THE STRATIFIED ADAPTIVE COMPUTERIZED ABILITY TEST

David J. Weiss

Research Report 73-3

Psychometric Methods Program
Department of Psychology
University of Minnesota

September 1973

Prepared under contract No. N00014-67-A-0113-0029
NR No. 150-343, with the Personnel and
Training Research Programs, Psychological Sciences Division
Office of Naval Research

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*

University of Minnesota
Department of Psychology

2a. REPORT SECURITY CLASSIFICATION

unclassified

2b. GROUP

3. REPORT TITLE

The Stratified Adaptive Computerized Ability Test

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Technical Report

5. AUTHOR(S) *(First name, middle initial, last name)*

David J. Weiss

6. REPORT DATE September 1973	7a. TOTAL NO. OF PAGES 45	7b. NO. OF REFS 17

8a. CONTRACT OR GRANT NO.

N00014-67-A-0113-0029

b. PROJECT NO.

NR 150-343

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

Research Report 73-3
Psychometric Methods Program

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

Personnel and Training Research
Programs, Office of Naval Research

13. ABSTRACT

The stratified adaptive (stradaptive) test is described as a strategy for tailoring an ability test to individual differences in testee ability. Stradaptive test administration is controlled by a time-shared computer system. The rationale of the method is described as it derives from Binet's strategy of ability test administration and findings concerning peaked tests from modern test theory. The essential elements of stradaptive testing which are considered include the differential entry point, branching rules, and individualized termination criteria. Different methods of scoring the stradaptive test are discussed, as are the implications of individual differences in consistency of test responses within the stradaptive test record. A number of examples of the results of live stradaptive testing are presented and discussed. Implications of additional data derived from stradaptive test response records are considered and related to other psychometric concepts.

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
testing						
ability testing						
computerized testing						
adaptive testing						
sequential testing						
branched testing						
individualized testing						
tailored testing						
programmed testing						
response-contingent testing						
automated testing						
stradaptive test						

Contents

THE STRATIFIED ADAPTIVE COMPUTERIZED ABILITY TEST

Since the development of the first group ability test over a half-century ago, paper and pencil tests have dominated ability testing. Paper and pencil tests, which represent one strategy of measuring human abilities, consist of a limited number of test items organized in a specified manner which are presented to all testees in the same way. Testees proceed through the test items in approximately the order in which they are printed in the test booklet. The paper and pencil test is thus a highly standardized testing strategy which was developed to permit one administrator to test large numbers of testees simultaneously. However, the group paper and pencil test has a number of deficiencies (Weiss & Betz, 1973) which make it desirable to investigate other strategies of administering ability tests.

The availability of time-shared computer systems now makes it possible to implement a variety of new strategies for measuring abilities. Interactive computer systems, in which the testee can be presented with test items by the computer and respond to them on a typewriter keyboard, or by means of a light-pen, permit the psychometrician to develop ways of adapting, or tailoring, test items to each individual's estimated ability level. This is accomplished as a result of the computer's capacity to receive the testee's response to a test item, evaluate that response, consult a pre-determined set of rules to determine the next item to be administered, and to administer the chosen next item. In a time-shared computer system, one computer can administer such adaptive ability tests essentially simultaneously to a large number of testees.

In adaptive testing it is the "pre-determined set of rules" governing the choice of the next test item to be administered that differentiate the various strategies of computerized ability testing. In paper and pencil testing each item is administered in succession whether a testee answers an item correctly or incorrectly. In adaptive testing, choice of the next item to be administered is contingent upon whether the testee's response to a previous item, or a set of previous items, was correct or incorrect. A number of different strategies, or decision rules for choice of subsequent test items, have been proposed to implement adaptive testing (Weiss & Betz, 1973). Among these are two-stage, pyramidal, flexilevel, Bayesian and maximum likelihood approaches for tailoring or adapting a test to individual differences among testees.

While each of these available adaptive testing strategies has its advantages and unique characteristics (Weiss, 1973), logical considerations suggest that additional ways of moving a testee through an item pool might be desirable. This paper proposes one such new method, describes its rationale, and presents some examples based on actual computerized testing.

"Peaked" Ability Tests

A peaked ability test is one in which all test items are very similar in difficulty. In the extreme case of peakedness, an ability test would have all items of the same level of difficulty. Thus, item difficulty would have no variance. Since this ideal condition is rather difficult to achieve in practice, operational peaked ability tests tend to have very low variances of their item difficulties, reflecting a set of test items distributed over a very narrow range of difficulty. The smaller the item difficulty variance, the greater the peakedness. When the range of the distribution of item difficulties in a test approaches the range of ability measured by that test, and there are an equal number of items at each level of difficulty, the distribution of item difficulties is said to be rectangular. Most commercial ability tests have distributions of item difficulties which lie between the extremes of the completely peaked test and the rectangularly distributed ability test. These tests tend to have item distributions which are approximately normally distributed across the ability continuum.

In a series of theoretical papers comparing completely peaked ability tests (i.e., tests composed of items of equal difficulty) with tests "administered" under a variety of adaptive testing strategies, Lord (1970; 1971a,b,c) reached one consistent conclusion: in terms of the precision of measurement, or the capability of responses to a set of test items to reproduce accurately the "true ability" of hypothetical testees, the peaked test always provided more precise measurement than an adaptive test of the same length when the testee's ability was at the point at which the test was peaked. As the testee's ability deviated from the point at which the test was peaked, the measurement efficiency (i.e., the number of test items required to achieve a given degree of precision) of the peaked test diminished more rapidly than that of the adaptive tests. Figure 1 illustrates Lord's general finding in this series of studies. As Figure 1 shows, at some point on the ability continuum, usually plus or minus .50 to 1.0 standard deviations, the efficiency of the adaptive test becomes higher than that of the peaked test.

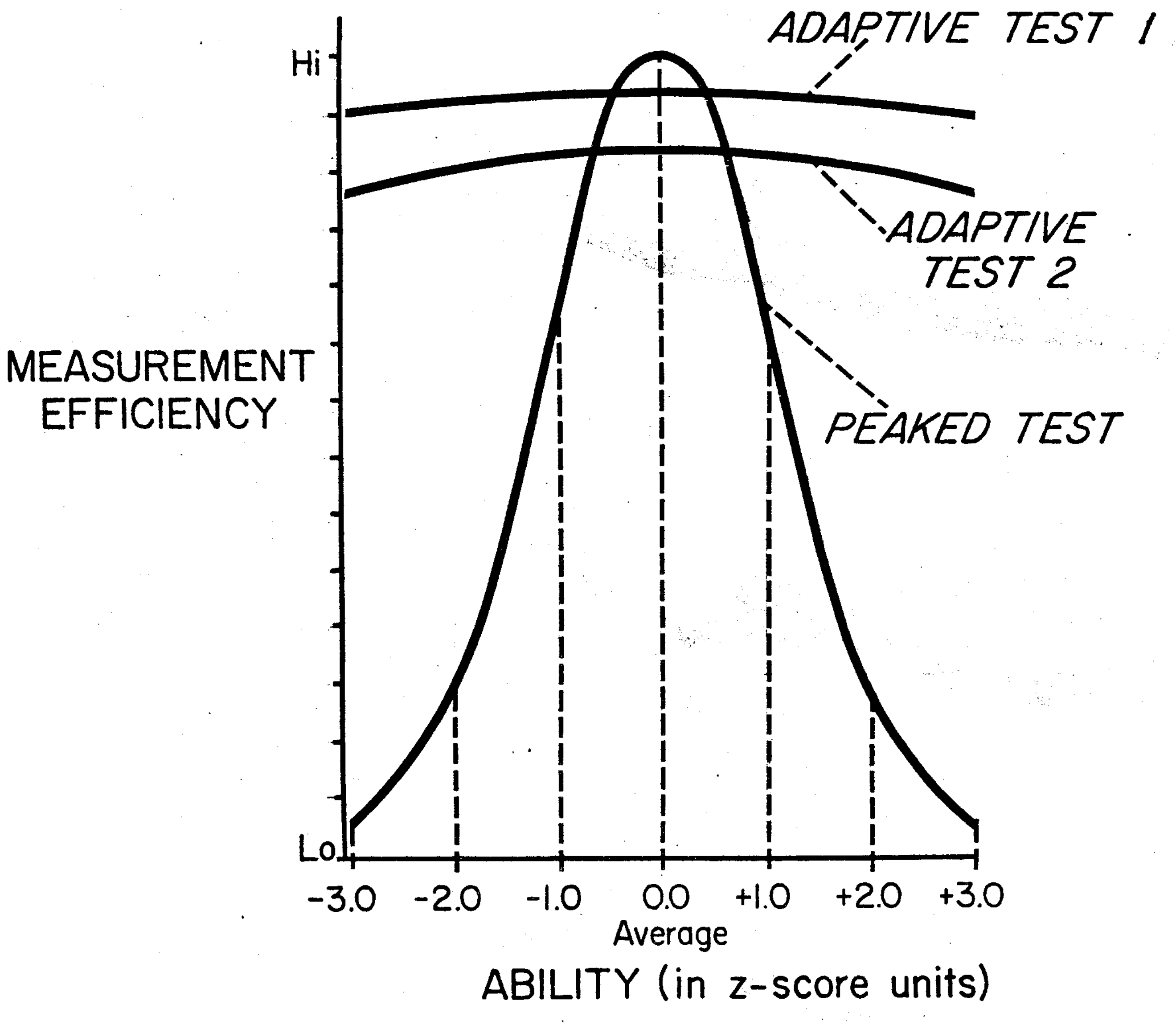


Figure 1. Efficiency of measurement as a function of ability level (after Lord, 1970; 1971a,b,c)

With increasing distance from the peaked point, the adaptive tests become more and more efficient in comparison to the peaked test. However, Lord's theoretical results did show that peaked tests can provide greater measurement efficiency than all adaptive tests studied thus far for up to about 70% of a population normally distributed around the peaked point of the test.

While Lord's theoretical analyses reflect an ideal set of conditions (i.e., all test items are of equal difficulty and equal discrimination), they are important enough not to be easily dismissed. Interpreted in another way, Lord's findings indicate that peaked tests provide most accurate measurement when the ability of the individual being measured is exactly equal to the difficulty level at which the test is peaked. His analysis is supplemented by the findings of information theory (e.g., Hick, 1951) which indicate that test items provide most information when the probability of a correct answer to a given test item is .50 for any individual. Thus, a test comprised of all items of .50 difficulty _for an individual_ would provide the most information about that individual's true ability level, and in Lord's terms, the most precise test score for him.

The important aspect of these findings from both test theory and information theory is that the test must be peaked at the individual's ability level for measurement to be most accurate. But ability level is not known in advance; it is the test's function to measure ability level. The typical solution to this problem is to peak tests at the estimated ability level of some _group_ of testees. Thus, a test designed to measure the abilities of college freshmen is peaked at the average ability level for college freshmen. Since testees always vary in ability, however, the precision of measurement of any individual's ability estimate derived from a peaked test will depend on the distance of his ability from the estimated mean ability of the group, as shown in Figure 1. Thus, the individual whose ability is at the group mean will have a test score of maximum precision. But individuals whose ability deviates from that mean will obtain ability estimates which are less precise, with precision decreasing with increasing distance from the mean. For individuals below the estimated mean ability level of the group, the test items will be too difficult. For these testees the probability of correctly answering the items will be less than .50; the items thus will provide less information on their true ability level. For individuals above the estimated mean ability level, the items will be too easy. Thus, their probability of a correct response will be greater than .50 and again, the test items will provide less information about the ability levels of those testees.

Following the administration of a peaked test, it is possible to tell if the test was appropriate for any given individual. If the test is peaked with items of average difficulty for a group of subjects, the difficulties of the items will be p = .50, i.e., half the group will have answered each item correctly. The appropriateness of that peaked test for any individual can be determined by the proportion of total items taken that he/she has answered correctly. A peaked test can be thought of as being most appropriate for an individual if he gets about half the items correct. Under these circumstances each item provides maximum information on that testee and his score has maximum precision. If an individual answers none of the items on a test correctly (or, if guessing is possible, operates at a chance level) or answers most or all the items in the test correctly, the test was inappropriate for that individual (Lord, 1971c). However, under conventional ability test administration procedures (i.e., paper and pencil tests), the appropriateness or inappropriateness of a test for any given individual can not be determined until after the test has been administered. For many uses of test information, such <u>post hoc</u> determination of appropriateness is too late; the obtained ability estimates may have associated with them very large errors which seriously reduce their utility in practical situations and frequently result in invalid uses of such test scores for practical decisions.

Binet's Testing Strategy

Recognition that a single peaked test may not be appropriate for a given testee seems to have been implicit in Binet's early work in individual testing. That work resulted in the Stanford-Binet Scales (Terman and Merrill, 1960), which are still acknowledged by many as the "standard" of ability measurement. Binet's approach to ability measurement, rather than depending on a single test peaked at the average ability level of the children whose ability it was measuring, used a series of tests organized around the concept of "mental age." Test items at each of the "mental age" levels were peaked around a given mental age, and there was little overlap between mental ages. Items were included in a peaked "mental age" test if about 50% of the norm group of that chronological age gave correct answers to those items. In other words, the items in the test labelled "mental age 8.0", for example, would be those items answered correctly by approximately 50% of those aged exactly 8.0 years who were part of the norm group. A similar rationale was used to construct the tests peaked at each other "mental age" comprising the Binet test. The Stanford-Binet can thus be characterized not as one test but as a series of tests, each peaked at a given mental age and providing most accurate measurement for individuals at that mental age.

Binet's test administration procedure implicitly recognizes that peaked tests which do not permit the testee to obtain about half correct and half incorrect answers provide little information about his ability and therefore should not be administered to him. In administering the Stanford-Binet, the administrator estimates an "entry point" into the hierarchy of mental age peaked tests. The usual entry point consists of that mental age closest to the testee's chronological age; thus, the testee whose chronological age is 8 years, 1 month, will likely start with the test peaked at the 8.0 year level. The administrator is allowed flexibility, however. If it is hypothesized on the basis of prior information that the child is "bright" for his age, the 8 year 1 month child might be started at the 9.0 mental age test; conversely, the child who is expected to be "less bright" might be started at the test peaked at age 6.5.

Following determination of the "entry point" on the scaled peaked tests, the administrator administers the items of the entry-point peaked test and then moves to tests of lesser difficulty. Items are scored as test administration proceeds, with the administrator searching first for the testee's "basal age" and then for his "ceiling age." Binet's basal age is the peaked test at which the individual answers all test items correctly. These data provide no information on an individual's ability except that it is likely not to be lower than that mental age. Thus, it is assumed that if the testee were administered items from tests peaked at mental ages below the obtained basal age, he would provide correct answers to all of those items. If this assumption is correct, those items also will provide no information on the testee's ability level (they would all be too easy), thus nothing would be gained by administering them. The "basal age" therefore defines a "floor" below which further ability testing is unfruitful.

Similarly, the "ceiling age" provides an upper limit beyond which further testing is unnecessary and, in terms of testee motivation (e.g., frustration), might even reduce the accuracy of the test score. The "ceiling age" identifies the peaked test at which the testee obtains all incorrect answers. Like the basal age test, in terms of information theory the test responses provide no information. The ceiling age simply indicates that the individual's ability is somewhere below that level, but it does not indicate where on the ability continuum the individual is likely to be located. It is also assumed that all peaked tests above the ceiling age will likely produce the same results as the ceiling age test, i.e., all responses would be incorrect, and therefore the tests would provide no information on the testee's ability level.

Once the administrator has determined a testee's basal age, testing proceeds through tests of higher difficulty until the ceiling age is identified. It is the peaked tests within the limits defined by the basal and ceiling ages that will likely provide meaningful information on a testee's ability level. The totality of test items between any testee's basal and ceiling ages will provide accurate measurement for that individual; for another testee with different basal and/or ceiling levels a different set of test items will provide maximum information on his ability level. If the test is properly unidimensional for a given individual, and administration conditions are optimal, the proportion correct at each mental age level from the basal age through the ceiling age should show a regular decrease. If there were a very large number of mental age peaked tests between the basal and ceiling ages, proportion correct on these tests would vary from 1.00 at the basal age, through a test on which the individual answers approximately .50 of the items correctly, to .00 correct at the ceiling age. It will be noted that the area between the basal and ceiling ages includes a peaked test (at least theoretically) of maximum measurement efficiency, i.e., a peaked test on which the individual answers 50% of the items correctly.

Assuming that the item pool is relevant for each individual (i.e., they are from the culture on which the test was normed) and that it is unidimensional for each testee, the Stanford-Binet is the only test which has this characteristic--measurement of any individual's ability is confined to that area of the ability continuum which provides, over all test items administered, maximum average information per test item. The Stanford-Binet should, therefore, provide scores of more nearly constant precision of measurement than tests which do not have this adaptive feature--the capability of "searching out" the individual's ability level among a series of scaled peaked tests. Perhaps it is this characteristic of the Binet tests which has made them the standard of comparison for other ability tests.

Thus, by adapting selection and administration of peaked tests to the individual being measured, Binet's concept of ability testing seems to anticipate Lord's later theoretical findings concerning the efficiency of peaked tests. The individual administration of the Binet tests, however, introduces other sources of score variance which attribute error to the measurements obtained (Weiss & Betz, 1973). In addition to the unreliability due to scoring, administrator effects such as sex and race and other characteristics of the administrator and surrounding conditions serve to offset the increases in precision of measurement gained from the adaptive strategy of test administration.

With the current availability of time-shared computers for use as test administration devices, it is now possible to minimize the effects of the administrator variables which affect test scores, and at the same time utilize Binet's insights, with some improvements, in the ability measurement process. The stratified adaptive (STRADAPTIVE) computerized test is proposed as a means of obtaining ability test scores with nearly constant precision across a wide-ranging group of testees, building on the logic of Binet's test administration procedure and implementing Lord's theoretical findings and those available from information theory.[1]

The STRADAPTIVE Test

The stradaptive test, like Binet's testing strategy, operates from a pool of items stratified by difficulty level, or organized into a set of scaled peaked tests. Each testee begins at a difficulty level estimated to correspond to his ability level, also following Binet's strategy. By using any of a number of branching procedures, the stradaptive test moves the testee through items of varying levels of difficulty in search of a region of the item pool which will provide maximum information about his ability level. The branching process leads to the identification of a "basal stratum" and a "ceiling stratum". Testing can be terminated when the ceiling stratum is reached. Each of these characteristics of stradaptive testing is considered below in detail.

[1] The term "stradaptive" is used rather than "stratified" to differentiate this approach from Cronbach's (Cronbach, Gleser, Nanda & Rajaratnam, 1972) conception of stratified tests, which are based on the idea of sampling test items from a stratified universe in which test items are classified by content, task, or difficulty.

Item Pool Structure

The stradaptive test requires an item pool stratified by the difficulty levels of the constituent test items. A stratified item pool is one in which items are organized into a series of tests peaked at different difficulty levels. The pool should be known or assumed to be unidimensional. It will be shown below, however, that unidimensionality of the pool might not be evident for some testees; but the pool should be unidimensional for most testees in order to provide the most constant precision of measurement. The steps in developing an item pool for a stradaptive test include the following:

1. Administer a large number of items measuring the same ability to a large group of subjects. The subjects should be representative of the wide-ranging population for which the stradaptive test is intended. The size of the original item pool will depend on the quality of the items used and the target size of the final stratified item pool. While the optimal size of the stradaptive item pool is yet to be determined, adequate results have been obtained with about 200 items in the final pool. Likewise, no information is as yet available on the required number of subjects in the norming item pool. Naturally, a larger norming group will result in more stable item parameter estimates.

2. Derive item discrimination and item difficulty estimates for the items administered to the norming group. These parameters can be either traditional item parameters (proportion correct, item-total score correlations) or parameters derived from modern test theory using normal ogive item assumptions or logistic item functions (Lord & Novick, 1968). Items with very low discriminations should be eliminated.

3. Organize the item pool into a number of independent strata by difficulty level, where each stratum is in effect, a peaked test of some number of items. There should be no overlap in item difficulties between the strata. The number of strata developed from an item pool, or the number of peaked tests available, depends on the size of the original item pool. The larger the number of strata the more likely the obtained ability tests will have equal precision across a group of testees of wide-ranging ability, since the peaked tests

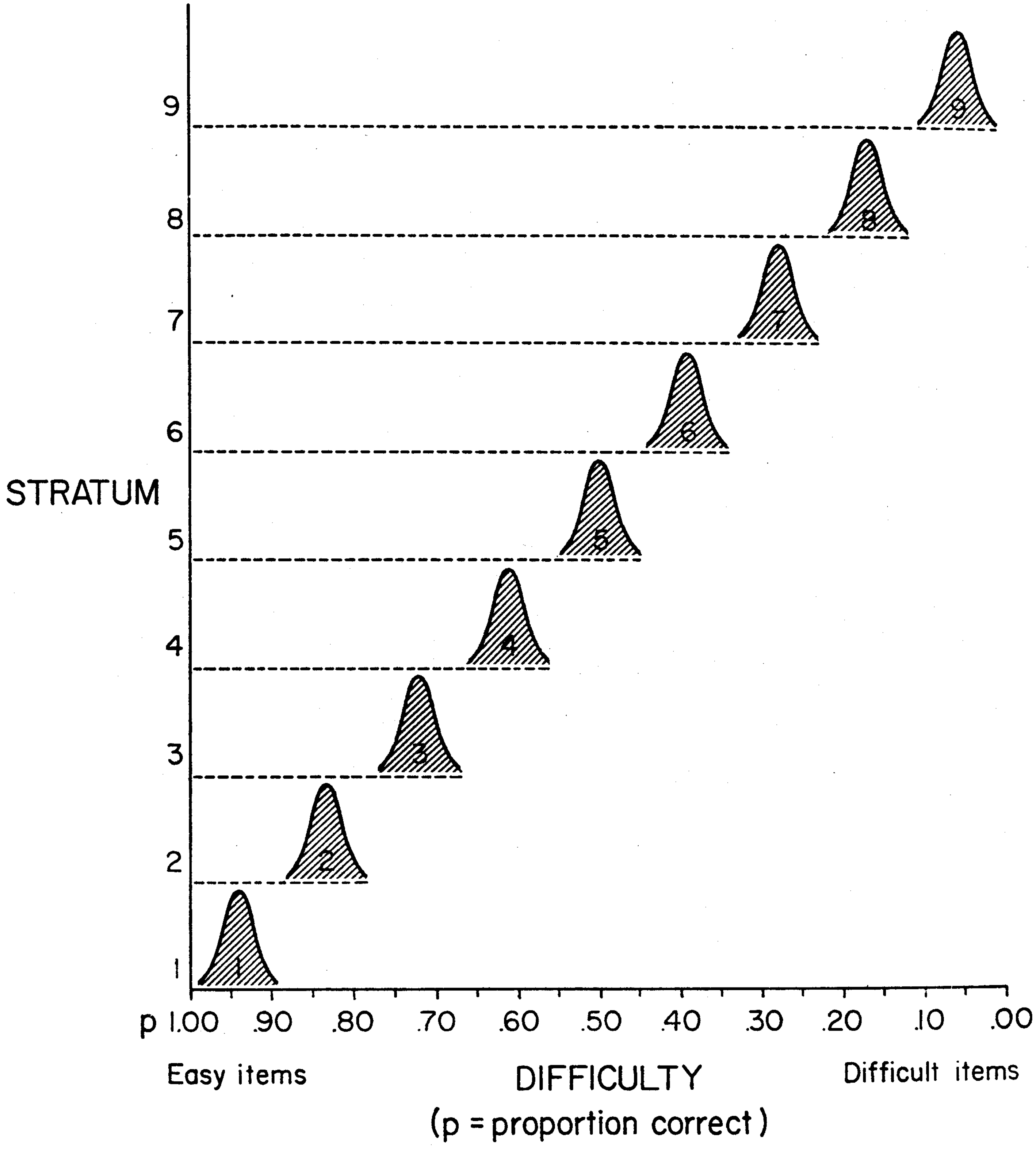


Figure 2. Distribution of items, by difficulty level, in a Stradaptive Test

are more likely to exactly match each testee's ability level. A minimum of nine or ten strata seems to be appropriate, since that number of strata seems to provide a good range of coverage of abilities without requiring very large item pools. The question is, of course, open for considerable further investigation.

The number of items at each stratum will vary with both the size of the original item pool and with the number of strata to be developed. A minimum of ten to fifteen items at any given stratum appears to be appropriate. There need not be an equal number of items at the various strata; experience suggests that the middle and lower difficulty strata might require more items than those at the upper extremes.

4. The items within each stratum should be arranged in decreasing order of item discrimination, if item discrimination indices were derived from analyses on the total norming group, as differentiated from indices computed on sub-groups based on ability levels. Since at the earlier stages of testing (i.e., the first few items at each stratum) items must discriminate across a wider range of abilities, item discriminations based on a group of wide-ranging ability will be more appropriate. On the other hand, at the later stages of testing when testing is confined to only a narrow range of abilities (i.e., within 2 or 3 of the available strata), items need not be able to discriminate on a group of wide-range ability. Rather, item discriminations should be based on discrimination indices derived from closely contiguous levels of ability. Thus, items with relatively low discrimination indices on the total group might be capable of discriminating between contiguous strata at the later stages of testing (Paterson, 1962; Bryson, 1971).

The result of this process of structuring the item pool is shown diagrammatically in Figure 2. The hypothetical stradaptive item pool shown in Figure 2 contains nine strata. Each stratum consists of a subset of items peaked around a different difficulty level, with the difficulty level increasing with each successive stratum. Thus, stratum 1 consists of a sub-set of very easy items distributed approximately normally around a difficulty level of p = .94, with items varying in difficulty from p = .99 to p = .89; stratum 1, therefore, represents a very easy peaked test. Stratum 2 consists of a set of items peaked at a difficulty level slightly higher than those of stratum 1; stratum 2 items are peaked at about p = .83 and vary from p = .88 to p = .78. Stratum 9 is a difficult test with items varying in difficulty from p = .01 to p = .11 and peaked at p = .06. Note that the item distributions in Figure 2 do not overlap between strata.

Table 1 shows an operational stradaptive item pool. The pool consists of 229 items grouped into 9 difficulty strata. The number of items at each stratum varies from 10 at stratum 9 (the most difficult peaked test) to 36 at strata 2 and 3. Items were selected from a larger pool of about 500 items on which normal ogive transformations of item discriminations (a) and difficulties (b) had been previously computed using estimates of Lord's (Lord & Novick, 1968) normal ogive item parameters. To construct the item pool, the range of item difficulties from +3.00 standard deviations to -3.00 standard deviations was divided into 9 equal parts. All items from the larger pool were included in the stradaptive item pool if their normal ogive discrimination parameters were a = .30 or above (with the exception of the tenth item at stratum 9 which was included to increase the number of items at that stratum to 10).[2]

The 9 strata in Table 1 are essentially nine peaked tests varying in average difficulty from -2.65 to +2.62. The most difficult peaked test (stratum 9) is composed of 10 items peaked at b = 2.62, varying from the most difficult item at b = 3.11 to the easiest item in that stratum at b = 2.32. Stratum 8 is a slightly less difficult peaked test with average b = 2.01 and with the 15 items varying in difficulties from b = 2.31 to b = 1.65. Within each stratum items are ordered by discrimination; for stratum 9 the first item has a discrimination of a = .84, and the last item at that stratum has a discrimination of a = .21. Similar patterns are obvious for the other strata. The greater number of items at the middle and lower level of difficulties reflects the composition of the original item pool from which these items were selected. However, in actual testing with the stradaptive test it has become evident that successful testing for many subjects requires the availability of a larger pool of items at the middle and lower ranges of difficulty.

Operationalizing the Stradaptive Test

Entry point. The stradaptive test permits the use of differential entry points for beginning testing for different individuals. While it is not necessary to use

[2] A further exception is item 19 at stratum 4, which has a discrimination of .27; that item was included in the pool by error.

Table 1

Item difficulties (b) and discriminations (a), based on normal ogive parameter estimates, for an operational Stradaptive Test item pool

	(easy)							Stratum								(difficult)	
	1		2		3		4		5		6		7		8		9	
Item Difficulties																		
Hi	-2.39		-1.64		-1.01		-0.34		0.33		.98		1.63		2.31		3.11	
Lo	-2.98		-2.32		-1.63		-1.00		-0.28		.34		1.00		1.65		2.32	
Mean	-2.65		-1.92		-1.29		-0.63		.02		.65		1.33		2.01		2.62	
No. of items	35		36		36		30		25		19		23		15		10	
Item Number Within Stratum	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a
1	-2.42	3.00*	-1.99	1.76	-1.51	1.40	-.70	1.82	-.05	1.31	.73	.98	1.07	.72	1.89	.85	2.95	.84
2	-2.42	3.00	-1.78	1.54	-1.23	1.35	-.73	.92	.14	1.07	.34	.91	1.49	.62	2.03	.64	2.47	.48
3	-2.45	3.00	-2.22	1.52	-1.08	1.23	-.52	.86	-.13	.98	.65	.77	1.33	.60	1.93	.57	2.61	.43
4	-2.45	3.00	-1.68	1.46	-1.33	1.16	-.68	.86	.15	.97	.79	.70	1.54	.58	2.31	.54	2.86	.42
5	-2.72	3.00	-1.87	1.43	-1.34	1.02	-.59	.83	-.08	.91	.79	.63	1.11	.56	1.79	.50	2.35	.42
6	-2.72	3.00	-1.92	1.23	-1.10	.99	-.75	.82	.16	.86	.49	.56	1.40	.55	2.04	.49	2.67	.42
7	-2.72	3.00	-1.88	1.14	-1.42	.92	-.57	.77	-.21	.86	.42	.55	1.17	.52	1.79	.49	2.32	.38
8	-2.66	1.79	-2.13	1.10	-1.21	.91	-.85	.75	-.25	.86	.98	.52	1.30	.52	1.88	.45	2.37	.34
9	-2.54	1.59	-1.64	1.08	-1.06	.89	-.47	.71	.21	.86	.37	.50	1.38	.51	2.07	.43	3.11	.32
10	-2.81	1.48	-2.22	1.07	-1.34	.89	-.40	.68	.16	.83	.46	.49	1.44	.49	2.13	.42	2.50	.21
11	-2.46	1.29	-1.67	1.02	-1.31	.87	-.90	.67	-.23	.81	.46	.48	1.31	.44	2.31	.40		
12	-2.78	1.26	-1.71	.99	-1.10	.77	-1.00	.67	.30	.78	.65	.48	1.25	.43	1.65	.39		
13	-2.47	1.16	-2.26	.98	-1.55	.77	-.69	.66	.08	.76	.78	.45	1.00	.42	1.82	.36		
14	-2.43	1.01	-2.21	.96	-1.07	.76	-.81	.66	-.28	.75	.71	.44	1.00	.40	2.26	.35		
15	-2.86	1.01	-1.66	.93	-1.43	.76	-.56	.66	.24	.66	.65	.43	1.26	.39	2.18	.34		
16	-2.94	.96	-1.65	.92	-1.40	.75	-.58	.66	.33	.53	.62	.41	1.36	.37				
17	-2.83	.94	-1.65	.82	-1.15	.73	-.84	.65	-.23	.43	.83	.37	1.24	.36				
18	-2.74	.93	-2.32	.80	-1.42	.71	-.85	.64	.09	.43	.75	.37	1.60	.35				
19	-2.89	.91	-1.80	.77	-1.63	.71	-.41	.27	.15	.42	.92	.37	1.21	.35				
20	-2.54	.88	-1.80	.76	-1.47	.67	-.94	.60	-.09	.41			1.47	.34				
21	-2.55	.79	-1.93	.74	-1.60	.66	-.41	.59	-.26	.40			1.61	.34				
22	-2.81	.74	-2.28	.70	-1.33	.62	-.89	.53	.08	.39			1.63	.32				
23	-2.50	.68	-1.83	.66	-1.04	.58	-.52	.48	.09	.37			1.36	.31				
24	-2.82	.67	-1.74	.63	-1.17	.57	-.58	.48	-.04	.35								
25	-2.54	.67	-1.70	.59	-1.27	.56	-.39	.40	.12	.32								
26	-2.50	.66	-2.19	.56	-1.07	.56	-.36	.40										
27	-2.51	.64	-1.89	.52	-1.02	.54	-.58	.40										
28	-2.39	.62	-2.20	.50	-1.01	.52	-.38	.38										
29	-2.58	.57	-1.71	.47	-1.31	.52	-.34	.32										
30	-2.98	.56	-2.21	.44	-1.30	.52	-.67	.30										
31	-2.73	.52	-2.08	.42	-1.19	.52												
32	-2.77	.50	-1.80	.42	-1.57	.49												
33	-2.68	.48	-1.82	.42	-1.26	.44												
34	-2.56	.44	-2.12	.41	-1.59	.38												
35	-2.95	.41	-1.92	.32	-1.35	.34												
36			-1.84	.31	-1.08	.32												

*Discriminations (a) were arbitrarily set to 3.00 when the biserial item-test correlation was .90 or higher.

differential entries, i.e., all testees can begin with the same test item, the differential entry point has at least two major advantages. First, beginning testing at different strata for different individuals might save time in testing in terms of the number of items administered to a given individual. Thus, if it is known or suspected that a given testee is likely to be high on the ability to be measured, say 1.5 standard deviations above the mean, it would be wasteful of the testee's time to begin testing with an item of average difficulty. Use of a differential entry point for this individual might save time by eliminating the administration of three or four unnecessary items. The time saving would increase as the individual's estimated ability deviated from an arbitrary fixed entry point.

The second major advantage of using a differential entry point for beginning testing involves the testee's motivation to continue testing or to do well. Beginning an individual of low ability at an item of median difficulty will almost insure that the first several items taken will be too difficult for him; a frustration or anxiety reaction might occur which could adversely affect his performance on the remainder of the test items. Conversely, administering items of median difficulty to an individual of high ability might cause a boredom or "irrelevance" reaction which could then affect his performance on the entire test.

It thus appears to be desirable to begin the stradaptive test at some point estimated to be approximately representative of the individual's ability level on the trait being measured. Two sources of entry point estimates are possible. First, the computer could have stored information on an individual which might be useful as entry point information. For example, if the stradaptive test is being used to measure verbal ability, such information as scores on other verbal ability tests, grades in English courses, grade point average, or simply number of years of formal schooling completed could be stored in the computer. Once the testee identifies himself to the computer by name or identification number, the computer would retrieve the appropriate information from his file and, based on known or estimated relationships between the prior information and test performance, determine the entry point on the ability continuum for that testee.

The testee himself is a second important source of entry point information. Rather than consulting actual records on the testee, it might be fruitful to ask testees for the information necessary to derive entry points.

Figure 3 shows two such entry point questions currently in use for stradaptive testing of verbal ability. The top half of Figure 3 is an entry point question for use with college students. In constructing the entry point estimate it was assumed that college grade point average (GPA) had a roughly positive and linear relationship with verbal ability. Individuals who answer in the first category, 3.76 to 4.00, enter the stradaptive test at stratum 9; individuals who indicate that their GPA's are between 2.51 and 2.75 enter the stradaptive test at stratum 4.

The bottom half of Figure 3 shows a different entry point question asked of the testee. This entry point information was developed for use with a group of inner-city high school students who could not be assumed to know their GPA and might also prove to be useful in a non-school testing situation. It is based on the assumption that the testee has a fairly good knowledge of his level of ability in comparison to his peers. Whether or not the testee can make a good estimate of his ability can be determined by the results of the stradaptive testing. The only effect of a poor estimate of a testee's entry point is that he will be administered a few more test items than would otherwise be necessary to measure his ability adequately. In any case, the stradaptive test is designed to converge upon the testee's level of ability regardless of the adequacy of the entry point. Thus, entry point information need only be very roughly related to the ability being measured.

Branching. The stradaptive test permits the use of virtually any branching rule for moving from an item at one stage to one at the next. Branching in the stradaptive test occurs between strata, therefore no pre-determined item branching network exists for the stradaptive test. The simplest branching rule is an "up-one/down-one" procedure. If a testee answers an item correctly, he is routed to an item at the next more difficult stratum; if he answers incorrectly he is routed to an item at the next easier stratum of difficulty. Other branching rules are also possible. For example, a correct response can lead to an item one stratum higher in difficulty, while an incorrect response can branch downward two strata. Such a rule might be adopted either where the opportunity for guessing may allow the testee to answer a number of items correctly solely by chance, or where it is desired to administer a very easy item (with a high probability of a correct answer for a given individual) following an incorrect response in order to prevent the testee from becoming discouraged.

Figure 3

Stradaptive Test Entry Point Questions

College Students

In which category is your cumulative GPA to date?

		Entry Stratum (not seen by student)
1.	3.76 to 4.00	9
2.	3.51 to 3.75	8
3.	3.26 to 3.50	7
4.	3.01 to 3.25	6
5.	2.76 to 3.00	5
6.	2.51 to 2.75	4
7.	2.26 to 2.50	3
8.	2.01 to 2.25	2
9.	2.00 or less	1

Enter the category (1 through 9) and press the return key.

Non-College Students

Everybody is better at some things than others.... Compared to other people, how good do you think your vocabulary is?

		Entry Stratum (not seen by testee)
Better than:	1 out of 10	1
	2 out of 10	2
	3 out of 10	3
	4 out of 10	4
	5 out of 10	5
	6 out of 10	6
	7 out of 10	7
	8 out of 10	8
	9 out of 10	9

Type in the number from 1 to 9 that gives the number of people you are better than (in vocabulary).

If it is desired to obtain a fairly quick estimate of the testee's "ceiling stratum" (i.e., the stratum at which he gets all items incorrect) the tester might use different branching rules at different stages of testing. At the earlier stages of testing, he might use an "up-two/down-two" rule in order to more quickly arrive at a narrower range of strata in which the testee's ability is likely to fall. Then, after perhaps the tenth stage of testing (i.e., ten items have been administered), the tester might adopt an "up-one/down-one" procedure which would concentrate item administration within the narrower range of strata (e.g., 2 or 3) estimated to include the testee's actual ability level.

The stradaptive test also allows for differential response option branching, as suggested by Bayroff (Bayroff, Thomas & Anderson, 1960). In this procedure, incorrect response alternatives in a multiple choice (or, for that matter, a free-response) test are graded in terms of the extent to which they show partial knowledge. A correct response always leads to the same upward branching decision. When an item is answered incorrectly, the step size of the downward branch (i.e., the number of strata branched over) is a function of the "incorrectness" of the chosen distractor. For example, a "very wrong" answer (e.g., a response given only by testees of very low ability) might lead to a downward branch of three steps; a response which is closer to being correct might result in branching two strata downward; while choice of the most plausible incorrect answer would branch the testee only one stratum down in difficulty. Such differential response option branching should permit more rapid identification of an individual's actual ability level, leading to a reduction in the time needed for the assessment of a particular ability.

For individuals whose abilities are at or near the highest or lowest stratum in the stradaptive item pool, there may be instances where items at higher or lower difficulty strata will not be available. In these cases, it will be necessary to administer successive items at the same stratum in place of the optimal items at higher or lower strata.

<u>Termination</u>. A unique feature of the stradaptive test is its individualized termination rule. In contrast to two-stage tests, all the pyramidal models, and the flexilevel test (see Weiss & Betz, 1973, for research on these strategies, and Weiss, 1973, for detailed descriptions of each), all of which administer a fixed and pre-determined number

of items to each individual testee, the stradaptive test permits the number of items administered to each testee to vary. While both Owen's (1969, 1970) Bayesian adaptive testing strategy and Urry's (1970) maximum likelihood strategy do permit an individualized number of test items, both of these strategies require restrictive assumptions about the hypothesized shape of the underlying ability distribution, and necessitate sophisticated mathematical calculations which might be difficult or time-consuming to implement on some computer systems. The stradaptive test, while retaining the individualized number of items, makes no assumptions about the shape of the ability distribution and requires no complex calculations.

As indicated above, the stradaptive test can be conceived of as a search for the peaked tests most appropriate for an individual testee. These peaked tests, which provice maximum information on a testee's ability level, can be identified, after the fact, as tests on which the testee answered about 50% of the items correctly, if guessing is not a factor. A peaked test is inappropriate if the testee answers all items correctly or all items incorrectly. Thus, the objective of the stradaptive test is to locate the region of the item pool in which measurement efficiency will be maximum for any individual.

This objective can be realized by a simple accounting procedure. Regardless of the branching rules used, the computer simply keeps track of 1) the number of items administered at each stratum and 2) the number of items answered correctly at that stratum. After each item has been answered, the ratio of these two values, or the proportion correct at each stratum, is computed. Prior to administering the next item, the termination criterion is checked to determine whether it has been met. If the criterion has been met, testing is stopped and the individual's response record is scored. If not, an additional item is selected using the branching rules previously chosen for testing. That item is administered and scored, the proportion of items correct at each stratum is computed, and the termination criterion again checked. Testing continues until the termination criterion is met.

One logical criterion for terminating stradaptive testing involves identifying the lowest (i.e., easiest) stratum at which the individual is answering at a chance level. Thus, the stradaptive test can be viewed as a search for the testee's "maximum" level of performance on that set of test items. In a multiple choice test the chance level is determined by 1/c, where c is the number of response choices in each test item. Thus, for 5-alternative multiple choice items, answering 1 (or zero) out of

5 items correctly at a given stratum would indicate chance responding. Using such a termination rule, then, testing would continue until a stratum is identified at which the testee has responded at chance or below, provided that, say, five items have been administered at that stratum. The last condition is necessary to avoid the situation where a testee answers the first one or two items at a given stratum incorrectly, but would answer correctly well above chance levels if administered enough items at that stratum. Variations in the minimum number of items required at any stratum before the proportion correct is used to check the termination criterion will probably result in stradaptive test scores with varying degrees of precision and stability. For example, requiring a larger number of items will probably result in fewer inappropriately early terminations, while decisions made on smaller numbers of items within a stratum might result in some artifactually early terminations after which further testing may have led to higher ability scores.

Conceptually, then, the tester can control the degree of precision of the ability estimates derived from stradaptive testing by manipulating the termination criterion in one of two ways. First, he can require that a larger number of items be administered at the ceiling stratum before the termination criterion is evaluated for an individual. Secondly, the tester can directly manipulate the confidence level of the termination decision. This can be accomplished by directly positing an hypothesis of a proportion of correct responses of, say, p = .20. The obtained proportion of correct responses (for any specified number of items) at a given stratum can then be tested against the hypothesized value by standard hypothesis test-procedures. This would involve either a binomial expansion given p, q and N (the number of items administered), or the computation of a confidence interval around the obtained proportion of correct responses using the same parameters. The alpha value associated with the test of hypothesis, or the confidence level of the confidence interval, could be chosen in advance by the tester as a way of controlling the precision of the obtained ability estimate. Testing would then continue until the data at any stratum failed to reject the hypothesis of chance responding (e.g., p = .20), or until the computed confidence interval included the hypothesized chance value. As the number of test items at the termination stratum increased, the power of the statistical test would also increase, thereby likely increasing precision of measurement and such practical criteria as test-retest stability of the ability estimates.

The proposed termination rule is applicable to multiple choice test items with a constant number of response choices, to true false test items, and to free-response test items. For four-choice test items, the pseudo-chance level is .25, for seven-choice items it is 1/7 or .14, and for true-false items it is .50. For free-response items, the termination criterion becomes the lowest stratum at which the individual answers no items correctly. Thus, when guessing can be completely ruled out, the stradaptive test would continue as long as an individual gets any items correct at strata of increasing difficulty. This termination criterion is identical to Binet's "ceiling age."

Implementation of the "lowest chance stratum" termination rule yields interesting results in actual stradaptive testing with an "up-one/down-one" branching rule. In general, for the majority of individuals these procedures identify a "basal stratum", i.e., a stratum at which all items are answered correctly, and a "ceiling stratum", i.e., the least difficult stratum at which the testee responds at a chance level. In between these two limiting strata, the proportion correct on each stratum will vary between 1.00 and the chance level (.20 or less) and will decrease fairly systematically from the basal to the ceiling stratum. This pattern is evident even when a relatively small number of items has been administered. Specific examples will be given below.

For some individual testees, inconsistency in their response records will occasionally cause the stradaptive pool to exhaust the supply of test items at some stratum. Thus, for a variety of reasons (e.g., motivation, fatigue, inappropriateness of the item pool for that testee), some individuals will fail to reach a termination criterion at a given stratum before exhausting the item pool at that stratum. When this occurs, the branching procedure can be modified to eliminate downward branching but to continue upward branching. Thus, following a correct response the testee would be presented with an item at the next higher stratum, but following an incorrect response an item at the same stratum would be administered if the next lower stratum is exhausted. This procedure will lead to a very rapid identification of the testee's ceiling stratum, at the expense of the probable positively reinforcing value of alternating difficult and easier test items.

Scoring

Since the stradaptive test adapts item presentation to characteristics of the individual being tested, the "number correct" score used almost universally for conventional tests is inappropriate. Number correct is inappropriate because the number of items administered to each individual will vary; some individuals reach termination in 11 or 12 items, while others require 30 or 40 items to safisfy the termination criterion. It might be expected, therefore, that determining the proportion of items correct for any testee would be an appropriate method of scoring the stradaptive test. Computing the proportion correct would account for individual differences in the number of items administered yet convey the same information as the number correct score.

However, this reasoning fails to take into account the fact that in the stradaptive test, item difficulties are tailored to the individual's ability level through the branching procedure. The end result of the branching procedure is to identify a subset of items on which the individual obtains about 50% correct responses. In the later stages of stradaptive testing, when the testing procedure begins to converge on an individual's ability level, each time an item is answered correctly the testee receives a more difficult item (at the next higher stratum). Because that item is likely to be too difficult for him, he will probably answer it incorrectly and will therefore receive an easier item. Since he is likely to get that item correct, the process will be repeated and the testee will approximately alternate between easier items and more difficult items until the termination criterion is reached. The proportion of items correct for an individual will, therefore, center around .50, with deviations from .50 due to inappropriate entry points, unusual testee-item pool interactions, guessing, or an item pool of inappropriate difficulty. Actual stradaptive testing results for over 300 testees show that the large majority of proportions correct vary from .40 to .60.

Since the number correct scores and their derivatives are inappropriate for stradaptive tests, new methods of scoring must be developed. Some methods that might prove satisfactory are suggested by the available research on pyramidal adaptive testing models (see Weiss & Betz, 1973, p. 20-35). Because of some similarities between the stradaptive models and the pyramidal tests (Weiss, 1973) some of these scoring methods can be applied to stradaptive testing. Other scoring methods are suggested by the logic of the stradaptive test itself, as it derives from Binet's approach to ability measurement. .

Following are a number of ways stradaptive tests can be scored. Most scoring methods assume that normal ogive difficulty parameters, or estimates thereof, have been computed for the items of the stradaptive test so that item difficulty data are on the same latent scale as ability estimates; in this way, item difficulties can be used to estimate the ability of persons correctly answering subsets of items. In using these parameters it is assumed that the items in the stradaptive item pool measure a single unidimensional continuum.

Highest item difficulty scores. These scoring methods are borrowed from the pyramidal testing models (e.g., Paterson, 1962; Bayroff & Seeley, 1967; Lord, 1970). They are all based on the "hurdle" conception of ability measurement; that is, the individual's ability level can be determined from the "height of the highest hurdle he can jump." The difficulty of an item is equivalent to the height of the hurdle; answering an item correctly implies jumping the hurdle. There are three variations of this score possible in the stradaptive test, with the third being unique to stradaptive testing:

1. Ability can be scored as the difficulty of the most difficult item answered correctly.

2. Since testing always terminates at an item at the ceiling stratum, ability can be measured as the difficulty of the "n+1th" item, or the item that would have been administered next if testing had not terminated. Thus, the individual who answers his final (nth) item correctly would obtain a higher ability estimate than the testee who answers the nth item incorrectly.

3. An individual's ability score can be conceived of as the difficulty of the most difficult item answered correctly below the testee's ceiling stratum.

A major weakness of these "highest item difficulty" scores is their probable unreliability, in terms of test-retest stability, if guessing is possible. Since in a multiple choice test it might be possible for a testee to obtain a correct answer above his true ability level solely by chance, the first two of these scoring methods would probably be unreliable. Method 2 would probably yield scores of somewhat lower reliability than method 1 since guessing would be more likely to occur on items at the testee's ceiling stratum. Method 3 is suggested as an alternative unique to the stradaptive test when guessing is expected to operate; since method 3 attempts to minimize the effects of chance successes, its results should be more stable than those of methods 1 or 2. When guessing is not possible, i.e., on free-response items, methods 1 and 3 will give similar results. Method 2 results will vary as a function of the adequacy of the termination rule.

Stratum scores. As indicated above, the stradaptive item pool can be considered to be a series of peaked tests graded in difficulty. Associated with each peaked test is a difficulty level, which can be characterized by the average difficulty of all items at a given stratum. That average diffculty level indicates the point on the underlying ability continuum at which each peaked test is peaked. It can, therefore, be used as an ability estimate for individuals in several ways, following the logic of scoring methods 1 through 3:

4. An individual's score is the difficulty level associated with the most difficult stratum at which he answered at least one item correctly.

5. The stradaptive test score can be determined from the difficulty level of the stratum of the n+1th item.

6. Test score is the difficulty level of the stratum just below the testee's ceiling stratum, i.e., the difficulty of the highest non-chance stratum reached.

These stratum scoring methods might result in somewhat more stable ability estimates than the "highest item" methods, since they would eliminate some of the variability due solely to variations in difficulties of specific items which would occur in methods 1 to 3. In using scoring methods 4 through 6, however, the number of possible scores will be equal only to the number of strata. Thus, when the number of strata is small, score variability will be severely decreased, leading to loss of information on individual differences and lowered correlations with other variables. The stratum scoring methods appear appropriate, therefore, only when the number of strata in the item pool is quite large (e.g., 25 or more).

Scoring method 6 also does not convey information on the proportion of items correct at the stratum just below the testee's ceiling stratum. At that highest non-chance stratum, one testee might answer 80% of the items correctly, while another might answer only 25% of the items correctly; using scoring method 6, both of these testees would obtain the same score even though their ability levels are probably different. It seems appropriate, therefore, to define an additional method of scoring, the "interpolated stratum difficulty score", which is designed to take account of the proportion correct data on individual testees at the highest non-chance stratum.

7. The interpolated stratum difficulty score can be defined as:

$$A = \bar{D}_{c-1} + S(p_{c-1} - .50)$$

where $\bar{D}_{c-1}$ is the average difficulty of the c-1th stratum, where c is the ceiling stratum. It is, therefore, the average difficulty of all items available at the testee's highest non-chance stratum, or the stratum just below his ceiling stratum.

p_{c-1} is the testee's proportion correct at the c-1th stratum.

and S is $\bar{D}_c - \bar{D}_{c-1}$, if p_{c-1} is greater than .50,

or $\bar{D}_{c-1} - \bar{D}_{c-2}$ if p_{c-1} is less than .50,

where $\bar{D}$ is the average difficulty of the designated stratum.

The interpolated stratum score assumes that the testee's ability lies at the mean of the difficulties of a peaked test (i.e., a stratum) if he answers exactly 50% of the items on that test correctly. If he answers very few of the items correctly, for example 25%, his ability is below the mean of that peaked test, tending toward the mean of the items at the next lower stratum. If the testee answers 80% of the items at a stratum correctly, his ability is above the mean of the peaked test and close to the lower range of ability measured by the items at the next most difficult stratum. Essentially, then, this scoring method interpolates the testee's ability level as a function of the distance between the relevant mean difficulties of the strata and the proportion of items answered correctly. In implementing the computations, if the cth or c-2th strata do not exist (i.e., are above or below the difficulties available in the item pool) the average difficulty of those hypothetical strata can be determined by adding or subtracting the constant or increment in difficulty between strata to the last actual average stratum difficulty available.

The interpolated stratum difficulty score, in addition to having the desirable characteristic of taking account of more of the information available from stradaptive testing, has the added advantage of increasing the range of scores possible over that available from the other stratum scoring methods.

Average difficulty scores. In an effort to compromise the probable unreliability of scoring methods 1-3 and the restricted range of methods 4-6, a number of average difficulty scores appear to be logically sound:

8. An individual's score can be determined as the average difficulty of all items answered correctly.

This method continues the "hurdle" analogy of ability scoring, but attempts to balance out chance factors by using an average. A major deficiency of this scoring method is that scores will be affected by inappropriate entry points. If the entry point is too low the testee will be presented with, and probably answer correctly, a number of items below his true ability level. His ability estimate will, therefore, be lower than it should be. An inappropriately high entry point will result in the administration of a number of items which are too difficult for a given testee. The administration of these difficult items might increase the probability of chance successes and thereby artifactually raise test scores based on this method of scoring.

9. Ability can be scored as the average difficulty of all items correct between (but not including) the basal stratum (100% correct) and the ceiling stratum (chance responding).

Thus, the "routing items", those items resulting from too high or too low an entry point, will not be scored in this method. Therefore, this scoring method will eliminate the problems inherent in method 8, and will probably result in more stable ability estimates. In order to use this method, however, the problem of individuals for whom a clear basal or ceiling stratum cannot be determined must be solved.

10. The stradaptive test can be scored by determining the average difficulty of items answered correctly at the highest non-chance stratum.

This method is the average difficulty analogue of method 3. It essentially identifies the peaked test of highest difficulty which is not inappropriate for a given testee, eliminating those that are too difficult and those that are too easy. It should give ability estimates with good variability and fairly high stability.

The variety of scoring methods available suggests a number of interesting research possibilities using stradaptive tests. Scoring methods may vary in terms of psychometric characteristics, such as stability, shape of resulting score distributions, or correlations with scores on other testing strategies. Scoring methods may also vary in terms of validity and/or utility, with some methods better predicting external criteria or being more useful in different kinds of situations. Only future research, using a variety of empirical, simulation, and theoretical studies will determine which scoring methods are best suited for particular purposes.

Consistency of Ability Estimates

The ten scoring methods described above, and others yet to be developed, all give "point estimates" of an individual's ability. Thus, they each return one value, based on some function of the difficulties of the items a testee has answered correctly, which indicates the point at which he falls on the underlying ability continuum. An analysis of the test records of individuals who have taken stradaptive tests shows additional information which reflects the consistency of the testee's response pattern. Such consistency data can be interpreted like data on the standard error of measurement; it indicates the range of confidence which can be attributed to a given ability point estimate. Individuals who are more consistent should have more stable ability estimates, while those who are less consistent should have less stable ability estimates. At present, this is only an hypothesis which will need empirical verification.

On stradaptive tests, individual differences occur in the number of strata between the basal stratum and the ceiling stratum. Thus, it is possible for some individuals to have the same score by one or more scoring methods (e.g., difficulty of the highest non-chance stratum), but the number of strata utilized in obtaining that score will differ widely. Some testees are consistent enough in their responses that their response records encompass only two or three strata. Other testees respond more inconsistently to the items, and their response records may encompass five or more strata between the basal and ceiling strata. Thus, the number of strata used by the testee can be a rough index of the consistency of his ability estimate, if items resulting from inappropriate entry points are eliminated. A related index would be the difference in average difficulties between the ceiling and basal strata.

A more meaningful consistency index might be the variance or standard deviation of the difficulties of the items answered correctly between the testee's basal and ceiling strata. This index would reflect more accurately the consistency of an individual's stradaptive test performance. It has the further advantage of being within the control of the tester. Since the variance is a mean, adding more items at or near the mid-point of the distribution of correct responses will reduce the variance. Reduction of this variance consistency estimate will occur then, by administering additional items at an individual's estimated ability level; since these items will have little or no deviation from his ability, the variance will continue to reduce with additional items. Testing could then continue in this fashion until a desired "standard error of measurement" was reached. At the same time that the variance reduction occurs by administering additional items, indicating greater confidence in the abilility estimate, the ability estimate itself should stabilize due to the greater number of items administered.

Individuals differ also in the number of items necessary to reach a termination criterion. In over 350 stradaptive tests administered to college students, the median number of items required to reach termination was 18; the shortest stradaptive test required only 9 items and the longest required 160 items. Individuals who required a larger number of items also utilized a larger number of strata. The number of items required for termination, therefore, is a rough indication of an individual's consistency of response. Only further research on the relationship of this additional individual differences variable with other consistency data and with other data external to the stradaptive testing procedure will determine its utility.

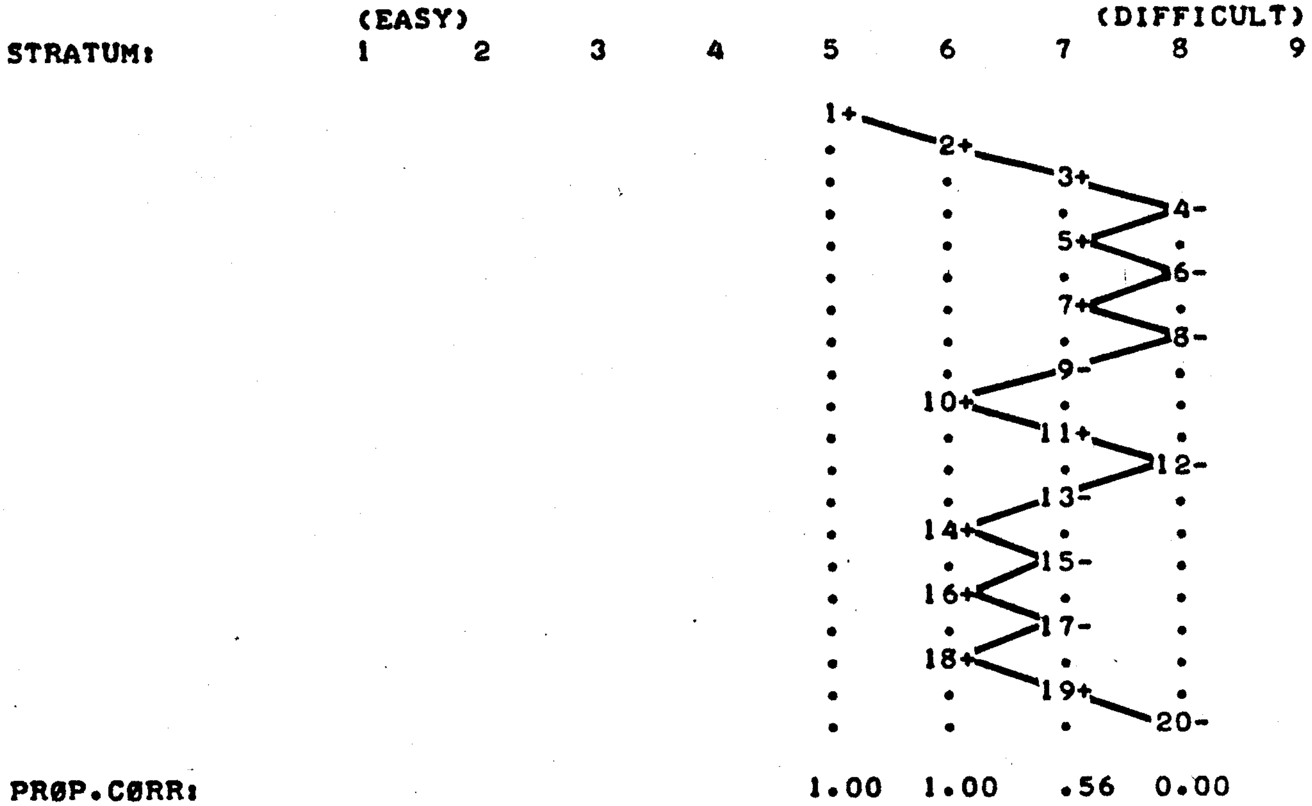
Illustrative Results from Stradaptive Testing

The previous sections have described the essential characteristics of the stradaptive test. However, to understand the method more completely, it is helpful to see the results of its application with actual testees. The following figures are graphical illustrations of the response records of a number of college students who took stradaptive tests.[3] The 9-stratum item pool used consisted of 229 5-response choice vocabulary items; the structure of the item pool is shown in Table 1. Entry point information was the student's report of his/her GPA as shown in Figure 3. An "up-one/down-one" branching rule was used. Termination occurred when a stratum was identified at which

[3] The stradaptive test administration program was written by Robert Swisher; the display program was written by David Vale.

Figure 4

```
                         REPØRT ØN STRADAPTIVE TEST

NAME: WILLIAM W.                                   DATE TESTED:  73/07/12

-----------------------------------------------------------------------
```



```
                 TØTAL PRØPØRTIØN CØRRECT=  .550

         SCØRES ØN STRADAPTIVE TEST

         1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT=  1.49

         2. DIFFICULTY ØF THE N+1 TH ITEM=  1.44

         3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT=  1.49

         4. DIFFICULTY ØF HIGHEST STRATUM
            WITH A CØRRECT ANSWER=  1.33

         5. DIFFICULTY ØF THE N+1 TH STRATUM=  1.33

         6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM=  1.33

         7. INTERPØLATED STRATUM DIFFICULTY=  1.37

         8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS=   .88

         9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN
            CEILING AND BASAL STRATA                    =  1.28

        10. MEAN DIFFICULTY ØF ITEMS CØRRECT
            AT HIGHEST NØN-CHANCE STRATUM=  1.28
```

Table 2

Number of items administered (N) and cumulative proportion correct (p) by stage, for William W.

	Stratum														
	1			5		6		7		8		9		Total	
Stage	N	p	...	N	p	N	p	N	p	N	p	N	p	N	p
1				1	1.00									1	1.00
2						1	1.00							2	1.00
3								1	1.00					3	1.00
4										1	0.00			4	.75
5								2	1.00					5	.80
6										2	0.00			6	.67
7								3	1.00					7	.71
8										3	0.00			8	.63
9								4	.75					9	.56
10						2	1.00							10	.60
11								5	.80					11	.64
12										4	0.00			12	.58
13								6	.67					13	.54
14						3	1.00							14	.57
15								7	.57					15	.53
16						4	1.00							16	.56
17								8	.50					17	.53
18						5	1.00							18	.56
19								9	.56					19	.58
20										5	0.00			20	.55

the proportion of correct responses was .20 or less, based on a minimum of five items completed at that stratum. Test items were presented to the student on a cathode-ray-terminal (CRT) with responses recorded through the CRT typewriter keyboard.

A typical response record. Figure 4 shows the stradaptive test performance of "William W.", a college sophomore. This test record is typical of the stradaptive test performance of college students. William was first presented with an entry point screen (Figure 3) and indicated that his cumulative grade point average to date was between 2.76 and 3.00. He thus began the stradaptive test at stratum 5. His answer to the first item was correct (indicated by a "+" in Figure 4), which branched him to the first available item in stratum 6. Correct answers to the second and third items resulted in his moving to stratum 8, where he received the first item from that more difficult peaked test. Since the stage 4 item was too difficult for him, his response was incorrect (-), and he branched downward to the first item in stratum 7. William then alternated between correct and incorrect responses for the items at stages 6 through 8, followed by an incorrect response to the stage 9 item. This returned him to stratum 6 for his tenth item. With a few minor deviations, William then essentially alternated between correct and incorrect responses from stages 11 through 20. Item 20 terminated the stradaptive test since the testing procedure had, at that point, located William's ceiling stratum; at stratum 8 William had answered all 5 items incorrectly.

Table 2 shows a complete "accounting" of William's stradaptive test performance. As the data in Table 2 indicate, tentative estimates of William's "basal" and "ceiling" strata were evident by stage 10; at that point he had 100% of the items correct at stratum 6, 75% correct at stratum 7 and none correct at stratum 8; his total percent correct at stage 10 was 60%. However, these percentages were based on only 2, 4, and 3 items respectively and therefore were not likely to be very stable. Since the termination criterion had not been met (i.e., 20% or less items correct based on 5 items administered at a stratum) the stradaptive test continued. As additional items were administered, William continued to answer all items at stratum 6 correctly, and at stratum 7 answered some items correctly and some incorrectly. By stage 19, he had completed the first 9 items available at stratum 7 and had answered 56% of those correctly. The final item administered (stage 20) was the fifth item at stratum 8, which he answered incorrectly.

The last column of Table 2 shows the proportion correct at each stage of the stradaptive test. That proportion shows a steady step-like decrease from 100% correct at stage 1 to 55% correct at stage 20. It is typical of stradaptive test performance for the proportion correct at the final stage to be near .50; in William's test performance the proportion correct stayed between .50 and .60 from stage 2 through termination.

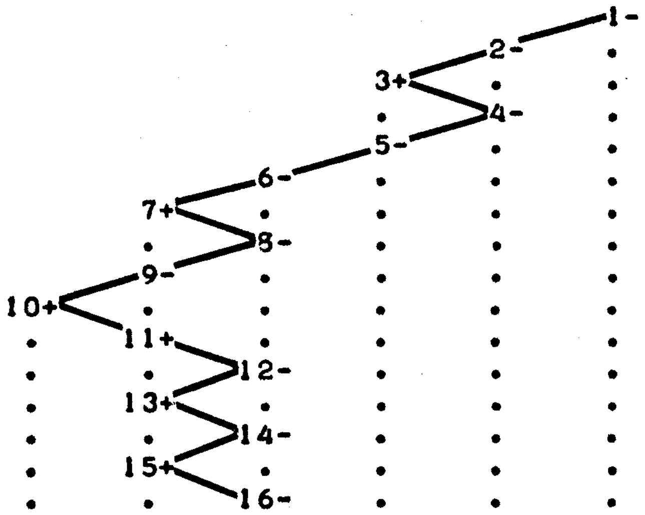
Figure 4 also shows stradaptive test scores for William, using the scoring methods described earlier. As might be expected, the "highest difficulty" scores produced the highest ability estimates, and methods 1 and 3 gave the same results since William answered no items correctly at or above his ceiling stratum. Methods 4, 5 and 6 gave identical results for similar reasons; with a different set of test responses, however, these results would differ. The "average difficulty" methods gave the lowest ability estimates as a group, since the averages were lowered by the inclusion of the less difficult items.

William's stradaptive test performance (Figure 4) is an example of a slightly low entry point. Because he entered at stratum 5, which was below his basal stratum 6, his response to the first item conveyed no information. However, it did serve to route him to the higher strata where testing was concentrated. Eliminating the first item administered from total proportion correct gives a proportion of .45 correct for William at the termination of testing.

<u>High entry point</u>. Occasionally an entry point is too high; an example is shown in Figure 5 for "Carol C." Carol reported her GPA to be in category 4, 3.01 to 3.25 (see Figure 3); this led to an entry at stratum 6. Her item responses quickly showed that the tests at strata 6, 5, 4, and 3 were too difficult for her. On the first six items Carol gave only one correct answer, an apparent "lucky guess" to a stratum 4 item. The routing procedure quickly brought Carol to strata 3, 2, and 1, which were composed of easier test items. Once she reached these strata her response pattern converged quickly on a region of the item pool in which she answered about 50% of the items correctly. Although her total proportion correct was only .375, eliminating the routing items due to the erroneous entry point (items 1 through 5), Carol obtained 5 correct answers out of 11 items in stages 6 through 10, for an effective proportion correct of .45. Disregarding the first 5 routing items, Carol's stradaptive test performance is similar to that of William's. In both cases the stradaptive test

Figure 5

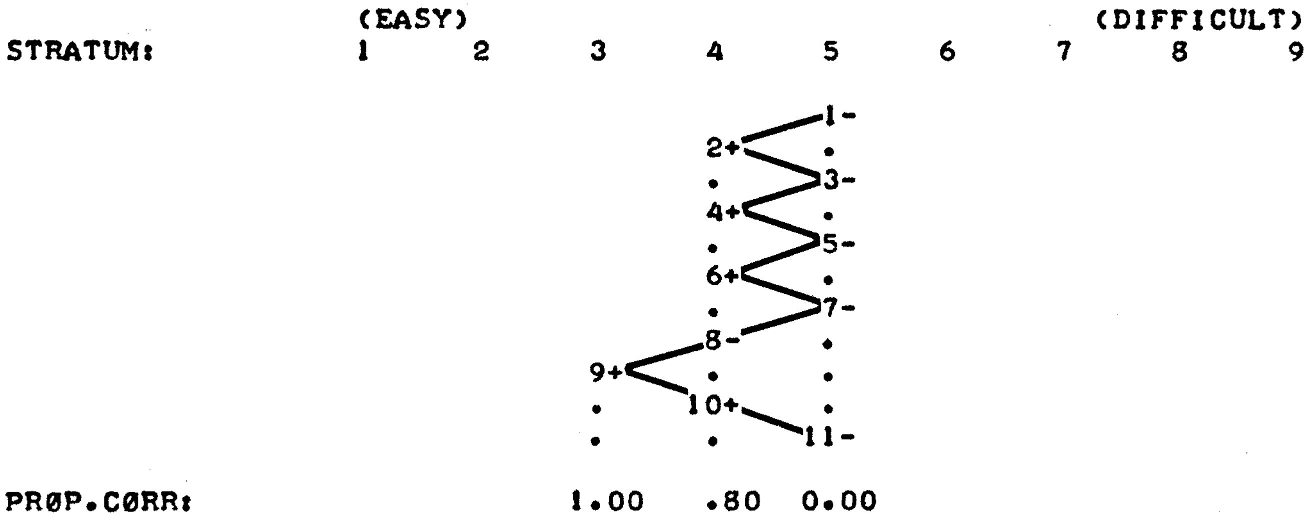
```
                         REPØRT ØN STRADAPTIVE TEST

NAME:   CARØL C.                                   DATE TESTED:  73/07/12
---------------------------------------------------------------------------

                  (EASY)                                        (DIFFICULT)
STRATUM:            1      2      3      4      5      6      7      8      9
```



```
PRØP.CØRR:          1.00   .80  0.00   .50  0.00  0.00

                      TØTAL PRØPØRTIØN CØRRECT=  .375


          SCØRES ØN STRADAPTIVE TEST

          1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT=  -.70

          2. DIFFICULTY ØF THE N+1 TH ITEM= -1.92

          3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT= -1.68

          4. DIFFICULTY ØF HIGHEST STRATUM
             WITH A CØRRECT ANSWER=  -.63

          5. DIFFICULTY ØF THE N+1 TH STRATUM= -1.92

          6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM= -1.92

          7. INTERPØLATED STRATUM DIFFICULTY= -1.73

          8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS= -1.81

          9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN
             CEILING AND BASAL STRATA                      = -1.94

         10. MEAN DIFFICULTY ØF ITEMS CØRRECT
             AT HIGHEST NØN-CHANCE STRATUM= -1.94
```

Figure 6

```
                     REPØRT ØN STRADAPTIVE TEST

NAME: JØHN J.                                  DATE TESTED:  73/04/09

-------------------------------------------------------------------------
```



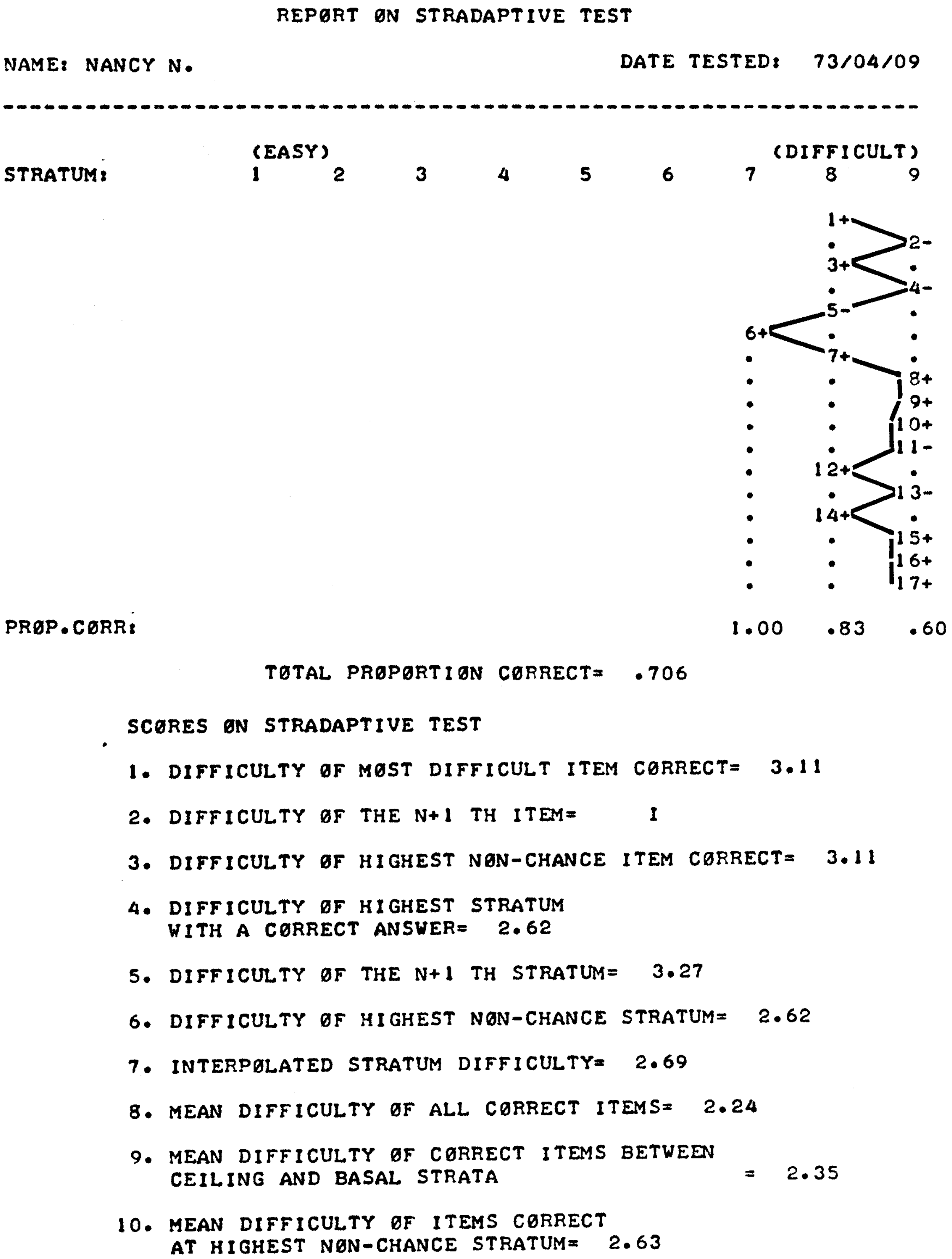
```
                   TØTAL PRØPØRTIØN CØRRECT=  .455

         SCØRES ØN STRADAPTIVE TEST

         1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT=  -.52

         2. DIFFICULTY ØF THE N+1 TH ITEM=  -.75

         3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT=  -.52

         4. DIFFICULTY ØF HIGHEST STRATUM
            WITH A CØRRECT ANSWER=  -.63

         5. DIFFICULTY ØF THE N+1 TH STRATUM=  -.63

         6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM=  -.63

         7. INTERPØLATED STRATUM DIFFICULTY=  -.44

         8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS=  -.81

         9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN
            CEILING AND BASAL STRATA                        =  -.63

        10. MEAN DIFFICULTY ØF ITEMS CØRRECT
            AT HIGHEST NØN-CHANCE STRATUM=  -.63
```

identified a ceiling stratum (none correct or chance responding) a basal stratum (all correct), and a peaked test in between on which the testee obtained an intermediate proportion correct. In Carol's case the optimal peaked test was at stratum 2, on which she obtained 80% correct responses, while William's optimal peaked test was at stratum 7, on which he obtained 56% correct responses. It is interesting to note that William's entry point was lower than Carol's, yet their terminal ability levels were quite the reverse.

Rapid convergence. When the entry point estimate is accurate, the stradaptive test record can be quite short. Figure 6 shows an actual test record for "John J.". John entered at stratum 5 and immediately began alternating between correct and incorrect responses through stage 8. An incorrect response at stage 8 led to the identification of the basal stratum (although based on only one item) at stratum 3. Finally, an incorrect response on the stage 11 item permitted John to reach the termination criterion in only 11 items, having identified stratum 5 as John's ceiling stratum. John's ability level lies in the vicinity of stratum 4 at which he answered 80% of the items correctly. Over all 11 items administered, John answered 5, or a proportion of .455, correctly.

Item pool too easy. Occasionally the stradaptive item pool is too easy, or too difficult, for a testee. Figure 7 shows the stradaptive test performance of "Nancy N.". Nancy entered at stratum 8, based on a GPA estimate in the range of 3.51 to 3.75, almost an A average. With the exception of the stage 6 item, at stratum 7, testing of Nancy was confined to the difficult peaked tests at strata 8 and 9. Seventeen items were administered to Nancy, with 10 of them at stratum 9, the stratum with the most difficult items in the stradaptive item pool. Since stratum 9 contained only 10 items, testing was terminated. It is obvious that further testing of Nancy would be unproductive even if additional items were available at stratum 9. Nancy answered 83% of the items correctly at stratum 8, and 60% correctly at stratum 9. Since it would be quite unlikely that stratum 9 could be her ceiling stratum (.20 or less correct), no purpose would be served by further testing. In this case, the stradaptive test simply indicates that Nancy's ability is very high, but it is unable to give an estimate of exactly how high it is since she is apparently "off the top" of the most difficult test in the stradaptive pool. However, her ability is probably not as high as the individual who would answer all items correctly at stratum 9. The latter individual would answer 100% of the

Figure 7



```
                        REPØRT ØN STRADAPTIVE TEST

NAME: NANCY N.                                  DATE TESTED:  73/04/09

--------------------------------------------------------------------------

                       (EASY)                                (DIFFICULT)
STRATUM:                1     2     3     4     5     6     7     8     9

                                                                  1+
                                                                  .     2-
                                                                  3+    .
                                                                  .     4-
                                                                  5-    .
                                                            6+    .     .
                                                            .     7+    .
                                                            .     .     8+
                                                            .     .     9+
                                                            .     .     10+
                                                            .     .     11-
                                                            .     12+   .
                                                            .     .     13-
                                                            .     14+   .
                                                            .     .     15+
                                                            .     .     16+
                                                            .     .     17+

PRØP.CØRR:                                                  1.00  .83   .60

                  TØTAL PRØPØRTIØN CØRRECT=  .706

         SCØRES ØN STRADAPTIVE TEST

         1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT=  3.11

         2. DIFFICULTY ØF THE N+1 TH ITEM=      I

         3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT=  3.11

         4. DIFFICULTY ØF HIGHEST STRATUM
            WITH A CØRRECT ANSWER=  2.62

         5. DIFFICULTY ØF THE N+1 TH STRATUM=  3.27

         6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM=  2.62

         7. INTERPØLATED STRATUM DIFFICULTY=  2.69

         8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS=  2.24

         9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN
            CEILING AND BASAL STRATA                  =  2.35

        10. MEAN DIFFICULTY ØF ITEMS CØRRECT
            AT HIGHEST NØN-CHANCE STRATUM=  2.63
```

items correctly, while Nancy answered only 60% correctly. Thus, the total proportion correct can be a rough indicator of the appropriateness of the stradaptive item pool for an individual. When that proportion, corrected for routing, is between .40 and .60, it indicates a test record appropriately adapted to the individual's ability level.

Two problems arose in computing scores for Nancy's stradaptive test performance. Scoring method 2, which determines score on the basis of the difficulty of the n+1th item could not be implemented for Nancy. Since she answered her last item correctly and it was the last item at stratum 9, the next item to be administered would have been an item at stratum 10. There were, however, only 9 strata in the stradaptive item pool. Thus, the difficulty of the n+1th item is indeterminate in Nancy's case, and an "I" is given on the computer report. A similar problem arose in computing the interpolated stratum difficulty score (method 7). Since Nancy answered 60% of the items correctly at stratum 9, her ability could be estimated to be above the mean difficulty of the stratum 9 peaked test ($z=2.62$, based on .50 correct). To compute the interpolated stratum difficulty score, the increment between the strata in the item pool, approximately .655, was added to the mean difficulty of stratum 9; Nancy's score was then interpolated into the interval between 2.62 and 3.27 by the formula given earlier.

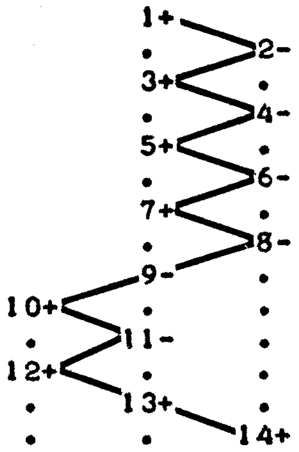
Consistent vs. inconsistent response records. As indicated above, stradaptive test records can reflect individual differences in consistency of test performance. Figures 8 and 9 contrast the test records of "Tom T." and "Dixie D". In both cases entry into the item pool was at about the same level of difficulty; Tom entered at stratum 6 while Dixie began at stratum 7. For the first 8 items, both Tom and Dixie alternated between items at strata 6 and 7, and both had moved to the easier items at stratum 5 by the 10th stage of testing. After two items at stratum 5, Tom recovered quickly to stratum 6 and reached the termination criterion after 14 items. Tom's basal stratum was stratum 5, and stratum 7 was his ceiling stratum. His highest non-chance stratum was stratum 6, at which he answered 71% of the items correctly.

Dixie's test performance, although similar to Tom's in the earlier stages of testing, diverged sharply after the twelfth item. At that point she began to answer easier items incorrectly, finally being presented with an item from stratum 3 at the 17th stage of testing. Dixie's response

Figure 8

REPØRT ØN STRADAPTIVE TEST

```
NAME: TØM T.                                      DATE TESTED:  73/07/02

------------------------------------------------------------------------

                        (EASY)                                (DIFFICULT)
STRATUM:                 1     2     3     4     5     6     7     8     9
```



```
PRØP.CØRR:                                       1.00  .71   .20

                        TØTAL PRØPØRTIØN CØRRECT=  .571
```

SCØRES ØN STRADAPTIVE TEST

1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT= 1.11

2. DIFFICULTY ØF THE N+1 TH ITEM= 1.89

3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT= .79

4. DIFFICULTY ØF HIGHEST STRATUM
   WITH A CØRRECT ANSWER= 1.33

5. DIFFICULTY ØF THE N+1 TH STRATUM= 2.01

6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM= .65

7. INTERPØLATED STRATUM DIFFICULTY= .80

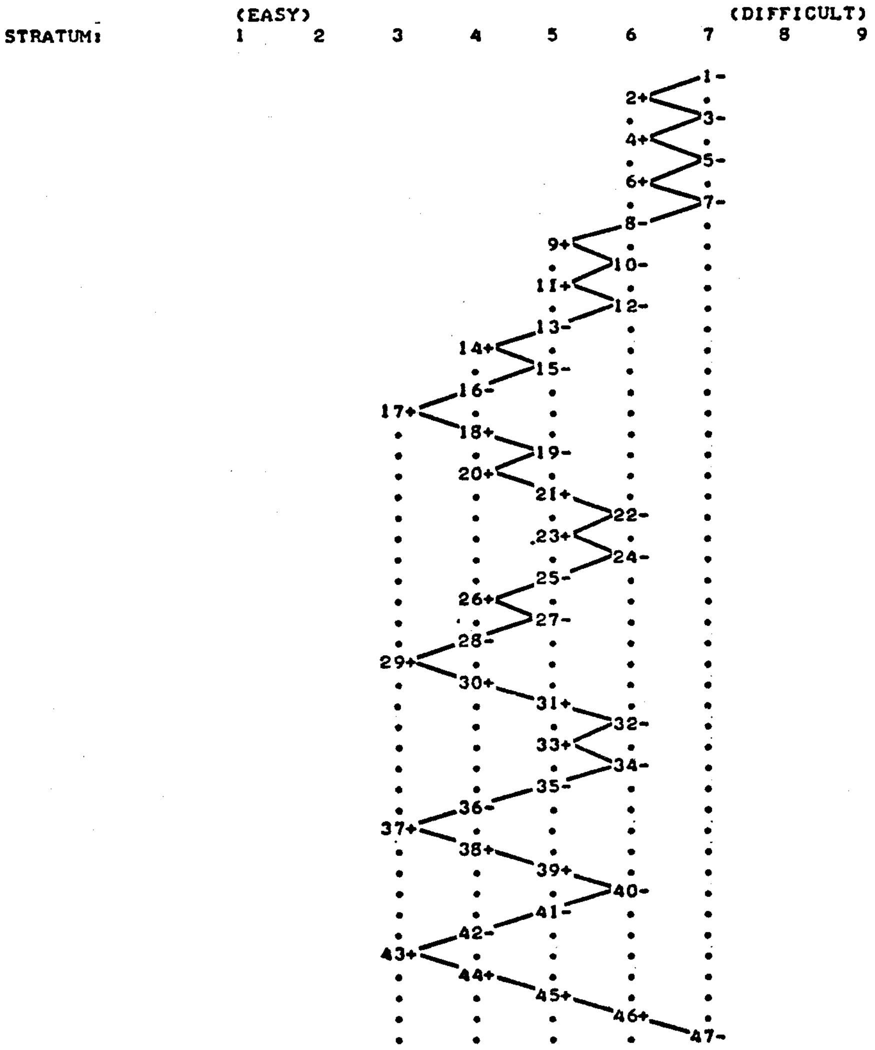
8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS= .52

9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN
   CEILING AND BASAL STRATA = .59

10. MEAN DIFFICULTY ØF ITEMS CØRRECT
    AT HIGHEST NØN-CHANCE STRATUM= .59

Figure 9

```
                    REPØRT ØN STRADAPTIVE TEST

NAME: DIXIE D.                               DATE TESTED:   73/04/09

----------------------------------------------------------------------
```



```
PRØP.CØRR:                    1.00   .64   .53   .33  0.00

              TØTAL PRØPØRTIØN CØRRECT=  .489
          SCØRES ØN STRADAPTIVE TEST

          1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT=   .73

          2. DIFFICULTY ØF THE N+1 TH ITEM=   .78

          3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT=   .73

          4. DIFFICULTY ØF HIGHEST STRATUM
             WITH A CØRRECT ANSWER=   .65

          5. DIFFICULTY ØF THE N+1 TH STRATUM=   .65

          6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM=   .65

          7. INTERPØLATED STRATUM DIFFICULTY=   .54

          8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS=  -.30

          9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN
             CEILING AND BASAL STRATA                    =  -.09

         10. MEAN DIFFICULTY ØF ITEMS CØRRECT
             AT HIGHEST NØN-CHANCE STRATUM=   .59
```

record then shows a series of wide swings between items at stratum 3 and those at stratum 6. While many testees converge on strata that are contiguous, Dixie's responses seem to show a convergence somewhere between strata 3 and 6. Thus, ability estimates derived from Dixie's stradaptive testing are likely to be less precise than those from Tom's responses. Dixie finally worked her way back up to stratum 7 after 47 items to satisfy the termination criterion.

Dixie's testing thus used five of the available nine strata, while Tom used only three. For both Tom and Dixie the ceiling stratum was stratum 7, but while Tom's basal ability was at stratum 5, Dixie's was at stratum 3. Stratum 6 was the highest non-chance stratum for both, but Tom's ability is probably closer to that of stratum 7 than to stratum 5, since he answered 71% of the items correctly at stratum 6. Dixie's, however, is more toward stratum 5, since she answered only 33% correctly at stratum 6. The difference is reflected by the interpolated stratum difficulty scores of .80 and .54 for the two testees, respectively. These two response records show how stradaptive test performance can differ in terms of both number of items administered and the number of strata used for ability determination.

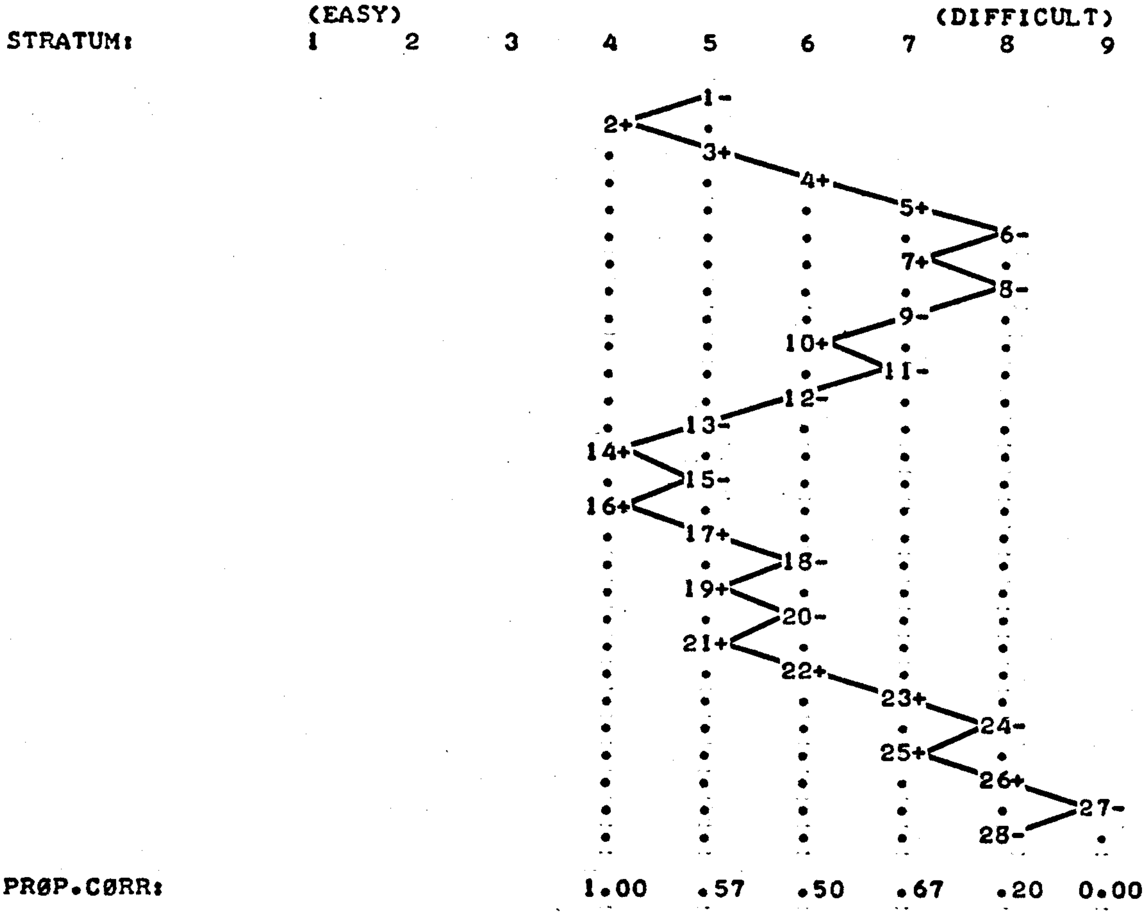
Another example of inconsistent stradaptive test performance is shown in Figure 10. This test record, for "Carl C.", shows a range of fluctuation even wider than that of Dixie D. (Figure 9). Carl seemed to answer almost optimally (i.e., about 50% correct) on the three peaked tests of strata 5, 6, and 7. His performance fluctuated rather consistently from strata 4 through 8, and he even attempted one item (27) at stratum 9, following a probable lucky guess at stratum 8. Carl's basal stratum was stratum 4(100% correct) and his ceiling stratum was stratum 8 (20% correct). Between these two he answered slightly more than 50% of the items correctly, with an overall proportion correct of .54. Carl's inconsistent performance on the stradaptive test stands in sharp contrast to that of, say, John J. (Figure 6), whose very consistent response record covered only three strata, and who reached the termination criterion in only 11 items. The utility of this information on individual differences in consistency of performance on the stradaptive test will be determined only through further research. Logically, however, it seems that such information could be used to derive individualized "standard errors of measurement."

Implications of Proportion Correct Data

The data in Figures 4 through 10 illustrate an interesting characteristic of stradaptive test records. For

Figure 10

```
                    REPØRT ØN STRADAPTIVE TEST

NAME: CARL C.                             DATE TESTED:  73/07/12

------------------------------------------------------------------
```



```
                TØTAL PRØPØRTIØN CØRRECT=  .536

     SCØRES ØN STRADAPTIVE TEST

     1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT=  2.31

     2. DIFFICULTY ØF THE N+1 TH ITEM=  1.17

     3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT=  1.49

     4. DIFFICULTY ØF HIGHEST STRATUM
        WITH A CØRRECT ANSWER=  2.01

     5. DIFFICULTY ØF THE N+1 TH STRATUM=  1.33

     6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM=  1.33

     7. INTERPØLATED STRATUM DIFFICULTY=  1.44

     8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS=   .47

     9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN
        CEILING AND BASAL STRATA                    =   .60

    10. MEAN DIFFICULTY ØF ITEMS CØRRECT
        AT HIGHEST NØN-CHANCE STRATUM=  1.27
```

most individuals completing a stradaptive test, the proportion of correct responses at the various strata decreases as the difficulty of the strata increases. These results are summarized in Figure 11, which plots the proportion of correct responses at each stratum. With the exception of the plots for Carl C. and Carol C., these plots resemble item trace lines (Lord & Novick, 1968). The steepness of the slope can be interpreted as an index of the consistency of responses of the individual and the capability of the item pool to "discriminate" that individual's ability level. The point of inflection of the curve (i.e., the point on the horizontal axis at which the testee answers 50% of the items correctly) could be interpreted as the "difficulty" of the item pool for the individual, or his position on the latent ability continuum.

Reasoning analogically from item characteristic curve theory, non-regular item characteristic curves, such as those for Carl C. and Carol C., might indicate item pool-testee interactions which are inappropriate. Thus, both Carol and Carl might not be interacting with the item pool on a unidimensional continuum. In order to get a more accurate ability estimate for such testees, it might be necessary to multidimensionally scale their response patterns to obtain subsets of test items (if possible) on which they responded in unidimensional fashion, as indicated by their test response "trace lines." Thus, Carl and Carol's response records might be analyzed by appropriate scaling methods to find the intra-individual probabilistic Guttman-type scales underlying their response patterns.

The "trace line" plots for John J., Tom T. and William W. approximate the classic step function Guttman-type trace line. Dixie D.'s trace line plot is very similar to the normal ogive probabilistic analogue of the Guttman trace line. Future research based on stradaptive tests with a large number of strata may lead to mathematization of these trace line ideas, which in turn may lead to greater utility for this type of test data.

It is interesting to note that the stradaptive test performance of many testees results in a Guttman-like scaling of the testee's performance with respect to the item pool. Since the stradaptive test developed from the testing rationale originally proposed by Binet, it follows that perhaps Binet's ability testing logic had embedded in it an unarticulated primitive version of Guttman's ideas and the present-day derivates of modern test theory as derived from latent trait theory.

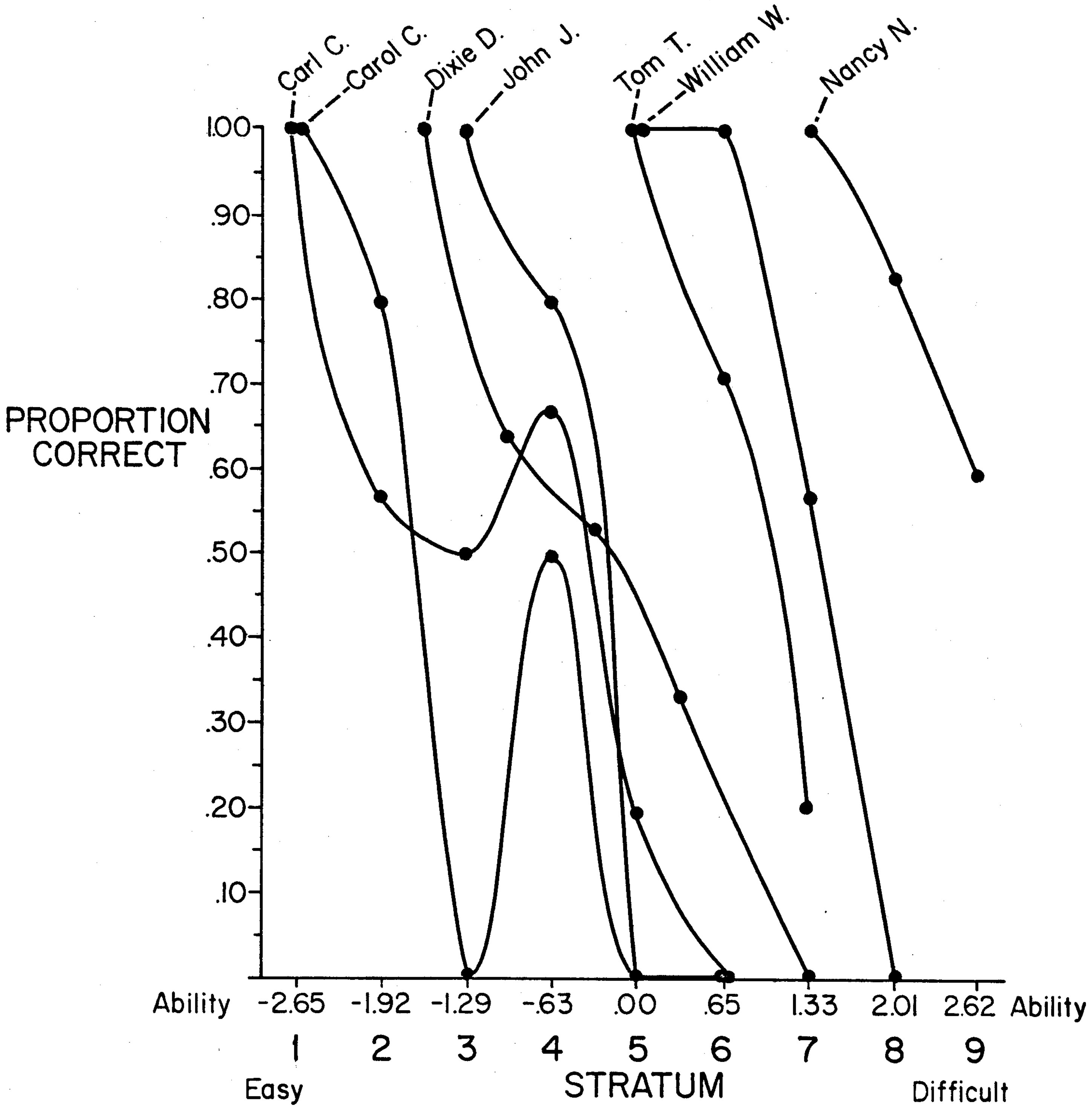


Figure 11. Proportion correct at each stratum, by individual

Conclusions

The stradaptive test is an operational computer-based testing model which draws simultaneously from Binet's pioneering work in ability measurement and from ideas in modern test theory. The testing procedure makes no restrictive assumptions about the nature of underlying ability distributions (beyond those involved in norming the item pool), and its implementation does not require complicated mathematical calculations. The procedure is also flexible with respect to size and composition of the item pool, branching rules, termination rules, and scoring methods. Data derived from the stradaptive test response record, including number of items completed, range of difficulties used, patterns of movement through the item pool, and various other methods of measuring a testee's interaction with a specified item pool appear to have promise as new sources of information derivable from ability testing.

The availability of the stradaptive testing strategy poses many new research questions. Among these are the optimal characteristics (e.g., size, number of strata) of the stradaptive item pool, methods of selecting and placing items in the pool, variations in branching rules, applications of stochastic models to the branching process, variations in step size, effects of various termination rules, the reliability and utility of the various scoring methods proposed and those yet to be developed, methods of expressing an individual's consistency or the accuracy of test scores, methods of controlling the accuracy of test scores within the stradaptive framework, and relationships of stradaptive scores and ability estimates to those derived from other adaptive strategies. These research questions should be studied by a variety of approaches, including live testing empirical studies, simulation studies, and theoretical studies, with the results of each approach supporting and nourishing research using the other approaches.

References

Bayroff, A. G., Thomas, J. J. & Anderson, A. A. Construction of an experimental sequential item test. Research Memorandum 60-1, Personnel Research Branch, Department of the Army, January 1960.

Bayroff, A. G. & Seeley, L. C. An exploratory study of branching tests. U. S. Army Behavioral Science Research Laboratory, Technical Research Note 188, June 1967.

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

Cronbach, L. G., Gleser, G. C., Nanda H. & Rajaratnam, N. *The dependability of behavioral measurements: theory of generalizability for scores and profiles*. New York: Wiley, 1972.

Hick, W. E. Information theory and intelligence tests. *British Journal of Psychology, Statistical Section*, 1951, *4*, 157-164.

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

Lord, F. M. Robbins-Munro procedures for tailored testing. *Educational and Psychological Measurement*, 1971, *31*, 3-31. (a)

Lord, F. M. A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement*, 1971 *31*, 805-813. (b)

Lord, F. M. A theoretical study of two-stage testing. *Psychometrika*, 1971, *36*, 227-241. (c)

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

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

Owen, R. J. Bayesian sequential design and analysis of dichotomous experiments with special reference to mental testing. Unpublished paper, 1970.

Paterson, J. J. An evaluation of the sequential method of psychological testing. Unpublished doctoral dissertation, Michigan State University, 1962.

Terman, L. M. & Merrill, M. A. Stanford-Binet Intelligence Scale. Boston: Houghton Mifflin, 1960.

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

Weiss, D. J. Strategies of computerized ability testing. Research Report 73-x, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1973 (in preparation).

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