A Comparison of Free-Response and Multiple-Choice Forms of Verbal Aptitude Tests

William C. Ward
Educational Testing Service

Three verbal item types employed in standardized aptitude tests were administered in four formats—a conventional multiple-choice format and 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.

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 that are provided. A free-response test of behavioral science problem solving, for example, was found to have a very low correlation with a test employing similar problems presented in a machine-scorable (modified multiple-choice) format; it differed from the latter in its relations to a set of reference tests for cognitive factors (Ward, Frederiksen, & Carlson, 1980). Comparable differences were obtained between free-response and machine-scorable tests employing nontechnical problems, which were designed to simulate tasks required in making medical diagnoses (Frederiksen, Ward, Case, Carlson, & Samph, 1981).

There is also suggestive evidence that the use of free-response items could make a contribution in standardized admissions testing. The open-ended behavioral science problems were found to have some potential as predictors of the professional activities and accomplishments of first-year graduate students in psychology; the Graduate Record Examination Aptitude and Advanced Psychology tests are not good predictors of such achievements (Frederiksen & Ward, 1978).

Problem-solving tasks like these, however, provide very inefficient measurement. They require a large investment of examinee time to produce scores with acceptable reliability, and they yield complex responses, the evaluation of which is demanding and time consuming. It was the purpose of the present investigation to explore the effects of an open-ended format with item types like those used in conventional examinations. The content area chosen was verbal knowledge and verbal reasoning, as represented by three item types—Antonyms, Sentence Completion, and Analogies.

The selection of these item types has several bases. First, their relevance for aptitude assess-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 6, No. 1, Winter 1982, pp. 1–11
© Copyright 1982 Applied Psychological Measurement Inc.
0146-6216/82/010001-11$1.55

ment needs no special justification, given that they make up one-half of present verbal ability tests such as the Graduate Record Examination (GRE) and the Scholastic Aptitude Test (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 tests. Second, such items produce reliable scores with relatively short tests. Finally, open-ended 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 free-response problems whose responses may be several sentences in length and may embody two or three complex ideas. Although not solving the difficulties inherent in the use of open-ended tests in large-scale testing, therefore, 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 with essentially the same meaning as a target word (Heim & Watts, 1967; Traub & Fisher, 1977; Vernon, 1962). All found high correlations across formats, but only Traub and Fisher attempted to answer the question of whether the abilities measured in the two formats were identical or only related. They concluded that the format does affect the attribute measured by the test and that there was weak evidence of a factor specific to open-ended verbal items. Unfortunately, they did not have scores on a sufficient variety of tests to provide an unambiguous test for the existence 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 was thus 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.

Method

Description of the Tests

Three item types were employed. Antonyms items (when given in the standard multiple-choice format) required the examinee to select the one of five words that was most nearly opposite in meaning to a given word. Sentence Completions required the identification of the one word which, when inserted into a blank space in a sentence, best fit the meaning of the sentence as a whole. Analogies, finally, called for the selection of the pair of words best expressing a relationship similar to that expressed in a given pair.

Three formats in addition to the multiple-choice one were used. For Antonyms, for example, the "single-answer" format required the examinee to think of an opposite and to write that word in an answer space. The "multiple-answer" format was still more open-ended; the examinee was to think of and write up to three different antonyms for each word given. Finally, the "keylist" format required the examinee to think of an opposite, to locate this word in a 90-item alphabetized list, and to record its number on the answer sheet. This latter format was included as a machine-scorable surrogate for a truly free-response test.

With two exceptions, all open-ended items were ones requiring single-word answers. The exceptions 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.

Instructions for each test paraphrased closely those employed in the GRE Aptitude Test, except as dictated by the specific response requirements of each format. With each set of instructions was given one sample question and a brief rationale for the answer or answers suggested. Except for the multiple-choice tests, two or three fully acceptable answers were given for each sample question.

The tests varied somewhat in number of items and in time limits. Each multiple-choice test consisted of 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 items. On the basis of extensive pretesting, 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 on the order of .7.

Test Administration

Subjects were 315 paid volunteers from a single state university. Slightly more than two-thirds were undergraduate juniors and seniors. The small number (13%) for whom GRE Aptitude Test scores were obtained were a somewhat select group, with means of 547, 646, and 616 on the Verbal, Quantitative, and Analytic scores, respectively. It appears that the sample is a somewhat more able one than college students in general but probably less select than the graduate school applicant pool.

Each student participated in one 4-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 systematic variations of this order were employed to permit an assessment of and adjustment for possible practice or fatigue effect. Each of the four large groups tested, including 51 to 60 subjects, received tests in one of these sequences; the remainder of the sample, tested in groups of 30 to 40, 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. Answers in one set were judged fully acceptable, while those in the second were of marginal appropriateness. An example of the latter would be an Antonyms response that identified the negative evaluation implied by a word but failed to capture an important nuance or the force of the evaluation. 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 to permit computer generation of several different test scores, depending on the credit to be given to marginally acceptable answers.

Preliminary scoring keys were checked for completeness by an examination of about 20% of the answer sheets. Most of the tests were then scored by a single highly experienced clerk and checked by her supervisor. Two tests, however, presented more complex scoring problems. For both single-answer and multiple-answer Analogies, the scoring keys consisted of rationales and examples, rather than a complete list of possible answers. Many scoring decisions therefore involved a substantial exercise of judgment. A research assistant scored each of these tests, and the author scored 25 answer sheets of each independently. Total scores derived from the two scorings correlated .95 for one test and .97 for the other.

Results

Preliminary Results

Quality of the data. No instances were found in which subjects appeared not to take their task seriously. Three answer sheets were missing or spoiled; sample mean scores were assigned for these. On 32 occasions a subject failed to attempt at least half the items on a test; but no individual subject was responsible for more than two of these. 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 was number correct minus one-fourth the number incorrect. Two of the keylist tests were simply scored for number correct. 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.

For the remaining tests, scores were generated in several ways. In one, scoring credit was given only for answers deemed fully acceptable; in a second, the same credit was given to both fully and marginally acceptable answers; and in a third, marginal answers received half the credit given to fully acceptable ones. This third approach was found to yield slightly more reliable scores than either of the others and was therefore employed for all further analyses.

Test order. Possible differences among groups receiving the tests in different orders were examined in two ways. One analysis was concerned with the level of performance; another considered the standard error of measurement, a statistic that incorporates information about both the standard deviation and the reliability of a test score and that indicates the precision of measurement. In neither case were there systematic differences associated with the order in which the tests were administered. Order was therefore disregarded in all further analyses.

Test difficulty. Test means and standard deviations are shown in Table 1. Most of the tests were of middle difficulty for this sample; two of the keylist tests were easy, whereas multiple-choice Antonyms was very difficult. Means for the multiple-answer tests were 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% of the examinees attempt all items and if virtually everyone attempts at least three-fourths of the items. By these criteria only one of the tests, multiple-answer Analogies, had any problems with speededness: About 75% of the sample reached the last item, but 14% failed to attempt the 12th item, which represents the three-fourths point. For all the remaining tests, 95% 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 ranged from .45 to .80, with a median of .69. There were no differences in reliabilities associated with the response format of the test—the medians ranged from .68 for multiple-choice tests to .75 for multiple-answer forms. There were differences associated with item type; medians were .75 for Antonyms, .71 for Sentence Completions, and .58 for Analogies. The least reliable of all the tests was 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.

Relations among the Tests

Correlations among tests. Zero-order correlations among the 12 tests are shown in the upper part of 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. Correlations corrected for

Table 1
Descriptive Statistics for Tests

Test	Mean	S.D.	Maximum Possible Score	Percent Complet-ing	Relia-bility
Multiple-Choice					
Sentence Completion	9.65	4.29	20	99.7	.68
Analogies	8.60	3.35	20	98.7	.45
Antonyms	4.86	4.34	20	98.4	.69
Keylist					
Sentence Completion	15.96	2.33	20	96.2	.62
Analogies	15.02	3.03	20	99.0	.70
Antonyms	11.48	3.34	20	97.5	.73
Single-Answer					
Sentence Completion	7.99	3.30	20	90.5	.73
Analogies	7.27	2.74	20	84.4	.57
Antonyms	6.88	3.44	20	98.1	.75
Multiple-Answer					
Sentence Completion	24.02	7.11	54	90.8	.80
Analogies	13.31	5.23	45	74.6	.59
Antonyms	17.03	5.67	45	97.1	.75

Table 2
Zero-Order and Attenuated Correlations Among Tests[a]

Test		1	2	3	4	5	6	7	8	9	10	11	12
Multiple-Choice													
Sentence Comp.	1		52	54	56	59	66	64	47	69	65	36	58
Analogies	2	92		46	44	48	52	48	44	50	47	33	45
Antonyms	3	79	83		43	52	56	54	40	61	53	29	59
Keylist													
Sentence Comp.	4	86	83	66		55	54	58	43	53	52	38	53
Analogies	5	86	86	75	83		60	59	47	59	58	33	57
Antonyms	6	94	91	79	80	84		58	44	68	60	37	61
Single-Answer													
Sentence Comp.	7	91	84	76	85	83	79		52	64	67	39	61
Analogies	8	75	87	64	72	74	68	81		47	49	41	52
Antonyms	9	97	88	85	78	81	92	86	72		68	41	63
Multiple-Answer													
Sentence Comp.	10	88	80	71	74	78	79	88	73	88		50	62
Analogies	11	57	64	45	63	51	56	59	71	62	73		42
Antonyms	12	81	77	82	78	79	82	82	80	84	80	63	

[a] Decimal points omitted. Zero-order correlations are presented above the main diagonal, while correlations corrected for attenuation are presented below.

attenuation are shown in the lower part of the table; the correction is based on coefficient alpha reliabilities. The correlations range from .45 to .97 and have a median of .80.

These coefficients indicate that the various tests share a substantial part of their true variance, but they do not permit a conclusion as to whether there are systematic differences among them. Three analyses that address this question are presented below.

Factor analyses. A preliminary principal components analysis produced the set of eigenvalues displayed in Table 3. The first component was very large, accounting for 57% of the total variance, while the next largest accounted for only 7% 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.

Table 3
Principal Components of the Correlation Matrix

Component	Eigenvalue
I	6.80
II	.83
III	.63
IV	.59
V	.57
VI	.48
VII	.43
VIII	.42
IX	.37
X	.30
XI	.29
XII	.28

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 4. The two factors were highly correlated ($r = .67$). Ten of the 12 scores had their highest loading on Factor I, one (single-answer Analogies) divided about equally between the two, and only one (multiple-answer Analogies) had 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 Analogies tests is less clear. The second factor is small (just under 5% 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 only 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 ($p < .01$).

Factor 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 ($N = 141$). For each group a one-factor solution was indicated. 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

Table 4
Factor Pattern for Two-Factor Analysis

Test	Factor I	Factor II	Communality
Multiple-Choice			
Sentence Completion	.84	-.05	.65
Analogies	.59	.06	.40
Antonyms	.79	-.12	.51
Keylist			
Sentence Completion	.57	.15	.47
Analogies	.74	.01	.55
Antonyms	.86	-.10	.65
Single-Answer			
Sentence Completion	.68	.16	.63
Analogies	.35	.37	.44
Antonyms	.85	-.03	.69
Multiple-Answer			
Sentence Completion	.59	.29	.66
Analogies	.06	.63	.45
Antonyms	.65	.17	.60

a one- or a two-factor solution; the two-factor solution, however, showed all tests having their highest loading on the first factor and only multiple-answer Analogies approaching an equal division of its variance between the two factors.

Thus, there was no strong evidence for the existence of a second factor in the data. There were weak indications that the multiple-answer Analogies test and, to a much lesser extent, the single-answer Analogies test provided 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 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 5.

Each row in the upper part of the table provides the average of (1) all those correlations that represent relations for a single item type as measured in different formats and of (2) 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 was some variance associated with trait for both Sentence Completion and

Table 5

Multitrait-Multimethod Summary of Average Correlations

Trait or Method	Monotrait-Heteromethod	Heterotrait-Heteromethod
Trait		
Sentence Completion	.86	.80*
Analogies	.75	.75
Antonyms	.85	.79*
	Monomethod-Heterotrait	Heteromethod-Heterotrait
Method		
Multiple-Choice	.86	.81
Keylist	.82	.79
Single-Answer	.80	.80
Multiple-Answer	.73	.73

*By Mann-Whitney U Test, the two entries in a row are significantly different at the 5% level of confidence.

Antonyms item types (by Mann-Whitney *U* test, $p < .05$). Analogies tests did not, however, relate to one another any more strongly than they related 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 did not approach significance ($p > .10$). For the truly open-ended response formats, there were no differences whatsoever. Like the factor analyses, this approach to correlational comparisons showed no tendency for open-ended tests to cluster according to the response format; to the slight degree that any differences were found, they represented 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 could 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 were .97 for Sentence Completion, .88 for Analogies, and 1.05 for Antonyms. It appears that relations across the two kinds of formats did not differ from 1.0, except for error in the data, for two item types. Analogies tests may fail to share some of their reliable variance across open and closed formats but still appear to share most of it.

Relationships with Background Variables

Students completed a questionnaire dealing with their academic background, accomplishments, and interests. Included were questions concerning (1) plans for graduate school attendance and advanced degrees, (2) undergraduate grade-point average overall and in the major field of study, (3) preferred career activities, (4) self-assessed skills and competencies within the major field, and (5) 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; these were level of degree planned, self-reported grade-point average (both overall and for the major field of study), and the choice of writing as the individual's single most preferred professional activity. No systematic differences in correlations associated with item type or response format were evident.

Information was also available on the student's gender and year in school. No significant correlations with gender were obtained. Advanced students tended to obtain higher test scores, again with no evidence of differences among the tests in the magnitude of the relations.

GRE Aptitude Test scores were available for a small number of students ($N = 41$). 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, having medians of .36 and .47, respectively. Here also there were no systematic differences associated with item types or test formats.

These results, like the analyses of correlations among the experimental tests, suggest that response format has little effect on the nature of the attributes measured by the item types under examination.

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. These open-ended items, however, provide little new information. There was no evidence whatsoever for a general factor associated with the use of a free-response format. There was strong evidence against any difference in the abilities measured by Antonyms or Sentence Completion items as a function of the response format of the task. Only Analogies presented some ambiguity in interpretation, and there is some reason to suspect that that difference should be attributed to the slight speededness of the multiple-answer Analogies test employed.

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. (1981) suggests a number of possibilities. In the problem-solving work, subjects had to read and to 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 verbal aptitude items, in contrast, are much more self-contained. The examinee has only to deal with the meaning 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 solution is appropriate. Thus, the verbal tests might be described as "well-structured" and the problem-solving tests as "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 that is to be produced.

For those 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 cutting 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 the 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.

References

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, *56,* 81-105.

Donlon, T. F. *An exploratory study of the implications of test speededness.* (GRE Board Professional Report GREB No. 76-9P). Princeton NJ: Educational Testing Service, 1980.

Frederiksen, N., & Ward, W. C. Measures for the study of creativity in scientific problem-solving. *Applied Psychological Measurement,* 1978, *20,* 1-24.

Frederiksen, N., Ward, W. C., Case, S. M., Carlson, S. B., & Samph, T. *Development of methods for selection and evaluation in undergraduate medical education* (Final Report to the Robert Wood Johnson Foundation). Princeton NJ: Educational Testing Service, 1981.

Goldberg, L. P., & Werts, C. W. The reliability of clinicians' judgments: A multitrait-multimethod approach. *Journal of Counseling Psychology,* 1966, *30,* 199-206.

Heim, A. W., & Watts, K. P. An experiment on multiple-choice versus open-ended answering in a vocabulary test. *British Journal of Educational Psychology,* 1967, *37,* 339-346.

Jöreskog, K. G. A general method for analysis of covariance structures. *Biometrika,* 1970, *57,* 239-251.

Simon, H. A. The structure of ill-structured problems. *Artificial Intelligence,* 1973, *4,* 181-201.

Steel, R. G. D., & Torrie, J. H. *Principles and procedures of statistics.* New York: McGraw-Hill, 1960.

Traub, R. E., & Fisher, C. W. On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement,* 1977, *1,* 355-369.

Vernon, P. E. The determinants of reading comprehension. *Educational and Psychological Measurement,* 1962, *22,* 269-286.

Ward, W. C., Frederiksen, N., & Carlson, S. B. Construct validity of free-response and machine-scorable forms of a test. *Journal of Educational Measurement,* 1980, *17,* 11-29.

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. Ledyard Tucker provided extensive advice on the analysis and interpretation of results. This research was supported by a grant from the Graduate Record Examination Board.

Author's Address

Send requests for reprints or further information to William C. Ward, Senior Research Psychologist, Educational Testing Service, Princeton NJ 08541, U.S.A.