AN EMPIRICAL INVESTIGATION OF WEISS' STRADAPTIVE TESTING MODEL

BRIAN K. WATERS
U.S. Air Force Human Resources Laboratory

This study[1] investigated the validity and utility of the stratified adaptive ("stradaptive") computerized testing model proposed by Weiss and colleagues in the Psychometric Methods Program, University of Minnesota. Weiss and his associates have reported the theoretical development of the stradaptive model (Weiss, 1973; DeWitt and Weiss, 1974; McBride and Weiss, 1974) including some examples of individual results. To date, no full empirical studies of the model have been published.

The Stradaptive Testing Model

Lord's 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 ± .5 standard deviations from the mean, the adaptive test requires fewer items for comparable measurement efficiency.

Lord suggests that an "ideal" testing strategy would present a sample of items to each subject comprising a peaked test with a .50 probability of a correct answer for examinees of the particular subject's true ability (P_c = .50). The catch, of course, is that the true ability of the subject is unknown; the estimation of which 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 hypothetical *average* ability level subject. This procedure worked well for examinees near the center of the ability continuum, but less efficiently near the extremes.

Weiss' stradaptive model extends the Binet rationale 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 subject's initial item is based upon a previously obtained ability estimate or the subject's own estimation of his ability on the dimension being assessed.

Figure 1 depicts a nine-strata distribution of items in a hypothetical stradaptive item pool.

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

The Item Bank

Verbal analogy test items were used in this study selected from the SCAT Series II.[2] This test series provided a single-format, unidimensional test with extensively-normed item parameter estimates. The item format was easily stored in a computer item file, being short and standard for all 244 items.

Item pool data received from Educational Testing Service contained five 50-item verbal analogy tests, Forms 1A, 1B, 1C, 2A and 2B of the SCAT Series II examinations. These tests had been nationally normed on a sample 3133 twelfth grade students in October 1966. *P*–values and biserial correlations on 249 items were provided by ETS. These values were transformed into normal ogive item parameters.

Table 1 shows the actual distribution of items used in this experiment. The final pool included 244 items grouped into 9 strata according to normal ogive item difficulty parameters as shown in Table 1.

The nine strata in Table 1 are essentially nine peaked tests, varying in average difficulty from -2.12 to +1.91. Stratum 9, the most difficult peaked test, for example, was composed of 19 items ranging from b_g = 1.27 to b_g=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.

Subject Pool. 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

[1] 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 c/o AFHRL/FT, Williams AFB, AZ 85224.

[2] Test materials from SCAT Series II Verbal Ability tests were adapted and used with the permission of Education Testing Service. The author of this paper gratefully acknowledges the help of ETS in the pursuit of this research.

(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.

Table 2 depicts linear vs stradaptive group test statistics on the 12V scores.

As can be seen in Table 2, the random assignment of subjects to linear or stradaptive testing groups did a good job in equating the groups on the ability continuum as presented.

Testing continued until a subject's ceiling stratum was identified. for this study, the ceiling stratum was defined as the lowest stratum in which 25% or less of the items measured by the Florida 12th Grade Verbal test.

Since SCAT-V published results had shown significantly different difficulty levels between the five forms, linear subtest scores were normalized within their separate distributions and then pooled into a linear total score distribution for comparison with stradaptive results.

CRT Testing

A computer program described by DeWitt and Weiss (1973) was adapted to fit the FSU Control Data Corporation 6500 computer.

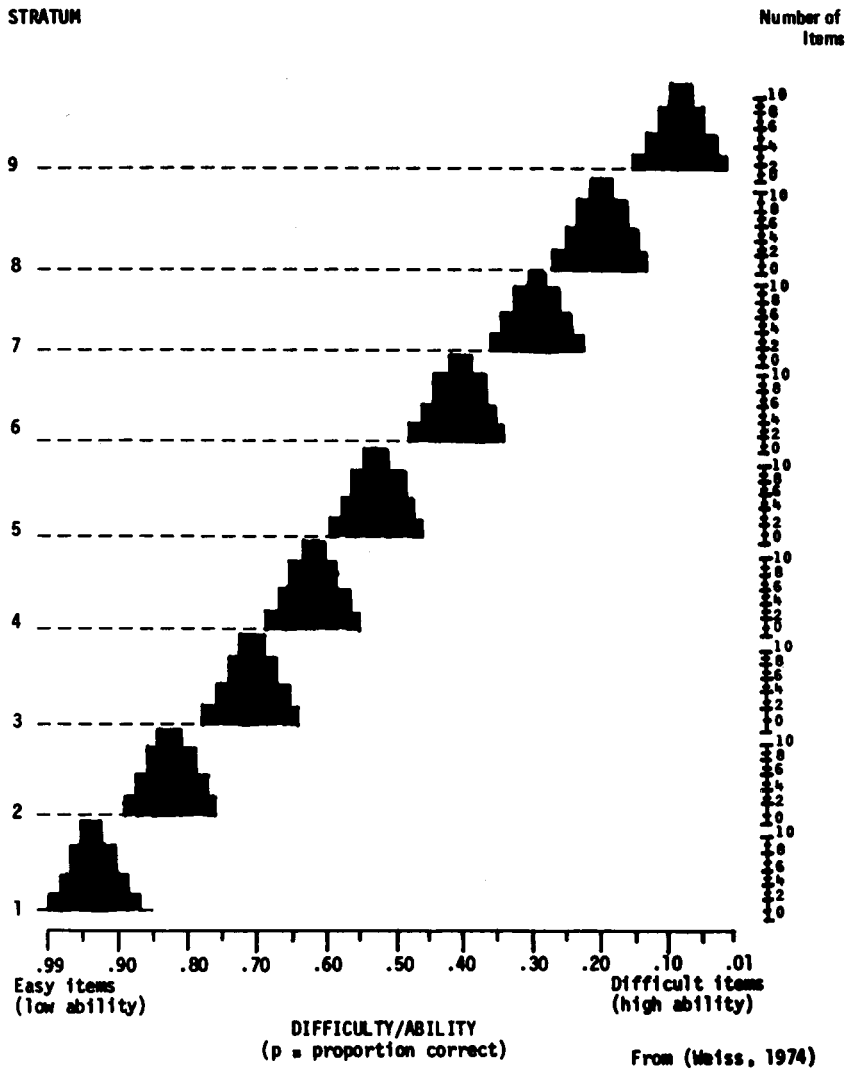

Figure 1. Distribution of items, by difficulty level, in a Stradaptive Test

TABLE 1

Item Difficulties (b) and Discriminations (a), Based on Normal Ogive Parameter Estimates, for the Stradaptive Test Item Pool

	Stratum (easy) 1		2		3		4		5		(difficult) 6		7		8		9	
Item Difficulties																		
High	-1.94		-1.46		- .90		- .49		- .10		.25		.67		1.34		3.68	
Low	-3.57		-1.91		-1.40		- .88		- .44		- .10		.27		.71		1.27	
Mean	-2.12		-1.68		-1.13		- .68		- .25		.04		.44		.95		1.91	
No. of Items	20		26		33		39		31		28		26		22		19	
Item Number Within Stratum	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a
1	-2.08	.48	-1.87	.50	- .90	.56	-.62	.79	- .19	.29	.20	.59	.38	.53	1.25	.46	1.76	.49
2	-1.97	.45	-1.74	.69	-1.05	.83	-.49	.69	- .33	.41	.25	.59	.31	.34	.76	.64	1.69	.44
3	-2.07	.42	-1.70	.95	-1.34	.42	-.72	.77	- .11	.50	.00	.44	.43	.53	1.19	.56	1.61	.44
4	-2.27	.64	-1.91	.52	-1.11	1.17	-.65	.73	- .17	.50	.24	.34	.63	.39	.81	.45	2.91	.49
5	-1.97	.86	-1.50	.77	-1.39	.63	-.88	.49	- .11	.53	.09	.69	.65	.66	1.13	.36	3.69	.28
6	-2.17	.36	-1.79	.59	- .92	.50	-.49	.61	- .16	.33	-.10	.61	.34	.49	.87	.34	1.57	.29
7	-2.31	.41	-1.47	.86	-1.06	.42	-.80	.50	- .16	.83	.09	.71	.30	.71	.71	.48	1.60	.33
8	-2.03	.41	-1.83	.55	-1.31	.44	-.69	.75	- .11	.52	.12	.81	.59	.53	.88	.53	1.34	.42
9	-2.13	.48	-1.68	.58	-1.22	.95	-.55	.52	- .44	.69	.11	.52	.28	.64	.88	.49	1.83	.52
10	-3.57	.30	-1.52	.93	-1.08	.52	-.80	.55	- .42	.55	.00	.39	.38	.52	.79	.61	1.27	.77
11	-2.03	.50	-1.69	.83	-1.19	.79	-.57	.33	- .21	.39	.00	.41	.29	.61	1.24	.30	2.29	.44
12	-2.63	.39	-1.69	.41	- .95	1.01	-.84	.68	- .35	.59	.21	.53	.62	.56	.71	.71	1.33	.33
13	-1.95	.25	-1.65	.71	-1.37	.53	-.86	.83	- .24	.79	.13	.68	.53	.55	.91	.26	1.91	.40
14	-1.95	.56	-1.56	.64	-1.31	.64	-.76	.59	- .10	.55	-.08	.77	.29	.37	1.06	.49	1.27	.42
15	-2.31	.63	-1.90	.69	-1.40	.75	-.54	.46	- .42	.75	.13	.44	.27	.68	1.24	.33	1.91	.27
16	-2.50	.53	-1.51	.88	- .90	.36	-.53	.73	- .41	.66	.00	.71	.56	.45	1.01	.56	2.94	.25
17	-2.03	.50	-1.88	.59	-1.04	.68	-.83	.58	- .16	.83	-.05	.56	.67	.46	.75	.79	1.94	.41
18	-2.36	.61	-1.83	.90	- .97	.81	-.51	.58	- .30	.58	.13	.44	.40	.59	1.34	.37	2.13	.27
19	-1.95	.81	-1.80	.36	-1.09	.68	-.62	.79	- .31	.34	.14	.66	.32	.53	.95	.25	1.33	.37
20	-2.03	.71	-1.55	.61	- .91	.77	-.86	.55	- .31	.45	.05	.64	.30	.73	.75	.66		
21			-1.65	.45	-1.02	.75	-.64	.68	- .18	.68	-.91*	.97	.29	.48	.79	.46		
22			-1.78	.68	-1.18	.46	-.85	.46	- .33	.64	-.06	.50	.66	.64	.94	.53		
23			-1.50	.77	-1.35	.45	-.59	.77	- .35	.69	.12	.77	.37	.66				
24			-1.46	.63	-1.17	.58	-.53	.41	- .18	.48	.06	.50	.56	.70				
25			-1.46	.49	-1.07	.27	-.65	.66	- .44	.52	.10	.55	.50	.68				
26			-1.90	.79	- .95	.66	-.75	.73	- .16	.81	.00	.45	.56	.39				
27					-1.36	.98	-.54	.88	- .23	.49	-.04	.88						
28					-1.27	.71	-.60	1.07	- .19	.44	.07	.36						
29					-1.39	.88	-.74	.61	- .37	.79								
30					- .90	.71	-.61	.64	- .14	.66								
31					-1.30	.69	-.83	.81	- .18	.48								
32					-1.38	.36	-.75	.73										
33					-1.21	.45	-.60	.59										
34							-.88	81										
35							-.77	.48										
36							-.49	.33										
37							-.65	.33										
38							-.76	.40										
39							-.73	.83										

*This item was misassigned to stratum 6 rather than 3. Fortunately, no subjects reached the item in the Stradaptive Pool.

TABLE 2

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

GROUP	# SUBJECT	MEAN	STD DEV	STD ERR	KURTOSIS	SKEWNESS
LINEAR	46	33.26	5.30	.855	.44	.70
STRADAPTIVE	53	34.06	6.12	.841	.36	-.03

P_r (μ lin = μ str)= > .05

P_r (σ^2 lin = σ^2 str) = > .05

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

Testing continued until a subject's ceiling stratum was identified. For this study, the ceiling stratum was defined as the lowest stratum in which 25% or less of the items attempted were answered correctly, with a constraint that at least 5 items be taken in the ceiling stratum. The 25% figure reflects the probability of getting an item right by random guessing on a 4-option multiple choice test. Once a subject's ceiling stratum was defined, the program looped back to the examinee'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, alternate forms were taken by all subjects who reached termination criterion on the first test.

A maximum of 120 items per subject was established, as pre-study trial testing suggested that subjects became saturated beyond this point.

Termination Rules. Weiss had two versions of his stradaptive testing computer program. Version one, which was used in this study, presented another item in the same stratum when a subject skipped an item.

The author of this study was unaware of the existence of the second branching strategy program prior to completion of data collection. However, Weiss' program procedure of ignoring skipped items in determining test termination was questioned. It appeared that valuable information was being lost when the Weiss procedure was followed.

It was reasonable to expect that a subject would omit an item *only* when he felt he had no real knowledge of the correct answer. Thus, investigation of test termination based upon omits counted as wrong answers was judged appropriate.

Weiss had set 5 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 4 would have upon the effectiveness of the stradaptive strategy.

These two questions of the handling of omits and the variation in the constraint on the termination of testing created the following three methods for comparisons:

Termination Method 1:
Omits ignored/constraint = 5 items
Termination Method 2:
Omits = wrong/constraint = 5 items
Termination Method 3:
Omits = wrong/constraint = 4 items

Data was collected using Termination Method 1 and then rescored 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 subject 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 linear 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 "5." 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 constant pattern for items 10 through 19 when she reached stopping rule criterion and her first test terminated.

The testing algorithm then selected the 6th item in stratum 5 (her ability estimate) to commence her second test. (The subject was totally unaware of this occurrence as no noticeable time delay occurred between her 19th and 20th items).

At the conclusion of her 31st item, this subject reached 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 for both tests. The interested reader may gain a more thorough understanding of the scoring methods used in this model by tracing this subject's ability estimate scores through Table 1.

RESULTS AND DISCUSSION

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 linear test, indicating an improved selection of items for the stradaptive subject. Table 3 shows the result of this comparison. It clearly indicates significantly different distributions of test difficulty. The stradaptive test was far more difficult than the linear test, with a smaller variance.

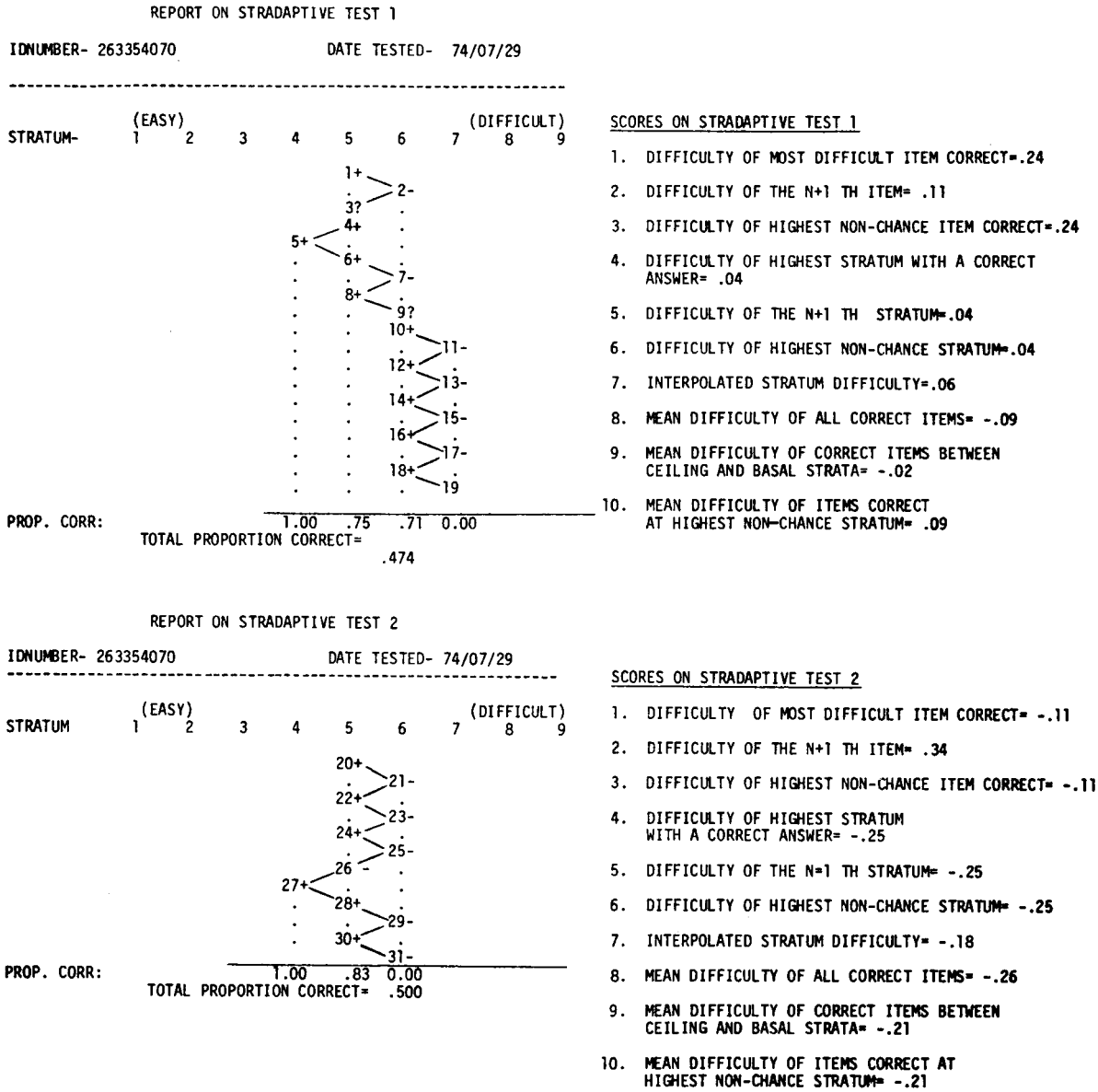

Figure 2. Example of stradaptive testing report.

TABLE 3

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

GROUP	# SUBJECTS	(P_c)	STD DEV	STD ERR	KURTOSIS	SKEWNESS
LINEAR	47	.752	.123	.018	-.87	-.39
STRADAPTIVE	55	.584	.084	.011	5.14	1.97

*P_r (μ Str = μ Lin) = < .0001

**P_r (σ^2 Str = σ^2 Lin) = < .05

Linear Test Reliability. Making the standard assumptions underlying the one factor random effects analysis of variance (ANOVA), the estimated reliability coefficient of the total scores is shown in Table 4 for the linear examinees.

The internal consistency reliability estimate for the linear test was .776 for a test of an average of 48.4 items in 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. Using Feldt's (1965) test, $Pr\,(p_{scat} = p_{lin}) = < .05$.

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

1. Testing mode (CRT vs paper and pencil)
2. Elimination of 6 of the 250 items from the original item pool.
3. Restriction of range in subject pool for this experiment.

The latter 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. It might also be mentioned that Stanley noted that intraclass item correlation is a lower bound to the reliability of the average item.

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.

Weiss' Scoring Method 8 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 summarizes these results.

Table 6 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 linear *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 ($p = < .05$) better than the linear *KR*-20 estimate *prior* to being stepped-up by the Spearman-Brown formula $Pr\ (.675 < p_{20} < .858) = .95$. Thus, the 19, 26, and 31 item stradaptive tests all proved more reliable than the 48 item linear test.

A comparison of the linear internal-consistency reliability coefficients (r_{tx}) and the stradaptive parallel-forms reliability estimates (r_{xx}) in Table 6 must be considered

TABLE 4

Analysis of Variance for Linear Test Person by Item Matrix

SOURCE	df	SUM OF SQUARES	MEAN SQUARES
Persons	46	37.57	.817
Error	2229	408.55	.183
Total	2275	446.12	

$r_{tx\ (lin)} = 1 - .183/.817 = .776$

TABLE 5

Analysis of Variance of Scoring Method 8
of Stradaptive Test Person-By-Item Matrix

TERMINATION		SOURCE	df	SUM OF SQUARES	MEAN SQUARES
	1	Persons	54	191.941	3.555
		Error	1675	588.253	.351
		Total	1729		(r_{20} = .901)
	2	Persons	54	178.870	3.312
		Error	1401	470.442	.336
		Total	1455		(r_{20} = .899)
	3	Persons	54	155.841	2.886
		Error	1001	366.447	.366
		Total	1055		(r_{20} = .873)

only tentatively since they are different kinds of estimates of the true reliability. 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 p. 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 reliabilities, number of subjects, and number of items involved in this experiment.)

Linear Test Validity. The correlation of obtained linear scores with the Florida 12th Grade Scores was .477, which was significantly lower than the published SCAT-V:SAT-v correlation of .83 ($p = < .01$). As with the linear reliability, this difference most likely resulted from subject homogeneity.

Stradaptive Test Validity. The validity coefficients of the stradaptive scoring under the three termination rules is shown in Table 7. Validity was estimated by the correlation between the test scores and 12V scores. None of the validity coefficients in Table 7 were significantly different from the linear validity coefficient of .477, although stradaptive validity coefficients were consistently higher than the linear indices.

Number of Items. Table 8 shows the difference in number of items presented for the linear and the three termination methods of the stradaptive test. The consistency in average number of items presented per subject 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, possibly due to the 60-item limit.

Item Latency. It was hypothesized that mean item latency would be higher for stradaptive subjects since they would have to "think" about each item as it was near the limit of their ability. Table 9 reflects the results of this comparison.

The hypothesis of no differences between item latencies was rejected. For the subjects in this experiment, the average stradaptive item required approximately 11% longer than the average linear 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 (CM), central processor (CP), permanent file storage (MS), data transmittal between the

TABLE 6

Comparison of Scoring Method 8 Parallel Form Reliability
with KR-20 Reliability Over Three Termination Rules Stepped Up to 50 Items

		TERMINATION RULES		
		1	2	3
		(N = 12)	(N = 28)	(N = 38)
Parallel Forms	r_{xx}(raw)	.892	.688	.732
	r_{xx}(50)	.929	.806	.903
		(N = 55)	(N = 55)	(N = 55)
KR-20	p_{20}(raw)	.901	.899	.873
	p_{20}(50)	.935	.943	.947
		$\overline{K}_1$ = 31.45	$\overline{K}_2$ = 26.47	$\overline{K}_3$ = 19.2

$\overline{K}_i$ = average number of items under termination rule 1.

TABLE 7

Comparison of Validity Coefficients of Scoring
Method 8 under Three Termination Rules

Termination Rule	N	r_{cx}	r_{cx}*
1	64	.536	.585
2	80	.536	.693
3	91	.499	.626

r_{cx} = Correlation between criterion measure (12V)

r_{cx}* = r_{cx} corrected for attenuation

CRT's and the computer, line printing (LP), and punch card output for 102 subjects. Data files were punched-out as they were created to assure that data would not be lost in case of hardware malfunction.

In the present study, 6 CRT's were kept on and tied to the computer continuously for 14 hours a day for 3 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.

The cost of actually testing each individual came to less than 2¢ per subject for CM, CP, MS and LP time. The vast majority of the costs cited above involve 42 hours on continual tie-in to the computer, the "unnecessary" punching 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 costs less than conventional testing with 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, as it indicates that future

TABLE 8

Comparison of Average Number of Items for Linear Test and Three Termination Methods of Alternate-form Stradaptive Tests

	# SUBJECTS	AVG # ITEMS	STD DEV # ITEMS	# SUBJECTS	AVG # ITEMS	STD DEV # ITEMS
LINEAR	47	48.43	.99			
		TEST 1			TEST 2	
	# SUBJECTS	AVG # ITEMS	STD DEV # ITEMS	# SUBJECTS	AVG # ITEMS	STD DEV # ITEMS
STRADAPTIVE						
Method 1	55	31.46	18.03	38	30.92	12.54
Method 2	55	26.94	16.76	41	21.98	13.10
Method 3	55	19.20	14.06	47	18.19	11.34

TABLE 9

Comparison of Distributions of Item Latency Between Linear and Stradaptive Groups

GROUP	# ITEMS	MEAN # SEC/ITEM	STD DEV
LINEAR	2276	35.999	12.062
STRADAPTIVE	1730	40.047	13.219

$Pr\ (\mu\ str = \mu\ lin) = < .001$

$Pr\ (\sigma^2\ str = \sigma^2\ lin) = < .001$

research into adaptive testing of any kind should take this variable into consideration 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 tacitly assumed that item latency was consistent across testing strategies. This study indicated this assumption to be false.

It is recommended that future stradaptive experimental studies should consider both stradaptive branching models with a comparison of results from variation in the minimum number of items in the ceiling stratum. A comparison between variable number of stage strategies and fixed number of stage strategies is desirable.

As suggested in previous research, adaptive testing may reach "peak" efficiency at between 15 and 20 items. A comparison of stradaptive test statistics for example with $k = 10$, 15, 20 and 25 items with linear testing should investigate this hypothesis. Once the stradaptive data is collected under the variable strategy, the fixed item statistics can be determined by grading the stradaptive test after "K" items and then "starting" the subject's second test at the first item of the entry point level.

Following the same logic which led to termination of a subject's testing when five items in a row in the highest stratum had been correctly answered, the missing of five items in a row of *any* stratum should provide immediate ceiling stratum definition. The probability of this occurrence would be less than .05 for a properly normed item pool. In the case of the present study, 13 of the 55 stradaptive subjects would have terminated a stradaptive test an average of 12.1 times earlier than termination method 1, with no effect upon the other 42 subjects. The resulting stradaptive test statistics obtained from the implementation of this suggestion have not been calculated, except that the change would have reduced the average number of items presented under termination method 1 to 28.4 from 31.45 (9.7%).

Further research is recommended into adaptive testing in which both the number of stages and step-size are variable. The Bayesian strategies and Urry's model (1970) are examples of this category of adaptive measurement and further model development seems appropriate.

Research is indicated with comparisons between adaptive models as well as the traditional design of comparing adaptive methods with conventional methods. Weiss' ongoing research is beginning this work, but more is needed. The traditional comparison assumes that conventional test statistics are the criterion that an adaptive testing procedure should try to duplicate. Lord, Green, Weiss 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.

One objective of this study was the attempt to estimate the degree to which the violation of the assumptions of the one-factor ANOVA model affected KR-20 reliability estimates. The assumption that items are independent of one another is clearly violated in any adaptive testing procedure. The extent of the effect this violation causes is unknown, yet most previous research in adaptive testing has only considered ANOVA KR-20 estimates.

The results from this study do not permit definitive statements on this question. Nevertheless, the three KR-20 estimates were consistently higher than the 3 parallel-forms reliabilities. Cleary & Linn's (1969) Monte Carlo study indicated that r_{20} provided better parameter estimation than parallel-forms reliability estimates, so one must question whether the higher p estimates are not the result of the dependency between items. Perhaps the only way this question can be validly investigated is through a Monte Carlo study of adaptive testing with p known and the two methods compared, for estimating p.

Green (1970) 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."[3]

The stradaptive testing model appears to be one such important change.

---

[3] Green, B.F., Jr., In Holtzman (Ed.), p. 194.

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,* L 1,25-27.

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

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.* #3, 153-160.

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, University of Minnesota, 1974.

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.

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

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

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