A COMPARISON OF FREE-RESPONSE AND MULTIPLE-CHOICE FORMS OF VERBAL APTITUDE TESTS

William C. Ward

GRE Board Professional Report GREB No. 79-8P
ETS Research Report 81-28

January 1982

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


Baird, L. L. The Relationship Between Ratings of Graduate Departments and Faculty Publication Rates. GREB No. 77-2aR, November 1980.


Centra, J. A. How Universities Evaluate Faculty Performance: A Survey of Department Heads. GREB No. 75-5bR, July 1977. ($1.50 each)

Centra, J. A. Women, Men and the Doctorate. GREB No. 71-10R, September 1974. ($3.50 each)


Clark, M. J. Program Review Practices of University Departments. GREB No. 75-5aR, July 1977. ($1.00 each)


Hartnett, R. T. Sex Differences in the Environments of Graduate Students and Faculty. GREB No. 77-2bR, March 1981.


Lannholm, G. V. and Parry, M. E. Programs for Disadvantaged Students in Graduate Schools. GREB No. 69-1R, January 1970.


Rock, D., Werts, C. An Analysis of Time Related Score Increments and/or Decrements for GRE Repeaters across Ability and Sex Groups. GREB No. 77-9R, April 1979.


Schrader, W. B. GRE Scores as Predictors of Career Achievement in History. GREB No. 76-1bR, November 1980.

Schrader, W. B. Admissions Test Scores as Predictors of Career Achievement in Psychology. GREB No. 76-1aR, September 1978.


Wiltsey, R. G. Doctoral Use of Foreign Languages: A Survey. GREB No. 70-14R, 1972. (Highlights $1.00, Part I $2.00, Part II $1.50).

Witkin, H. A.; Moore, C. A.; Oltman, P. K.; Goodenough, D. R.; Friedman, F.; and Owen, D. R. A Longitudinal Study of the Role of Cognitive Styles in Academic Evolution During the College Years. GREB No. 76-10R, February 1977 ($5.00 each).
A Comparison of Free-Response and Multiple-Choice Forms of Verbal Aptitude Tests

William C. Ward

GRE Board Professional Report GREB No. 79-8P

ETS Research Report 81-28

January 1982

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board. An article based on this report will appear in Applied Psychological Measurement.

Copyright ©1982 by Educational Testing Service. All rights reserved.
Abstract

Three verbal item types employed in standardized aptitude tests were administered in four formats—conventional multiple-choice along with three formats requiring the examinee to produce rather than simply to recognize correct answers. For two item types, Sentence Completion and Antonyms, the response format made no difference in the pattern of correlations among the tests. Only for a multiple-answer open-ended Analogies test were any systematic differences found; even the interpretation of these is uncertain, since they may result from the speededness of the test rather than from its response requirements. In contrast to several kinds of problem-solving tasks that have been studied, discrete verbal item types appear to measure essentially the same abilities regardless of the format in which the test is administered.
Acknowledgments

Appreciation is due to Carol Dwyer, Fred Godshalk, and Leslie Peirce for their assistance in developing and reviewing items; to Sybil Carlson and David Dupree for arranging and conducting test administrations; to Henrietta Gallagher and Hazel Klein for carrying out most of the test scoring; and to Kirsten Yocum for assistance in data analysis. Dr. Ledyard Tucker provided extensive advice on the analysis and interpretation of results.
A Comparison of Free-Response and Multiple-Choice Forms
of Verbal Aptitude Tests

Tests in which an examinee must generate answers may require different abilities than do tests in which it is necessary only to choose among alternatives provided. Open-ended tests of behavioral science problem solving, for example, have been found to possess some value as predictors of professional activities and accomplishments early in graduate training—areas in which the GRE Aptitude Test and Advanced Psychology Test are not good predictors (Frederiksen & Ward, 1978). Moreover, scores based on the open-ended measures had very low correlations with scores from similar problems presented in machine-scorable form, and differed from the latter in their relations to a battery of reference tests for cognitive factors (Ward, Frederiksen, & Carlson, 1980). Comparable results were obtained using nontechnical problems, in which the examinee was given several opportunities to acquire information and generate explanatory hypotheses in the course of a single problem (Frederiksen, Ward, Case, Carlson, & Samph, 1980).

The importance of the kind of response required by a test is also suggested by a voluminous literature on "creativity." "Divergent" tests, in which the examinee must produce one or more acceptable answers from among a large number of possibilities, measure something different from "convergent" tests, which generally require the recognition of the single correct answer to a question (e.g., Guilford, 1956, 1967; Torrance, 1962, 1963; Wallach & Kogan, 1965).

These results suggest that the addition of open-ended items might make a contribution in standardized aptitude assessment. Such items would at the least increase the breadth of abilities entering into aptitude scores and could potentially improve the prediction of aspects of graduate and professional performance that are not strongly related to the current tests. However, the work discussed involves measures that are quite distant from the kinds of items typically used in standardized tests. The divergent thinking measures involve inherently trivial tasks (name uses for a brick, name words beginning and ending with the letter "t", for example) that would lack credibility as components of an aptitude test. The problem-solving measures have greater face validity but provide very inefficient measurement, in terms of both the examinee time
required to produce reliable scores and the effort required to evaluate the performance.

It is the purpose of the present investigation to explore the effects of the response format of a test, using item types like those employed in conventional examinations. The content area chosen is that of verbal knowledge and verbal reasoning, as represented by three item types—Antonyms, Sentence Completions, and Analogies.

The choice of these item types is motivated by several considerations. First, their relevance for aptitude assessment needs no special justification, given that they make up one-half of present verbal ability tests, such as the GRE and SAT. Thus, if it can be shown that recasting these item types into an open-ended format makes a substantial difference in the abilities they measure, a strong case will be made for the importance of the response format in considering the mix of items that enter into aptitude assessments. Moreover, divergent forms of these item types require only single-word or, in the case of Analogies, two-word answers. They should thus be relatively easy to score, in comparison with open-ended item types whose responses are often several sentences in length and may embody two or three complex ideas. While not solving the problems inherent in the use of open-ended tests in large-scale testing, they would serve to some degree to reduce their magnitude.

Surprisingly, no published comparisons of open-ended and multiple-choice forms of these item types are available. Several investigators have, however, examined the effects of response format on Synonyms items—items in which the examinee must choose or generate a word meaning essentially the same thing as a target word (Vernon, 1962; Heim & Watts, 1967; Traub & Fisher, 1977). All found high correlations across formats, but only the last of these studies was designed to answer the question of whether the two formats measured identical or only related abilities. Traub and Fisher concluded that format does affect the attribute measured by the test and found weak evidence favoring the existence of a factor specific to open-ended verbal items. Unfortunately, there were not enough scores, or scores on a sufficient variety of tests, to provide an unambiguous demonstration of a verbal production factor.

The present study was designed to allow a factor-analytic examination of the influence of response format. Each of three item types was given in each of four formats, varied in the degree to which they require production of answers. It would thus be possible to examine the fit of the data to each of two "ideal" types of factor structure: one in which only item-type factors would be
found, indicating that items of a given type measure essentially the same thing regardless of the response format; and one involving only format factors, indicating that the response requirements of the task are of greater importance than are differences in the kind of knowledge tested.

A brief description of the four formats is required as a basis for outlining several specific comparisons to be made. The Antonyms item type will serve as an example. There is, first, the conventional multiple-choice format; the examinee is asked to choose the one of five words that is most nearly opposite in meaning to a given word and to write its letter in an answer space. Second is the "single-answer" format. The examinee is to think of an opposite and to write this word in a space provided. The third format, "multiple-answer," is still more open-ended; the examinee produces and writes up to three different antonyms for the given word. Finally, the "keylist" format requires the examinee to think of an opposite, locate this word in a long numbered and alphabetized list, and record its number on the answer sheet.

The use of these formats allows several illuminating contrasts to be made. First, the comparison of the multiple-choice format with the open-ended ones provides the critical test of format effects. A failure to distinguish between these two types of test would indicate that the item type, not the response format, was the sole important consideration in determining what abilities influence test performance.

Second, the single-answer format, in which the examinee must write one appropriate response, provides the most interesting contrast from a theoretical perspective. Assuming that the extreme cases (conventional and multiple-answer formats) can be distinguished, there are two alternative bases on which to predict how this format relates to those extremes. The first is the possibility that response generation, rather than recognition, is the important factor; if so, this format would relate more closely to the multiple-answer one. The latter format, however, requires more than simply producing one response; the examinee must break whatever set is established in developing the first response in order to consider further possibilities. Thus, flexibility of thinking is important. If flexibility is a critical determinant of performance, the single-answer format would instead relate more closely to the multiple-choice form of the test.

Finally, the keylist format is of pragmatic interest. Do examinees respond to such items as they do to conventional
limited-alternative multiple-choice items, or does the very large number of response alternatives force a strategy in which answers are generated as in a truly open-ended test and then sought in the master list? If the latter is the case, measurement of divergent abilities could be incorporated in a testing program without major operational difficulties, since this response format can easily be adapted for scoring by machine.

Method

Test Development

Instructions. Appendix A shows instructions and sample items for each of the twelve item-type-by-response-format combinations employed. Instructions for each test paraphrase closely those employed in the GRE Aptitude Test, except as dictated by the specific response requirements of each format. With each set of instructions is given one sample question along with a brief rationale for the answer or answers suggested. Except for the multiple-choice tests, two or three fully acceptable answers are suggested for each sample question.

Test specifications. Assembly of a form of a standardized aptitude test requires meeting a number of constraints concerning the set of items. Some of the specifications employed in GRE tests are not applicable to open-ended tests; for example, the requirement that Analogies items be balanced as to whether they are concrete, mixed, or abstract; or as to whether their terms are overlapping or nonoverlapping—these classifications depend on the distractors supplied for a multiple-choice item as well as on the stem. Other specifications could only be approximated roughly within the limitations on the number of items that could be developed and pretested and the small numbers of subjects, generally on the order of 30, who supplied pretest data for each item. The content areas from which items were drawn, for example, were as varied as possible, but no formal classification and balancing of contents across formats were attempted. Items were chosen for "reasonable" difficulty level and range in pretests without attempting to equate these figures across tests. The GRE specification that Antonyms items be balanced in terms of the percentages of adjectives, nouns, and verbs employed was, however, followed.

One additional constraint was imposed to facilitate answering and scoring the open-ended items—that is (with two exceptions, described below), all items should be ones
permitting a single word as an answer. GRE Sentence Completion items can contain either one or two blanks to be filled in from the alternatives provided, and all three item types conventionally permit several words or a phrase as a correct answer. Allowing for several-word answers would have proven cumbersome for the keylist format; they were therefore avoided in all formats.

The exceptions to this restriction were the single-answer and multiple-answer Analogies tests. Here the examinee was required to produce pairs of words having the same relationship to one another as that shown by the two words in the stem of the question. Requiring two words made these tests more directly comparable to the multiple-choice Analogies than they would otherwise have been, since that test requires recognition of the similarity of relationship between one pair of words and another pair. It does, however, decrease the parallelism between these tests and the Keylist Analogies test, in which three terms were given and the examinee was asked to supply the fourth. There is some evidence, however, that the use of two-terms-given or three-terms-given analogies in itself may not be important; Guilford and Hoepfner (1971) show that the two kinds of items presented in multiple-choice form load on the same factor, identified in Guilford's model as Cognition of Semantic Relations.

Item writing. The initial plan for test development called for the development of a pool of about 100 multiple-choice items of each item type, with review by ETS test development staff to assure that all were items which would be acceptable within an operational examination. Items would be chosen at random from this set for assignment to each response format, to assure that the various test forms were equivalent to one another in all respects except for their response requirements.

However, it quickly became apparent that items suitable for use in one response format are generally not suitable in others. The requirements for acceptable open-ended items vary from one item type to another, but, in all three types, differ in some important way from those for acceptable multiple-choice items.

With Antonyms, the major problem is item difficulty. The first open-ended items pretested were drawn from the pool of multiple-choice items, after those had been reviewed by a member of test development staff. Few of these words were of reasonable difficulty level in an open-ended form, even with a liberal scoring that allowed partial credit for many responses
that would not have been acceptable as keys for multiple-choice forms of the items. A second Antonyms pretest used words drawn from vocabulary lists in SAT practice guides; again, many were unacceptably difficult. The final pretesting included a number of still easier words, including some vocabulary words drawn from an eighth-grade English text.

With Sentence Completion, just the opposite problem is found. Many items that are appropriate in multiple-choice form, where the difficulty level of the alternatives to be considered is under the control of the item writer, become trivially easy in an open-ended form; very simple, common words are often fully acceptable completions of the meaning of the sentence. It is necessary to add constraints to require discrimination among fine shades of meaning or implication of the sentence (without, however, turning the item into a test of vocabulary) to produce items with even moderate difficulty.

Analogies, finally, present problems of specificity. A multiple-choice Analogies stem need not involve words that stand in only one strongly defensible relationship to one another; interpretation of the relationship between the two given words can depend in part on the relationships expressed in the list of alternatives provided. With open-ended responding, however, it is necessary to control the relationship tightly in order to be able to score responses with any confidence. Otherwise, the scorer is faced with a problem more difficult than that of the examinee--it becomes necessary to induce the relationship that the examinee believed to be exemplified in both pairs before the appropriateness of the response can be judged.

There are, moreover, differences among the three nonmultiple-choice formats in their requirements. For example, desirable keylist items are ones for which there are only a few possible good answers--the keylist from which the examinee determines the number of the answer he has produced must include all the best possible answers. The best multiple-answer items, in contrast, are ones for which many different good answers can be produced.

With all these constraints, item preparation became, in effect, the development and pretesting of a separate pool of items for each open-ended format; very few items were found that could reasonably have been included in more than one test format.
Pretesting. Pretest subjects were paid volunteers from two local colleges, tested in groups of 15 to 30. A total of 270 individuals completed a one-hour testing session before final choices of items were made.

Item review. With a few exceptions, all items were reviewed by an expert in the development of verbal aptitude tests before inclusion in the final forms of the tests. Multiple-choice items used were ones considered appropriate for inclusion in an operational test. Preliminary scoring keys were agreed upon for open-ended items; for the multiple-answer forms, at least three answers to each item were judged to be fully acceptable responses.

Test assembly. Based on testing program experience with the time requirements and reliabilities of discrete verbal item types, each multiple-choice test included 20 items to be completed in 12 minutes. Slightly longer times (15 minutes) were allowed for forms including 20 single-answer or 20 keylist items. The multiple-answer forms allowed still more time per item--15 minutes for 15 Antonyms or Analogies items or for 18 Sentence Completion. It was anticipated that these time limits would be adequate to avoid problems of test speededness and that the number of items would be sufficient to produce scores with reliabilities of .70 or better--not high enough for operational testing, but quite acceptable for the purposes of a research study.

A 90-word alphabetized list was developed to accompany each keylist test. The list included approximately three or four answers for each test item, three answers to the sample item, and a few fillers.

Test Administration

Sample. Subjects were 315 paid volunteers from a single state university. It had been intended to limit the sample to undergraduate seniors who had taken (or would take) the GRE Aptitude Test in the current academic year. However, too few seniors were available to supply an adequate sample, and many who indicated their intention to take the Aptitude Test within the year did not in fact do so. Of the sample actually obtained, slightly more than two-thirds were undergraduate juniors and seniors. The small number (13%) for whom Aptitude Test scores were obtained were a somewhat select group, with means of 547, 646, and 616 on the verbal, quantitative and analytical scores,
respectively. It appears that the sample was a somewhat more able one than college students in general, but probably less selected than the GRE applicant pool.

Testing sessions. Each student participated in one four-hour testing session. Included in the session were 12 tests, representing all combinations of the three item types with four response formats, and a brief questionnaire relating to the student's academic background, accomplishments, and interests.

The tests were presented in a randomized order, subject to the restriction that no two successive tests should employ either the same item type or the same response format. Four variations of this order were employed to permit an assessment of and adjustment for possible practice or fatigue effects: the first group received tests in the order 1 through 12; the second received the reverse of this order; the third received tests 7-12 followed by 1-6; and the fourth received the reverse of the third order. Each of the first four large groups tested (including 51-60 subjects each) received tests in one of these orders; the remainder of the sample all were given tests in the first of the four orders.

Scoring

For each of the open-ended tests, scoring keys were developed that distinguished two degrees of appropriateness of an answer. One set of answers was judged fully acceptable. These were answers that would have been acceptable keys for a multiple-choice form of an item to be used in a selection test. The second set consisted of answers that were of marginal appropriateness; for example, an Antonyms response that identified the negative evaluation implied by a word, but that failed to capture an important nuance or the force of the evaluation.

There were few marginally acceptable responses for items given in the keylist format, since only 90 words were available from which to produce responses. It was determined, through a trial scoring, that partial credits were unnecessary for two of the keylist tests (Antonyms and Analogies). Responses to the remaining open-ended tests were coded in such a way as to permit computer generation of several different test scores, depending on the credit to be given to marginally acceptable answers.
Before scoring was begun, the preliminary scoring keys were modified to accommodate any fully or marginally acceptable answers suggested by an examination of the data provided by about 60 subjects. During scoring, the scorers noted any additional answers that might receive credit, and further revisions to the keys were made as necessary. Items varied appreciably in the number of questionable answers they elicited and in the difficulty of decisions as to whether an answer should receive any credit. On the average, nearly two hours was required per open-ended item for development and revision of the scoring keys. These keys still reflect primarily the judgments of one individual; review by several additional individuals would be required if such keys were to be developed for other than experimental use of open-ended items.

Given the availability of detailed scoring keys for the items, most of the tests could be scored clerically. A single highly experienced clerk scored the tests and her supervisor checked all her work; both individuals contributed suggestions leading to revisions in the scoring keys. Two of the tests, however, presented more complex scoring problems: For both single-answer and multiple-answer Analogies, the scoring keys consisted of statements of the relationships between the two words making up the stem of the item, along with examples, rather than an ostensibly complete list, of fully and marginally acceptable answers. Many scoring decisions therefore involved a substantial exercise of judgment. A research assistant scored each of these tests, and the author scored 25 protocols of each independently. Total scores over the tests for the two scorings correlated .95 for one test and .97 for the other.

Scorers kept records of the time spent to complete scoring of each test. For the single-answer forms, the time required for scoring ranged from 27 to 35 hours to score 315 tests, or approximately six minutes per test, in addition to several hours required for checking. For the multiple-answer forms, the time ranged from 50 to 70 hours.

Results

Preliminary Results

Quality of the data. In any group testing in which subjects are asked to perform at a high level over the course
of a long testing session, the first concern is with the quality of the data. No cases were found in which subjects appeared not to take their task seriously. There were only three instances of missing or spoiled tests. (Sample mean scores were substituted for these.) Over all tests and subjects, there were a total of 32 instances in which a subject failed to complete at least half of a test; two individual subjects were each responsible for two of these instances, and the remainder involved individuals for which this occurred only once. It appeared that data from all subjects were of acceptable quality.

Score derivation. The three multiple-choice tests were scored using a standard correction for guessing: for a five-choice item, the score is number right minus one-fourth the number wrong. Two of the keylist tests were simply scored for number right. It would have been possible to treat those tests as 90-alternative multiple-choice tests and to apply the guessing correction, but the effect on the scores would have been of negligible magnitude. A subject who generated 10 wrong answers on a 20-item test, for example, would have received a corrected score only about 0.5 points lower than his number-right score.

For the remaining open-ended tests, the coding of fully and marginally acceptable answers presents several options for score generation. It is possible to score strictly, giving credit only for answers deemed fully acceptable; to score very leniently, giving the same credit to both categories of answers; or to give differential weights to answers depending on their quality. These three schemes were contrasted; for the third option mentioned, the weighting was one point for fully acceptable answers and one-half point for marginal answers. The basis for the comparison was the internal consistency reliability (coefficient alpha) of the total score for the test.

The strict scoring produced the least reliable scores of any method. Over the seven tests for which the comparison was made, the median difference in coefficients was .07, comparing strict with intermediate scoring, and .05, comparing it with lenient scoring. Differential weighting and lenient scoring produced results much more similar to one another (a median difference of .01 favoring the intermediate scoring); in five of seven comparisons, the intermediate scoring produced slightly more reliable results. This method is therefore employed in all further analyses.
Test order. Differences in performance among groups receiving the tests in different orders could have arisen either because of factors related directly to order (for example, practice or fatigue effects) or because the students who were tested in different sessions differed in some respect (students were permitted to select whichever session was most convenient for them). Two analyses were performed to discover whether there were group differences. In each case the data were treated as involving five groups of subjects—the initial four groups, which received tests in the counterbalanced order described earlier, and a fifth group, which was actually an amalgam of all the remaining subjects and which received tests in the same order as did the first group.

The first analysis was an examination of possible differences in the mean level of performance. For each test a one-way analysis of variance was conducted. Eleven of the 12 tests failed to show group differences significant at the 5 percent level of confidence. A significant difference was obtained for the multiple-answer Analogies, but was not attributable to test order: By the least-significant-difference test (Steel & Torrie, 1960), the fifth group had lower mean performance than any of the others, but the first four groups (which included all test orders employed) did not differ from one another. Thus, if the difference is not a chance fluctuation, it indicates one instance of a lower level of performance associated with a specific group of subjects rather than with test order per se.

A second analysis examined possible differences in the standard error of measurement—a statistic that incorporates information about both the standard deviation and the reliability of a test score and indicates the precision of measurement. Here, since only one number was obtained from the performance of each group on each test, the analysis was a 12 (tests) by 5 (groups) analysis of variance. No difference associated with groups was found ($F < 1$).

Thus, there was no systematic difference in performance associated with the order in which the tests were administered, and little if any difference among the groups in which subjects were tested. Group membership was therefore ignored in all further analyses.

Test difficulty. Test means and standard deviations are shown in Table 1. Most of the tests ranged from slightly difficult to average difficulty level; two of the keylist tests
### Table 1

**Test Statistics**

<table>
<thead>
<tr>
<th>Type of Test</th>
<th>Type of Item</th>
<th>Mean</th>
<th>S.D.</th>
<th>Maximum Possible Score</th>
<th>Percent Completing</th>
<th>Reliability</th>
<th>Median Item</th>
<th>Biserial</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multiple-Choice</strong></td>
<td>Sentence Completion</td>
<td>9.65</td>
<td>4.29</td>
<td>20</td>
<td>99.7</td>
<td>.68</td>
<td>.40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Analogies</td>
<td>8.60</td>
<td>3.35</td>
<td>20</td>
<td>98.7</td>
<td>.45</td>
<td>.32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Antonyms</td>
<td>4.86</td>
<td>4.34</td>
<td>20</td>
<td>98.4</td>
<td>.69</td>
<td>.39</td>
<td></td>
</tr>
<tr>
<td><strong>Keylist</strong></td>
<td>Sentence Completion</td>
<td>15.96</td>
<td>2.33</td>
<td>20</td>
<td>96.2</td>
<td>.62</td>
<td>.39</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Analogies</td>
<td>15.02</td>
<td>3.03</td>
<td>20</td>
<td>99.0</td>
<td>.70</td>
<td>.40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Antonyms</td>
<td>11.48</td>
<td>3.34</td>
<td>20</td>
<td>97.5</td>
<td>.73</td>
<td>.41</td>
<td></td>
</tr>
<tr>
<td><strong>Single-Answer</strong></td>
<td>Sentence Completion</td>
<td>7.99</td>
<td>3.30</td>
<td>20</td>
<td>90.5</td>
<td>.73</td>
<td>.42</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Analogies</td>
<td>7.27</td>
<td>2.74</td>
<td>20</td>
<td>84.4</td>
<td>.57</td>
<td>.34</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Antonyms</td>
<td>6.88</td>
<td>3.44</td>
<td>20</td>
<td>98.1</td>
<td>.75</td>
<td>.42</td>
<td></td>
</tr>
<tr>
<td><strong>Multiple-Answer</strong></td>
<td>Sentence Completion</td>
<td>24.02</td>
<td>7.11</td>
<td>54</td>
<td>90.8</td>
<td>.80</td>
<td>.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Analogies</td>
<td>13.31</td>
<td>5.23</td>
<td>45</td>
<td>74.6</td>
<td>.59</td>
<td>.37</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Antonyms</td>
<td>17.03</td>
<td>5.67</td>
<td>45</td>
<td>97.1</td>
<td>.75</td>
<td>.48</td>
<td></td>
</tr>
</tbody>
</table>
were easy, while multiple-choice Antonyms was very difficult. Means for the multiple-answer tests are low in relation to the maximum possible score, but represent one to one-and-a-half fully acceptable answers per item.

**Test speededness.** Tests such as the GRE Aptitude Test are considered unspeeded if at least 75 percent of the examinees attempt all items and virtually everyone attempts at least three-fourths of the items. By these criteria only one of the tests, multiple-answer Analogies, has any problems with speededness: About 75 percent of the sample reached the last item, but 14 percent failed to attempt the twelfth item, which represents the three-fourths point. For all the remaining tests, 95 percent or more of the subjects reached at least all but the final two items. Table 1 shows the percent of the sample completing each test.

**Test reliability.** Reliabilities (coefficient alpha) are also shown in Table 1. They range from .45 to .80, with a median of .69. There are no differences in reliabilities associated with the response format of the test—the medians range from .68 for multiple-choice tests to .75 for multiple-answer forms. There are differences associated with item type; medians are .74 for Antonyms, .71 for Sentence Completion, and .58 for Analogies. The least reliable of all the tests is the multiple-choice Analogies. The differences apparently represent somewhat less success in creating good analogies items rather than any differences inherent in the open-ended response formats.

**Item biserials.** Biserial correlations were computed between each item and the total test score of which it was a part. All the biserials were positive, and they were of similar magnitude across the various formats and item types. Again the lowest coefficients were those for several of the Analogies tests. The median correlation for each test is shown in the last column of Table 1.

**Relations among the Tests**

**Correlations among tests.** Zero-order correlations among the 12 tests are shown in Table 2. The correlations range from .29 to .69, with a median of .53. The seven lowest coefficients in the table, and the only ones below .40, are correlations involving the multiple-answer Analogies test.
Table 2

Zero-Order Correlations

<table>
<thead>
<tr>
<th></th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
<th>(9)</th>
<th>(10)</th>
<th>(11)</th>
<th>(12)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multiple-Choice</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>.52</td>
<td>.54</td>
<td>.56</td>
<td>.59</td>
<td>.66</td>
<td>.64</td>
<td>.47</td>
<td>.69</td>
<td>.65</td>
<td>.36</td>
<td>.58</td>
</tr>
<tr>
<td>Analogies</td>
<td>.46</td>
<td>.44</td>
<td>.48</td>
<td>.52</td>
<td>.48</td>
<td>.44</td>
<td>.50</td>
<td>.47</td>
<td>.33</td>
<td>.45</td>
<td></td>
</tr>
<tr>
<td>Antonyms</td>
<td>.43</td>
<td>.52</td>
<td>.56</td>
<td>.54</td>
<td>.40</td>
<td>.61</td>
<td>.53</td>
<td>.29</td>
<td>.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Keylist</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>.55</td>
<td>.54</td>
<td>.58</td>
<td>.43</td>
<td>.53</td>
<td>.52</td>
<td>.38</td>
<td>.53</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analogies</td>
<td>.60</td>
<td>.59</td>
<td>.47</td>
<td>.59</td>
<td>.58</td>
<td>.33</td>
<td>.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antonyms</td>
<td>.58</td>
<td>.44</td>
<td>.68</td>
<td>.60</td>
<td>.37</td>
<td>.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Single-Answer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>.52</td>
<td>.64</td>
<td>.67</td>
<td>.39</td>
<td>.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analogies</td>
<td>.47</td>
<td>.49</td>
<td>.41</td>
<td>.52</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antonyms</td>
<td>.68</td>
<td>.41</td>
<td>.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Multiple-Answer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>.50</td>
<td>.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analogies</td>
<td>.41</td>
<td>.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antonyms</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Correlations corrected for attenuation are shown in Table 3; the correction is based on coefficient alpha reliabilities. The correlations range from .45 to .97 and have a median of .80. The lowest coefficient, other than ones for correlations involving multiple-answer Analogies, is .64.

Factor analyses. The structure of the set of correlations was examined in several ways. We will consider first the results of several factor analyses.

A preliminary principal components analysis produced the set of eigenvalues displayed in Table 4. The first component is very large, accounting for 57 percent of the total variance, while the next largest accounts for only 7 percent of the variance. By one rule of thumb for determining number of factors, that of the number of eigenvalues greater than 1.0, there is only a single factor represented in these results. By another, that of differences in magnitude of successive eigenvalues, there is some evidence for a second factor, but none at all for more than two.

It was originally planned to use a confirmatory factor analytic approach to the analysis (Jöreskog, 1970) in order to contrast two idealized models of test relations—one involving three item-type factors and one involving four response-format factors. In view of the results of the principal components analysis, however, either of these would clearly be a distortion of the data. It was decided, therefore, to use an exploratory factor analysis, which could be followed by confirmatory analyses comparing simpler models if such a comparison seemed warranted from the results. The analysis was a principal axes factor analysis with iterated communalities.

A varimax (orthogonal) rotation of the two-factor solution produced unsatisfactory results—10 of the 12 scores had appreciable loadings on both factors. The results of the oblimin (oblique) rotation for two factors are presented in Table 5. The two factors are highly correlated (r = .67). Ten of the 12 scores have their highest loading on Factor I, one (single-answer Analogies) divides about equally between the two, and only one (multiple-answer Analogies) has its principal loading on the second factor.

For two item types, Sentence Completion and Antonyms, these results leave no ambiguity as to the effects of response format: The use of an open-ended format makes no difference in the attribute measured by the test. The interpretation for the
Table 3

Correlations Corrected for Attenuation

<table>
<thead>
<tr>
<th></th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
<th>(9)</th>
<th>(10)</th>
<th>(11)</th>
<th>(12)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multiple-Choice</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>.92</td>
<td>.79</td>
<td>.86</td>
<td>.94</td>
<td>.91</td>
<td>.75</td>
<td>.97</td>
<td>.88</td>
<td>.57</td>
<td>.81</td>
<td></td>
</tr>
<tr>
<td>Analogies</td>
<td>.83</td>
<td>.83</td>
<td>.86</td>
<td>.91</td>
<td>.84</td>
<td>.87</td>
<td>.88</td>
<td>.80</td>
<td>.64</td>
<td>.77</td>
<td></td>
</tr>
<tr>
<td>Antonyms</td>
<td>.66</td>
<td>.75</td>
<td>.79</td>
<td>.76</td>
<td>.64</td>
<td>.85</td>
<td>.71</td>
<td>.45</td>
<td>.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Keylist</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>.83</td>
<td>.80</td>
<td>.85</td>
<td>.72</td>
<td>.78</td>
<td>.74</td>
<td>.63</td>
<td>.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analogies</td>
<td>.84</td>
<td>.83</td>
<td>.74</td>
<td>.81</td>
<td>.78</td>
<td>.51</td>
<td>.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antonyms</td>
<td>.79</td>
<td>.68</td>
<td>.92</td>
<td>.79</td>
<td>.56</td>
<td>.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Single-Answer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>.81</td>
<td>.86</td>
<td>.88</td>
<td>.59</td>
<td>.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analogies</td>
<td>.72</td>
<td>.73</td>
<td>.71</td>
<td>.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antonyms</td>
<td>.88</td>
<td>.62</td>
<td>.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Multiple-Answer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>.73</td>
<td>.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analogies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antonyms</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4

Principal Components of Correlation Matrix

<table>
<thead>
<tr>
<th>Component</th>
<th>Eigenvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>6.80</td>
</tr>
<tr>
<td>II</td>
<td>.83</td>
</tr>
<tr>
<td>III</td>
<td>.63</td>
</tr>
<tr>
<td>IV</td>
<td>.59</td>
</tr>
<tr>
<td>V</td>
<td>.57</td>
</tr>
<tr>
<td>VI</td>
<td>.48</td>
</tr>
<tr>
<td>VII</td>
<td>.43</td>
</tr>
<tr>
<td>VIII</td>
<td>.42</td>
</tr>
<tr>
<td>IX</td>
<td>.37</td>
</tr>
<tr>
<td>X</td>
<td>.30</td>
</tr>
<tr>
<td>XI</td>
<td>.29</td>
</tr>
<tr>
<td>XII</td>
<td>.28</td>
</tr>
</tbody>
</table>
Table 5
Factor Pattern for Two-Factor Analysis

<table>
<thead>
<tr>
<th>Test</th>
<th>I</th>
<th>II</th>
<th>Communality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple-Choice</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>.84</td>
<td>-.05</td>
<td>.65</td>
</tr>
<tr>
<td>Analogies</td>
<td>.59</td>
<td>.06</td>
<td>.40</td>
</tr>
<tr>
<td>Antonyms</td>
<td>.79</td>
<td>-.12</td>
<td>.51</td>
</tr>
<tr>
<td>Keylist</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>.57</td>
<td>.15</td>
<td>.47</td>
</tr>
<tr>
<td>Analogies</td>
<td>.74</td>
<td>.01</td>
<td>.55</td>
</tr>
<tr>
<td>Antonyms</td>
<td>.86</td>
<td>-.10</td>
<td>.65</td>
</tr>
<tr>
<td>Single-Answer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>.68</td>
<td>.16</td>
<td>.63</td>
</tr>
<tr>
<td>Analogies</td>
<td>.35</td>
<td>.37</td>
<td>.44</td>
</tr>
<tr>
<td>Antonyms</td>
<td>.85</td>
<td>-.03</td>
<td>.69</td>
</tr>
<tr>
<td>Multiple-Answer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>.59</td>
<td>.29</td>
<td>.66</td>
</tr>
<tr>
<td>Analogies</td>
<td>.06</td>
<td>.63</td>
<td>.45</td>
</tr>
<tr>
<td>Antonyms</td>
<td>.65</td>
<td>.17</td>
<td>.60</td>
</tr>
</tbody>
</table>
two genuinely open-ended Analogies tests is less clear. The second factor is small (just under 5 percent of the common factor variance), and it is poorly defined, with only one test having its primary loading on that factor. Moreover, the one test that did load heavily on factor II was also the one test in the battery that was at all speeded. There is a reasonable interpretation of factor II as a speed factor (Donlon, 1980); the rank-order correlation between factor II loadings and the number of subjects failing to attempt the last item of a test was .80 ($p < .01$).

The poor definition of the second factor makes its location problematic. An expert consultant suggested a graphic orthogonal rotation as most appropriate. This rotation leaves the pattern of factor loadings unchanged, but weakens the already weak second factor—it leaves the multiple-answer Analogies test with approximately equal loadings on factors I and II and no other test with a loading as high as .3 on the second factor.

Several additional factor analyses were undertaken to check the applicability of these results to subgroups of the sample. First, analyses were conducted separately for high and low ability subjects. To avoid problems of correlated error variance, the sample was divided at the median on the basis of performance on one test (keylist Antonyms), and that test was excluded from the analysis. Results differ somewhat for the two groups. For high ability subjects, the tests tended to divide along the lines of item type, with all Antonyms tests, and all Sentence Completion tests except for the keylist format, loading factor I, while all Analogies tests loaded factor II along with the Sentence Completion keylist format. For low ability subjects, all tests had their highest loading on factor I, and only the multiple-answer Analogies had an appreciable loading on factor II. Thus, for neither group is there evidence suggesting the separation of tests on the basis of response format, except for the repetition of the finding from the total group analysis of something distinct about multiple-answer Analogies.

Analyses were also performed taking into account the academic level of the student. The sample included two groups large enough to be considered for separate analyses—seniors ($N = 75$) and juniors ($N = 141$); on all 12 tests there were small differences in the direction both of higher means and of higher variances for seniors. Analyses (courtesy of Dr. Ledyard Tucker) were performed using uniterated squared multiple correlations in the diagonal. For both groups a one-factor
A combined analysis was also carried out after adjusting for mean and variance differences in the data for the two groups. The eigenvalues suggested either a one- or a two-factor solution; the two-factor solution, however, showed all tests having their strongest loading on the first factor and only multiple-answer Analogies approaching an equal division of its variance between the two factors.

Thus, there is no strong evidence for the existence of a second factor in the data. There are weak indications that the multiple-answer Analogies test and, to a much lesser extent, single-answer Analogies provide somewhat distinct measurement from the remainder of the tests in the battery; evidence is clear that Sentence Completion and Antonyms item types measure the same attribute regardless of the format in which the item is administered.

Multitrait-multimethod analysis. The data may also be considered within the framework provided by multitrait-multimethod analysis (Campbell & Fiske, 1959). Each of the three item types constitutes a "trait," while each of the four response formats constitutes a "method." The data were analyzed following a scheme suggested by an article by Goldberg and Werts (1966). All the correlations relevant for each comparison were corrected for attenuation and then averaged, using Fisher's r-to-z transformation. Results are summarized in Table 6.

Each row in the upper part of the table provides the average of (a) all those correlations that represent relations for a single item type as measured in different formats and of (b) all those correlations that represent relations between that item type and other item types when the two tests employ different response formats. Thus, for the Sentence Completion item type, the entry in the first column is an average of all six correlations among Sentence Completion scores from the four formats. The entry in the second column is an average of 24 correlations— for each of four Sentence Completion scores, the six correlations representing relations to each item type other than Sentence Completion in each of three formats. The lower part of the table is organized analogously; it provides for each response format a comparison of average correlations within format with those between formats for all test pairs involving different item types.

Results in the upper part of the table show that there is some variance associated with trait for both Sentence Completion and Antonyms item types (by Mann-Whitney U test, p < .05).
Table 6
Multitrait-Multimethod Presentation of Average Correlations

<table>
<thead>
<tr>
<th>Trait</th>
<th>Monotrait-Heteromethod</th>
<th>Heterotrait-Heteromethod</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence Completion</td>
<td>.86</td>
<td>.80*</td>
</tr>
<tr>
<td>Analogies</td>
<td>.75</td>
<td>.75</td>
</tr>
<tr>
<td>Antonyms</td>
<td>.85</td>
<td>.79*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Monomethod-Heterotrait</th>
<th>Heteromethod-Heterotrait</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple-Choice</td>
<td>.86</td>
<td>.81</td>
</tr>
<tr>
<td>Keylist</td>
<td>.82</td>
<td>.79</td>
</tr>
<tr>
<td>Single-Answer</td>
<td>.80</td>
<td>.80</td>
</tr>
<tr>
<td>Multiple-Answer</td>
<td>.73</td>
<td>.73</td>
</tr>
</tbody>
</table>

*By Mann-Whitney U Test, the two entries in a row are significantly different at the 5% level of confidence.
Analogy tests do not, however, relate to one another any more strongly than they relate to tests of other item types.

The lower part of the table shows differences attributable to response format. There is an apparent tendency toward a difference in favor of stronger relations among multiple-choice tests than those tests have with tests in other formats, but this tendency does not approach significance ($p > .10$). For the truly open-ended response formats, there are no differences whatsoever. Like the factor analyses, this approach to correlational comparisons shows no tendency for open-ended tests to cluster according to the response format; to the slight degree that any differences are found, they represent clustering on the basis of the item type rather than the response format employed in a test.

Correlations corrected for "alternate forms" reliabilities. The multitrait-multimethod correlational comparison made use of internal consistency reliability coefficients to correct correlations for their unreliability. Several interesting comparisons can also be made using a surrogate for alternate forms reliability coefficients. The battery, of course, contained only one instance of each item-type-by-response-format combination, so that no true alternate form examinations can be made. It may be reasonable, however, to consider the two truly open-ended forms of a test, multiple-answer and single-answer, as two forms of the same test given under "open" conditions, and the two remaining forms, multiple-choice and keylist, as two forms of the same test given under "closed" conditions. On this assumption, relations across open and closed formats for a given item type can be estimated by the average of the four relevant correlations and corrected for reliabilities represented by the correlations within open and within closed formats.

The corrected correlations are .97 for Sentence Completion, .88 for Analogies, and 1.05 for Antonyms. It appears that relations across the two kinds of formats do not differ from 1.0, except for error in the data, for two item types. Analogy tests may fail to share some of their reliable variance across open and closed formats, but still appear to share most of it.

An analogous procedure can be used to estimate relations across item types. For example, the relation between Sentence Completion and Analogies is estimated by averaging the 16 correlations between the two (four Sentence Completion tests times four Analogies tests); the correlation is corrected on
the basis of estimated reliability of Sentence Completion (the average of correlations among Sentence Completion tests) and of Analogies tests. The estimated true correlation between Sentence Completion and Analogies is .98, that between Sentence Completion and Antonyms is .96, and that between Analogies and Antonyms is .94. Thus there is no indication that any one of the item types stands out as differentiating tests from one another.

Correlations with Background Variables

Students completed a questionnaire dealing with their academic background, accomplishments, and interests. Included were questions concerning (a) plans for graduate school attendance and advanced degrees, (b) undergraduate grade-point average overall and in the major field of study, (c) preferred career activities, (d) self-assessed skills and competencies within the major field, and (e) independent activities and accomplishments within the current academic year. Correlations were obtained between questionnaire variables and scores on the 12 verbal tests.

Most of the correlations were very low; only four of the questions produced a correlation with any test as high as .20. Results for these questions are presented in Table 7.

Level of degree planned was coded 3 for doctorate, 2 for master's, and 1 for bachelor's or no answer. This variable had low positive correlations with all of the tests.

Self-reported grade-point average, overall and for major field alone, had significant correlations with test scores in all but one instance. Most of the correlations fell between .20 and .30; no systematic differences associated with item type or response format of the tests was evident.

The final questionnaire variable showing correlations as large as .20 was one of six in which the students indicated what would be their single most preferred professional activity. An interest in writing correlated significantly with scores from 8 of the 12 tests.

Information was also available on the student's sex and year in school. No significant correlations with sex were obtained; results for year in school, also presented in Table 7, show superior performance for more advanced students, significantly so for 8 of the 12 tests.
Table 7

Correlations with Background Variables

<table>
<thead>
<tr>
<th>Test</th>
<th>Level of Degree Planned</th>
<th>Overall GPA</th>
<th>Major GPA</th>
<th>Interest in Writing</th>
<th>Year in School</th>
<th>GRE-Verbal</th>
<th>GRE-Quantitative</th>
<th>GRE-Analytic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multiple-Choice</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>.11</td>
<td>.32*</td>
<td>.28*</td>
<td>.21*</td>
<td>.22*</td>
<td>.72*</td>
<td>.40*</td>
<td>.35</td>
</tr>
<tr>
<td>Analogies</td>
<td>.11</td>
<td>.27*</td>
<td>.25*</td>
<td>.10</td>
<td>.12</td>
<td>.50*</td>
<td>.32</td>
<td>.35</td>
</tr>
<tr>
<td>Antonyms</td>
<td>.14</td>
<td>.19*</td>
<td>.22*</td>
<td>.24*</td>
<td>.21*</td>
<td>.64*</td>
<td>.35</td>
<td>.31</td>
</tr>
<tr>
<td><strong>Keylist</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>.11</td>
<td>.22*</td>
<td>.20*</td>
<td>.09</td>
<td>.13</td>
<td>.59*</td>
<td>.33</td>
<td>.56*</td>
</tr>
<tr>
<td>Analogies</td>
<td>.11</td>
<td>.26*</td>
<td>.29*</td>
<td>.19*</td>
<td>.19*</td>
<td>.55*</td>
<td>.25</td>
<td>.48*</td>
</tr>
<tr>
<td>Antonyms</td>
<td>.21*</td>
<td>.30*</td>
<td>.24*</td>
<td>.18*</td>
<td>.26*</td>
<td>.74*</td>
<td>.47*</td>
<td>.46*</td>
</tr>
<tr>
<td><strong>Single-Answer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>.15*</td>
<td>.29*</td>
<td>.29*</td>
<td>.28*</td>
<td>.19*</td>
<td>.59*</td>
<td>.25</td>
<td>.37</td>
</tr>
<tr>
<td>Analogies</td>
<td>.13</td>
<td>.20*</td>
<td>.14</td>
<td>.13</td>
<td>.08</td>
<td>.58*</td>
<td>.37</td>
<td>.54*</td>
</tr>
<tr>
<td>Antonyms</td>
<td>.14</td>
<td>.32*</td>
<td>.29*</td>
<td>.21*</td>
<td>.17*</td>
<td>.73*</td>
<td>.37</td>
<td>.48*</td>
</tr>
<tr>
<td><strong>Multiple-Answer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>.16*</td>
<td>.36*</td>
<td>.34*</td>
<td>.17*</td>
<td>.28*</td>
<td>.71*</td>
<td>.46*</td>
<td>.63*</td>
</tr>
<tr>
<td>Analogies</td>
<td>.17*</td>
<td>.20*</td>
<td>.20*</td>
<td>.12</td>
<td>.14</td>
<td>.52*</td>
<td>.54*</td>
<td>.49*</td>
</tr>
<tr>
<td>Antonyms</td>
<td>.21*</td>
<td>.35*</td>
<td>.30*</td>
<td>.22*</td>
<td>.17*</td>
<td>.60*</td>
<td>.26</td>
<td>.26</td>
</tr>
</tbody>
</table>

*Significantly different from zero at the 1% level of confidence. Ns range from 308 to 315, except for GRE scores, for which they are 41.
Also included in the table are correlations with GRE Aptitude Test scores, obtained with the students' permission from GRE files. Scores were available for only 41 of the subjects. Correlations with the GRE verbal score were substantial in magnitude, ranging from .50 to .74 with a median of .59. Correlations with the GRE quantitative and analytical scores were lower but still appreciable; the medians are .36 for the quantitative score and .47 for analytical.

These results provide no evidence of systematic differences in relationships across test formats. Like the analyses of correlations among the experimental tests, they suggest that response format has little effect on the nature of the attributes measured by these item types.

Discussion

This study has shown that it is possible to develop open-ended forms of several verbal aptitude item types that are approximately as good, in terms of score reliability, as multiple-choice items and that require only slightly greater time limits than do the conventional items. The free-response tests do, however, present some unique problems for both test development and scoring. In general, it is not desirable to use the same item stems for tests to be given in different formats; for example, much easier vocabulary must be used in an open-ended Antonyms test than is appropriate for a multiple-choice one. (As an illustration of the differences, consider the set of items receiving mean scores equivalent to correct responses by approximately 40 percent of the sample for two Antonyms tests. For the multiple-choice test, there were five words at this difficulty level: aver, dolor, execrate, fecund, and sanguine. For the single-answer test, there were three: contempt, expedite, and robust.) And in scoring the tests, more reliable scores are obtained when partial credit is given for marginally appropriate answers than when only the best possible answers are counted. The more liberal scoring is apparently sensitive to partial knowledge possessed by examinees, which enables them to produce answers reflecting some aspects of the meaning of a word even when they lack appreciation of its nuances.

Scoring the open-ended tests is time consuming. For tests in which the examinee was required to produce and write a single answer for each item, for example, scoring required on the average 38 percent as much time as the examinees spent in taking
the test. Scoring is also not an entirely mechanical process--
even after careful initial preparation of scoring keys and
checking against actual responses by 20 percent of the sample,
occasional new answers appeared that were judged fully acceptable.

Scoring problems, however, can be overcome by use of the
keylist format as a surrogate for truly open-ended tests. Only
one of the three keylist tests employed required the use of
partial credit scoring, and even for that test it seems likely
that sets of items could be assembled that would not require
such scoring. The keylist tests were hand-scored in the present
study, but could easily be designed so that the examinee would
grid the number of his response on a machine-scorable answer
sheet.

The major finding of the study, however, is that open-ended
versions of these item types add little information distinct
from that provided by multiple-choice items. Relations among
the tests were examined in several ways--through factor analysis,
by use of the multitrait-multimethod approach to correlational
comparisons, and with an analysis that approximates an alternate-
forms comparison. When correlations across test formats were
corrected for attenuation on the basis of internal consistency
reliability coefficients, the estimates of true correlations
were on the order of .80; when they were corrected using
"alternate forms," the estimates were still higher. For two
item types, Antonyms and Sentence Completion, in fact, those
latter estimates were indistinguishable from 1.0.

There is no evidence, therefore, for a general factor
attributable to clustering of open-ended tests across item
type, and there is strong evidence against any difference in
the attribute measured by Antonyms or Sentence Completion items
as a function of the response requirements of the test. Only
in the case of Analogies is there some possibility of differences
associated with an open-ended format.

The strength of even these differences is open to some
question. One approach to factor analysis identified two
factors, with 10 of the 12 tests employed loading only the
first factor, single-answer Analogies dividing about equally
between the two factors, and multiple-answer Analogies loading
only the second. Another approach produced weaker differences--
only multiple-answer Analogies had an appreciable loading on
the second factor, and even that test divided its variance
between the two about equally.
Furthermore, there are several possible interpretations of the basis on which the multiple-answer Analogies test differs from the remainder of the battery. This test was the most difficult to score of any that were employed, requiring (along with single-answer Analogies) complex judgments rather than a largely clerical approach to scoring; and it was the only test in the battery that was at all speeded, according to testing program criteria. The possibility that the second factor is a speed factor is strengthened by the finding of a high correlation between loadings on factor II and the number of subjects who failed to complete each test.

The general lack of differentiation seen in relations among the tests also receives some support from relations between these tests and information on the subjects' academic background. Few of the questionnaire variables showed any substantial relations with performance on the verbal tests; where such relations were found, the pattern was one of relatively uniform relations irrespective of test format.

It is clear that an open-ended response format was not in itself sufficient to determine what these tests measured. Neither the requirement to generate a single response, nor the more difficult task of producing and writing several different answers to an item, could alone change the abilities that were important for successful performance. What, then, are the characteristics of an item that will measure different attributes depending on the response format employed? A comparison of the present tests with those employed in the earlier problem-solving research of Ward et al. (1980) and Frederiksen et al. (1980) suggests a number of possibilities. In the problem-solving work, subjects had to read and comprehend passages containing a number of items of information relevant to a problem. They were required to determine the relevance of such information for themselves and often to apply reasoning and inference to draw conclusions from several items of information. Moreover, they needed to draw on information not presented—specialized knowledge concerning the design and interpretation of research studies, for the behavioral science problems, and more general knowledge obtained from everyday life experiences, for the nontechnical problems. Finally, subjects composed responses that often entailed relating several complex ideas to one another.
The discrete verbal item types, in contrast, are much more self-contained. The examinee has only to deal with the meanings of one word, of a pair of words, or at most of the elements of a short sentence. In a sense, the statement of the problem includes a specification of what information is relevant for a solution and of what kind of a solution is appropriate. Thus the verbal tests might be described as "well-structured," while the problem-solving tests are "ill-structured" problems (Simon, 1973). The verbal tests also, of course, require less complex responses—a single word or at most a pair of words.

Determining which of these features are critical in distinguishing tests in which an open-ended format makes a difference will require comparing a number of different item types in multiple-choice and free-response formats. It will be of particular interest to develop item types that eliminate the confounding of complexity in the information search required by a problem with complexity in the response to be produced.

For programs concerned with standardized aptitude testing, the present results indicate that one important component of existing tests amounts to sampling from a broader range of possible test questions than had previously been demonstrated. The discrete verbal item types presently employed by the GRE and other testing programs appear to suffer no lack of generality because of exclusive use of a multiple-choice format; for these item types at least, use of open-ended questions would not lead to measurement of a noticeably different ability that cuts across the three item types examined here. It remains to be seen whether a similar statement can be made about other kinds of questions employed in standardized tests, and whether there are ways in which items that will tap "creative" or "divergent thinking" abilities can be presented so as to be feasible for inclusion in large-scale testing.

Further, it may be worthwhile to explore the possibility that addition of the keylist format of one or more verbal item types would make a contribution to existing aptitude tests. While such items would not lead to measurement of different abilities, they could measure the same abilities using concepts drawn from a somewhat different pool. They would therefore increase the number of items potentially available for testing a given population and might facilitate differentiation among very high ability candidates without the use of the extremely rare and difficult vocabulary that is often needed to achieve discrimination with multiple-choice items.
References


Appendix A

Instructions and Sample Items
Sentence Completion

Multiple Choice Form

Time - 12 minutes
20 questions

Directions: Each of the sentences below has one blank space, the blank indicating that a word has been omitted. Beneath the sentence are five lettered words. Choose the lettered word that, when inserted in the sentence, best fits in with the meaning of the sentence as a whole. Mark your answer by writing its letter in the space provided.

Sample Question:

1. Laura presented her viewpoint so _______ that no one knew how strongly she really felt about it.

   D (A) tangentially    (B) euphoniously    (C) aptly
     (D) dispassionately (E) supportively

Laura failed to show the strength of her feelings. The correct answer is (D): she presented her viewpoint dispassionately.

Turn to the next page and begin work.
Sentence Completion

Keylist Form

Time - 15 minutes
20 questions

Directions: Each of the sentences below has one blank space, the blank indicating that a word has been omitted. Think of the one word that, when inserted in the sentence, best fits in with the meaning of the sentence as a whole. Locate this word on the sheet entitled Sentence Completion Keylist. Mark your answer by writing its number on the line below the sentence. If your first answer does not appear in the list, try to think of a different answer.

Sample Question:

1. The government, which appeared to be in control, was in reality ------- to determinants in the marketplace.

79

Rather than being in control, the government was subordinate to determinants in the marketplace. The word subordinate is number 79 in the Keylist; therefore that number is entered in the blank space.

Note that there are several good answers to this item. The blank space could have been filled with number 5 (ancillary), number 7 (answerable), or number 78 (subject).

Turn to the next page and begin work.
Sentence Completion

Single Answer Form

Time - 15 minutes
20 questions

Directions: Each of the sentences below has one blank space, the blank indicating that a word has been omitted. Think of the one word that, when inserted in the sentence, best fits in with the meaning of the sentence as a whole. Write your answer on the line below the sentence.

Sample Question:

1. The doctors had to determine the ______ dosage of the drug for the patient: enough to control his hyperactive excesses but not enough to make him lethargic.

   optimal

   The problem was to find a dosage that would be just enough but not too much: the optimal dosage.

   Note that there are several good answers to this question. The blank space could have been filled with proper.

   Turn to the next page and begin work.
Sentence Completion

Multiple Answer Form

Time - 15 minutes
18 questions

Directions: Each of the sentences below has one blank space, the blank indicating that a word has been omitted. Think of three single words each of which, when inserted in the sentence, fits in with the meaning of the sentence as a whole. Write your answers on the lines below the sentence.

Sample Question:

1. One of Poe's great talents as a storyteller was the relating of _______ incidents in such a way that the reader smiles even while shuddering at the revelations.

   gruesome, horrifying, macabre

To make the reader shudder, the incidents must be gruesome, horrifying, macabre.

Turn to the next page and begin work.
Analogies

Multiple Choice Form

Time - 12 minutes
20 questions

Directions: In each of the following questions, a related pair of words is followed by five lettered pairs of words. Select the lettered pair that best expresses a relationship similar to that expressed in the original pair. Mark your answer by writing its letter in the space provided.

Sample Question:

1. JESTER:AMUSING  (A) villain:reactionary
                             (B) protagonist:melodramatic
                             (C) vassal:experienced
                             (D) oaf:awkward
                             (E) pauper:insensitive

       D

A jester is expected to be amusing. The correct answer is (D): an oaf is expected to be awkward.

Turn to the next page and begin work.
Analogies

Keylist Form

Time - 15 minutes
20 questions

Directions: In each of the following questions, a related pair of words is followed by a third word and a blank space. Think of a word that will complete the analogy—that is, a word that has the same relation to the third word as the second word has to the first. Locate this word on the sheet entitled Analogies Keylist. Mark your answer by writing its number in the blank space. If your first answer does not appear in the list, try to think of a different answer.

Sample Question:

1. sermon:lecture sacrament

A sermon is a religious lecture. A sacrament is a religious ceremony. The word ceremony is number 5 in the Keylist; therefore that number is entered in the blank space.

Note that there are several good answers to this item. The blank space could have been filled with number 70 (rite).

Turn to the next page and begin work.
Analogies

Single Answer Form

Time - 15 minutes
20 questions

Directions: In each of the following questions, a related pair of words is followed by a pair of blank spaces. Fill in the blank spaces with two words that have a relationship similar to that expressed in the original pair of words.

Sample Question:

1. thermometer:temperature

A thermometer is an instrument for measuring temperature. The sample answer names another instrument and the characteristic it is designed to measure.

Turn to the next page and begin work.
Analogies

Multiple Answer Form

Time - 15 minutes
15 questions

Directions: In each of the following questions, a related pair of words is followed by three pairs of blank spaces. Fill in each pair of blank spaces with two words that have a relationship similar to that expressed in the original pair of words.

Sample Question:

1. nod:assent

<table>
<thead>
<tr>
<th>sneer</th>
<th>contumy</th>
</tr>
</thead>
<tbody>
<tr>
<td>frown</td>
<td>displeasure</td>
</tr>
<tr>
<td>applaud</td>
<td>appreciation</td>
</tr>
</tbody>
</table>

To nod is a way of expressing assent. Each sample answer names an action and the feeling that it is a way of expressing.

Turn to the next page and begin work.
Antonyms

Multiple Choice Form

Time - 12 minutes
20 questions

Directions: Each question below consists of a word printed in capital letters followed by five words lettered A through E. Choose the lettered word that is most nearly opposite in meaning to the word in capital letters. Since some of the questions require you to distinguish fine shades of meaning, be sure to consider all the choices before deciding which one is best. Mark your answer by writing its letter in the space provided.

Sample Question:

1. PROMULGATE

   C (A) distort (B) demote (C) suppress (D) retard (E) discourage

   Promulgate means to make known or public by open declaration. The correct answer is (C): suppress means to prohibit publication or to keep from public knowledge.

Turn to the next page and begin work.
Antonyms

Keylist Form

Time - 15 minutes
20 questions

Directions: Each question below consists of a word printed in capital letters followed by a blank space. Think of the word that is most nearly opposite in meaning to the word in capital letters. Locate this word on the sheet entitled Antonyms Keylist. Mark your answer by writing its number in the blank space. If your first answer does not appear in the list, try to think of a different answer.

Sample Question:

1. DEPLORABLE

Deplorable means wretched or lamentable. A good antonym is praiseworthy. The word praiseworthy is number 65 in the Keylist; therefore that number is entered in the blank space.

Note that there are several good answers to this question. The blank space could have been filled with number 53 (laudable) or number 17 (commendable).

Turn to the next page and begin work.
Antonyms

Single Answer Form

Time - 15 minutes
20 questions

Directions: Each question below consists of a word printed in capital letters followed by a blank space. Think of the word that is most nearly opposite in meaning to the word in capital letters. Write your answer in the blank space.

Sample Question:

1. PIVOTAL
   incidental

   Pivotal means vitally important or crucial. A good antonym is incidental.

   Note that there are several good answers to this question. The blank space could have been filled with peripheral or unimportant.

   Turn to the next page and begin work.
Antonyms

Multiple Answer Form

Time - 15 minutes
15 questions

Directions: Each question below consists of a word printed in capital letters followed by three blank spaces. Think of as many words as you can, up to three, that are opposite or nearly opposite in meaning to the word in capital letters. Write your answers in the blank spaces.

Sample Question:

1. SUBTLE

   gross  obvious  dense

   Subtle has a number of different meanings, including delicate, obscure, and keenly insightful. Gross is a good antonym for the first meaning, obvious for the second, and dense for the third.

   Turn to the next page and begin work.
Boldt, R. R. Comparison of a Bayesian and a Least Squares Method of Educational Prediction. GREB No. 70-3P, June 1975.


Campbell, J. T. and Donlon, T. F. Relationship of the Figure Location Test to Choice of Graduate Major. GREB No. 75-7P, November 1980.


Donlon, T. F.; Reilly, R. R.; and McKee, J. D. Development of a Test of Global vs. Articulated Thinking: The Figure Location Test. GREB No. 74-9P, June 1978.


Pike, L. Implicit Guessing Strategies of GRE Aptitude Examinees Classified by Ethnic Group and Sex. GREB No. 75-1OP, June 1980.


Rock, D. A. The "Test Chooser": A Different Approach to a Prediction Weighting Scheme. GREB No. 70-2P, November 1974.

Sharon, A. T. Test of English as a Foreign Language as a Moderator of Graduate Record Examinations Scores in the Prediction of Foreign Students' Grades in Graduate School. GREB No. 70-1P, June 1974.


