DOCUMENT RESUME

ED 279 709	TM 870 121

AUTHOR	Ward, William C.; And Others
TITLE	Keylist Items for the Measurement of Verbal Aptitude. Research Report.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-86-28
PUB DATE	Jun 86
NOTE	42p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Aptitude Tests; *Construct Validity; Correlation; Difficulty Level; Factor Analysis; Higher Education; *Multiple Choice Tests; Reading Comprehension; Scoring; Standardized Tests; *Test Construction; *Test Format; *Test Items; Test Reliability; Undergraduate Students; Verbal Ability
IDENTIFIERS	Analogies; Antonyms; *Keylist Tests; Reasoning Tests; Speededness (Tests)

ABSTRACT

The keylist format (rather than the conventional multiple-choice format) for item presentation provides a machine-scorable surrogate for a truly free-response test. In this format, the examinee is required to think of an answer, look it up in a long ordered list, and enter its number on an answer sheet. The introduction of keylist items into standardized tests could potentially offer several important benefits, among them the construction of items requiring production rather than simply recognition of correct answers, ease of item development, and resistance to coaching. A number of questions had to be answered before the keylist format could be considered for use in operational tests. This study addressed several of the most important of these in an examination of two item types employed in verbal aptitude tests--Antonyms and Analogies--and administered in both keylist and multiple-choice formats. These item types were selected for two reasons: (1) there is evidence that multiple-choice forms of these item types are susceptible to coaching; and (2) prior work has shown the feasibility of developing keylist versions of them. Relations among tests employing different response formats were analyzed and their correlations with other measures of aptitude and achievement were compared. The analyses indicated that the format has little or no systematic effect on the construct validity of tests employing item types used in standardized tests of verbal aptitude. There was, in addition, little agreement among experienced test developers on the set of keys that should be supplied for each keylist item. Appendices include instructions and sample items for experimental tests and additional aptitude tests. (JAZ)

RESEARCH REPORT

KEYLIST ITEMS FOR THE MEASUREMENT OF VERBAL APTITUDE

William C. Ward
Dan Dworkin
Sybil B. Carlson

ETS

Educational Testing Service
Princeton, New Jersey
June 1986

Keylist Items for the Measurement of Verbal Aptitude

William C. Ward
Dan Dworkin
Sybil B. Carlson

June, 1986

Acknowledgements

Kirsten Yocom conducted the data analysis. Ledyard Tucker provided advice on the analysis and interpretation of results.

Appreciation is owed to the Graduate Record Examinations Board and to the College Entrance Examination Board for permission to modify and use copyrighted test materials.

Abstract

Two verbal item types employed in standardized aptitude tests were administered in a conventional multiple-choice format and in the keylist format, in which the examinee is required to think of an answer, look it up in a long ordered list, and write its number. The keylist format provides a machine-scorable surrogate for a truly free-response test. Its potential attractions include the increased acceptability of items given in a production rather than a recognition format, resistance to coaching based on "gaming" strategies for eliminating multiple-choice alternatives, and elimination of the need in item writing to produce plausible distractors for an item.

Relations among tests employing different response formats were analyzed and their correlations with other measures of aptitude and achievement were compared. As in several previous studies, these analyses indicated that the format has little or no systematic effect on the construct validity of tests employing item types used in standardized tests of verbal aptitude.

One of the purposes of the study was to determine the degree to which experienced test developers could agree on the set of keys that should be supplied for each keylist item. Agreement among reviewers was far from the near-perfect consensus that would be required for use of this format, perhaps because the two item types investigated, Antonyms and Analogies, represent tests dealing with word meanings taken largely out of context. Many English words can convey multiple shades of meaning and can be constrasted along multiple dimensions. Without the constraint imposed by context, the number of possibly acceptable answers can become unmanageably great, particularly if it is required that all acceptable keys be included in the list and that all that are included must be clearly defensible. Several suggestions were offered of situations in which variations on this format could appropriately be employed.

Keylist Items for the Measurement of Verbal Aptitude

The keylist format for item presentation provides a machine-scorable surrogate for a truly free-response test. In this format, the examinee is required to think of an answer, look it up in a long ordered list, and enter its number on an answer sheet.

The introduction of keylist items into standardized tests could potentially offer several important benefits. The first is an increased acceptability to examinees and critics resulting from the use of items that require production rather than simply recognition of correct answers. Whether or not the change in format would result in changes in the construct measured by the items, disparagers of "multiple-guess" questions are unlikely to be satisfied by any amount of evidence for the validity of tests that rely exclusively on multiple-choice items.

A second potential benefit is that of resistance to coaching, to the degree that coaching concentrates on "gaming" strategies for eliminating alternatives so as to increase the probability of a correct guess. While a well-developed test should not be susceptible to such coaching, White and Zammarelli (1981) demonstrated the possible importance of these strategies. They developed formal rules that yielded nearly perfect performance on two commonly used figural reasoning tests, and showed that even untutored subjects were able to obtain better than chance results in choosing correct answers without exposure to the test questions.

A final benefit may be that of ease of item development. It is not necessary to spend the effort to produce plausible distractors when the keys for an item are embedded in a list comprised of keys for other items in the test. Offsetting this gain, however, is the need to include in the list all possible excellent answers for an item, a requirement that can easily be met for some item types but that would prove burdensome or worse for others.

A number of questions must be answered before the keylist format could be considered for use in operational tests. This study addressed several of the most important of these in an examination of two item types employed in verbal aptitude tests--Antonyms and Analogies. These item types were selected for two reasons: (1) because there is evidence that multiple-choice forms of these item types are susceptible to coaching (Alderman & Powers, 1980), and (2) because prior work has shown the feasibility of developing keylist versions of them (Ward, 1982).

One question concerns the comparability of psychometric characteristics of tests using the two formats. Earlier work was limited to comparisons of short tests based on different pools of items (Ward, 1982). The present study compared tests and individual items in a design in which equivalent groups of students completed tests in which the same item stems were presented in the two formats.

A second question concerns the similarity of what is measured by these item types when given in the keylist and multiple-choice formats. The earlier research suggested that they measure essentially the same aptitudes. That study, however, dealt only with relations among the tests, as indicated by correlational comparisons and factor analysis. The present study adds evidence concerning their construct similarity--the degree to which they have similar relations to measures of several additional aptitudes.

In addition, a small-scale assessment will be made of the degree to which experienced test development staff can agree on the appropriate keys for an item presented in the keylist format. Use of this format requires a procedure that will assure that all the best possible keys for an item are included on the list.

Method

Test Development

Items suitable for administration in both multiple-choice and keylist formats were drawn from disclosed forms of the GRE General Test and the Scholastic Aptitude Test. Some items were revised and additional items were written as needed; the multiple-choice versions of these were reviewed by experienced test development staff to assure that they were sound in content and that they conformed to ETS guidelines for style of presentation.

Multiple-choice Analogies items were prepared using the same format as that used in ETS testing programs: two terms were presented in the stem, and the examinee was required to identify the option that consisted of two terms embodying the same relation as that expressed in the stem. A more restrictive format was required for the keylist Analogies, so as to constrain the number of acceptable answers. These items were cast in a format in which three terms were given; the examinee was required to identify the appropriate fourth term to complete the analogy.

Many good multiple-choice items are not appropriate for use in keylist form. For example, an Antonyms stem for which a good key could be made simply by adding a negative prefix to the stem (CONCLUSIVE-INCONCLUSIVE) would be unacceptably easy. There are also problems with words that have multiple distinct meanings or that permit multiple dimensions of contrast; FAWN, for example, could be contrasted with IGNORE, an antonym for its meaning of "to show affection," or with DOMINEER, an antonym for its meaning of "to grovel." Such words are likely to have too many acceptable antonyms to be manageable.

Preparation of the list of acceptable keys for an item relied heavily on dictionary and thesaurus identifications of synonyms and antonyms, but was not straightforward. Often, for example, acceptable antonyms for a word were located by examining words identified as synonyms of, or as similar in meaning to, the one or two words that were given as antonyms. Many potential items had to be discarded after extensive study of their near-neighbors in meaning showed them to be unacceptably open in the space of possible answers.

Staff members experienced in the development of verbal aptitude tests offered their own keys for some of the keylist items as part of a small study of keying agreement that will be described in the section on results. Final decisions as to which stems and keys to include in the study, however, were made by the senior investigator and thus are subject to whatever limitations in comprehensiveness result from relying on one individual's judgments.

The final pool of items consisted of 72 Antonyms and 72 Analogies, each realized in both formats. Item stems of each type were randomly assigned to create two 36-item tests, each test to be administered as two separately timed 18-item sections. The keylist and multiple-choice versions of a test contained the same item stems in the same order.

For each keylist test section, a different keylist, consisting of words arranged in alphabetic order and numbered consecutively, was prepared. The lists contained an average of 4.1 acceptable answers for an item; the number of answers per item ranged from 1 to 8. Between 81 and 98 filler words were added to each list to bring it to the desired total of 165, approximately the maximum number of words that would fit comfortably on an 8 1/2 by 11 page printed from 12-pitch typewritten copy. Test booklets were prepared with the 18 items on one page and the corresponding keylist on a facing page, so that no page turning was needed to look up an answer. Appendix A provides instructions, a sample item, and a keylist for each item type, along with instructions and a sample item for the multiple-choice versions of the tests.

Three additional aptitude measures were prepared for use in the study. A test of Reading Comprehension, consisting of 25 questions based on four passages, was assembled using disclosed materials from GRE General Tests and the SAT. A Reasoning test was prepared, drawing on items from the Kit of Factor-Referenced Cognitive Tests (Ekstrom, French, Harman, & Dermen, 1976); it consisted of 15 Logical Diagrams items, which require that the examinee choose the Venn diagram that best illustrates the relationship between three given classes, and 20 Letter Sets, which require identification of the one of five sets of letters that does not fit the rule that describes the other four. The first is identified in the Kit as a test of Logical Reasoning, the second as one of Induction. Finally, a test of Divergent Thinking was constructed. It consisted of three Pattern Interpretations items (Ward, 1968), in which the examinee was to write possible interpretations of a simple abstract pattern, and of three Unexpected Results items (based on an unpublished variant by Ekstrom on the Guilford, 1959, Consequences Test), in which the examinee wrote possible consequences or results of an unlikely situation or event. Instructions and sample items are given in Appendix B.

A brief questionnaire dealing with the examinee's academic background and interests was also administered. Few correlations as high as .20 were found between test scores and variables derived from the questionnaire; therefore, no results involving the questionnaire are presented.

Test Administration

The design of the study is illustrated in Table 1. Each examinee completed two 18-item sections of each of four tests (Analogies and Antonyms, each presented in both the keylist and multiple-choice formats). Those in Group B received the same item stems in the same order as did those in Group A, but the latter received multiple-choice versions of the items which the former answered in keylist form, and conversely.

Students were tested in groups of 50 to 60. The two versions of the battery were spiraled so that half the students in each group responded to each version. A single session of approximately 3 1/2 hours was required for each group's testing.

Sample

Students were 286 paid volunteers from a single large state university. The sample was approximately evenly divided among freshmen, sophomores, juniors, and seniors. Nearly half the students indicated that they were majoring in the social sciences; the remainder were drawn, in decreasing order of numbers, from the natural sciences, biological sciences, and humanities. Sixty-six percent were female. Mean SAT scores, available on 76% of the sample, were 514 for V and 566 for M, substantially above the national average for college-bound high school seniors.

Scoring

All the tests were scored for number right, without correction for guessing. Examination of the Divergent Thinking test indicated that there were few instances of duplicate or inappropriate responses; the score for each item of this test was obtained simply by counting the number of answers without judgment as to their appropriateness.

The Reasoning and Divergent Thinking tests each comprised two different item types. Scores reflecting performance on each item type separately were showed no indication of differential relations with other variables. For simplicity of presentation, all results involving these tests are reported using total scores.

Table 1

Design of the Study

	Instrument	Group A	Group B	Time Limit (Minutes)
1.	Consent Form			--
2.	Antonyms	Multiple-Choice	Keylist	12
3.	Analogies	Keylist	Multiple-Choice	12
4.	Divergent Thinking			20
5.	Analogies	Multiple-Choice	Keylist	12
6.	Antonyms	Keylist	Multiple-Choice	12
	(Break)			(10)
7.	Reasoning			20
8.	Analogies	Multiple-Choice	Keylist	12
9.	Antonyms	Keylist	Multiple-Choice	12
10.	Reading Comprehension			25
11.	Antonyms	Multiple-Choice	Keylist	12
12.	Analogies	Keylist	Multiple-Choice	12
13.	Questionnaire			--

Results

Preliminary Results

Test speededness. None of the multiple-choice tests was speeded by ETS standards; 95% to 100% of the sample attempted the last item of each of the test sections. Four keylist test sections showed some indication of speededness, with between 27% and 56% not attempting the last item. However, at least 97% of the sample attempted the 14th item, representing the three-quarters point, on all but one of the test sections. The latter was an Analogies keylist test section on which 89% attempted at least the 14th item. Thus none of the tests was seriously speeded.

Ten students in Group A and 15 in Group B failed to answer one-half the questions on two or more keylist sections. Their data were excluded from all analyses.

Descriptive statistics. Descriptive statistics for all tests administered are presented in Tables 2 and 3. Reliabilities for the total scores on the Antonyms and Analogies tests are based on test-retest correlations across the two sections of each test; all other reliabilities reported are coefficient alpha.

The two groups were very similar in both the level and the reliability of their scores on all the tests given. For both groups, three of the four experimental tests were moderately difficult, with scores averaging between 54% and 64% of the maximum possible. The Antonyms keylist tests were more difficult, with averages in groups A and B, respectively, of 36% and 40% of the maximum possible score. By the *t*-test for correlated means, however, most within-group comparisons of levels of performance across these tests failed to reach statistical significance. Reliabilities of the full-length experimental tests ranged from .62 to .84 with a median of .77; there were no systematic differences associated with test format (medians of .76 for multiple-choice tests and of .78 for keylist tests), and only suggestive differences associated with item type (medians of .73 for Analogies and .81 for Antonyms).

Correlations among experimental test scores

Zero-order correlations among scores derived from the experimental tests are shown in Table 4. Correlations for Group A are shown above the main diagonal, while those for Group B are below. The coefficients range from .57 to .79 with a median of .64, and do not differ systematically by group, item type, or response format.

Table 2

Descriptive Statistics for Group A

Test	Mean	S.D.	N	Reliability
Analogies - Multiple-Choice				
Section 1	11.00	3.23	133	.68
Section 2	11.14	2.87	133	.61
Total	22.14	5.43	133	.74
Analogies - Keylist				
Section 1	9.53	2.94	133	.68
Section 2	10.53	2.82	133	.63
Total	20.06	5.11	133	.73
Antonyms - Multiple-Choice				
Section 1	10.67	2.55	133	.53
Section 2	10.26	3.60	133	.75
Total	20.92	5.59	133	.78
Antonyms - Keylist				
Section 1	6.36	2.70	133	.63
Section 2	6.63	2.55	133	.64
Total	12.99	4.89	133	.84
Reading Comprehension	15.34	4.31	133	.74
Reasoning	26.26	4.70	133	.78
Divergent Thinking	32.81	8.42	133	.73
SAT - V	511.94	84.50	98	--
SAT - M	560.71	82.25	98	--

Table 3

Descriptive Statistics for Group B

Test	Mean	S.D.	N	Reliability
Analogies - Multiple-Choice				
Section 1	9.66	2.63	128	.55
Section 2	11.12	3.04	128	.77
Total	20.78	4.84	128	.62
Analogies - Keylist				
Section 1	10.32	3.59	128	.75
Section 2	9.26	3.85	128	.79
Total	19.58	6.69	128	.76
Antonyms - Multiple-Choice				
Section 1	10.56	3.18	128	.70
Section 2	10.27	2.91	128	.65
Total	20.84	5.59	128	.81
Antonyms - Keylist				
Section 1	7.77	3.34	128	.75
Section 2	6.55	3.31	128	.76
Total	14.31	6.10	128	.81
Reading Comprehension	15.38	4.05	128	.71
Reasoning	26.11	4.99	128	.81
Divergent Thinking	32.36	8.37	128	.77
SAT - V	515.80	83.44	100	--
SAT - M	570.10	84.85	100	--

Table 4

Zero-Order Correlations Among Experimental Test Scores

	Analogies		Antonyms	
	Multiple-Choice	Keylist	Multiple-Choice	Keylist
Analogies				
Multiple-Choice		.67	.63	.62
Keylist	.66		.57	.63
Antonyms				
Multiple-Choice	.62	.68		.68
Keylist	.63	.71	.79	

Correlations for Group A are presented above the main diagonal, while those for Group B are presented below.

Correlations corrected for attenuation are shown in Table 5. The correction is based on test-retest correlations across the two parts of each test. The corrected coefficients range from .76 to .98, with a median of .87; those for Group B tend to be somewhat greater than those for Group A.

These correlations can be examined within the framework provided by multitrait-multimethod analysis (Campbell & Fiske, 1959). Each item type constitutes a "trait," while each format for item presentation constitutes a "method." The data are presented in Table 6 following a scheme suggested by Goldberg & Werts (1966).

Each row in the upper part of the table provides a comparison of (1) the average correlation between tests employing the same item type but using different response formats and (2) the average correlation between tests that differ in both item type and format. Averages were obtained using Fisher's r to z transformation. While an appropriate test of statistical significance is not available, it appears that each of the two item types has some variance that is not shared with the other, and that the true relations across formats within an item type are nearly perfect.

The lower part of the table compares (1) the average correlation between tests employing the same response format but differing in item type and (2) the average correlation between tests differing in both item type and format. Here there is little difference between the two columns, suggesting that there is little or no distinct variance associated with the "method" (format) in which a test is presented.

Factor analysis

Another approach to the examination of relations of performance on different item types and formats is through factor analysis. For each group, a matrix of eight scores was analyzed--two item types times two response formats times two sections of each test. The analysis was a principal axes factor analysis, using Tucker's adjusted highest off-diagonal element without iteration as the communality estimate; the factor matrix was rotated to an oblimin (oblique) solution.

In the data for Group A, the first three factors accounted for 86.5%, 7.9%, and 5.5% of the common variance. The two-factor solution divided the tests by item type. The three-factor solution, shown in Table 7, further divided Antonyms tests between two factors, one representing the multiple-choice format and one representing the keylist format. Correlations among the factors ranged from .67 to .74.

For Group B, the first factor accounted for 96% of the common variance. As shown in Table 7, a meaningful second factor could not be extracted.

Table 5

True Score Correlations Among Experimental Test Scores

	Analogies		Antonyms	
	Multiple-Choice	Keylist	Multiple-Choice	Keylist
Analogies				
Multiple-Choice		.91	.83	.79
Keylist	.96		.76	.80
Antonyms				
Multiple-Choice	.87	.87		.84
Keylist	.89	.90	.98	

Correlations for Group A are presented above the main diagonal, while those for Group B are presented below.

Table 6

Multitrait-Multimethod Summary of Average Correlations

Trait or Method	Monotrait-Heteromethod	Heterotrait-Heteromethod
Trait		
Analogies	.94	.84
Antonyms	.95	.84
	Monomethod-Heterotrait	Heteromethod-Heterotrait
Method		
Multiple-Choice	.85	.84
Keylist	.86	.84

Table 7

Oblimin Factor Pattern for Experimental Tests

	Group A - Loadings			Group B- Loadings	
	I	II	III	I	II
Analogies - Multiple-Choice					
Section 1	.55	.31	-.09	.61	.09
Section 2	.80	.02	-.02	.50	.36
Analogies - Keylist					
Section 1	.68	-.03	.10	.64	.24
Section 2	.68	-.04	.12	.63	.27
Antonyms - Multiple-Choice					
Section 1	-.02	.64	.13	.91	-.21
Section 2	.13	.75	.07	.79	-.03
Antonyms - Keylist					
Section 1	-.00	.17	.72	.77	-.05
Section 2	.15	.00	.76	.93	-.12

No explanation for the difference between the two groups is available. These results are, however, consistent with those from the multitrait-multimethod analysis in that most of the common variance in the set of tests is shared across item types and formats, and there is very little systematic variance associated with the response format.

Correlations with other variables

A third approach to the comparison of response formats is to examine correlations of the experimental tests with other measures of aptitude and achievement. Correlations are shown in Table 8. The experimental test scores showed substantial correlations with the SAT-V and the test of Reading Comprehension; with one exception, they showed moderate relations with Reasoning and with SAT-M; and all had near-zero relations with Divergent Thinking. There is no evidence of differential relations to these measures either for corresponding scores based on different response formats or for scores based on different item types within a response format.

Difficulties of individual items

The results discussed thus far deal primarily with intact tests rather than the individual items of which they are composed. Analyses were also performed to compare the difficulties of individual items across the two response formats; for each item, the proportion of examinees answering the item correctly in each format was contrasted by *t*-test.

Overall, Antonyms items were consistently more difficult in the keylist format than when given as multiple-choice items. Fifty-six Antonyms items were significantly more difficult in keylist format, while only two were significantly more difficult in their multiple-choice version. The same tendency was found for Analogies items but was less pronounced; 25 items were significantly more difficult in keylist form, 13 when given as multiple-choice items.

The difference between the two item types is understandable. Keylist Antonyms items provide no cues to guide the examinee's effort to produce an answer. Analogies items, however, provide three of the four terms that constitute the completed item, thus making it possible to rule out some rationales that an examinee might otherwise entertain for the relation between the first two terms. In addition, an examinee may be able to answer some items without correctly identifying the relation between the given terms, relying instead on a weaker relationship such as "A is associated with B" and searching for a word that is, in some poorly defined way, associated with the third term given. A similar strategy could not be employed in dealing with the multiple-choice versions of these items, since all the options offered are likely to conform to such a generic relation.

Table 8

Correlations of Experimental Test Scores with Cognitive Variables

Score	Cognitive Variable				
	Reading Comprehension	Reasoning	Divergent Thinking	SAT-V	SAT-M
Group A					
Analogies					
Multiple-Choice	.51	.47	.09	.74	.27
Keylist	.51	.41	.12	.73	.34
Antonyms					
Multiple-Choice	.37	.17	.17	.70	.15
Keylist	.50	.35	.18	.70	.37
Group B					
Analogies					
Multiple-Choice	.52	.45	-.09	.58	.33
Keylist	.60	.34	-.01	.64	.33
Antonyms					
Multiple-Choice	.56	.34	.06	.75	.33
Keylist	.58	.41	-.01	.71	.36

Items that were significant.y easier in keylist form were examined to determine whether some obvious characteristic distinguished them from the remainder. The two Antonyms items have two characteristics that might be important: First, the multiple-choice form of the item has one distractor that, while unambiguously incorrect, was chosen by a large percent of the examinees. Second, the key for the multiple-choice version of the item was not the answer given by most examinees who answererd the item correctly in keylist form; it was chosen by 12% of those examinees on one item and by none on the other.

Analogies items that were significantly easier in keylist form were not distinguished from other Analogies items in the number of keys offered in the keylist form; the mean number was 3.8, as compared with 4.1 for all items, and the range was from one to seven keys. They were not extreme in difficulty level, averaging 67% correct in keylist form and 52% as multiple-choice items, compared with an overall mean percent correct of 56% for all Analogies keylist items and 62% for all Analogies multiple-choice items. Their biserial correlations were also not extreme, having an average of .58 for the keylist items and .47 for multiple-choice, compared with .58 for all Analogies keylist items and .52 for all Analogies multiple-choice items. About half of these items had one strong distractor in their multiple-choice form, but the remainder did not.

There were, however, two suggestive differences. First, there were nine Analogies questions for which only one response was correct in the keylist form. Four of these were items that were significantly easier in the keylist form. Thus, 31% of items significantly easier in this form, but no items significantly more difficult in this form, had only one correct answer. It may be that, when there is only one strong answer available for an item, the distractors presented in the multiple-choice version attract some examinees who would have been able to generate the answer themselves if not distracted.

Second, all those items that had more than one possible key in the keylist version were examined to determine whether the second term of the key used in the multiple-choice version was the answer given by most examinees who answered the item correctly in the keylist form. For the nine items that were significantly easier in keylist form, the multiple-choice key was the most popular correct keylist answer 22% of the time; while for the 25 items that were significantly more difficult in the keylist form, the multiple-choice key was the most popular correct answer 68% of the time. Here, it may be that when the key to a multiple-choice item involves vocabulary that a knowledgeable examinee would have been likely to use spontaneously, the multiple-choice item is easier because it requires recognition rather than production of a relationship. When the multiple-choice item involves vocabulary that an examinee would not spontaneously use, however, unknown or uncommon vocabulary is a source of difficulty for some examinees who recognize the relationship represented in the item and who would have been able to generate appropriate answers using different words.

It goes without saying that these are ad hoc speculations; design of sets of items controlling properties that might affect the relative difficulty of the two formats would be necessary to demonstrate that these or other factors have reliable effects.

Study of keying agreement

A small study was carried out during the test development phase of the investigation to determine the degree to which expert test developers would agree on the appropriate keys for an item. Eighty Antonyms and 80 Analogies items, believed to be appropriate for administration in both response formats, were tentatively selected to make up the experimental tests. Twenty-five item stems from each set were chosen randomly and submitted to four test developers for independent keying. The task was defined as that of producing all possible excellent keys for these items. The list was to contain only single-word answers, not multiple-word descriptions. If a stem could reasonably be taken as of either of two parts of speech, all keys appropriate for both interpretations were to be included. It was suggested that between one and six to eight keys might be appropriate for each item, but no limit was imposed on the number of keys that could be given.

The result of this exercise was a clear demonstration of the richness of the English language. For Antonyms items, four experts offered an average of 13.5 keys per item, with a range from as few as four to as many as 23. On average, 4.4 keys per item were offered by two or more individuals, while the remaining 9.1 were idiosyncratic. An example of a reasonably typical set of results would be the keys offered as antonyms of EUPHONIOUS: CACOPHONOUS (suggested by four reviewers), DISCORDANT (4), DISSONANT (3), INHARMONIOUS (2), and the following suggested by one reviewer each: GRATING, HARSH, HARSH-SOUNDING, JARRING, RASPY, RAUCOUS, STRIDENT, UNHARMONIOUS, and UNMELODIC.

The Analogies keying had a similar outcome. A mean of 15.5 keys was offered per item, with a range from three to 36. On average, 4.8 keys per item were offered by at least two reviewers, while 10.7 were given by only one.

Differences among individuals in their interpretation of the task were evident. Some limited their lists to words that they believed to provide excellent keys, while others explicitly included all words that might reasonably be considered. The extremes are illustrated by individuals' lists of Antonyms keys; one reviewer provided a total of 66 keys, or 2.6 per item, while another provided 201, or 8.0 per item. The range was even greater for Analogies, where one reviewer provided an average of 2.3 and another provided 9.4 keys per item.

One outcome of this exercise was that the supposedly near-final set of items for inclusion in the main test administration was revised extensively in an attempt to tighten the rationales for Analogies items and to limit items of both types to those with small numbers of clearly good keys. A second was a decision to attempt, on a very limited scale, to determine whether a larger group of reviewers could be induced to produce better agreement. Here, 14 test development staff were given two Antonyms and two Analogies items drawn from those reviewed in the previous stage, with a list for each item of all the keys that had been proposed by one or more reviewers. They were asked to mark each word they considered to be an excellent key for the item, adding new keys only "if you must," and to spend no more than two or three minutes on each item.

Twelve of the 14 were willing to accept the lists they were given, involving between 11 and 17 potential keys per item; one proposed two additional keys and one proposed 13. The possible keys that were checked by the majority of this group did tend to agree with those that were most popular in the first review; for example, at least 11 of 14 respondents marked as acceptable antonyms for EUPHONIOUS the three choices that were most often offered by the initial four reviewers. However, all 13 alternatives listed for this item were judged acceptable by at least two of those undertaking the task.

It appears to be possible to elicit good but not perfect agreement on a moderate number of answers that are acceptable; it may not, however, be possible to obtain good agreement that a large number of alternative answers can be excluded.

Moreover, there is some evidence for consistent individual differences in the number of alternatives judged acceptable. One indication of this is given by counting the number of keys each individual offered over the two Antonyms items and the number offered over the two Analogies items. The correlation between these totals was .67. Further, when the results are arranged in the matrix shown in Table 9, it appears that the pattern of endorsements of alternatives resembles a Guttman scale: Individuals who accept few alternatives tend largely to accept only those that are very popular; as the number accepted increases, it does so by progressively including endorsements of less and less popular choices.

Table 9

Endorsements of Potential Antonyms for Euphonious

Alternatives in Decreasing Order of Number of Endorsements Received	Respondents in Decreasing Order of Number of Alternatives Endorsed													
Dissonant	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Discordant	1	1	1	1	1	1	1	1	1	1	1	1		1
Cacophonous	1	1	1	1	1	1	1		1	1	1	1		
Jarring	1	1	1	1		1	1	1						
Unharmonious		1	1	1		1		1		1	1			
Unmelodious	1	1	1	1		1	1							
Inharmonious	1	1		1		1	1							
Strident	1		1	1	1				1					
Raucous	1	1		1	1									
Harsh-sounding	1	1	1										1	
Harsh	1		1		1									
Grating	1	1			1									
Raspy	1	1												

Note: A "1" indicates that the respondent accepted the word shown as an antonym for Euphonious, while a blank indicates that the word was not accepted.

Discussion

Three sources of evidence support the conclusion that there was little or no systematic difference between the keylist and multiple-choice formats in the aptitudes that contribute to test performance. First, a multitrait-multimethod analysis of correlations corrected for test reliability revealed no variance associated with format. Second, factor analysis showed for one group of examinees only a weak format factor, specific to one item type and accounting for less than 6% of the common variance, whereas for the second group no format factor could be identified. Finally, correlations with additional aptitude and achievement measures showed a very similar pattern of relationships for corresponding tests using the two formats.

These results are consistent with those obtained by Ward (1982), in comparing Antonyms, Analogies, and Sentence Completion items given in four formats, including the present two and two that were truly open-ended. Comparable results were also found in a study (Ward, Dupree, & Carlson, 1986) that contrasted free-response and multiple-choice versions of Reading Comprehension items drawn from standardized admissions tests. Taken together, these studies indicate strongly that the use of a multiple-choice format has little consequence in terms of the constructs underlying performance on the kinds of items that are typically employed in standardized tests of verbal aptitude.

It is also clear that the keylist format employed in the present study, as well as the more open formats employed in the two earlier investigations, can yield tests with reasonable psychometric properties. Reliabilities of tests using the keylist and multiple-choice formats were very similar, as were their correlations with the SAT-V and with a conventional measure of Reading Comprehension.

Factors affecting the relative difficulty of comparable items given in the two formats are undoubtedly complex. Some speculations about these were offered, but studies designed explicitly to isolate the processes examinees employ in solving the items and the reasons for failure would be required to permit confident conclusions.

From a test development perspective, the limited evidence available suggests that the keylist format as employed here is unlikely to warrant serious consideration for introduction into standardized tests of verbal aptitude. The difficulty of producing items with a sufficiently constrained set of acceptable keys, and the inability to obtain even an approximation to perfect consensus on keys among experienced test developers, both diminish the possibility.

However, there may be instances in which versions of the keylist format do merit consideration. The present study employed item types that deal with word meanings largely without context; and in that situation the multiplicity of parts of speech, shades of meaning, and dimensions of contrast to which English words are susceptible was an abundantly evident source of difficulty in producing exhaustive lists of keys. Item types in which there is sufficient context to constrain the range of possibilities more tightly might prove more amenable to use in this format.

Moreover, a fundamental problem with the format, in the view of a number of those who reviewed the current test materials, is that it would only be acceptable if they could be confident that all acceptable keys had been identified. In this view, an examinee who thought of an acceptable response but did not find the word in the keylist would be unfairly penalized, even if consensus could be reached that, say, the six best possible alternatives were present on the list.

This conclusion may be an appropriate one if the format is to be presented as a free-response one, using keylists too long to make recognition of a match between stem and option an effective approach to solving an item. A variant of the format using shorter lists, however, could be employed, with explicit instructions that the task is to identify the best available match in the list rather than to generate an answer and then locate it in the list. This format has proven effective in classroom testing (Carlson, 1985), and offers two of the potential benefits over the standard multiple-choice format that were proposed in introducing this report: freedom from coachability based on "gaming" approaches to the elimination of alternatives [illegible] reduction in the need to write plausible distractors for a stem. [illegible] arise from the use of keys to other items in a set as the alternatives within which to embed the key to a given item.

References

Alderman, D. L., & Powers, D. E. (1980). The effects of special preparation on SAT-Verbal scores. American Educational Research Journal, 17, 239-251.

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

Carlson, S. B. (1985). Creative classroom testing. Princeton, NJ: Educational Testing Service.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). Kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service.

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

Guilford, J. P. (1959). Personality. New York: McGraw-Hill.

Ward, W. C. (1968). Creativity in young children. Child Development, 39, 737-754.

Ward, W. C. (1982). A Comparison of free-response and multiple-choice forms of verbal aptitude tests. Applied Psychological Measurement, 6, 1-11.

Ward, W. C., Dupree, D., & Carlson, S. B. (1986). A comparison of free-response and multiple-choice questions in the assessment of reading comprehension. Princeton, NJ: Educational Testing Service.

White, A. P., & Zammarelli, J. E. (1981). Convergence principles: Information in the answer sets of some multiple-choice intelligence tests. Applied Psychological Measurement, 5, 21-27.

Appendix A

Instructions and Sample Items for Experimental Tests

Analogies

Multiple Choice Form

Time - 12 minutes
18 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

   D

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

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

Begin work.

Analogies

Keylist Form

Time - 12 minutes
18 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 22

A sermon is a religious lecture. A sacrament is a religious ceremony. The word ceremony is number 22 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 121 (rite).

Turn to the next page and begin work.

Analogies Keylist

1. absorbent
2. abuse
3. accompany
4. anger
5. applause
6. arbitration
7. arena
8. attitude
9. automation
10. banjo
11. bark
12. belief
13. bench
14. beverage
15. blame
16. bondage
17. boredom
18. break
19. calculation
20. cease
21. censure
22. ceremony
23. chaos
24. classification
25. clay
26. clean
27. clear
28. concern
29. conclusion
30. condemnation
31. courage
32. courtroom
33. cowardice
34. crime
35. criticism
36. crushed
37. cunning
38. decease
39. decorum
40. delay
41. designation
42. desire
43. devious
44. diabolical
45. dialect
46. dialogue
47. die
48. direction
49. disapproval
50. diva
51. dividend
52. drudgery
53. duress
54. dusty
55. education
56. educe
57. emotion
58. energy
59. evolve
60. expire
61. exploit
62. fear
63. feeling
64. fiddle
65. figure
66. fine
67. fluctuate
68. glowing
69. ground
70. health
71. hot
72. identification
73. imagination
74. include
75. insidious
76. insult
77. integer
78. interest
79. interlude
80. intermission
81. interval
82. intonation
83. involvement
84. judgment
85. kind
86. language
87. leisure
88. libretto
89. lull
90. maestro
91. malfeasance
92. misbehavior
93. misconduct
94. monotony
95. name
96. nomenclature
97. number
98. opinion
99. outlook
100. passenger
101. pause
102. perish
103. petition
104. pickle
105. pitiful
106. population
107. powdery
108. profanity
109. prologue
110. quality
111. quantity
112. rash
113. rebuff
114. rebuttal
115. regularity
116. rehabilitate
117. rehearsal
118. rendition
119. response
120. responsibility
121. rite
122. routine
123. rule
124. scheming
125. science
126. script
127. seed
128. sin
129. sly
130. small
131. sociology
132. solution
133. speech
134. spirit
135. stick
136. submissiveness
137. succumb
138. tedium
139. tempered
140. terminology
141. text
142. theology
143. time
144. timidity
145. tissue
146. tribunal
147. tumult
148. university
149. valuable
150. vegetable
151. verdict
152. vice
153. viewpoint
154. viola
155. violin
156. vocabulary
157. warfare
158. warming
159. water
160. weakness
161. weight
162. whistle
163. whole
164. witty
165. wrongdoing

Antonyms

Multiple Choice Form

Time - 12 minutes
18 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.

Begin work.

Antonyms

Keylist Form

Time - 12 minutes
18 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 130

Deplorable means wretched or lamentable. A good antonym is praiseworthy. The word praiseworthy is number 130 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 108 (laudable) or number 23 (commendable).

Turn to the next page and begin work.

Antonyms Keylist

1. abnormality
2. abundant
3. accomplishment
4. activate
5. aggravate
6. alien
7. alienation
8. altercate
9. amiable
10. amicable
11. amputate
12. anesthetic
13. anonymity
14. available
15. awkward
16. beneficial
17. breach
18. calm
19. center
20. challenging
21. clarify
22. close
23. commendable
24. competent
25. complacent
26. completion
27. concealment
28. congenial
29. contend
30. convert
31. cordial
32. core
33. covering
34. damage
35. debase
36. deceit
37. deliberateness
38. depletion
39. depress
40. describe
41. deserted
42. deteriorate
43. diligence
44. direct
45. disaffection
46. disagree
47. discontinuous
48. disequilibrium
49. displace
50. dispute
51. disseminate
52. dissent
53. distance
54. distinctive
55. distort
56. divert
57. dog-eared
58. eager
59. early
60. easygoing
61. effective
62. elucidate
63. emotion
64. empty
65. encourage
66. enhance
67. enlighten
68. estrangement
69. even
70. exactness
71. exhausted
72. exotic
73. expand
74. expedite
75. explicitness
76. farness
77. flaw
78. foreign
79. forget
80. formal
81. fresh
82. friendly
83. grim
84. growth
85. harm
86. hasten
87. heart
88. hesitate
89. homely
90. hurry
91. identify
92. illuminate
93. imbalance
94. impair
95. improvidence
96. incongruity
97. infertile
98. infirmity
99. intensify
100. interior
101. intermittent
102. interrupted
103. invidious
104. involuntary
105. irregular
106. judge
107. lateness
108. laudable
109. learning
110. lily-livered
111. linear
112. lively
113. lower
114. maintain
115. malevolent
116. malleable
117. mild
118. minimize
119. naturalized
120. noise
121. noncommittal
122. noticeable
123. objective
124. oppose
125. patient
126. periodic
127. petulant
128. placid
129. pliable
130. praiseworthy
131. refreshed
132. refuse
133. reliable
134. remoteness
135. renewed
136. retrieve
137. rigid
138. secure
139. serene
140. shadow
141. similar
142. sink
143. smooth
144. speak
145. speed
146. stern
147. stiff
148. straight
149. straightforward
150. taut
151. tight
152. toughness
153. tranquil
154. trite
155. trough
156. unconcern
157. undesirability
158. unproductive
159. unquestioning
160. urban
161. vitalized
162. voluptuous
163. vulgar
164. watchful
165. worsen

Appendix B

Instructions and Sample Items for Additional Aptitude Tests

READING COMPREHENSION

Directions: This test consists of four reading passages, each followed by questions based on its content. After reading a passage, choose the best answer to each question on the basis of what is stated or implied in the passage. Mark your answer by writing its letter in the space provided.

There are 25 questions to be answered in 25 minutes.

TURN TO THE NEXT PAGE AND BEGIN WORK

REASONING

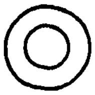
This test consists of two different kinds of questions that measure skill in logical reasoning. For each kind of question there is a page of instructions before the test items are presented. You will have 20 minutes to complete the entire test. Plan to spend about half your time on each kind of question.

TURN TO THE NEXT PAGE AND BEGIN WORK

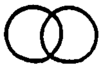
LOGICAL DIAGRAMS

For these questions you are to choose from five diagrams the one that illustrates the relationship among three given classes better than any of the other diagrams offered.

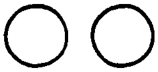
There are three possible relationships between any two different classes:



indicates that one class is completely contained in the other, but not vice versa.



indicates that neither class is completely contained in the other, but the two do have members in common.



indicates that there are no members in common.

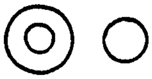
Note: The size of the circles does not indicate relative size of the classes.

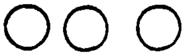
Example:

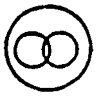
Sample Answer

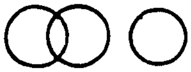
Birds, pets, trees

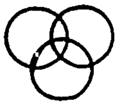
D

(A)


(B)


(C)


(D)


(E)


The correct answer, (D), shows that one of the classes (trees) has no members in common with the other two. (No trees are either birds or pets, and no birds or pets are trees). (D) also shows that the other two classes have some members in common, but neither is completely included in the other (some birds are pets and some pets are birds, but there are birds that are not pets and there are pets that are not birds).

On the page of test questions, the five possible choices for all the questions are given at the top of the page.

GO ON TO THE NEXT PAGE

LETTER SETS

Each question consists of five sets of letters. There is a rule that makés four of the sets of letters alike in some way. You are to find the set that is different and does not fit the rule. Mark the answer space to indicate which set is different.

Examples:	(A)	(B)	(C)	(D)	(E)	ANSWER
1.	NOPQ	DEFL	ABCD	HIJK	UVWX	B
2.	NLIK	PLIK	QLIK	THIK	VLIK	D

In Example 1, four of the sets have letters in alphabetical order. The second set, DEFL, does not fit this rule, so a B has been entered in the answer space. In Example 2, four of the sets contain the letter L. The fourth set, THIK, does not fit the rule, so a D has been entered in the answer space.

Note that the rules are not based on the sounds of sets of letters, the shapes of letters, or whether letter combinations form words or parts of words.

GO ON TO THE NEXT PAGE

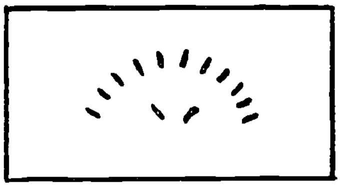
DIVERGENT THINKING

This test consists of two different kinds of questions that measure skill in divergent thinking. For each kind of question there is a page of instructions before the test items are presented. You will have 20 minutes to complete the entire test. Plan to spend about half your time on each kind of question.

TURN TO THE NEXT PAGE AND BEGIN WORK

PATTERN INTERPRETATIONS

In these problems you are to think of possible interpretations for simple abstract patterns. Here is an example:



Possible interpretations:

porcupine

rising sun

man hearing a hair-raising story

explosion

track left by a bear with mutated fo

Write down all the different things you can think of that each complete pattern might be.

GO ON TO THE NEXT PAGE

UNEXPECTED RESULTS

In these problems you are given an unlikely situation or event, and are asked to think of its possible consequences or results. Write as many different results as you can. Try to think of results other than the obvious or expected ones.

Example:

What would happen if one year no birds flew south for the winter?

many birds would die of cold or starvation
sales of birds feed would increase
many cats would get fat

Write down all the different consequences you can think of.

GO ON TO THE NEXT PAGE