Adaptive Verbal Ability Testing in a Military Setting

James R. McBride
Navy Personnel Research and Development Center

Since January 1976 all military services have used a common battery of mental tests for enlisted personnel selection and classification: the Armed Services Vocational Aptitude Battery (ASVAB). The battery includes 12 subtests of cognitive aptitudes. These subtests are necessarily short; they are usually scored by hand; the raw scores are manually converted into service-specific scaled scores using conversion tables; and the scale scores are manually recorded and manually transcribed into permanent individual personnel records.

The U.S. Marine Corps has identified some difficulties with the ASVAB testing program. Now that the ASVAB has supplanted service-specific classification test batteries, a single test battery must serve all the special testing needs of the four services. In many cases, ASVAB subtests are excessively difficult for Marine Corps selection and classification purposes; this can result in inefficient and inaccurate classification. There has been some compromise of ASVAB test security: Test booklets and answer keys have been stolen. This problem, if uncontrolled, could seriously degrade the validity of the tests for classification purposes. The manual nature of the test scoring, score conversion, and score recording procedures provides opportunity for clerical error, and it is believed that such errors may have resulted in numerous accession errors.

The Marine Corps formulated an operational requirement to lessen or eliminate the impact of the problems discussed above. Computer-administered adaptive testing (CAT) was identified as one potential solution to all of these problems. In an adaptive test, test difficulty is tailored dynamically to the ability level of the individual examinee; in principle, then, CAT eliminates the problem of excessive test difficulty and should yield scores that promote accurate selection and classification decisions. CAT addresses the test security problem by eliminating printed booklets and scoring keys and by administering an individually tailored set of test items to each examinee. Additionally, since CAT automates test administration, test scoring and recording are automated as well, thereby eliminating human clerical error from the testing system.

Recognizing the potential of CAT for selection and classification testing, the Marine Corps tasked NPRDC with investigating the feasibility of CAT as part of a program of phased research and development related to military personnel accessioning.

Purpose

The research reported here was intended to assess the feasibility of using computerized adaptive testing (CAT) in a Marine Corps recruit/applicant population and, at the same time, to verify the claimed merits of CAT as a psychological measurement technique. These two research issues could only be addressed by administering adaptive tests to appropriate examinee samples. The capability to do this had to be developed--equipment identified, software written, and large banks of test items assembled and calibrated using item characteristic curve (ICC) models. After this development was completed, a pilot study involving verbal ability tests was conducted. This report describes the pilot study of the feasibility and psychometric merits of an adaptive procedure for measuring verbal ability.

Background

Group-administered paper-and-pencil "objective" ability tests date back to World War I, when the introduction of the Army Alpha test signalled an era of vast improvements in the administrative efficiency of psychological testing. The price paid for this efficiency was loss of flexibility, since all examinees had to answer a common set of test questions. The psychometric effect of this was not too serious, provided that a test was designed to have a difficulty level appropriate to its intended application or that a test was sufficiently long to overcome minor design deficiencies. For persons whose ability level was not near the target difficulty level of the test, however, the paper-and-pencil test was not a particularly accurate or precise measuring instrument.

The psychological tests used by the armed services for selection and classification are group-administered paper-and-pencil tests. Such tests, as just discussed, lack the flexibility to measure well over a wide range of ability. In order to achieve that flexibility, the difficulty level of the test would have to be chosen to fit individual ability levels. Since individual ability levels are not known prior to testing, this is not practical; however, it can be accomplished using an adaptive test in which test items are chosen sequentially on the basis of the examinee's performance. This sequential item choice can best be accomplished using automated test administration, for example, by having the test administered at an interactive computer terminal.

The historical development of computer-administered adaptive testing was reviewed by Weiss and Betz (1973) and by Wood (1973). Weiss surveyed a variety of alternative adaptive testing methods (1974) and summarized a number of potential advantages of CAT over conventional paper-and-pencil tests (1975). Despite those potential advantages, most research into adaptive testing had been at the basic research level, until 1975 when the U.S. Civil Service Commission began moving toward early 1980s implementation of computer-based adaptive administration of its PACE examination (Gorham, 1975).

The U.S. Civil Service Commission's implementation plans were based on research conducted by Urry and his colleagues (e.g., Urry, 1977). Urry chose to adopt a Bayesian sequential adaptive testing procedure proposed by Owen (1969, 1975) and demonstrated that the procedure could achieve satisfactory levels of measurement reliability in substantially less than half the number of items required of a conventional test; in one instance he estimated that an adaptive test was equivalent in reliability to a conventional test fivc times as long (Urry, 1977). It is this efficiency of measurement which has motivated most psychometric interest in adaptive testing, although test users have often been more attracted by its practical advantages, which were discussed above.

Marine Corps interest in CAT for personnel selection and classification testing resulted from dissatisfaction with certain aspects of the joint service paper-and-pencil testing battery. Subtests used for selection decisions were also used as a basis for personnel classification and assignment to specialized training; a test designed for one of these purposes would likely be inappropriate for the other, and this might result in disproportionate numbers of selection or assignment errors. Clerical errors in the manual scoring and score recording processes were felt to be another serious source of accessioning errors; and the effects of test compromise were inevitable with the use of the same test battery over a period of several years.

Recognizing that computerized test administration could eliminate scoring and clerical errors and that adaptive testing could substantially reduce test compromise, Marine Corps Headquarters tasked NPRDC with evaluating the feasibility of CAT for testing Marine recruits. The purpose of this paper is to report the results of the first in a series of studies investigating both the feasibility and the utility of CAT in comparison with a conventional test design.

The study was designed in part to address three research questions: (1) Is computer-based testing of military recruits administratively feasible? (2) Is a computer-administered adaptive test more reliable than a conventional test, holding test length constant? (3) If so, what is an appropriate length criterion for an adaptive test?

These questions were motivated by the results of previous research done elsewhere. The first question--that of administrative feasibility--seems trivial but is not. Interviews with military testing personnel indicated some misgivings about the ability of military recruits to use relatively sophisticated automated testing equipment, such as CRT computer terminals. This potential man-machine interface problem is the analogue of administrative difficulties encountered years earlier with paper-and-pencil tailored tests. For example, Seeley, Morton, and Anderson (1962) found that a substantial proportion of their military examinees did not successfully follow instructions on an experimental sequential item test; this experience may have caused a five-year lapse in military research on tailored or adaptive testing. Olivier (1974) had a similar experience using a paper-and-pencil flexilevel test in a sample of high school students.

The question of the advantages of adaptive tests over conventional ones in terms of reliability has a clear and positive theoretical answer: Holding test length and all else constant, a good tailored test design is superior, provided that highly discriminating test items are available (Urry, 1970).

This theoretical advantage is not always corroborated in empirical investigations. For instance, Bryson (1971) questioned the advantage of tailored testing over certain methods of conventional test design; Olivier (1974) failed to find an advantage for the flexilevel tests he used; and the results reported by Weiss and his colleagues have been less than unanimous in favor of adaptive tests. All these results are in contrast with those of Urry (1977), who reported that for his sample of 57 Civil Service job applicants an adaptive verbal ability test achieved an 80% reduction (compared to a conventional test) in the test length required to attain any of several specified levels of reliability. Urry's result was extraordinary. The only cloud over it is that it was based on indirect evidence: The conventional test reliabilities were based on Spearman-Brown equation adjustments to the reliability obtained in an independent sample, and the tailored test reliability was merely assumed, not rigorously verified.

Previous research into the reliability, validity, and efficiency of adaptive tests has often been inconclusive because of design flaws or nuisance factors. The major problem has been the lack of suitable means for estimating the adaptive test's reliability without making dubious assumptions. Another problem has been the general failure to match adaptive and counterpart conventional tests in item quality, with an unfair advantage usually in favor of the adaptive test. The research reported here was intentionally designed to remove those two problems--to provide credible indices of reliability that are appropriate for both test types and to provide a fair comparison by matching item quality across the test types. With those two problem sources eliminated, there is hope for an unequivocal comparison between adaptive and conventional test designs.

Method

The general method used was that of equivalent tests administered to independent examinee groups. One group took two equivalent computer-administered adaptive tests. The other group took two equivalent conventional tests, also administered by computer. In order to control for item quality, both test types were made up of items from the same source--a common pool of 150 verbal ability items, which had previously been calibrated in large samples of Marine recruits, using ICC methods.

Research Design

Each examinee was randomly assigned to one of the two treatment groups--Group A or C. Group A took two 30-item adaptive verbal ability tests, followed by a 50-item criterion test of word knowledge. Group C took two 30-item conventional verbal ability tests, followed by the same criterion test. All tests were administered at a computer terminal. Figure 1 is a schematic representation of the research design.

Observations. For each examinee who completed the tests, the following data were observed and automatically recorded:

1. Elapsed time for the testing session;
2. Elapsed time to complete pretest instructions;
3. Number of errors made during the instructions;
4. Number of times the proctor was called;

Figure 1
The Research Design for Administration
of the Experimental and Criterion Tests

	Tests				
Treatment	Adaptive		Conventional		
Group	Form 1	Form 2	Form 1	Form 2	Criterion
A	X	X			X
C			X	X	X

5. Raw item scores (correct/incorrect);
6. Cumulative raw score after each item;
7. Latent trait ability estimates (experimental tests only);
8. Bayes posterior variance of the ability estimate after each item; and
9. Criterion test raw score.

The format for these observations is schematized in Figure 2.

Figure 2
Example of Examinee Record (Abbreviated)

	Raw Score		Ability Estimate		Posterior Variance	
Form:	1	2	1	2	1	2
Stage						
1	0	0	-.69	-.73	.548	.533
2	1	1	-.36	-.37	.401	.394
3	2	2	-.10	-.20	.332	.318
4	2	3	-.30	.02	.248	.266
5	3	4	-.14	.25	.229	.213
6	4	5	.01	.48	.193	.210
7	4	6	-.17	.65	.160	.184
8	5	6	-.05	.45	.145	.143
9	5	6	-.22	.26	.124	.115
10	6	7	-.15	.33	.115	.107
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
30	20	21	.59	.97	.053	.048

Criterion score	27
Total time	57.3 minutes
Instruction time	8.5 minutes
Instruction errors	1
Proctor calls	0

Independent variables. For the comparisons between the adaptive and conventional testing methods there were two independent variables: (1) test type (adaptive versus conventional) and (2) test length (5, 10, 15, 20, 25, 30 items).

Within the adaptive testing method, the test termination rule was treated as an independent variable for some analyses: Tests were terminated (1) at a fixed test length (5, 10, ..., 30 items) or (2) at a specified posterior variance (variable length). The number-of-items termination rule resulted, of course, in a test of predetermined length; and the posterior variance rule resulted in a variable length test, depending on the number of items required to attain specified levels of the Bayes posterior variance.

Dependent variables. Measures of the dependent variables were formed from the individual observations. The dependent variables included:

1. Testing time;
2. Instruction time;
3. Number of keyboard errors;
4. Number of proctor calls;
5. Alternate tests reliability coefficient after 5, 10, ..., 30 items; and
6. Test-criterion correlation after 5, 10, ..., 30 items.

Procedure

Items. The 150 items in the pool were calibrated using Urry's ancillary estimation method and were selected according to the prescriptions given by Urry (1977): All ICC slope parameters exceeded .80. The average value of the discrimination (*a*) parameter was 1.24; item difficulty (location, or *b*) parameters ranged from -2.0 to +2.0; and there were no items with a pseudo-guessing (*c*) parameter greater than .30.

Examinees. Male Marine recruits reporting for duty at the Marine Corps Recruit Depot, San Diego, were the examinees. They were tested one at a time at a Burroughs TD832 terminal controlled by a Burroughs B1717 time-sharing minicomputer system. Assignment to groups (Group A or C) was randomized. Two hundred one examinees completed the tests--96 of these took the adaptive tests and 105 took conventional tests.

Tests. The conventional tests administered to Group C were rectangular tests spanning the difficulty range of the item pool. This broad range of difficulty was chosen in order to simulate the psychometric design of the verbal tests used in the ASVAB. Two 30-item equivalent forms--Form 1 and Form 2--were constructed from the 150-item pool. Items were chosen to be as highly discriminating as possible, consistent with the broad difficulty range. The two forms were constructed to be "weakly parallel" (Samejima, 1977), i.e., to have approximately equal test information functions. Within each form, the 30 items were sorted into five difficulty levels, then arranged in descending order of discriminating power within each level. The first five items in each form were the most discriminating items at their respective difficulty levels; items 6 through

10 were the second most discriminating items at each level; and so on. This arrangement resulted in two 30-item tests consisting of a sequence of six 5-item subsets each. This design was intended to permit meaningful analysis of the psychometric properties of rectangular conventional tests of lengths of 5, 10, 15, 20, 25, and 30 items. In order to equalize any effects due to test length, fatigue, or other extraneous factors, the two conventional tests were administered in counterbalanced item order, i.e., the two 30-item tests were administered as one 60-item test in the following order:

Item sequence:	1	2	3	4	5	6	7	8	...
Test Form:	1	2	2	1	2	1	1	2	...

The two 30-item adaptive tests were based on Owen's (1969, 1975) Bayesian sequential tailored testing procedure. For each examinee and each test form an initial normal prior distribution of ability was assumed, with mean 0 and variance 1.0. The test form (either 1 or 2) was counterbalanced for each examinee in a manner identical to that of the conventional tests: 12212112.... Both forms of the Bayesian test--Form 1 and Form 2--drew items from the same 150-item pool; counterbalancing the order of administration here served the added purpose of equalizing item quality across the two forms. The two adaptive tests were independent of each other except for their use of a common item pool.

The criterion test was formed by concatenating two obsolete operational test forms measuring word knowledge. This resulted in a 50-item test expected to be a highly reliable and fairly broad-range test of an important facet of verbal ability.

Results and Discussion

Feasibility

Data pertaining to the feasibility of using computer terminals to administer tests to military recruits are summarized in Table 1. Mean testing time was 61.0 minutes for the adaptive test group versus 50.4 minutes for the conventional test group. These were the mean times to answer 110 items--60 items from either the adaptive or the conventional alternate forms, followed by 50 criterion test items common to both groups. The adaptive tests required about 11 more seconds per item, or as much as 39% longer to answer than the conventional tests. Some or all of this difference may have been due to computations required for adaptive item selection, but this result does agree generally with Waters' (1977) finding that an adaptive test required significantly longer examinee processing per item than a similarly administered conventional test. In the present study, however, the observed time difference may be due in large part to idiosyncrasies of the computer system; if so, differences of the size reported here would not be expected if a faster computer were used to control and to administer the adaptive tests.

Instruction time averaged 9.5 minutes for the adaptive test group and 10.3 minutes for the conventional group; overall, the instructions required an average of 9.9 minutes. During this time, the examinees were familiarized with the CRT and keyboard by means of a programmed instructional sequence with special

Table 1
Testing Time and Examinee Error Summary for
Computer-Administered Test Sessions

	Group and Test		
	A	C	
Data	Adaptive	Conventional	Overall
Number of examinees	96	105	201
Mean time (minutes)			
Total	70.5	60.7	
Instructions	9.5	10.3	9.9
Testing	61.0	50.4	
Errors			
Procedural errors	25	30	55
Proctor calls	5	12	17

Note. Each session consisted of programmed instruction, 60 experimental test items, and a 50-item criterion test.

branching following procedural errors and with an audible call to the proctor if the examinee had difficulty correcting an error. Errors and proctor calls were counted. As the table indicates, there were 55 errors in all, in 201 test sessions; in only 17 cases was the proctor called. This amounts to about one procedural error per 4 test sessions and to a requirement for proctor intervention about one time per 12 test sessions.

Psychometric Characteristics

Reliability. Table 2 summarizes reliability and criterion validity data for both the adaptive and conventional alternate forms tests at lengths of 5, 10, 15, 20, 25, and 30 items.

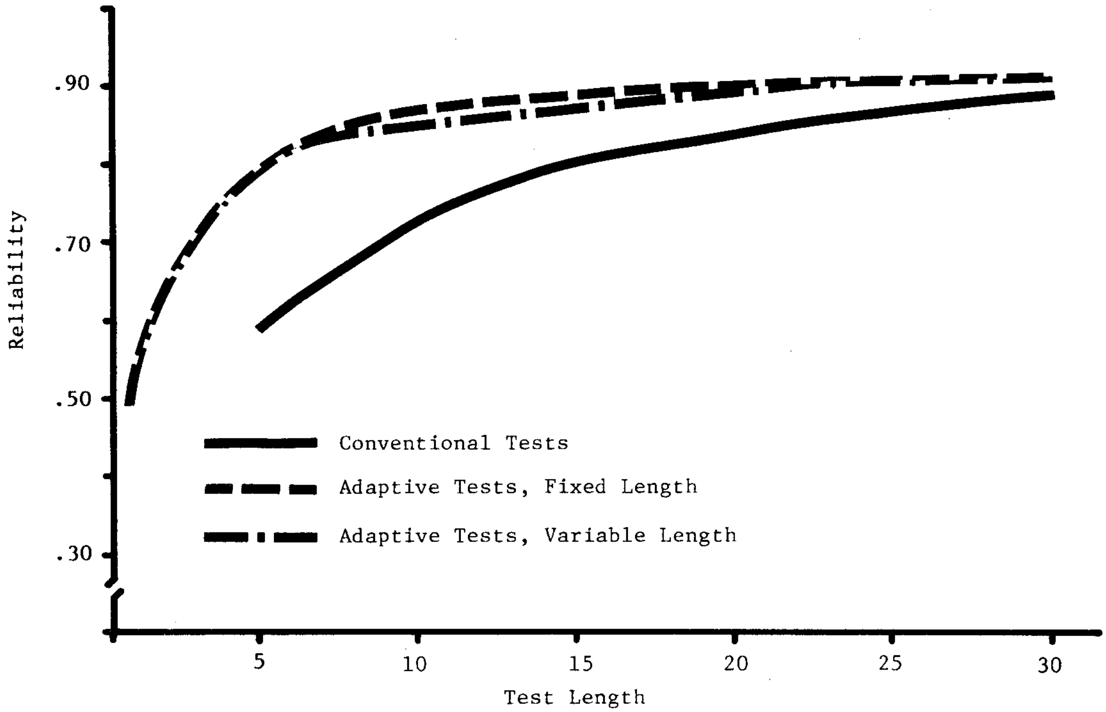
Table 2
Psychometric Characteristics of the Computer-Administered
Verbal Ability Tests as a Function of Test Type and Test Length

Psychometric Characteristic and Test		Test Length					
	N	5	10	15	20	25	30
Reliability							
Adaptive	96	.79	.87	.88	.90	.91	.91
Conventional	105	.59	.73	.80	.83	.86	.89
Validity							
Adaptive	93	.77	.82	.83	.84	.85	.85
Conventional	103	.73	.81	.84	.85	.85	.87
Relative efficiency		2.70	2.50	1.90	1.80	1.70	1.30

Reliability was operationalized as the correlation between scores on alternate forms at a given test length. The scoring procedure used was the same for both test types--latent ability estimation using the sequential estimation formulae developed by Owen (1969). From the table it is clear that the adaptive tests had substantially higher reliability coefficients than the conventional tests for any given test length. Viewing these data another way, it can be seen that the adaptive test reliability at a 5-item test length was practically equivalent to the conventional test's reliability at 15 items; similarly, the adaptive test's reliability at a length of 10 was superior to that of the conventional test at a length of 25.

Figure 3 contains a graphic comparison of the adaptive and conventional tests in terms of alternate forms reliability as a function of test length. Analysis of Table 2 and Figure 3 indicates that in terms of test length required to attain a given level of reliability, the adaptive tests had a substantial advantage over the conventional tests. This advantage was essentially the same for both fixed length and variable length stopping rules; there was no apparent advantage to variable length, as opposed to fixed length, within the adaptive testing method.

Figure 3
Alternate Forms Reliability Plotted as a Function of
Test Length for the Conventional and Adaptive Tests



<u>Relative efficiency</u>. Thus, the adaptive tests achieved specific levels of reliability more efficiently than the conventional tests. How much more effi-

ciently is indicated in row 3 of the table, labeled "relative efficiency." These data, based on the Spearman-Brown equation, estimate for each test length how much the conventional tests would have to be lengthened to attain the reliability of the adaptive tests. For example, the adaptive test reliability at 5 items, .79, was estimated to be equivalent to that of a conventional test 2.70 times as long, or 13.5 items in length. Notice that the relative efficiency of these adaptive tests always exceeds unity but diminishes as test length increases. Thus, the adaptive tests are more advantageous, at least in terms of relative efficiency, at fairly short test lengths. At lengths of 10 or fewer items, these adaptive tests were at least 2.5 times as efficient as the conventional tests. At lengths of 15 and more, however, the advantage, although still appreciable, is not quite so striking.

Validity. The advantage of adaptive tests was not so clear when the validity of the two test types is compared. Validity was operationalized as the correlation between test scores and the examinee's raw score on the concurrently administered 50-item Word Knowledge test. From their superior reliability, it would be expected that the adaptive tests would also be superior in validity at any constant test length. As Table 2 indicates, the adaptive tests had higher validities at test lengths up to 10 items; at lengths of 15 and up, however, the conventional tests had slightly higher validity. None of the validity differences was statistically significant at the .05 level.

Conclusions

Based on the data reported above, several conclusions are offered with regard to the feasibility and psychometric merits of adaptive aptitude testing of Marine recruits.

1. Testing Marine recruits with CRT terminals is feasible from both practical and human engineering standpoints. Embedded programmed instructions can effectively teach the recruits the use of the testing terminals. The number of proctors or attendants required to supervise and to assist in the testing room appears to be acceptably small.

2. Striking psychometric efficiency was demonstrated for the adaptive tests of verbal ability used in this study. It appears that in military personnel testing applications, well-constructed short adaptive tests can achieve high levels of measurement reliability with less than half the number of items required using conventional testing procedures.

3. There is no apparent psychometric advantage to the intuitively appealing notion of variable-length adaptive tests, at least for the adaptive testing method used here.

4. Short fixed-length adaptive tests of about 10 items per examinee seem to be sufficiently reliable for personnel testing purposes. The adaptive tests achieved a minimally satisfactory reliability level (.80) in just 5 items; additional test lengths beyond 10 items did not yield psychometric returns proportional to the added administration time required.

REFERENCES

Bryson, R. A comparison of four methods of selecting items for computer-assisted testing (Technical Bulletin STB 72-8). San Diego, CA: Naval Personnel and Training Research Laboratory, December 1971.

Gorham, W. A. Opening remarks. In W. A. Gorham (Chair), Computers and testing: Steps toward the inevitable conquest (PS-76-1). Symposium presented at the 83rd annual convention of the American Psychological Association, Chicago, 1975. Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, September 1976. (NTIS No. PB 261 694)

Olivier, P. An evaluation of the self-scoring flexilevel testing model. Unpublished doctoral dissertation, Florida State University, 1974.

Owen, R. J. A Bayesian approach to tailored testing (Research Bulletin 69-92). Princeton, NJ: Educational Testing Service, 1969.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Samejima, F. Weakly parallel tests in latent trait theory with some criticisms of classical test theory. Psychometrika, 1977, 42, 193-198.

Seeley, L. C., Morton, M. A., & Anderson, A. A. Exploratory study of a sequential item test (Technical Research Note 129). Washington, DC: U.S. Army Personnel Research Office, December 1962.

Urry, V. W. A monte carlo investigation of logistic test models. Unpublished doctoral dissertation, Purdue University, 1970.

Urry, V. W. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

Waters, B. K. An empirical investigation of the stratified adaptive computerized testing model. Applied Psychological Measurement, 1977, 1, 141-152.

Weiss, D. J., & Betz, N. E. Ability measurement: Conventional or adaptive? (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1973. (NTIS No. AD 757788)

Weiss, D. J. Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis: University of Minnesta, Department of Psychology, Psychometric Methods Program, December 1974. (NTIS No. AD A004270)

Weiss, D. J. Computerized adaptive ability measurement. Naval Research Reviews, 1975, 28, 1-18.

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

ACKNOWLEDGMENTS

This paper was also presented at the 1979 Conference of the Military Testing Association, San Diego, October 1979. Any opinions expressed are those of the author, and not necessarily those of the Department of Defense or the Department of the Navy.