An Empirical Investigation of the Stratified Adaptive Computerized Testing Model

Brian K. Waters[1]
Air Force Human Resources Laboratory

This study empirically investigated the validity and utility of the stratified adaptive computerized testing model (stradaptive) developed by Weiss (1973). The model presents a tailored testing strategy based upon Binet IQ measurement theory and Lord's (1972) modern test theory. Nationally normed School and College Ability Test Verbal analogy items (SCAT-V) were used to construct an item pool. Item difficulty and discrimination indices were re-scaled to normal ogive parameters on 244 items. One hundred and two freshmen volunteers at Florida State University were randomly assigned to stradaptive or conventional test groups. Both groups were tested via cathode-ray-tube (CRT) terminals coupled to a Control Data Corporation 6500 computer. The conventional subjects took a SCAT-V test essentially as published, while the stradaptive group took individually tailored tests using the same item pool. Results showed significantly higher reliability for the stradaptive group, and equivalent validity indices between stradaptive and conventional groups. Three stradaptive testing strategies averaged 19.2, 26.5, and 31.5 items per subject as compared with 48.4 items per conventional subject. A 50% reduction from conventional test length produced equal precision of measurement for stradaptive subjects. Item latency comparisons showed that those in the stradaptive group required significantly longer per item (about 11%) than conventional group members. It is recommended that testing time rather than number of items be used as a dependent variable in future adaptive testing research.

Lord's (1970, 1972) theoretical analysis of adaptive testing versus conventional testing makes one point very clear: a peaked test provides more precise measurement than an adaptive test of the same length *when the testee's ability is at the point at which the conventional test is peaked.* At some point on the ability continuum, generally beyond $\pm$ .5 standard deviations from the mean, the adaptive test requires fewer items for comparable measurement efficiency.

Lord (1970, p. 178) suggests that an "ideal" testing strategy would present to each testee a sample of items comprising a peaked test with a .50 probability of a correct answer for examinees of the particular testee's true ability ($P_c = .50$). Of course, the true ability of the subject is unknown; the estimation of ability level is, in fact, the desired outcome of the measurement procedure.

Traditionally, this problem has been circumvented by peaking the test at $P_c = .50$ for the hy-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 1, No. 1 Winter 1977 pp. 141–152

[1]The views expressed herein are those of the author and do not necessarily reflect the views of the United States Air Force and/or the Department of Defense.

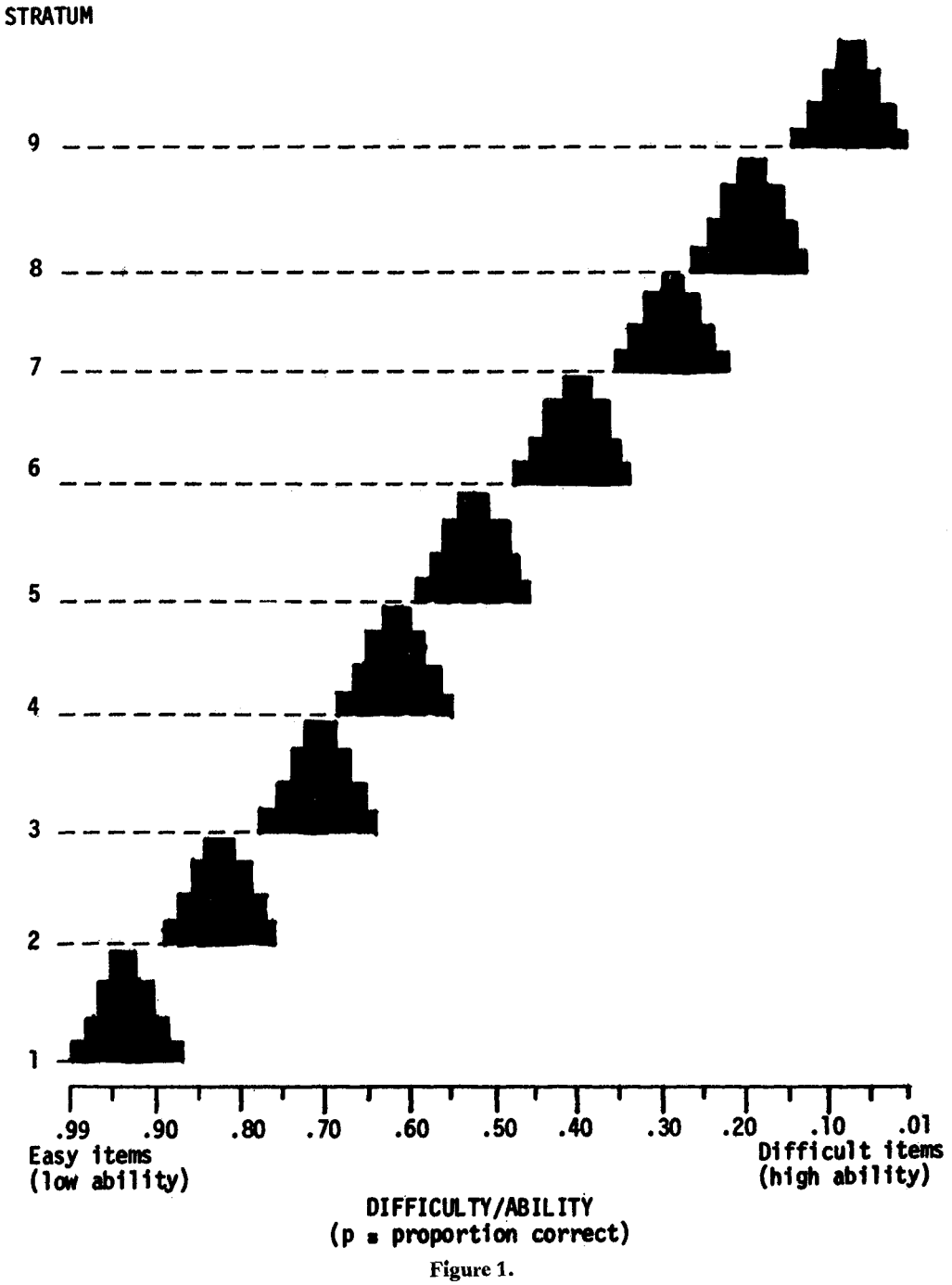

Figure 1.

Distribution of Items, by Difficulty level, in a Hypothetical Stradaptive Test.

pothetical *average* ability level subject. This procedure worked well for examinees near the center of the ability continuum, but less efficiently near the extremes (Weiss, 1973).

Weiss' (1973) stradaptive model extends the rationale for ability measurement originally proposed by Binet (Terman & Merrill, 1937) to computer-based ability measurement. A large item pool is necessary with item parameter estimates based upon a large sample of subjects from the same population as potential examinees. Items are scaled into peaked levels (strata) according to item difficulty. A testee's initial item is based upon a previously obtained ability estimate or his/her own estimation of ability on the dimension being assessed.

Figure 1 depicts the distribution of item difficulties in a hypothetical nine strata stradaptive item pool.

As in the Binet tests, the testee's basal and ceiling strata are defined, and testing ceases when the ceiling stratum has been determined. A testee's score is a function of the difficulties of the items answered correctly, utilizing various scoring strategies (Weiss, 1973).

Method

Item Bank

Verbal analogy test items used in this study were selected from Educational Testing Service's (ETS) SCAT Series II. This test series provided single-format items with extensively-normed item parameter estimates. Item pool data received from ETS contained five 50-item verbal analogy tests, forms 1A, 1B, 1C, 2A, and 2B of SCAT II. These tests had been nationally

Table 1

Summary Statistics of Normal Ogive Parameter Estimates for Stradaptive Test Item Pool

	Number of	Item Difficulty (b)				Item Discrimination (a)			
Stratum	Items	Mean	S.D.	Low	High	Mean	S.D.	Low	High
1	20	-2.22	.373	-3.57	-1.94	.52	.159	.25	.86
2	26	-1.69	.155	-1.90	-1.46	.67	.165	.36	.95
3	34	-1.15	.177	-1.40	- .90	.66	.215	.27	1.17
4	39	- .68	.126	- .88	- .49	.63	.168	.33	1.07
5	31	- .25	.111	- .44	- .11	.57	.153	.29	.83
6	27	.10	.071	- .10	.25	.58	.145	.34	.88
7	26	.44	.142	.27	.67	.55	.112	.34	.73
8	22	.95	.201	.71	1.25	.48	.146	.25	.79
9	19	1.91	.651	1.27	3.69	.40	.122	.25	.77

normed on a sample of 3133 twelfth grade students. P-values and biserial correlations on 249 items were provided by ETS. These values were transformed into normal ogive item parameters (Lord & Novick, 1968), with five items removed from the item pool due to excessively high or low difficulty values or low item discrimination values. Table 1 shows the actual distribution of item difficulties and discriminations used in this study.

The nine strata summarized in Table 1 are essentially nine peaked tests, varying in average difficulty from −2.22 to +1.91. Stratum 9, the most difficult peaked test, for example, was composed of 19 items ranging from b = 1.27 to b = 3.68. In this study, items were randomly ordered within strata, unlike in Weiss' model, in order to permit an alternate-forms reliability coefficient to be calculated for stradaptive examinees. As is typical in educational and psychological research, the concentration of more difficult items contains the lower discrimination values. A correlation between b_g and a_g of −.31 reflects this problem.

Subjects

One hundred and two incoming freshmen to Florida State University were tested in late July 1974. Ninety-nine of the subjects had Florida Twelfth Grade (12V) Verbal Scores or 12V estimates derived from ACT or CEEB verbal scores to serve as criteria for the validity investigation of the stradaptive test scores.

Subjects were randomly assigned to conventional or stradaptive groups, and those in the conventional group were further randomly assigned to five subgroups corresponding to the five versions of the SCAT-V. As each subject entered the testing area, the test proctor assigned him the next test listed on the randomized testing order schedule. Schematically, the research design is depicted in Table 2.

A comparison of outcomes 0_2 through 0_5 would indicate the effectiveness of the randomization process in equating subtest assignment. Assuming no significant differences between these outcomes, comparisons between 0_6 through 0_{10} could then be made. Since SCAT-V published results had shown significantly dif-

Table 2
Number of Testees Assigned to each Testing Group

Testing Group	Number of Subjects	Criterion Variable: 12V Score	Criterion Variable: SCAT-V Score	
Conventional Test Group				
1A	8	0_1^*	0_6	
1B	7	0_2	0_7	
1C	9	0_3	0_8	
2A	13	0_4	0_9	
2B	10	0_5	0_{10}	
Conventional Total	47	0_{11}	0_{12}	
Stradaptive Test Group	55	0_{13}	0_{14}	0_{15}^{**}

*0_i = outcome for group i.

**0_{15} = stradaptive parallel second test.

ferent difficulty levels between the five forms, it was planned that conventional subtest scores would be normalized within their separate distributions and then pooled into a conventional total score distribution (0_{12}) for comparison with the stradaptive results (0_{14} and 0_{15}).

The independent variables for the comparisons in this study were conventional or stradaptive group, termination rule, 12V score, and scoring method. Dependent variables included test scores, item latency, number of items, standard errors (and/or reliability), and validity.

Table 3 shows conventional vs. stradaptive group test statistics on the 12V scores. The random assignment of subjects to conventional or stradaptive testing groups resulted in no significant differences in means or variances of the criterion 12V scores.

CRT Testing

A computer program described by DeWitt and Weiss (1973) was adapted for the Florida State University Control Data Corporation 6500 computer.

Testing sequence. The testees estimated their ability using the procedures described by DeWitt and Weiss. The first item the stradaptive testee received was the first item in the stratum commensurate with his ability estimate. The testee then was branched to the first item in the next higher or lower stratum depending upon whether the initial response was correct or incorrect. If the testee entered a question mark (?), the next item in the same stratum was presented.

Testing continued until each person's ceiling stratum was identified. For this study, the ceiling stratum was defined as the lowest stratum in which 25% or fewer of the items attempted were answered correctly, with a constraint that at least five items be taken in the ceiling stratum. The 25% figure reflects the probability of getting an item correct by random guessing on a four-option multiple choice test. Once a testee's ceiling stratum was defined, the program looped back to that person's ability estimate stratum and commenced a second stradaptive test with item selection continuing down the item matrix from where the first test ended. Since items were randomly positioned within each stratum, parallel or alternate forms were taken by all testees who reached the termination criterion on the first test.

A maximum of 60 items per test per subject was established, since prestudy trial testing suggested that subjects became saturated beyond 120 items.

Termination rules. It was reasonable to expect that a testee would omit an item *only* when he felt he had no real knowledge of the correct answer. Thus, investigation of test termination

Table 3

Comparison of Distributions of Conventional and Stradaptive Group Florida 12th Grade Verbal Scores

Group	Number of Subjects	Mean	Std Dev	Std Err	Kurtosis	Skewness
Conventional	46	33.26	5.30	.855	.44	.70
Stradaptive	53	34.06	6.12	.841	.36	-.03

based upon omissions counted as wrong answers was judged appropriate.

Weiss had set five items in the ceiling stratum as the minimum constraint upon termination. A secondary goal of the present study was to determine what effect the reduction of this constraint to four would have upon the effectiveness of the stradaptive strategy.

These two questions of the handling of omissions and the variation in the constraint on the

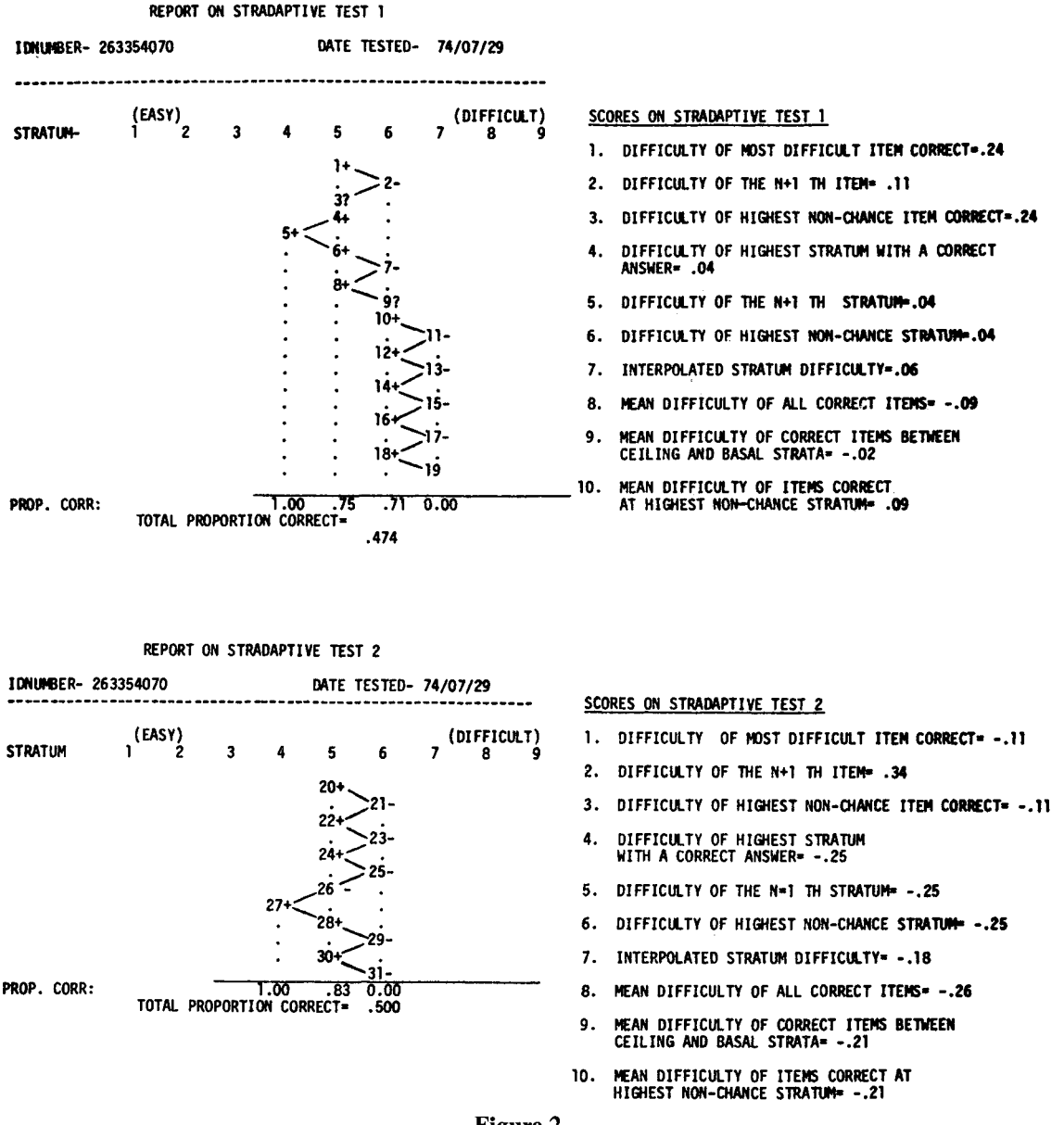

Figure 2.

Example of Stradaptive Testing Output.

termination of testing created the following three methods for comparisons:

Termination Method 1: Omissions ignored; stop testing at minimum of five items.

Termination Method 2: Omissions = wrong; stop testing at minimum of five items.

Termination Method 3: Omissions = wrong; stop testing at minimum of four items.

Data were collected using termination Method 1 and then re-scored using Methods 2 and 3. This was possible since no indication of the termination of the first test was given to the subject and since items were randomly ordered within strata. Once test termination was reached using termination Method 2 or 3, the next item taken by the testee in his entry point stratum acted as the start of a parallel-forms test under the termination rule used.

Of course, Method 2 required fewer items than Method 1, and Method 3 considerably fewer than Method 2. The thrust of this investigation, then, was to determine the relative efficiency of the three methods in comparison with one another and with conventional testing after equalizing test length using the Spearman-Brown prophecy formula.

Stradaptive test output. Figure 2 provides an example of a stradaptive test report from this experiment. A "+" next to an item indicates a correct response; a "–" an incorrect response, and "?" shows that the subject omitted the item.

The examinee in Figure 2 estimated her ability as "average". Hence, her first item was the first item in the 5th stratum. She correctly answered this question but missed her second item, and after responding somewhat inconsistently for the first nine items, "settled down" with a very consistent pattern for items 10 through 19 when she reached criterion, and her first test was terminated.

The testing algorithm then selected the sixth item in stratum 5 (her ability estimate) to commence her second test. (The subject was totally unaware of this occurrence since no noticeable time delay occurred between her 19th and 20th items.)

At the conclusion of her 31st item, this subject reached the termination criterion for her second test, was thanked for her help in this research project, and given her score of 15 correct answers out of 31 questions with a percentage correct of 48.4%. The scores for this subject are shown in Figure 2 for both tests.

Results and Discussion

Proportion Correct

Test theory suggests that measurement efficiency is maximized at $P_c = .50$ for a given test group. It was hypothesized that the stradaptive test strategy would more nearly approach this standard than the conventional test, indicating an improved selection of items for the stradaptive subject. Table 4 shows the result of this comparison. It clearly indicates significantly different distributions of test difficulty. The stradaptive test was far more difficult than the conventional test, with a smaller variance. The mean P_c for the stradaptive test was closer to the desired value of .50.

Reliability

Conventional test reliability. Making the standard assumptions underlying the one factor random effects analysis of variance (ANOVA), the estimated reliability coefficient of the total scores was .776 for the conventional examinees for a test of an average of 48.4 items in length. Because six items were removed from the original item pool, the five subtests varied in length from 48 to 50 items, resulting in the 48.4 average test length. Stepped up to 50 items via the Spearman-Brown prophecy formula, this estimate becomes .782. The reported reliability of the original SCAT-V tests was .87. This difference in reliabilities was statistically significant ($p < .05$) using Feldt's (1965) test.

It can be assumed that the difference between these reliabilities was caused by one or more of three factors:

1. Testing mode (CRT vs. ETS paper and pencil).

Table 4

Comparison of Difficulty Distributions (P_c) for Conventional and Stradaptive Groups

Group	Number of Subjects	P_c	Std Dev	Std Err	Kurtosis	Skewness
Conventional	47	.752*	.123*	.018	-.87	-.39
Stradaptive	55	.584	.084	.011	5.14	1.97

* $p < .05$

2. Elimination of six of the 250 items from the original item pool.
3. Restriction of range in subject pool for this experiment.

The last factor most likely caused the decrease in the reliability of the test scores. The homogeneity of the subjects would yield a relatively small amount of between-person variance, which would lower the reliability estimate.

Stradaptive total-test reliability. Using Stanley's (1971) procedure, it was possible to estimate the internal-consistency reliability of the person-by-item stradaptive test matrix. Of the 244 items in the stradaptive pool, only 133 items were actually presented to the subject pool in this experiment. Stradaptive Scoring Method 8, mean difficulty of all items answered correctly, provided the only set of stradaptive test scores wherein a person's total test score was a linear function of his item scores. Hence, this scoring method was used to estimate internal-consistency reliability.

Table 5 shows the parallel forms and KR-20 reliability estimates for the three termination rules used in this study. Direct comparisons can be made between the stradaptive KR-20 values and the .782 conventional KR-20 estimate. According to Feldt's (1965) approximation of the distribution of KR-20, all of the estimates of the stradaptive test reliability are significantly ($p \leq .05$) better than the conventional KR-20 estimate *prior* to being stepped-up by the Spearman-Brown formula. Thus, the 19-, 26-, and 31-item stradaptive tests all proved more reliable than the 48-item conventional test.

The unequal N's in Table 5 for the parallel forms reliability estimates were a result of the varying lengths of the tests under the three termination rules. Parallel reliabilities were calculated only when two complete tests were given to a stradaptive subject prior to the 60-item constraint.

A comparison of the conventional test internal consistency reliability coefficients (r_{tx}) and the stradaptive parallel-forms reliability estimates (r_{xx}) in Table 5 must be considered only tentative since they are different kinds of reliability estimates. The sampling distribution of r_{xx} is known and that of r_{tx} has been approximated by Feldt (1965). Cleary & Linn (1969) compared standard errors of both indices with generated data of known rho. They found the standard error of KR-20 to be somewhat smaller than that of the parallel-test correlation (approximately .05 vs. .04 in the range of reliabil-

Table 5

Comparison of Parallel Forms Reliability with KR-20 Reliability Stepped-up to 50 Items Using Three Termination Rules

	Termination Rule		
	1	2	3
Parallel Forms	(N=12)	(N=28)	(N=38)
r_{xx} (raw)	.892	.688	.732
r_{xx} (50)	.929	.806	.903
KR-20	(N=55)	(N=55)	(N=55)
r_{20} (raw)	.901	.899	.873
r_{20} (50)	.935	.943	.947
Average Number of Items Administered	31.45	26.47	19.20

ities, number of subjects, and number of items involved in the present study). Thus, it is expected that the KR-20 estimates would be more "stable" estimates of rho than the parallel forms estimates.

Validity

Conventional test. The correlation of obtained conventional scores with the Florida 12th Grade Scores was .477, which was significantly lower than the published SCAT-V:SAT-V correlation of .83 ($p \leq .01$). As with the conventional reliability, this difference most likely resulted from subject homogeneity.

Stradaptive test. The validity coefficients of the stradaptive scoring under the three termination rules are shown in Table 6. Validity was estimated by the correlation between the test scores and 12V scores. None of the validity coefficients in Table 5 was significantly different

Table 6

Comparison of Validity Coefficients (Correlation with 12V Scores) of Scoring Method 8 under Three Termination Rules

Termination Rule	N	r_{cx}
1	64	.536
2	80	.536
3	91	.499

from the conventional test validity coefficient of .477, although stradaptive validity coefficients were consistently higher than those for the conventional test.

Other Characteristics of the Stradaptive Test

Number of items. Table 7 shows the difference in number of items presented for the conventional test and the three termination methods of the stradaptive test. The consistency in average number of items presented per subject in the stradaptive test was surprisingly constant over the two parallel tests of termination methods 1 and 3. Method 2 did show a significant ($p < .05$) drop in the average number of items on the second test.

Item latency. It was hypothesized that mean item latency would be higher for stradaptive testees since they would have to "think" about each item as it was near the limit of their ability. Table 8 reflects the results of this comparison.

The hypothesis of no differences between item latencies was rejected ($p < .01$). For the subjects in this experiment, the average stradaptive item required approximately 11% longer than the average conventional test item.

Testing costs. No full cost analysis was planned for this study; however, computer costs were available for the three-day data collection. A total of $89.00 was spent over the entire period on the CDC 6500 computer. This total included core memory, central processor, permanent file storage, line printing, and punch card output for 102 subjects. (Data files were punched as they were created to assure that data would not be lost in case of hardware malfunction.)

In the present study, six CRTs were kept on and tied to the computer continuously for 14 hours a day for three days in order to be ready for subject-volunteers whenever they arrived. In any institutional implementation of computer-testing outside the experimental situation, exam time would be scheduled, thus minimizing telephone line transmittal costs.

A large proportion of the total cost cited above resulted from 42 hours of continual tie-in to the computer, the "unnecessary" punching

Table 7

Average Number of Items for Conventional Test and Three Termination Methods of Parallel Forms Stradaptive Tests

	Test 1			Test 2		
		No. of Items			No. of Items	
Type Test	N	Mean	S.D.	N	Mean	S.D.
Stradaptive Method 1	55	31.46	18.03	38	30.92	12.54
Stradaptive Method 2	55	26.94	16.76	41	21.98	13.10
Stradaptive Method 3	55	19.20	14.06	47	18.19	11.34
Conventional	47	48.43	.99			

Table 8

Distributions of Item Latency Between Conventional and Stradaptive Groups

		Seconds Per Item	
Group	No. Items	Mean	S.D.
Conventional	2276	35.999	12.062
Stradaptive	1730	40.047*	13.219

$p < .05$

out of all data, and the extensive file manipulations done by the author because direct access space became critically short during data collection. The latter factor required restorage of data files from direct to indirect file space.

This cost approximation could be compared with testing costs from the reader's experience. Without trying to define conventional testing costs per se, there is still little doubt that computer-based testing is competitive with conventional testing using the paper and pencil mode for any large-scale testing program.

Conclusions and Implications for Future Research

The results of this study favor further investigation of the stradaptive testing model. The model produced consistently higher validity coefficients than conventional testing with a significant reduction in the number of items from 48 to 31, 25 and 19 for the three stradaptive termination rules investigated in the study. The internal consistency reliability for the best stradaptive scoring methods was significantly higher than the conventional KR-20 estimate, and the stradaptive parallel-forms reliability estimates were consistently higher than conventional KR-20 estimates.

No prior research was found showing a comparison of item latency data between adaptive and conventional testing modes. Results in this study clearly indicate that subjects take significantly longer to answer items adapted to their ability level, about 11% longer in the present study. This is an important result, since it indicates that future research into adaptive testing of any kind should consider this variable when evaluating an adaptive test strategy. The net gain of the adaptive model is a function of the testing time needed to adequately measure a subject's ability, not the number of items presented to the subject. All prior research reviewed (See Waters, 1974, for review) tacitly assumed that item latency was consistent across testing strategies. This study indicated this assumption to be false.

As suggested in previous research (Lord, 1970), adaptive testing may reach "peak" efficiency at 15 to 20 items. A comparison of stradaptive test statistics for example with $k = 10$, 15, 20 and 25 items with conventional testing should investigate this hypothesis. Once the stradaptive data are collected under the variable strategy, the fixed item statistics can be determined by scoring the stradaptive test after "k" items and then "starting" the subject's second

test at the first item of the entry point level.

Research is indicated with comparisons between adaptive models as well as the traditional design of comparing adaptive methods with conventional methods. The traditional comparison assumes that conventional test statistics are the criterion that an adaptive testing procedure should try to duplicate. Lord (1970), Green (1970), Weiss (1973), and others have argued that improved measurement of the individual at all ability levels may be hidden by the use of classical test statistics such as validity and even reliability.

Green (1970, p. 194) stated that the computer has only begun to enter the testing business, and that as experience with computer-controlled testing grows, important changes in the technology of testing will occur. He predicted that "most of the changes lie in the future . . . in the inevitable computer conquest of testing." The stradaptive testing model appears to be one such important change.

References

Cleary, R. A., & Linn, R. L. A note on the relative sizes of the standard errors of two reliability estimates. *Journal of Educational Measurement*, 1969, *6*(1), 25-27.

DeWitt, L. J., & Weiss, D. J. A computer software system for adaptive ability measurement. *Research Report 74-1*, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974. (AD 773961).

Feldt, L. S. The approximate sampling distribution of Kuder-Richardson Reliability Coefficient Testing. *Psychometrika*, 1965, *30* (3), 357-370.

Green, B. F., Jr. Comments on tailored testing. In W. H. Holtzman (Ed.), *Computer-Assisted instruction, testing and guidance*. New York: Harper & Row, 1970.

Hoyt, C. Test reliability estimated by analysis of variance. *Psychometrika*, 1941, *6*, 153-160.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-Assisted instruction, testing, and guidance*. New York: Harper & Row, 1970.

Lord, F. M. Individualized testing and item characteristic curve theory. Educational Testing Service, ETS-RB-72-50, Princeton, N. J., November 1972.

Lord, F. M. & Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

McBride, J. R., & Weiss, D. J. A word knowledge item pool for adaptive ability measurement. *Research Report 74-2*, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974. (AD 781894).

SCAT Series II, *Cooperative school and college ability tests*, Princeton: Educational Testing Service, 1967.

Stanley, J. C. Reliability. In R. I. Thorndike (Ed.), *Educational Measurement*. Washington, D. C.: American Council on Education, 1971.

Terman, L. M., & Merrill, M. A. *Measuring Intelligence*. Boston: Houghton Mifflin, 1937.

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

Waters, B. K. An empirical investigation of the stradaptive testing model for the measurement of human ability. Air Force Human Resources Laboratory/Technical Training Division, TR 75-27, March, 1975.

Weiss, D. J. The stratified adaptive computerized ability test. *Research Report 73-3*. Psychometric Methods Program, Department of Psychology, University of Minnesota, September 1973. (AD 768376).

Weiss, D. J. Strategies of adaptive ability measurement. Research Report 74-5. Psychometric Methods Program, Department of Psychology, University of Minnesota, December, 1974. (AD A004270).

Acknowledgements

This paper is based on the author's doctoral dissertation conducted at Florida State University under the direction of Dr. Howard W. Stoker. Requests for copies of the dissertation should be sent to the author at the address below. Test materials from SCAT Series II Verbal Ability tests were adapted and used with the permission of Educational Testing Service. The author gratefully acknowledges the help of ETS in the pursuit of this research.

Author's Address

Brian K. Waters, AFHRL/FT, Williams AFB, AZ, 85224.