A BROAD-RANGE TAILORED TEST OF VERBAL ABILITY

FREDERIC M. LORD
Educational Testing Service

This report describes briefly a broad-range tailored test of verbal ability, appropriate at any level from fifth grade upwards, through graduate school. The test score places everyone at all levels directly on the same score scale.

In a tailored test, the items administered to an individual are chosen for their effectiveness for measuring him. Items administered later in the test are selected by computer, according to some rule based on the individual's performance on the items administered to him earlier. Improved measurement is obtained 1) by matching item difficulty to the ability level of the individual and 2) by using the more discriminating items in the available item pool. The matching of test difficulty to the individual's ability level is advantageous and desirable for psychological reasons. For references on tailored testing, see Wood (1973). Also Cliff (1975), Jensema (1974a, 1974b), Killcross (1974), Mussio (1973), Spineti and Hambleton (1975), Urry (1974a, 1974b), Waters (1974), Betz and Weiss (1974), DeWitt and Weiss (1974), Larkin and Weiss (1974), McBride and Weiss (1974), Weiss (1973, 1974), Weiss and Betz (1973).

The broad-range test consists of 182 verbal items. These were chosen from all levels of Cooperative Tests' SCAT and STEP, from the College Entrance Examination Board's Preliminary Scholastic Aptitude Test, and from the Graduate Record Examination. The choice was made solely on the basis of item type and difficulty level. There was no attempt to secure the best items by selecting on item discriminating power.

Two parallel forms of this 182-item tailored test were constructed. Only one of these forms is considered here.

Ideally there should be only one item type in each row, so that all examinees would take the same number of items of each type. The arrangement of Table 1 is an attempt to approximate this ideal using the items available. (Few if any hard items of types a and e were in the total pool; also few if any easy items of types b and c. Types a and b, also types c and e, seem fairly similar.)

TABLE 1

Broad-Range Verbal Test Items Arranged by Difficulty Level and Serial Number.
(a, b, c, d, e represent different verbal item types.)

Item Serial No.	(easy)←			Item Difficulty Level						→(hard)
	Grade Level: IV	V	VI VII	VIII		XII				
1			a	a	a	a	a	b		
2			e	e	e	e	c			
3			d	d	d	d	d	d		
4			e	e	e	e	c	c		
5			d	d	d	d	d	d		
6			a	a	a	a	b	b		
7			e	e	e	e	c			
8		d	d	d	d	d	d			
9			e	e	e	c	c	c		
10		d	d	d	d	d	d			
11			a	a	a	a	b	b	b	b
12		e	e	e	c	c	c	c		
13			d	d	d	d	d	d		
14		e	e	e	c	c	c	c	c	
15			d	d	d	d	d	d	d	
16			a	a	a	b	b	b	b	b
17		e	e	c	c	c	c	c	c	
18	d	d	d	d	d	d	d	d		
19		e	e	c	c	c	c	c	c	
20	d	d	d	d	d	d	d	d	d	d
21		a		a	a	b	b	b	b	b
22	e	e	c	c	c	c	c	c	c	c
23		d	d	d	d	d	d	d	d	d
24	e	e	c	c	c	c	c	c	c	c
25		d	d	d	d	d	d	d	d	d

The 182 items in a single form of the test are represented in Table 1, where they are arranged in columns by difficulty level. An individual answers just one item in each row of the table—a total of just 25 items. There are five verbal item types, denoted by a, b, c, d, e. Within each item type, the items in each column are arranged in order of discriminating power with the best items at the top.

The examinee starts with an item in the first row. The difficulty level of this item is determined by the examinee's grade level, or some other rough estimate of his ability. If he answers the first item correctly, he next takes an item in the second row that is harder than (to the right of) the first item. If he answers the first item incorrectly, he next takes an item in the second row that is easier than (to the left of) the first item.

He may continue with the third and subsequent rows, moving to the right after each correct answer, or to the left after each incorrect answer, until he has at least one right answer and at least one wrong answer. At this point, the computer uses item characteristic curve theory to compute the maximum likeihood estimate of the examinee's ability level. In effect, the computer asks: For what ability level is the likelihood of the observed pattern of responses at a maximum, taking into account the difficulty and other characteristics of the items administered up to this point? The ability level that maximizes this likelihood is the current estimate of the examinee's ability.

From this point on, the next item to be administered will be of the same item type as the item in the next row that best matches in difficulty the examinee's estimated ability level. Given this item type, we survey all items of this type and administer next the item that gives the most information at his estimated ability level.

After each new response by the examinee, his ability is reestimated. The item type of the next item is determined, as above, and the best item (not already used) of that type is chosen and administered. This continues until he has answered 25 items, one for each row of the table. The maximum likelihood estimate of his ability determined from his responses to all 25 items is his final verbal ability score. According to the item characteristic curve model, all such scores, for various examinees, are automatically on the same ability scale, regardless of which set of items was administered.

About thirty different designs for a broad-range tailored test of verbal ability were tried out on the computer, administering each one to a thousand or so simulated examinees. The final design was recently chosen and has not yet been implemented on the computer for administration to real flesh-and-blood examinees.

Consider first the effect of the difficulty level of the first item administered. The vertical dimension in Figure 1 represents the standard error of measurement of obtained test score on the broad-range tailored test, computed by a Monte Carlo study. Each symbol shows how the standard error of measurement varies with ability level (horizontal axis). The four symbols represent the results obtained with four different starting points. The points marked + were obtained when the difficulty level of the first item administered was near -1.0 on the horizontal scale--about fifth grade level. The small dots represent the results when the difficulty level of the first item was near 0--about

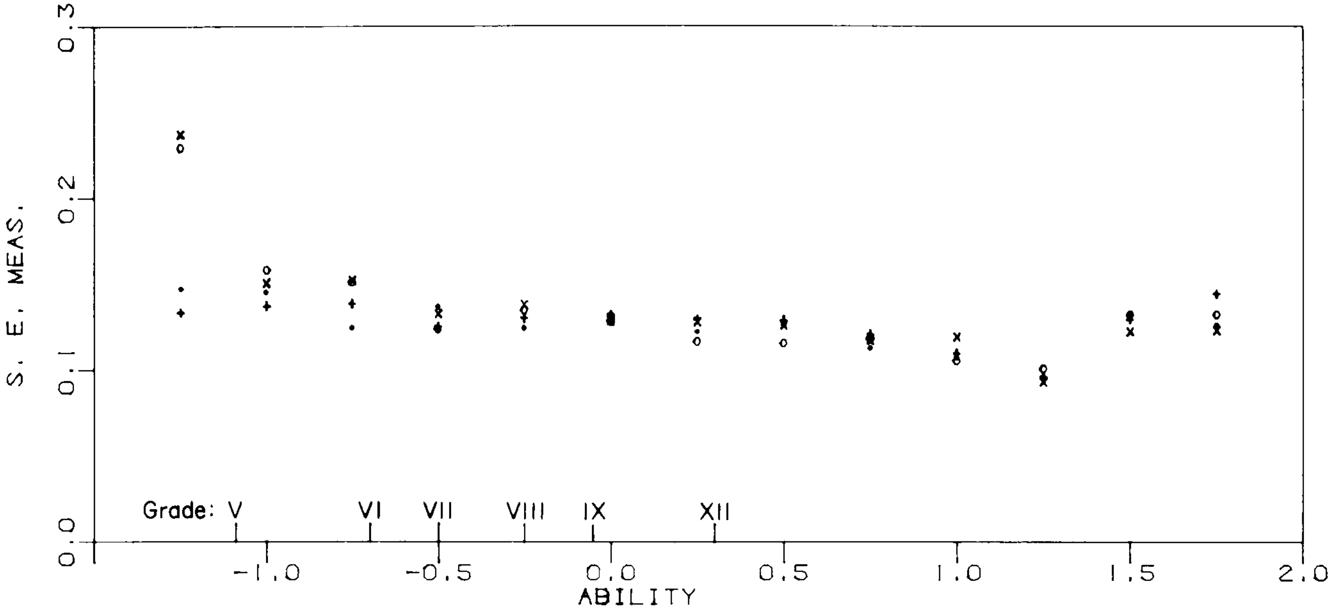

Figure 1. The standard error of measurement at 13 different ability levels for four different starting points for the 25-item broad-range tailored test.

ninth-grade level. For the hexagons, it was near 0.75--near the average verbal ability level of college applicants taking the College Entrance Examination Board's Scholastic Aptitude Test. For the points marked by an x, it was near 1.5. For any given ability level, the standard error of measurement varies surprisingly little, considering the extreme variation in starting item difficulty.

Various designs were also tried out with more columns or with fewer than the 10 columns shown in Table 1. A test with 20 columns, spanning roughly the same difficulty range as Table 1 but requiring 363 items, was found to be at least twice as good as the 10-column 182-item test of Table 1. The reason for this is not that the columns in Table 1 are too far apart, but mainly that selecting the best items (best for a particular individual) from a 363-item pool will give a much better 25-item test than selecting the same number of items from a smaller, 182-item pool. Still better tests could be produced by using still larger item pools, even though only 25 items are administered to each examinee.

It is important to compare the broad-range tailored test with a conventional test. Let us compare our broad-range tailored verbal test with the Preliminary Scholastic Aptitude Test of the College Entrance Examination Board. Figure 2 shows the information function for the Verbal score on each of three forms of the PSAT adjusted to a test length of just 25 items. Also the information function for the Verbal score on the broad-range tailored test, which administers just 25 items to each examinee. The tailored test shown in Figure 2 corresponds to the hexagons of Figure 1, since they represent the results obtained when the first item administered is at a difficulty level appropriate for average college applicants. The PSAT information functions are computed from estimated item parameters. For points spaced along the ability scale, the tailored test information function is estimated from the test responses of simulated examinees.[1]

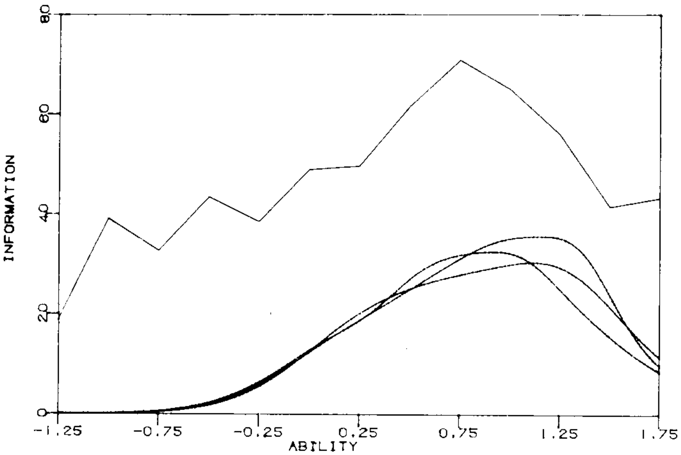

Figure 2. Information function for the 25-item tailored test, also for three forms of the Preliminary Scholastic Aptitude Test (dotted lines) adjusted to a test length of 25 items.

It is encouraging but not surprising to find that the tailored test is at least twice as good as a 25-item conventional PSAT at almost all ability levels. After all, at the same time that we are tailoring the test to fit the individual, we are taking advantage of the large item pool, using the best 25 items available within certain restrictions already mentioned concerning item type. It would, of course, be desirable to confirm this evaluation by extensive test administrations, using flesh-and-blood examinees instead of simulated examinees.

In conclusion, the writer would like to make an offer that should enable research workers and graduate students to conveniently design and build actual tailored tests and administer them to real examinees. On written request from suitably qualified individuals, he will provide estimated item parameters for the verbal items in any or all of the following Cooperative Tests:

SCAT II, Forms 1A, 2A, 2B, 3A, 3B, 4A (50 items each);

STEP II, Reading Test, Part I only, Forms 2A, 2B, 3A, 3B, 4A (30 items each);

SCAT I, Forms 2A, 2B, 3A, 3B (60 items each).

This represents a pool of 690 calibrated verbal items available for research or other purposes. (This offer expires when better methods for estimating item parameters have been developed—very soon, it is to be hoped.)

REFERENCES

Betz, N. E., & Weiss, D. J. Simulation studies of two-stage ability testing. Research Report 74-4. Minneapolis, Minn.: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974.

Cliff, N. Complete orders from incomplete data: interactive ordering and tailored testing. *Psychological Bulletin,* 1975, *82,* 289-302.

De Witt, L. J., & Weiss, D. J. A computer software system for adaptive ability measurement. Research Report 74-1. Minneapolis, Minn.: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974.

Jensema, C. J. An application of latent trait mental test theory. *British Journal of Mathematical and Statistical Psychology,* 1974, *27,* 29-48. (a)

Jensema, C. J. The validity of Bayesian tailored testing. *Educational and Psychological Measurement,* 1974, *34,* 757-766. (b)

Killcross, M. C. A tailored testing system for selection and allocation in the British Army. A paper presented at the 18th International Congress of Applied Psychology, Montreal, August 1974.

Larkin, K. C., & Weiss, D. J. An empirical investigation of computer-administered pyramidal ability testing. Research Report 74-3. Minneapolis, Minn.: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974.

[1] When the test score is an unbiased estimator of ability, the information function is simply the reciprocal of the squared standard error of measurement. A k -fold increase in information may be interpreted as the kind of increase that would be obtained by lengthening a conventional test k -fold.

Mc Bride, J. R., & Weiss, D. J. A word knowledge item pool for adaptive ability measurement. Reasearch Report 74-2. Minneapolis, Minn.: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974.

Mussio, J. J. A modification to Lord's model for tailored tests. Unpublished doctoral dissertation, University of Toronto, 1973.

Spineti, J. P., & Hambleton, R. K. A computer simulation study of tailored testing strategies for objective-based instructional programs. Unpublished manuscript, University of Massachusetts, 1975.

Urry, V. W. Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement,* 1974, *34,* 253-269. (a)

Urry, V. W. Computer assisted testing: the calibration and evaluation of the verbal ability bank. Technical Study 74-3. Washington, D.C.: Personnel Research and Development Center, U.S. Civil Service Commission, in preparation. (b)

Waters, B. K. An empirical investigation of the stradaptive testing model for the measurement of human ability. Unpublished doctoral dissertation, The Florida State University, 1974.

Weiss, D. J. The stratified adaptive computerized ability test. Research Report 73-3. Minneapolis, Minn.: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1973.

Weiss, D. J. Strategies of adaptive ability measurement. Research Report 74-5. Minneapolis, Minn.: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974.

Weiss, D. J., & Betz, N. E. Ability measurement: conventional or adaptive?Research Report 73-1. Minneapolis, Minn.: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1973.

Wood, R. Response-contingent testing. *Review of Educational Research,* 1973, *43,* 529-544.