48	.51
(N=2,980)	S D	.11	.09	.09
H 62	mean	.52	.51	.49
(N=2,985)	S D	.06	.11	.11
H 63	mean	.50	.47	.51
(N=2,980)	S D	.10	.09	.10
H 64	mean	.52	.50	.48
(N=2,980)	S D	.06	.11	.11

TABLE 10

Means and Standard Deviations of Equated Deltas and Biserial Correlations of the Matched Sets of Reading Comprehension Items

		Form		
		3GGR2	3GGR3	3GGR4
Equated Delta	Mean	11.7	11.8	11.7
	S D	1.8	1.9	2.0
Biserial Correlation	Mean	.53	.51	.54
	S D	.06	.11	.09
Reliability		.745	.745	.725

TABLE 11

Mean Deviations of Examinee and School Mean UGPA
For Each Major Field For Combinations
of Ethnic and Gender Groups

Group	Number of Under-Graduate Institutions	Humanities	Social Science	Biological Science	Physical Science
Black females	28	.056	.099	.010	-.021
Black males	8	.011	-.030	-.138	.350
Hispanic females	21	-.027	.062	-.126	.003
Hispanic males	19	.388	-.114	-.268	-.032
White females	292	.118	.030	-.099	.011
White males	296	.167	-.028	-.186	.025

4. The means of the mean residuals, called mean deviations, were determined for each combination of ethnic group and gender.

For example, Table 11 shows that White males majoring in Humanities tended to have a higher mean UGPA than did all students within the institution (the mean UGPA for White male humanities majors exceeded the overall institution mean UGPA by 0.167). White males majoring in the physical sciences also had a mean UGPA greater than the institution mean, although the difference was not as great as for humanities majors. Finally, both social science and biological science majors who were White males had mean UGPAs lower than the institution mean UGPA.

The most discernible pattern that can be seen in Table 11 is that examinees majoring in the biological sciences had the lowest relative UGPAs of all the groups, except for Black women. However, there is no clear or uniform ranking among the four major field areas across the ethnic-by-gender groupings.

Table 12 provides the mean R_o^2's and the increases in R^2 attributable to each item type for the six ethnic group-by-gender combinations for the matched/experimental section items. This analysis is quite similar to a within fields group analysis, with the increase in R^2 similar to the average R^2 for an item type across the four groups. The fields are predictors only in the limited sense that grades vary by field. In this table, the institutions upon which data are based all have at least 10 examinees for the particular ethnic-by-gender group studied. Table 12 shows that the R_o^2's for the minority groups are uniformly greater than the R_o^2's for the White groups. Thus, while there exists explanatory power of examinees' SR-UGPAs simply by knowing their major field areas, the percentage of variability of UGPA explained by major field area is much greater for all the minority groups than for the White groups. The magnitude of these R^2's attributable to major field area is surprisingly large.

The addition of each of the four item types results in improved prediction of SR-UGPA over that already attributable to differences in major field area for all ethnic and sex groups. There is no overall pattern to the increase in explanatory power according to item type across the groups. Each of the four item types has the highest increase in validity for at least one of the six groups, and the differences among the increases in validities are often small. For the White groups, the reading comprehension items provide the greatest increase in prediction. This is consistent with Ramist's findings for the SAT. It is also consistent with our intuitive feeling that the tasks required to answer reading comprehension items are more similar to the tasks required of students in higher education than are the tasks required to answer the other verbal item types. However, among the minority groups, the reading comprehension item type consistently has the second highest increase in validity. Across the six groups, analogies and sentence completions provide slightly lower increases in validity than do reading comprehension and antonyms.

The increase in validity of the total 64-item experimental score (R_v^2) is presented in the table for comparative purposes. The total score adds

TABLE 12

Summary of Means of R^2's and Increases in R^2's Due to Each Item Type For the Matched/Experimental Section Items

Group	Number of Schools	R_o^2	RC*	SC*	ANA*	ANT*	R_v^2
Black females	28	.204	.072	.054	.073	.070	.069
Black males	8	.163	.048	.043	.041	.092	.062
Hispanic females	21	.094	.096	.062	.079	.109	.138
Hispanic males	19	.112	.074	.101	.054	.068	.087
White females	292	.042	.067	.060	.049	.051	.081
White males	286	.049	.065	.055	.043	.046	.069

R_o^2 heads the column giving the proportion of variance accounted for by the major field area variables.

*RC heads the column of increases in R^2 (above R_o^2) attributable to the reading comprehension items. SC, ANA, and ANT head similar columns for sentence completion, analogy, and antonym items, respectively. R_v^2 heads a column of the increases in R^2 (above R_o^2) attributable to the total experimental verbal score.

most to the validity for Hispanic females and males and least to the validity for Black males, Black females, and White males.

Table 13 provides data similar to that in Table 12 for a slightly different population. There was concern that making comparisons among the groups in Table 12 might be confounded by the particular schools included in each sample because a different grouping of schools comprised each ethnic-by-gender sample. As a result, the analysis in Table 13 for each group and for White males was based on the same undergraduate institutions. Thus, Table 13 allows one to compare the increase in validity for each item type for White males to the increase in validity for each item type for each of the other groups in an analysis based on the same institutions. Comparing the results from Table 13 to those in Table 12 indicates no change in the general pattern.

Table 14 provides the mean squared multiple correlations of item type score with self-reported undergraduate grade point average excluding undergraduate major field area variables. If the primary source of differences in grades across major field areas is differences in grading standards, the effect of including the major field area variables would be to reduce noise in the criterion variable and thus increase the observed validities. If the primary source of differences in grades is actual differences in student performance, the effect of including the major field area variables would be to reduce the true variance in the criterion measure and thus decrease the observed validity. A comparison of Tables 12 and 14 shows that the proportion of variance accounted for by the item type scores and total score is higher when undergraduate major field area is included in the regression equation. For example, the variance accounted for by the 64-item experimental score alone for White males is .059, while the increase in validity of the 64-item experimental score after controlling for major field areas is .069. The patterns of item type validity within each sex-by-ethnic group are quite consistent in Tables 12 and 14.

These results suggest that it is appropriate to use the major field area variables because the primary source of differences in grades across major field areas appears to be coming from different grading standards. To some extent it appears not to matter which procedure is followed since the pattern of relative validities among the four item types is the same. For these reasons, the remainder of the analyses include the major field area variables.

Table 15 provides, for the subscores based on operational items, information similar to that found in Table 12 based on matched sets of items. A comparison of the results in Tables 12 and 15 will provide evidence about whether the differences in R^2 are similar for the subscores based on the matched and unmatched sets of items. Although the R^2's differ, the pattern of results in Table 15 is essentially similar to the pattern found in Table 12. The increases in validity for the minority groups were greater than the increases in validity for the White groups (except for Black males and females for reading comprehension and Black males for sentence completion). Also, the same pattern exists of higher increases for reading comprehension for White groups, but not for minority groups.

TABLE 13

Comparison of White Males to Other Ethnic-by-Gender Groups
For the Matched/Experimental Section Items

Group	Number of Schools	R_o^2	RC*	SC*	ANA*	ANT*
Black Females	22	.207	.075	.054	.082	.084
White Males	22	.049	.061	.051	.040	.044
Black Males	6	.177	.061	.061	.043	.135
White Males	6	.048	.055	.047	.038	.042
Hispanic Females	21	.105	.111	.108	.100	.169
White Males	21	.049	.065	.056	.042	.046
Hispanic Males	16	.112	.074	.103	.055	.071
White Males	16	.049	.062	.053	.041	.044
White Females	284	.042	.067	.060	.049	.051
White Males	284	.049	.065	.055	.043	.046

*RC heads the column of increases in R^2 (above R_o^2) attributable to the reading comprehension items. SC, ANA, and ANT head similar columns for sentence completion, analogy, and antonym items, respectively. R_v^2 heads a column of the increases in R^2 (above R_o^2) attributable to the total experimental verbal score.

TABLE 14

Summary of Means of R^2's* Due to Each Item Type
For the Matched/Experimental Section Items
Excluding Major Field Area Variables

Group	Number of Schools	R_{rc}^2	R_{sc}^2	R_{ana}^2	R_{ant}^2	R_t^2
Black Females	28	.067	.047	.048	.029	.068
Black Males	8	.039	.036	.033	.060	.022
Hispanic Females	21	.071	.041	.040	.046	.070
Hispanic Males	19	.061	.097	.047	.048	.079
White Females	292	.059	.051	.040	.043	.071
White Males	286	.054	.045	.033	.035	.059

*R_{rc}^2 heads the column of proportion of variance accounted for by the reading comprehension items when the major field area variables are not included as independent variables. R_{sc}^2, R_{ana}^2, and R_{ant}^2 head similar columns for sentence completion, analogy, and antonym items, respectively. R_t^2 heads the column of proportion of variance accounted for by the total experimental verbal score when the major field area variables are not included as independent variables.

TABLE 15

Summary of Means of R^2's and Increases in R^2's
For the Operational Section Items

Group	Number of Schools	R_o^2	RC*	SC*	ANA*	ANT*
Black Females	28	.204	.069	.097	.076	.070
Black Males	8	.163	.040	.029	.088	.098
Hispanic Females	21	.094	.106	.079	.141	.078
Hispanic Males	19	.112	.087	.067	.072	.086
White Females	292	.042	.072	.058	.046	.050
White Males	286	.049	.068	.053	.048	.048

*RC heads the column of increases in R^2 (above R^2_o) attributable to the reading comprehension items. SC, ANA, and ANT head similar columns for sentence completion, analogy, and antonym items, respectively. R^2_v heads a column of the increases in R^2 (above R^2_o) attributable to the total experimental verbal score.

R_o^2 heads the column giving the proportion of variance accounted for by the major field area variables.

Comparing the results from the two tables shows that for the two White groups, the differences in R^2's were negligible. For the combinations of minority ethnic groups by gender, there were noticeable differences in R^2's, although there was no apparent pattern. These results indicate that the item type validities may be confounded by the differences in difficulty and item-test correlations for minority examinees, but not necessarily for White groups. Alternatively, these differences may result from the smaller sample sizes for the minority groups.

Table 16 provides, for each of the six groups, the mean R_o^2's, the increase in R^2 attributable to adding the total 64-item experimental score (R_v^2), and the change in R_v^2 resulting from the deletion of each of the four 16-item experimental subscores from the 64-item total experimental score. Thus, for Black females, the percentage of variability of UGPA explained by knowing the students' major field is .204 (R_o^2), and the increase in R^2 contributed by the total experimental score is .069 (R_v^2). The change in R_v^2 that results from deleting the 16 reading comprehension items from the total experimental score is .000. Similarly, the change in R_v^2 for Black females that results from deleting the 16 sentence completion items is -.001, the change from deleting the analogy items is -.004, and the change from deleting the antonym items is .006.

Overall, Black females had the highest proportion of variance predicted from major field and total experimental score, .273. White females and White males had the lowest proportion of variance predicted by major field and total score, .123 and .118, respectively. As in Table 13, the minority groups have a larger proportion of explained variance in UGPA than do the White groups.

The effect on R_v^2 of deleting any one of the item types in any of the analysis groups is small. Any set of three of the four item types is almost as valid as are all four together. Although some of the decreases are non-negligible fractions of the R_v^2 in certain cases, all the decreases are trivial values of no practical importance. For example, the greatest loss in R_v^2 across the groups results from the deletion of the reading comprehension item type, which causes a decrease in R_v^2 of 11 to 13% for four of the six groups, no change in one group, and an increase of 12% in R_v^2 for the remaining group. The actual change in R_v^2 for reading comprehension ranges from -.018 to .008. The smallest loss in R_v^2 results from the deletion of the analogy item type, which causes a decrease in R_v^2 of 9% for Hispanic females and 6% for Black females, an increase in R_v^2 of 19% for Black males and 23% for Hispanic males, and no change for the other two groups.

Discussion

The purpose of this study was to investigate the concurrent validity of the item types in the verbal measure of the GRE General Test for various ethnic and gender subgroups. To do this, we looked at correlations of item type subscores with self-reported undergraduate grade point averages. As an aid to understanding these data, we also looked at the

TABLE 16

Effect on R_v^2 of Deleting Each of the Four Experimental Subscores From the Total Experimental Score

Group	R_o^2	R_v^2	$R_o^2 + R_v^2$	Change in R_v^2 from Deleting			
				RC*	SC*	ANA*	ANT*
Black Females (N=28)	.204	.069	.273	.000	-.001	-.004	.006
Black Males (N=8)	.163	.062	.225	-.007	.008	.012	-.007
Hispanic Females (N=21)	.094	.138	.232	-.018	-.022	-.012	-.021
Hispanic Males (N=19)	.112	.087	.199	.008	-.015	.020	-.003
White Females (N=292)	.042	.081	.123	-.010	-.004	.000	-.001
White Males (N=286)	.049	.069	.118	-.008	-.003	.000	.000

*RC heads the column of changes in R_v^2 from deleting the reading comprehension items from the total experimental score. SC, ANA, and ANT head similar columns for sentence completion, analogy, and antonym items, respectively.

statistical characteristics of the item types in existing editions of the GRE General Test and in the experimental tests used in this study. Overall, the results do not suggest any specific revisions to the item type composition of the verbal measure of the GRE General Test.

The review of the item type statistics based on operational test forms (Tables 3-7) illustrates the potential problems with drawing conclusions about item type validity from operational data. The number of items per item type varies from 14 to 22, and the reliability and biserial correlation by item type varies, as does the difficulty of the item type. In operational forms, the sentence correction item type is generally the easiest. The mean difficulty of the analogy item type varies more than any other item type from form to form. Antonyms tend to be the item type with the highest average item difficulty and the greatest variability of item difficulty within an edition. Reading comprehension items tend to vary least in difficulty within an edition. The data in Tables 8-10 suggest that we were at least somewhat successful in minimizing the statistical differences among item type scores in the experimental sections.

One unexpected finding of this study is the substantial proportion of variance in UGPA that is accounted for by the students' major field areas. In Table 12, one finds that the proportion of R^2 accounted for by major field area ranges from a low of .042 for White females to a high of .204 for Black females. Major field area accounts for about twice as much of the variance in the Hispanic groups and three to four times as much of the variance in the Black groups than it does in the White groups. These are very large differences in proportions of variance, especially considering the relatively modest proportions of variance in UGPA typically found to be predicted by test scores in validity studies. Table 16 shows that, for Black females and males and for Hispanic males, the addition to R^2 attributable to adding the total 64-item experimental score to major field area was less than the R^2 attributable to major field area alone. This is true even though the increase in R^2 due to adding the total experimental score to the major field area variable is higher for Hispanic males than it is for either of the White groups.

These results suggest that great caution should be used in designing validity studies using undergraduate grades as criterion measures. Evidence cited by Willingham (1985) suggests that upper division grades do differ by field and that these differences are due to variations in grading practices. For the 300 institutions in this study, the grades for the last two years of undergraduate school across major field areas appear not to be interchangeable. It would seem that researchers studying the relative under- or overprediction of grades for different subgroups in a population would also need to consider whether the groups being compared are really alike in their course-taking patterns.

The item type validity patterns are not consistent across groups. For White females and White males, the reading comprehension item type is the most valid and the sentence completion item type the next most valid, while analogies and antonyms tend to add slightly less to the prediction equation. For the White group, this same finding is consistent in the operational and experimental data. However, different item types have

the highest concurrent validities in the minority subgroups. In the experimental data, the reading comprehension item type does consistently have the second highest increase in validity of all the item types in the minority subgroups, even though there is no one item type that is consistently best for all minority groups. Across the six groups, analogies and sentence completions provide slightly lower increases in validity than do the other two item types.

The results for the White groups--that the reading comprehension and sentence completion item types are more valid than the analogy and antonym item types--are consistent with the results obtained by Ramist (1981) from predictive validity studies of the SAT and by Wilson (1984, 1986) in predictive validity studies of the GRE. Most of the students in those studies were White. In the present study, the higher validity of the reading comprehension and sentence completion item types for the White groups was evident in the data from both the item type subscores based on the operational items and the matched sets of experimental items. This suggests that, for White students, the greater validities of reading comprehension and sentence completion do not result from incidental differences in difficulty and discrimination in operational test forms. Even when these factors are controlled, the analogy and antonym item types are less valid for White students. Thus, it is possible that these differences in validity for White students may be due to differences in the skills being measured by the different item types. It should be noted, however, that the differences in validity among the four item types for White students are not great.

The results for the minority groups do not follow this same pattern, although reading comprehension items were consistently the second most valid item type in the experimental data for minority students. Overall, the differences in validity among the four item types seem to be greater for minority groups than for White groups.

In addition to investigating how much each item type contributed individually to the concurrent validity, the effect of eliminating any one item type from the total score was compared. It appears that any set of three of the four item types is almost as valid as are all four together. This occurs because the four item types are so highly correlated. This finding was consistent for all ethnic-by-gender group combinations. Across the six groups, deletion of the reading comprehension item type resulted in the greatest loss in validity and deletion of the analogy item type resulted in the least loss in validity. However, these results are not consistent for all groups. Hispanic examinees show the greatest loss in validity from the deletion of the sentence completion item type. For the two Black groups, deletion of the different item types showed no consistent pattern of gains or losses in validity.

The results of this study suggest that all the verbal item types studied contribute to the concurrent validity of the verbal measure. Differences among the item types are small. Although all the item types contribute to the concurrent validity, there is a great deal of commonality among them. As a result, little concurrent validity is lost by deleting any one item type. The results suggest that, of the four item types, reading

comprehension may be slightly more consistently valid than the other item types across all groups. The analogy item type may contribute slightly less than the other item types to the concurrent validity. These differences are small and do not suggest any specific revisions to item type composition of the verbal measure of the GRE General Test. However, the results do suggest that dropping any one of the four item types would not have serious implications for the validity of the verbal score. Of course, a high reliability is necessary on a test that is used to make decisions about individuals. Thus, even if an item type were dropped, the current number of items in the test would need to remain approximately the same to maintain the current level of reliability. As suggested in the introduction to this paper, any decision about the content of a test requires the consideration of a number of issues in addition to concurrent validity.

References

Angoff, W. H. (1984) Scales, Norms, and Equivalent Scores. Princeton, NJ: Educational Testing Service.

Braun, H. I. (1988) Empirical Bayes Methods: A Tool for Exploratory Analysis. In the Proceedings from the 1987 Invitational Multilevel Conference, Princeton, NJ (eds. R. D. Bock and L. Burstein). New York, NY: Academic Press.

Cowell, W. R., & Tessema, A. (1986a). Graduate Record Examinations General Test, Test Analysis of Form 3HGR3. Internal Reference Document (SR-86-110). Princeton, NJ: Educational Testing Service.

Cowell, W. R., & Tessema, A. (1986b). Graduate Record Examinations General Test, Test Analysis of Form 3HGR4. Internal Reference Document (SR-86-132). Princeton, NJ: Educational Testing Service.

Cowell, W. R., & Tessema, A. (1986c). Graduate Record Examinations General Test, Test Analysis of Form 3IGR1. Internal Reference Document (SR-86-141). Princeton, NJ: Educational Testing Service.

Cowell, W. R., & Tessema, A. (1986d). Graduate Record Examinations General Test, Test Analysis of Form 3IGR2. Internal Reference Document (SR-86-148). Princeton, NJ: Educational Testing Service.

Cowell, W. R., Tessema, A., & Black, W. (1987a). Graduate Record Examinations General Test, Test Analysis of Form 3IGR3. Internal Reference Document (SR-87-07). Princeton, NJ: Educational Testing Service.

Cowell, W. R., Tessema, A., & Black, W. (1987b). Graduate Record Examinations General Test, Test Analysis of Form 3IGR4. Internal Reference Document (SR-87-17). Princeton, NJ: Educational Testing Service.

Cowell, W. R., Tessema, A., & Black, W. (1987c). Graduate Record Examinations General Test, Test Analysis of Form 3JGR1. Internal Reference Document (SR-87-127). Princeton, NJ: Educational Testing Service.

Cowell, W. R., Tessema, A., & Black, W. (1987d). Graduate Record Examinations General Test, Test Analysis of Form 3JGR2. Internal Reference Document (SR-87-128). Princeton, NJ: Educational Testing Service.

Educational Testing Service (1986). 1986-87 GRE information bulletin. Princeton, NJ: Educational Testing Service.

Educational Testing Service (1987). 1987-1988 GRE information bulletin. Princeton, NJ: Educational Testing Service.

Goldman, R. D., & Widawski, M. H. (1976). Why college grade point average is difficult to predict. Educational and Psychological Measurement, 36, 381-390.

Klitgaard, R. (1985). Choosing elites. New York: Basic Books.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test score reliability. Psychometrica, 2, 151-160.

Ramist, L. (1981). Validity of SAT verbal subscores. Internal Educational Testing Service memorandum, February 12, 1981.

Rock, D. A., Werts, C. & Grandy, J. (1982). Construct validity of the GRE Aptitude Test across populations -- an empirical confirmatory study. GRE Board Professional Report (78-1P). Princeton, NJ: Educational Testing Service.

Schrader, W. B. (1984a). A survey of item and test analysis results for SAT-verbal item types. College Entrance Examinations Board Research Report No. 84-7). Princeton, NJ: Educational Testing Service.

Schrader, W. B. (1984b). The relation of SAT reading and vocabulary scores to measures of high school performance. College Entrance Examinations Board Research Report (No. 84-7). Princeton, NJ: Educational Testing Service.

Schrader, W. B. (1984c). The validity of SAT-verbal item types. College Entrance Examinations Board Research Report (No. 84-7). Princeton, NJ: Educational Testing Service.

Strenta, A. C., & Elliott, R. (1987). Differential Grading Standards Revisited. Journal of Educational Measurement, 24(4), 281-291.

Wilson, K. M. (1984). The relationship of GRE item-type part-scores to undergraduate grades. Graduate Record Examinations Board Report (81-22). Princeton, NJ: Educational Testing Service.

Wilson, K. M. (1986). The relations of scores based on GRE General Test item types to undergraduate grades: An exploratory study for selected subgroups. Graduate Record Examinations Board Report (83-19P). Princeton, NJ: Educational Testing Service.

Willingham, W. W. (1985). Success in college. New York: College Entrance Examination Board.

Notes

1. Two additional methods of matching were also considered. Subtests could be constructed by selecting matched sets of items from within an operational edition of the test, although this matching procedure would result in fewer items in each subtest (probably not more than eight). These operational subtests would therefore have lower reliabilities than the proposed experimental subtests. In addition, the operational subtests would not meet the content specifications in the overall test, and the matching would be on the item difficulty only, not on the item-test correlations.

The second method considered was to give the reading comprehension items in an experimental section of the test as well as the discrete verbal item types. The disadvantage of this approach was that examinees would take only one of the experimental pretests and would therefore have scores on either reading comprehension or one score for each of the three discrete item types. Thus, it would not have been possible to see if the matched item types were adding equally to the correlation with self-reported undergraduate grade point average.

Appendix A

Equating the Item-Type Subscores

The three operational forms used in this study, Forms 3GGR2, 3GGR3, and 3GGR4, were spiraled (packaged and distributed to examinees in alternating sequence) at domestic test centers at the October 1984 test administration. The four experimental sections, however, were not administered at that administration. Because the three forms were administered to randomly equivalent groups, the ability levels of the groups taking each of the three forms are assumed to be equivalent. To facilitate comparison among the various scores, the raw scores on Form 3GGR2 were converted to a scale with a mean of 50 and a standard deviation of 10. Then, the scores derived from the operational sections of Forms 3GGR3 and 3GGR4 were equated to the common scale by setting raw score means equal to the corresponding 3GGR2 scaled score standard deviation. This method (Angoff, 1984) is known as Design I (random groups--one test administered to each group). Scores are defined as equivalent if their standard-score deviates in their respective, randomly-equivalent groups are equal. These operational scores were used at subsequent administrations for linking the scores on the matched item-type subsets (i.e., matched on item difficulty and discrimination indices) to the common scale.

All four experimental sections were administered at each of three test administrations: June 1985, December 1985, and February 1986. Because the sample sizes were largest for the December 1985 administration, data for that administration were used to equate scores derived from the experimental sections. Raw scores derived from the operational sections were converted to scaled scores using the score conversion equations derived from the October 1984 data. Then, the raw scores derived from the experimental sections were equated to the common scale using the corresponding operational scores as an external anchor test. This method (Angoff, 1984) is known as Design IV (nonrandom groups--one test administered to each group, common equating test administered to both groups).

The score conversion equations based on the October 1984 and December 1985 data were used to convert all the raw scores to the common scale. These scores could then be pooled across administrations for the subsequent analyses.

Equating the Item-Type Subscores

Data from four test administrations were used in the scaling, equating, and subsequent analysis of item-type subscores for this study:

October 1984	Forms 3GGR2, 3GGR3, and 3GGR4 were spiraled at domestic test centers. Scores derived from the operational sections of Form 3GGR2 were scaled to a mean of 50 and a standard deviation of 10. Scores derived from the operational sections of Forms 3GGR3 and 3GGR4 were equated to the common scale by setting raw score means equal to the corresponding 3GGR2 scaled score mean and raw score standard deviations equal to the corresponding 3GGR2 scaled score standard deviation.
December 1985	Form 3GGR3 was administered with the four experimental sections. Raw scores derived from the operational sections were converted to scaled scores using the score conversion equations derived from the October 1984 data. Raw scores derived from the experimental sections were equated to the common scale using the corresponding operational scores as an external anchor test.
June 1985	Form 3GGR2 was administered with the four experimental sections. Raw scores derived from the operational and experimental sections were converted to scaled scores using the score conversion equations derived from the October 1984 and December 1985 data.
February 1986	Form 3GGR4 was administered with the four experimental sections. Raw scores derived from the operational and experimental sections were converted to scaled scores using the score conversion equations derived from the October 1984 and December 1985 data.