DOCUMENT RESUME

ED 128 404 TM 005 587

AUTHOR Weiss, David J.
TITLE Computerized Adaptive Ability Measurement.
SPONS AGENCY Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
PUB DATE [Sep 75]
CONTRACT N00014-67-A-0113-0029
NOTE 39p.; Paper presented at the Annual Conference of the Military Testing Association (17th, Fort Benjamin Harrison, Indiana, September 15-19, 1975); Also included in TM 005 585

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Ability; Ability Grouping; Achievement Tests; *Computer Oriented Programs; *Error Patterns; Feedback; *Individual Differences; *Response Style (Tests); Scores; Statistical Analysis; Test Bias; Test Construction; *Testing; Testing Problems; Test Interpretation
IDENTIFIERS Adaptive Testing; *Computer Assisted Testing

ABSTRACT

The general objective of a research program on adaptive testing was to identify several sources of potential error in test scores, and to study adaptive testing as a means for reducing these errors. Errors can result from the mismatch of item difficulty to the individual's ability; the psychological effects of testing and the test environment; the inability to extract enough information from the testee's response; deviations from unidimensionality; and an oversimplistic conceptualization of ability. Several different strategies of adaptive testing are discussed, along with the information level they yield, and the bias that can result from various scoring methods. In a discussion of the unidimentionality of test items, the consistency of the testee's response is analyzed. Finally, group differences are examined in terms of the psychological effects of receiving immediate feedback, especially on low ability groups. The author concludes that adaptive testing and immediate knowledge of results may be able to provide testing conditions more conclusive to each person's ability to demonstrate his/her fullest capacities in test performance. (Author/BW)

Computerized Adaptive Ability Measurement[1]

David J. Weiss
University of Minnesota

Adaptive Testing and Error Reduction

The general objective of our research program on adaptive testing is to identify several sources of potential error in test scores, and to study adaptive testing as a means for reducing these errors of measurement.

The first general source of error that we have been concerned with is the error that results from the mis-match of item difficulties in an ability test with the individual's ability. Obviously, the testee's ability is not known at the start of testing. But the different strategies of adaptive testing that have been proposed can be viewed as different ways of matching item difficulties with testee ability and sequentially estimating the testee's ability. Consequently, one of our major emphases is to determine the best, or at least better, ways of adapting item difficulties to individual abilities. Much of what I have to say today will be concerned with these various strategies of adaptive testing.

We are approaching this in two complementary ways. First, we have been doing live computerized testing. Since late 1972 we have tested more than 5,000 subjects on a variety of strategies of adaptive testing. But live testing cannot provide answers to all the questions concerning which strategies are best under which conditions, because there are too many questions to be answered. Therefore, we are using computer simulation to supplement and extend the results that we obtain from live testing.

The second main emphasis of our research is a concern with the psychological effects of adaptive testing. Here we are concerned with identifying the psychological aspects of testing and the test environment which can introduce error into test scores. These variables include guessing, test anxiety, boredom, frustration, lack of motivation, and racial or ethnic group effects.

---

[1]Early development work on this research was supported during 1969 and 1970 by grants from the General Research Fund of the Graduate School, University of Minnesota. Research reported in this paper was supported since early 1972 by Personnel and Training Research Programs, Office of Naval Research, Contract No. N00014-67-A-0113-0029, NR 150-343.

Portions of this paper were written by the project staff: Nancy Betz, Jim McBride, Brad Sympson and Dave Vale.

Special thanks are due to our project programmer, John Dewitt, without whom this research would have been nearly impossible.

Guessing can obviously artificially increase test scores; frustration, anxiety, motivation and other factors can result in test scores lower than true ability. All of these, therefore, are sources of error in test scores which are due to the psychological effects of testing.

We are also concerned with the psychological effects that will result from the man-machine interface. This, from our experience, is going to be an important problem in computerized adaptive testing. There are different kinds of computer systems on which we can implement adaptive testing and each of those computer systems has its positive and negative effects on testee behavior. There are different kinds of terminal devices for adaptive testing and each kind of terminal device displays in different ways and at different speeds. All of these variations in the man-machine interface are going to be new problems for us to consider in the years to come. Past research has demonstrated that answer sheets in paper and pencil testing sometimes had an effect on test scores. Similarly, research in adaptive testing will need to study different kinds of CRT's, different kinds of computer systems and different display speeds as part of the psychological effects of computerized testing. In the second half of today's presentation I will present some data relevant to the psychological effects of adaptive testing.

A third source of error that we are concerned with is error that results from not extracting enough information from a testee's response to a test item. To date most psychometric research has been concerned with binary or 0-1 scoring. But we can extract more information from a test response if we assign different scores to different incorrect response alternatives. Test responses can be even more informative if we use continuous responding, or probabilistic responding.

The fourth source of error that we are studying is the error that results from deviations from unidimensionality. Latent trait theory, as it is usually used in testing, is based on the assumption of unidimensionality, although there are multidimensional latent trait models being developed. But dimensionality that is defined on a group, such as the unidimensionality of latent trait theory, does not necessarily hold true for an individual. That is, dimensionality defined by factor analysis or other methods, when applied to an individual, assumes that the individual is the typical or average member of the group on which the dimensionality was defined. Thus, in the testing situation, when a set of "unidimensional" items is administered to an individual, the result may be a set of responses that are not unidimensionally determined.

Consequently, our research is concerned with individual-item pool interactions--the interaction of one individual with a set of "unidimensional" items. We are studying item response protocols of

this nature to determine if meaningful deviations from unidimensionality do occur for specific individuals. If they do, we will then develop interactive adaptive testing models that will take account of intra-individual multidimensionality. I'll have more to say about the dimensionality problem later.

A fifth kind of error that we plan to study in the future is the error that results from an over-simplistic conceptualization of ability. In the past fifty years, we have largely let the nature of our ability tests be determined by the restrictions imposed by the paper-and-pencil testing medium. Thus, many of our abilities are "static" abilities, such as verbal ability measured by the multiple-choice vocabulary test. But interactive computer systems permit us to break out of these shackles and measure abilities that are not measureable in paper-and-pencil formats. We should now be able to measure such abilities as reasoning, by following an individual's chain of decisions given a structured set of problem stimuli. Or, we will be able to measure memory abilities within a dynamic framework, or perceptual abilities, including perception of movement, using computer-controlled stimuli. The possibilities are endless, and the net result should be new kinds of ability measures which will likely be more meaningful and accurate for occupational prediction.

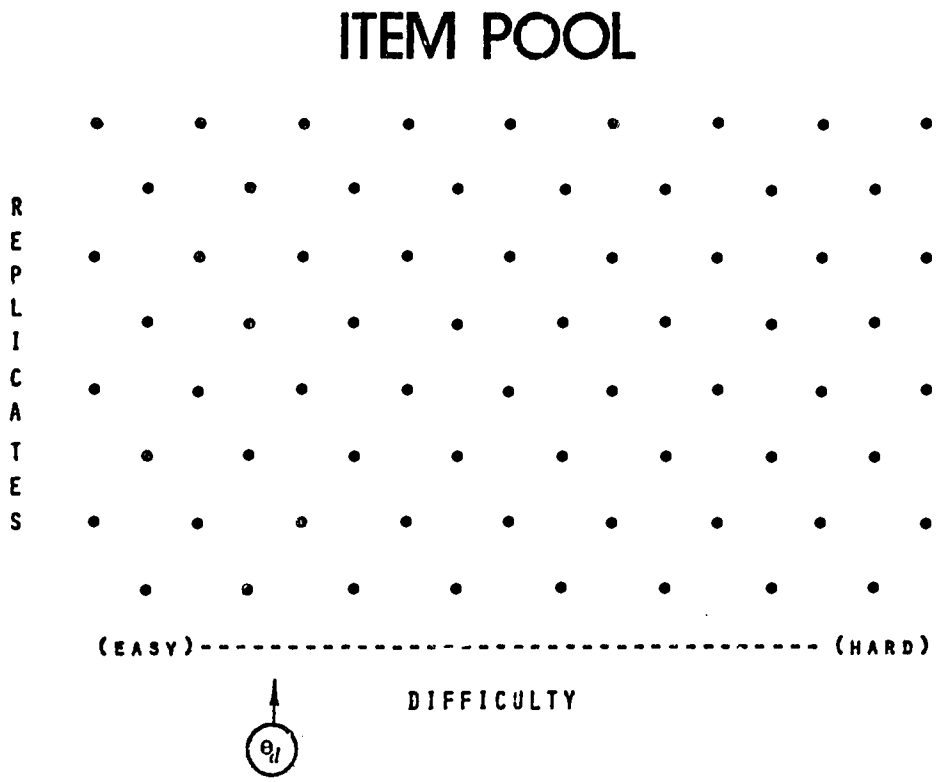
Strategies of Adaptive Testing

The bulk of our research during the last several years has been concerned with the first type of error. That is, we have been studying various methods for selecting items from a pre-calibrated pool, to match each individual's ability level as it is estimated during the process of testing. The basic premise of adaptive testing is this: an individual's ability level will be most accurately estimated when the items administered are as close to his/her ability level as possible. But, since ability is not known before testing--since that is the purpose of administering the test--we must choose items for each individual while testing is in progress. Thus, computerized adaptive testing uses an interactive computer system to administer tests; each item is chosen based on the testee's responses to previous items. Items are typically administered on a cathode-ray-terminal (CRT), and the testee responds on the CRT keyboard.

I will describe some strategies that have been used for selecting items in the framework of their evolution from the simple conventional test to complex adaptive or tailored testing models. To clarify the distinctions between some of the models we will follow the progress of a hypothetical low ability subject through a test administered under each strategy and note how his items are selected. We will further examine differences between strategies.

Figure 1 shows the item pool that will be used to describe the way the various testing strategies function. On the horizontal dimension we have 17 columns, each containing four items, ranging from very easy items at the left to very difficult items at the right. The vertical dimension represents replications of items at each difficulty level; all items in a column are equally difficult.

Figure 1



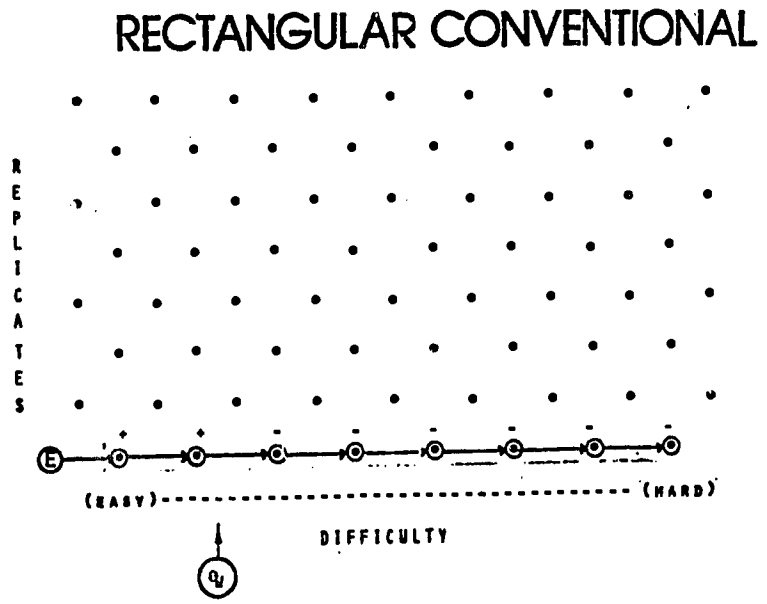
I will illustrate the various item selection strategies using eight items from this pool of 68. While an eight item test is convenient for illustration, eight items are too few for measurement of reasonable accuracy. Therefore, for evaluation of the strategies a 24-item test was used. Items for the 24-item test were chosen in a manner analogous to the way items were chosen for the illustrated eight-item test.

The results that I will present are from computer simulations. In order to make possible the analyses done for this presentation, some simplifying assumptions were made. First, it was assumed that a large pool of equally good items (i.e., items with equivalent

discriminating power) was available to chose from. Second, it was assumed that these were free-response items and, hence, guessing was not possible. Third, it was assumed that all tests were scored by a common technique, in this case, a Bayesian scoring procedure. Finally, to make comparisons between some strategies meaningful, it was assumed that a prior estimate of ability, correlating 0.5 with ability, was available.

One way to compose a test is to select a fixed set of items having a wide range of difficulties. Figure 2 shows such a rectangular conventional test. In this case, eight items equally spaced on the difficulty continuum were chosen from alternate columns ranging from the next to easiest to the next to most difficult columns. Our low ability

Figure 2

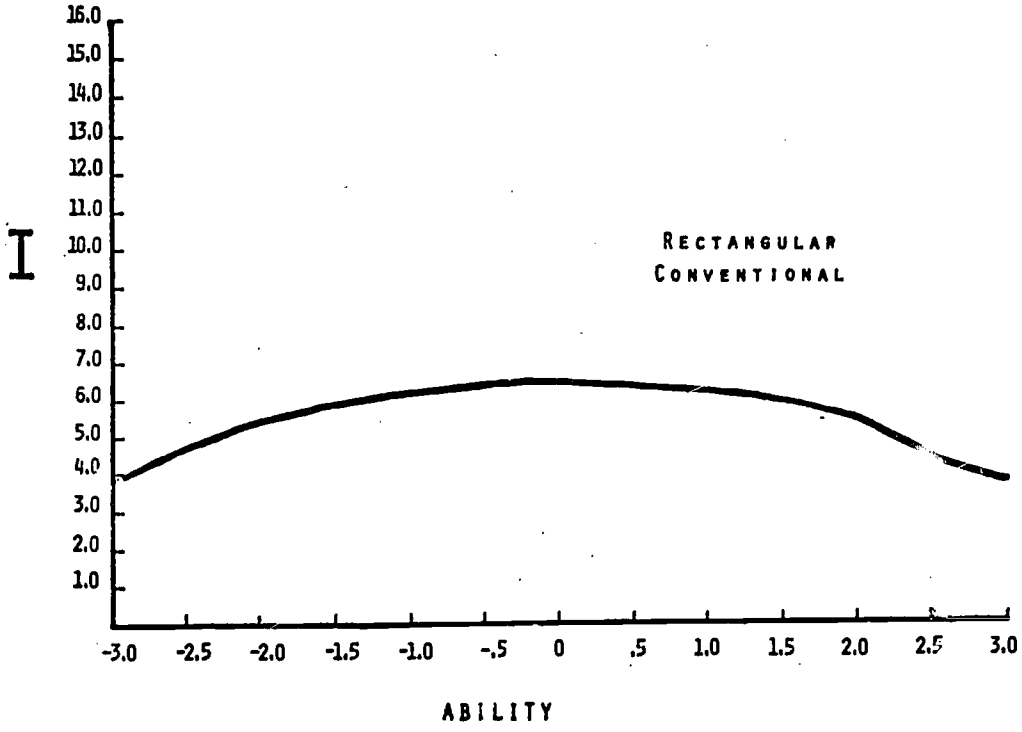


subject, produced the response record shown, with those items he answered correctly marked by a "+" and those he answered incorrectly indicated by a "-". The items in this test could have been administered in any order, but for clarity of presentation, we started at the left and worked toward the right.

The first item encountered was beneath the testee's ability level (Θ_d) and knowing the answer, he responded correctly. The second item was a bit more difficult but he still answered it correctly. The third item, being a bit above his ability was too difficult and he answered it incorrectly. Similarly, the fourth through eighth items were even more difficult and he answered all of them incorrectly.

Figure 3 shows an information curve produced by the rectangular conventional test. Information can be thought of as related to the precision of measurement produced by a test at a given level of ability, or as how well a test can discriminate between two contiguous ability levels. A good test produces an information function that is high

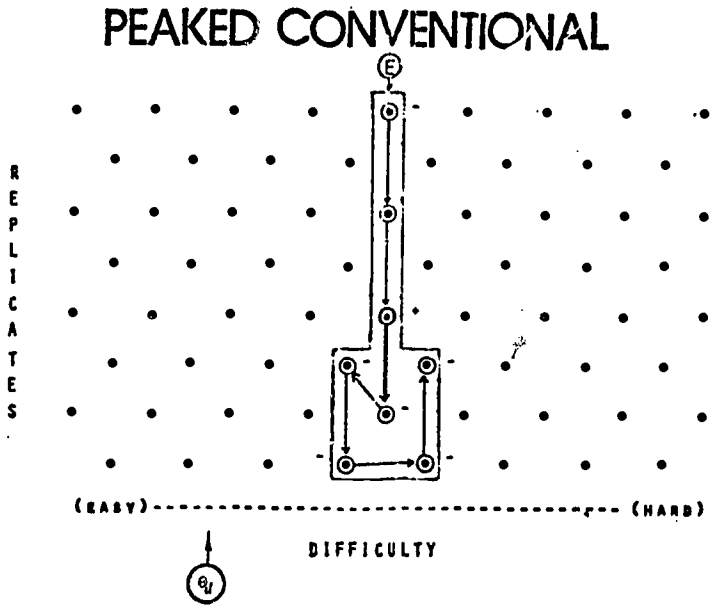
Figure 3



(i.e., provides precise measurement) and is flat (i.e., provides this high level of precision for all testees at all ability levels. Although not apparent from Figure 3, it will become obvious from comparisons with later results that the rectangular conventional test produces an information function that is fairly flat but somewhat low. It can be seen, however, that even this information function tapers off at the extremes indicating poorer measurement for testees where ability level is distant from the mean.

Instead of choosing items with a wide range of difficulty, we could instead choose items peaked at the center of the ability range and administer them to all testees. Figure 4 shows such a peaked conventional test. We chose the four items from the median difficulty column and two from each of the adjacent columns. Again, these items could have been administered in any order but we began at the top for clarity.

Figure 4

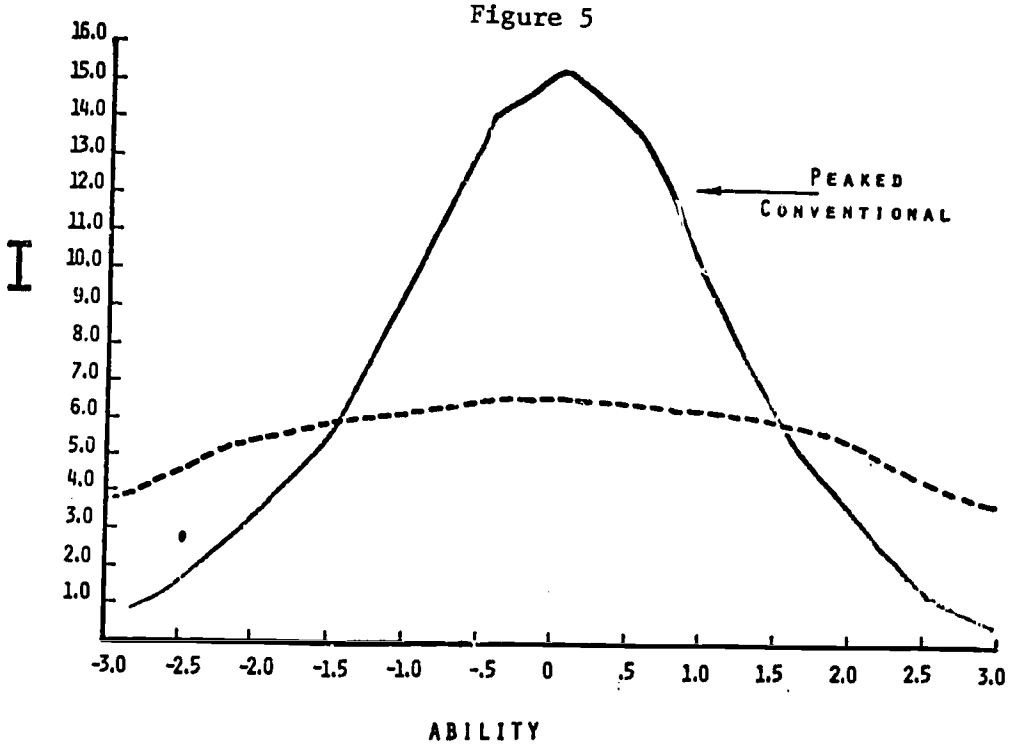


These items were intended for average ability testees and were all too difficult for our low ability testee. He missed the first item, the second item, and most of the rest of the items.

The information curve for the peaked conventional test (Figure 5) shows graphically what our testee felt as he took the test; the peaked conventional test provides good measurement for some testees but very poor measurement for others. As Figure 5 shows, the peaked conventional test produces precise measurement for individuals with abilities in the middle range but little information for extreme ability subjects. The peaked conventional test provides more information

about ability than does the rectangular conventional test within the range of ±1.5 standard deviations of ability but less outside of this range.

Figure 5



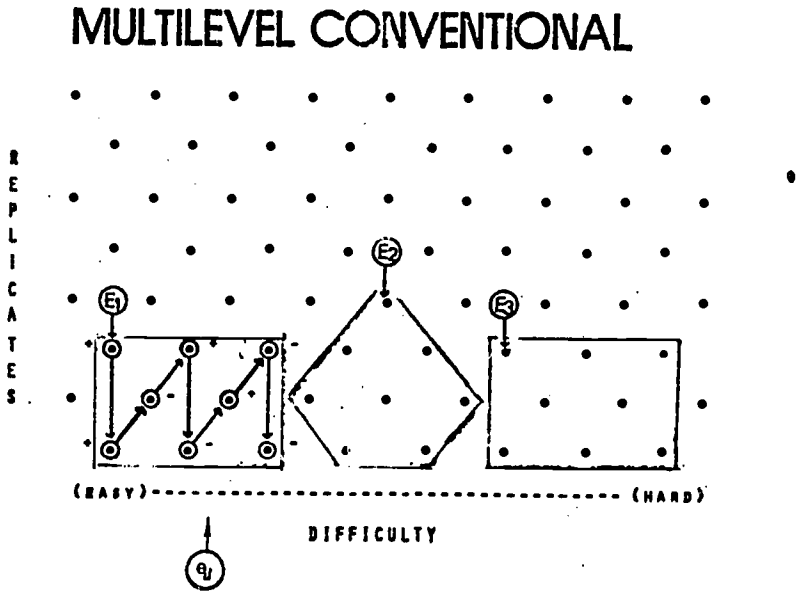
It seems that with a fixed set of items (i.e., a conventional test) we can please some of the people all of the time or all of the people some of the time but can't please all of the people all of the time. If, however, we could figure out a way to move a peaked ability test to the ability of each person being tested, we could please all of the people all of the time and provide a high level of information at all ability levels. If a testee's ability were known a priori, we would construct a test made up of those items with difficulties closest to his ability (i.e., items which he would be expected to answer correctly 50% of the time). But, if we knew his ability beforehand, we would have no reason to administer the test at all.

In practice we have, at best, a fallible prior estimate of the testee's ability and may want to administer items more or less

rectangularly distributed in a narrow range around his estimated ability. Some achievement tests use a prior ability estimate, such as grade in school, to determine which section of a test a testee should take.

Figure 6 illustrates such a test. Knowing that a testee ranked at the 27th percentile in his grade school graduating class, if this were a high school freshman achievement test, we might use this prior information to start him at the easiest entry point (E_1). Or, if we had a testee with straight A's in grade school, we might start him at the high entry point (E_3). Given a prior ability estimate, therefore, it is possible to adapt the test to the individual within the framework of a conventional test. But if prior information is not

Figure 6

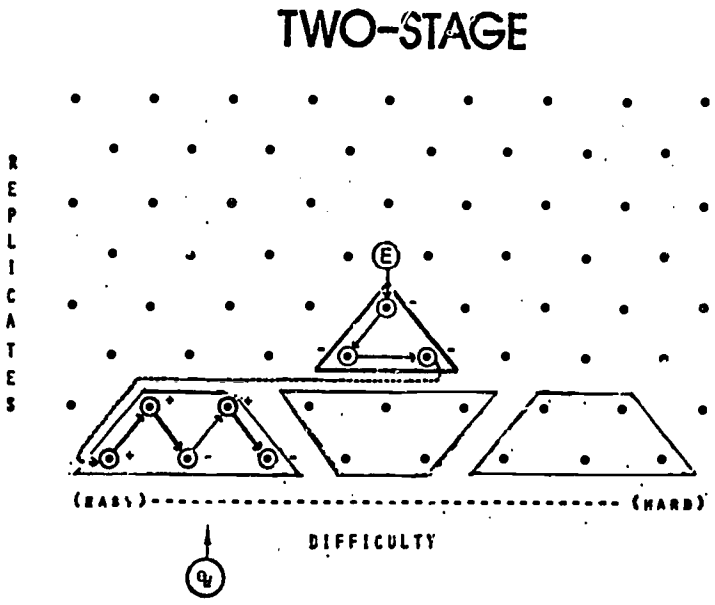


available, we have to use a test that tailors item difficulty in its absence. One possible strategy for doing this is the two-stage testing strategy which is like the previous test but generates its own prior ability estimate.

In a two-stage test, a testee is first administered a short routing test and, on the basis of his score on that test, is branched

to a measurement test of more appropriate difficulty. Figure 7 shows a two-stage test. A testee takes a three-item routing test and one of three five-item measurement tests. Our low ability testee answered all three of the routing test items incorrectly as they were too difficult for him. Since this suggested that his ability was low, he was branched to the easiest measurement test where he answered three out of the five items correctly.

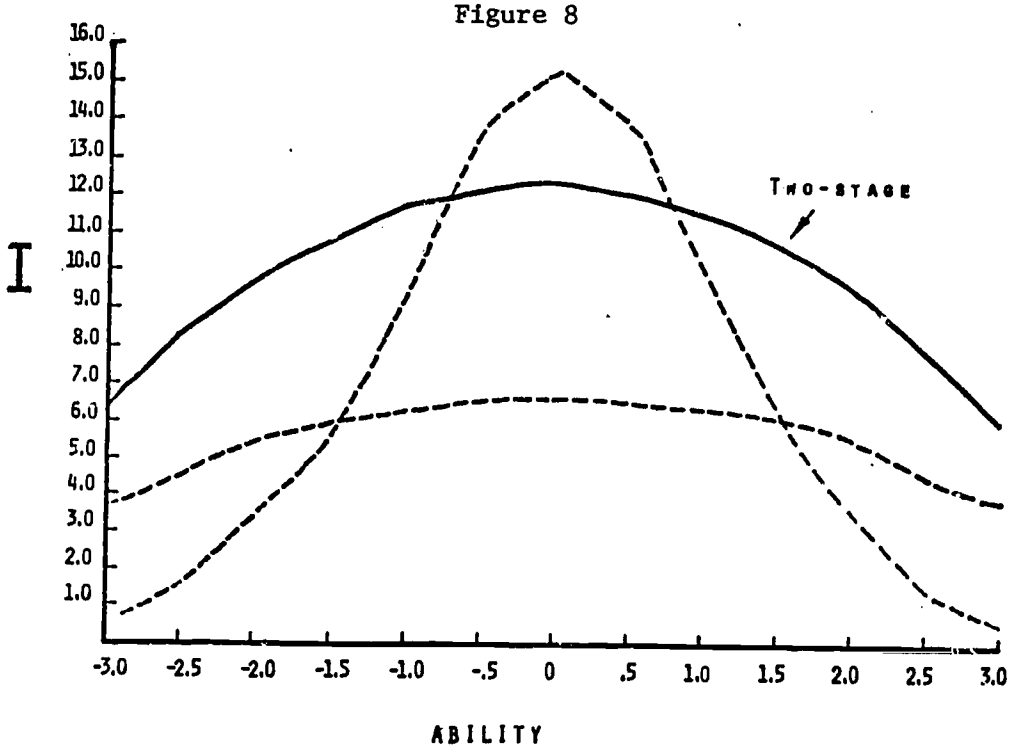
Figure 7



As Figure 8 shows, this two-stage test yields an information curve that is at all points higher than the rectangular conventional test and higher than the information curve of the peaked conventional test except in the center. So this two-stage test provides more precise measurement than the rectangular conventional test at all ability levels and more precise measurement than the peaked conventional test at most ability levels.

One problem with the two-stage testing strategy is that if a testee's ability is between the difficulties of two adjacent measurement tests, there is no measurement test of appropriate difficulty. A solution to this problem is available in the form of

Figure 8

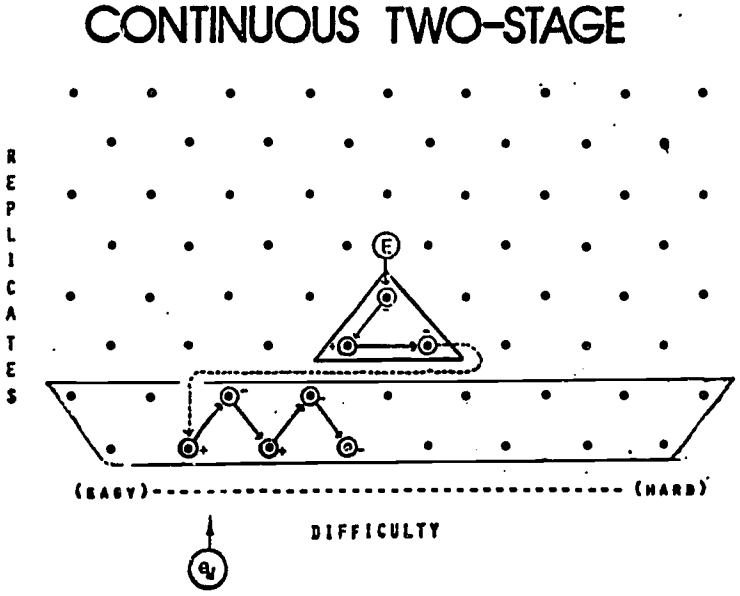


the continuous second stage two-stage test (Figure 9), a variant of the previous two-stage test. As in the standard two-stage test, the testee is first administered the routing test. Then, on the basis of the score on that test, he is branched to a measurement test. But instead of using one of a series of pre-structured measurement tests, a measurement test is individually composed for that individual using items closest to his ability estimate, plus items on either side. Given our restricted circumstances, the information curve of the continuous two-stage test would be very similar to that of the standard two-stage test and will not be shown here.

Another problem inherent in the two-stage procedure is that of misrouting. The measurement test decision is based on a short and fallible routing test and thus may be incorrect. There are two solutions to the misrouting problem: One is to route more; the other is to route less (i.e., not at all). An example of the latter strategy is the flexilevel test (Figure 10). For this test the potential item set is the same as the potential measurement test item set of

the continuous two-stage test. But, rather than taking a routing test, each testee starts with the median difficulty item of the

Figure 9



item set and following each correct response is branched to the next more difficult unadministered item. Following an incorrect response, he is branched to the next less difficult unadministered item.

In the case illustrated, the testee missed the first three items and was branched appropriately downward until he reached the third item below the median, an item slightly above his ability level. Knowing the answer, he answered it correctly and was branched to the first item above the median, which he answered incorrectly. He was branched to the fourth item below the median item and continued oscillating between easy and difficult items until he had answered eight items.

The information curve for the flexilevel test is shown in Figure 11. Although the flexilevel test solves the problem of misrouting, the

Figure 10

FLEXILEVEL

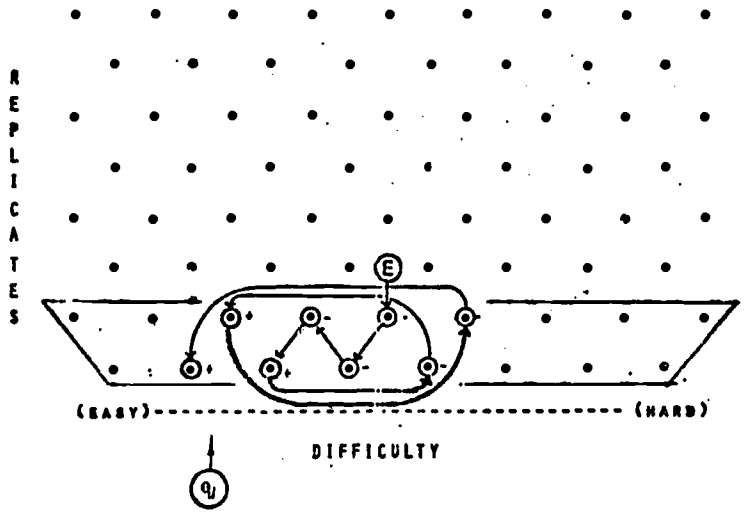
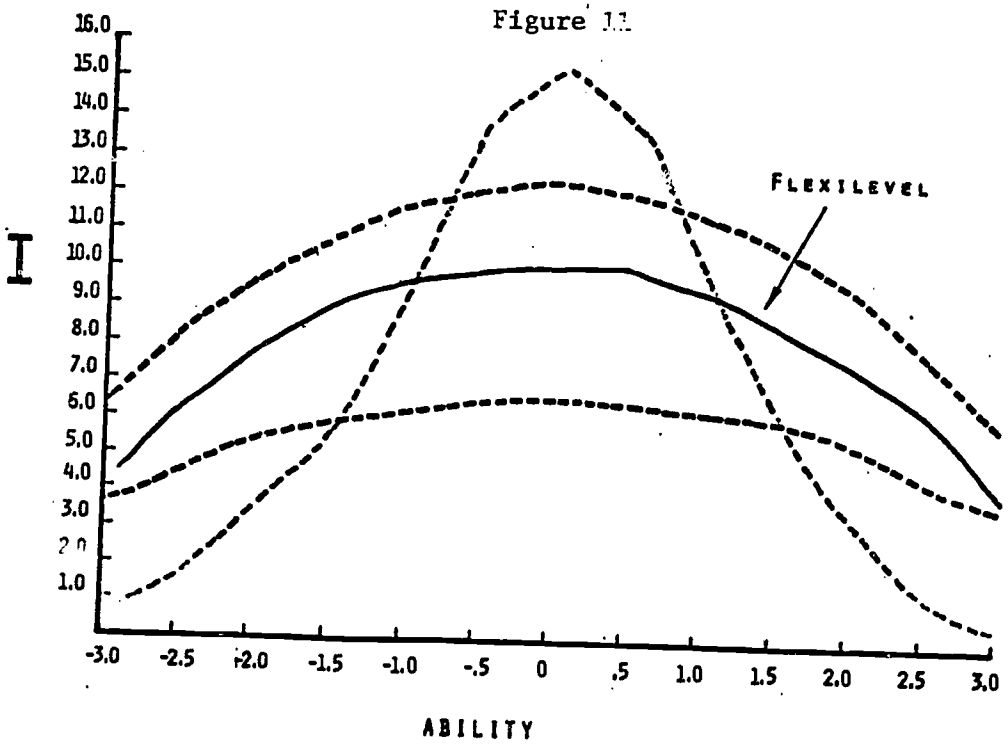


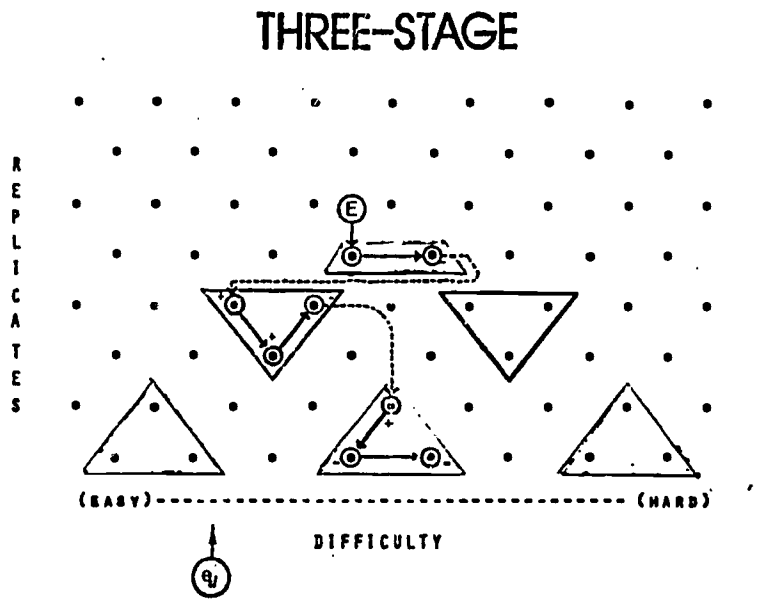
Figure 11



information it provides is always less than that provided by the two-stage test.

Figure 12 shows an example of the other solution to the problem of misrouting, the three-stage test (sometimes referred to as the double-routing two-stage test). In this strategy, an individual takes one routing test which routes him to a second routing test which routes him to a measurement test. Errors resulting from the first routing can be ameliorated by the second routing.

Figure 12



Carrying the idea of multiple routing to its logical extreme, and using one item per stage, results, in this case, in the eight-stage test or, in the general case, the pyramidal test. In this strategy (Figure 13), a testee starts with a median difficulty item and is branched after each item. A less difficult item is administered following an incorrect response, and a more difficult item is administered following a correct response.

The information curve for this test (Figure 14) shows it to provide more information than any of the strategies discussed thus far, except in the middle ability range where it is slightly surpassed by the peaked conventional test. It should be noted, however, that the information

Figure 13

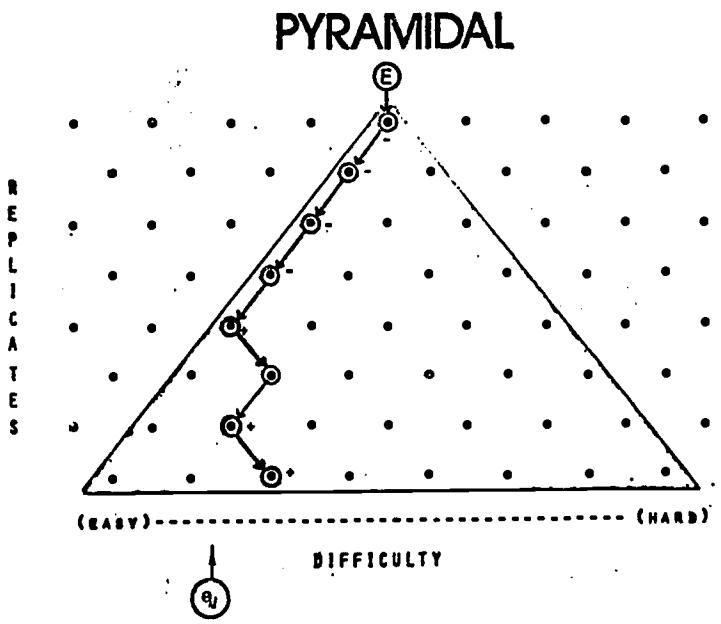
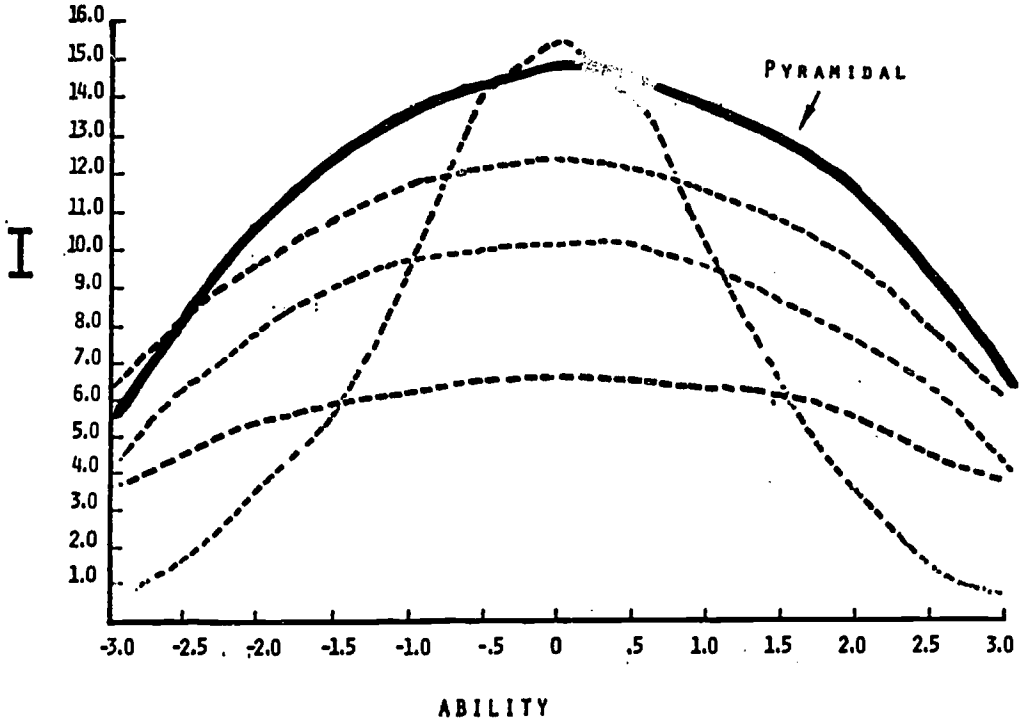


Figure 14

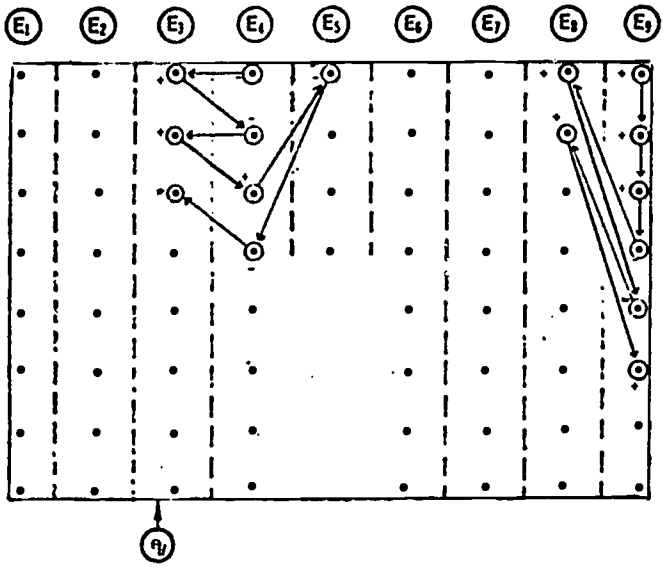


curve is far from flat. Less than half of the amount of information provided at the middle range of ability is provided at the extremes of this information curve, three standard deviations from the mean.

The previously discussed adaptive tests have been developed for the situation in which prior ability information was not available and are not capable of using it when it is available. Now that we have reached the top of the pyramid, so to speak, we can make use of prior information by extending the pyramidal structure to allow entry at several points. A direct extension is unable to handle branching for some extreme ability testees, however, so a modified extension of the pyramidal structure is used by the stratified-adaptive (stradaptive) testing strategy shown in Figure 15. Two changes beyond a direct extension are observed: 1) items are grouped into strata consisting of items of possibly slightly different difficulty; and 2) branching is between strata with the item selected being the first unadministered item in a stratum.

Figure 15

STRADAPTIVE

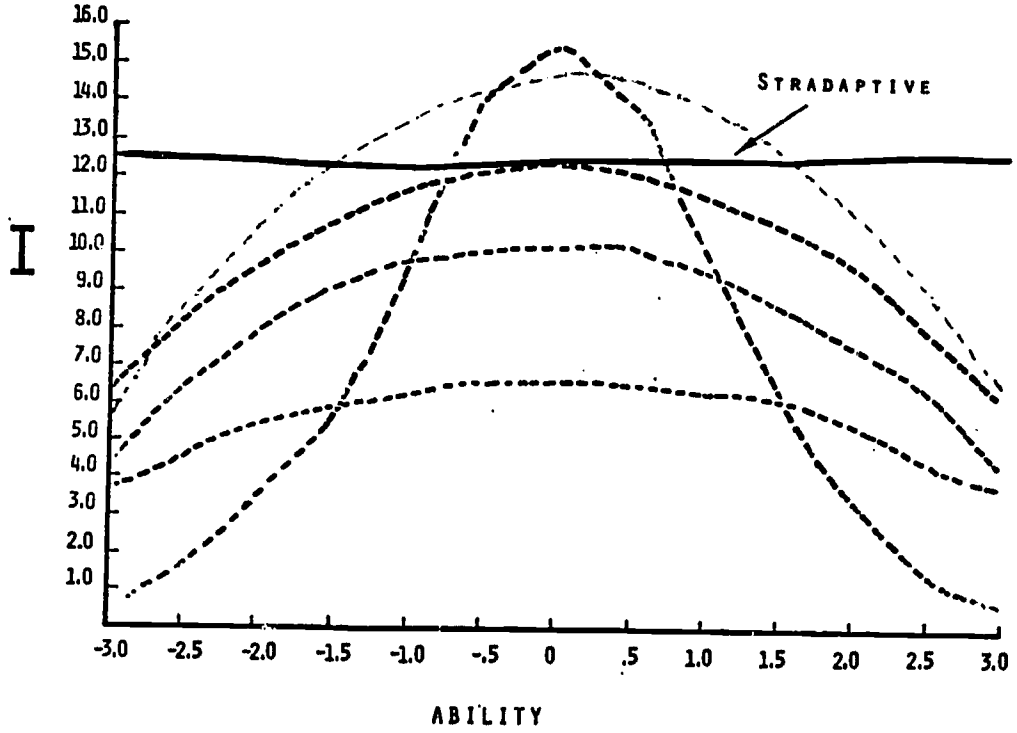


The testee started at the fourth entry point. He missed the first item in stratum four, was branched to the first item in stratum three,

got this item correct, and alternated between these two strata until his fifth item. He answered the fifth item, which was in the fourth stratum, correctly and was branched to the first item in the fifth stratum. He incorrectly answered this and the next item and finished with his eighth item in the third stratum.

Branching to the first item in a stratum is of little value in a situation where all items are equally discriminating, but is useful when using a real item pool because all items will not be equally discriminating. This feature allows the most discriminating items to be put where they have the highest probability of being administered; as the first items to be administered in each stratum. The information curve for the stradaptive test (Figure 16) is almost flat indicating that it provides very equiprecise measurement. Its level is surpassed by several other strategies in the center, however.

Figure 16



The previous adaptive strategies are all among the fixed branching strategies. The branching has been a function solely of the testee's performance at the immediately preceeding stage. The variable branching

procedures calculate an ability estimate after each item and select as the next item the item best suited for an individual of that ability.

An example of the variable branching procedures is the Bayesian strategy, which is illustrated in Figure 17. On the basis of a prior ability estimate, which may be simply the mean ability of the population of testees, a first item is selected. On the basis of the response to that item and a prior ability distribution, which may consist simply of population parameters, a score is calculated and on the basis of that score, another item is selected. This procedure is repeated, each time selecting the one item in the pool which is closest in difficulty to the last ability estimate.

Figure 17

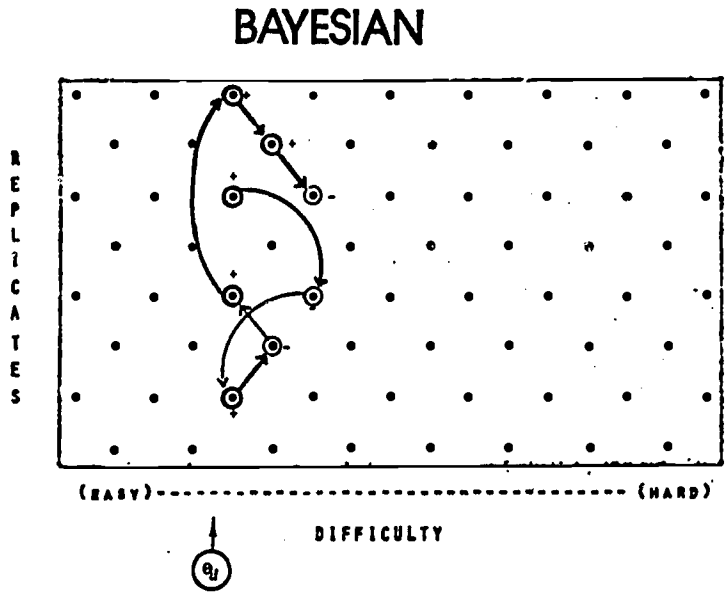


Figure 18 shows a report derived from live testing with a Bayesian adaptive test. In this figure, an "X" indicates a correct response to an item, while an "O" indicates an incorrect response; the "E" indicates the entry ability estimate, based on a rough prior estimate of the testee's ability level. The dotted lines on either side of these symbols indicate the standard deviation of the ability estimate, a value analagous to the standard error of measurement for that ability estimate. Note how the ability estimate itself (i.e., the E,X, or O) changes after each item response. Note also that the range of change

in the ability estimate decreases as testing proceeds. This illustrates the convergence nature of the Bayesian process. Similarly, the error of the ability estimate decreases after each item response, with the amount of decrease reducing at each stage of the testing procedure.

Figure 18

REPORT ON BAYESIAN TEST

X=CORRECT 0=INCORRECT ?=NO RESPONSE .
ERROR BAND PLOTTED IS + AND - STANDARD DEVIATION

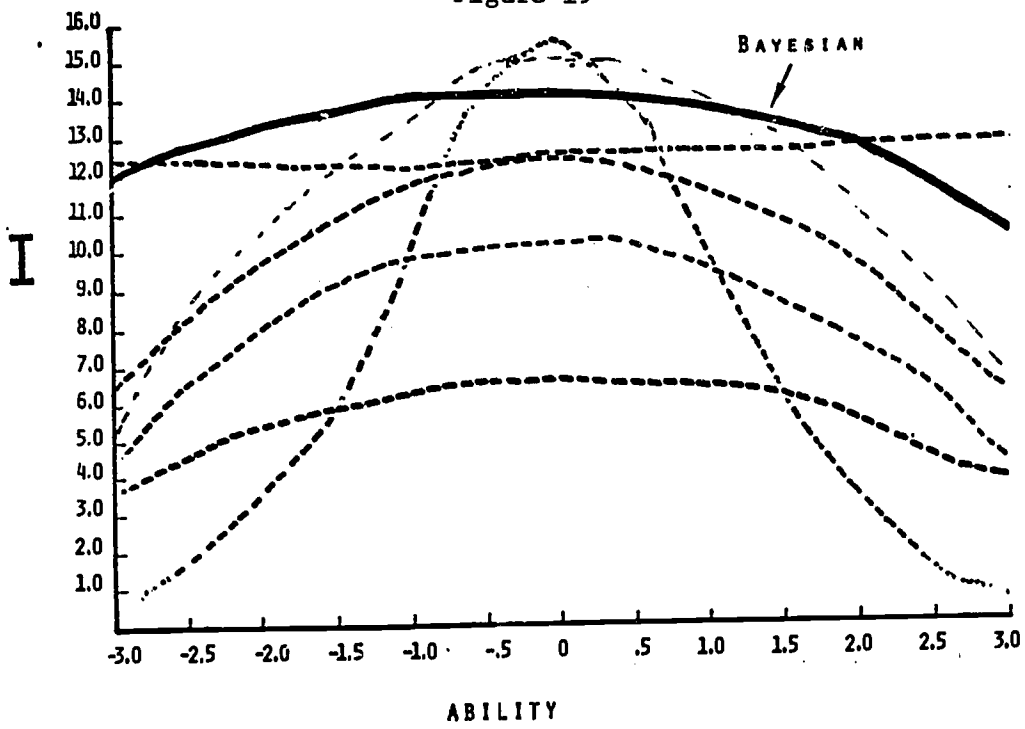
	LOW ABILITY LEVEL HIGH										POSTERIOR ABILITY	
STAGE	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	EST	SD
0						E						
1						X					-.65	1.87
2						0					.19	1.30
3						X					-.85	.94
4						0					-.36	.76
5						0					-.73	.66
6						0					-1.07	.65
7						0					-1.59	.50
8						X					-1.42	.44
9						X					-1.29	.40
10						X					-1.20	.39
11						0					-1.30	.36
12						0					-1.40	.36
13						X					-1.31	.35
14						X					-1.24	.33
15						X					-1.17	.33
16						0					-1.23	.32
17						0					-1.28	.32
						0					-1.35	.30

17 ITEMS WERE ADMINISTERED

The information curve from the Bayesian testing procedure is shown in Figure 19. It is slightly higher than the stradaptive test's information curve and nearly as flat, although it drops more in the tails. The peaked conventional test and the pyramidal test still provide more information in the center of the ability distribution.

If the evaluation of adaptive testing strategies were as simple as this presentation, however, our research would be unnecessary. This evaluation was very limited in a number of ways, which seriously restrict the generalizability of these findings.

Figure 19

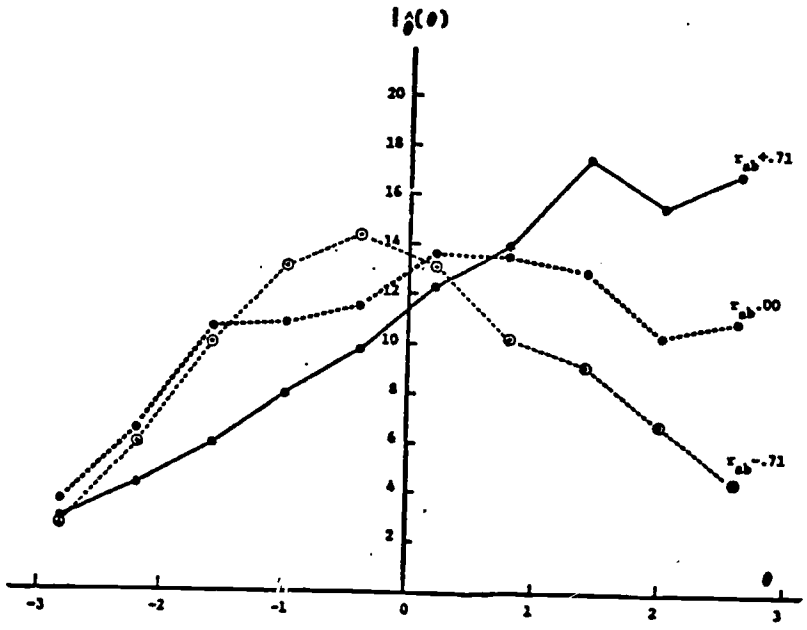


First, the information curves were calculated using a response model which may not accurately portray response tendencies of real subjects. For example, on multiple-choice tests some testees will guess. And guessing will affect the accuracy of measurement for some testees. Figure 20 shows information curves for the Bayesian adaptive test when random guessing is introduced into the response model. The curve labelled r_{ab}.00 is analagous to the data previously shown for the Bayesian strategy except that random guessing was allowed. Note that the information curve in Figure 20 is not horizontal, as it was in Figure 19. Rather, the Bayesian test provides decidedly poorer measurement for testees below mean ability (<0) than it does for those with higher abilities. Thus, comparisons of testing strategies will change as the response model changes.

A second limitation of these results is that they were based on an unrealistic item pool. First, the item pool included only 68 items;

real item pools for adaptive testing will require about 200 items per ability. Secondly, the item pool consisted of items with equal and

Figure 20



Smoothed curves of the information functions of the Bayesian sequential test under three different item pool difficulty-by-discrimination configurations.

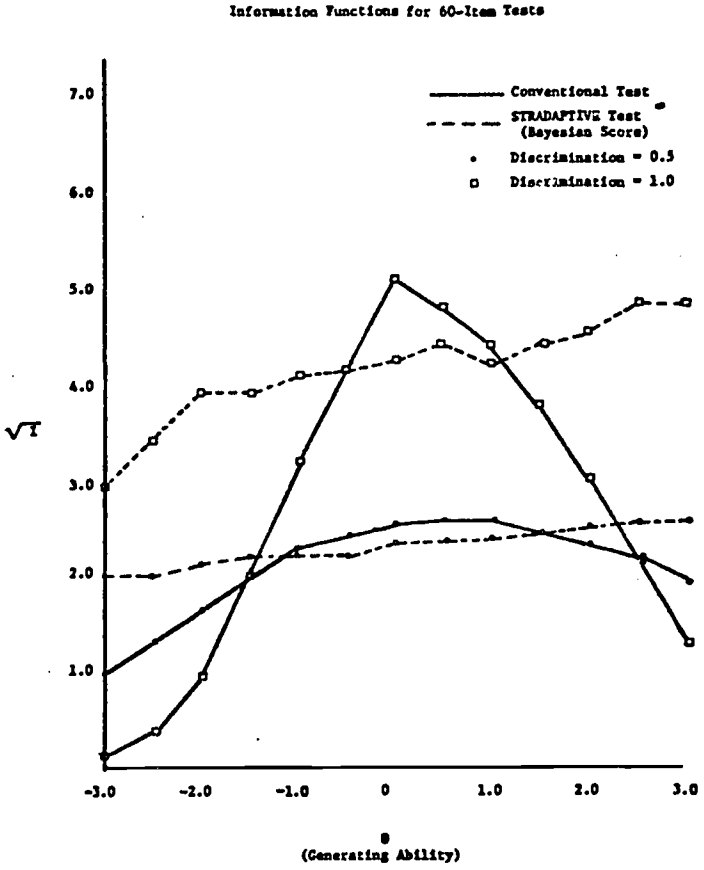
high discriminations. If an item pool consists of items whose discriminations are correlated with their difficulties, as is usually found in real item pools, the information curves will also change shape.

For example, the other two information curves shown in Figure 20 were derived from the Bayesian test using an item pool in which the more discriminating items were of higher difficulty (r_{ab}+.71) and another in which the more discriminating items were less difficult (r_{ab}-.71). As Figure 20 shows, information curves under these, more realistic, item pool configurations are far from horizontal. Under both item pool configurations, the Bayesian test loses its capability of providing measurement of equal precision throughout the ability range.

Not all adaptive testing methods are as seriously affected by

characteristics of the response model or the item pool as is the Bayesian strategy. Figure 21 shows information curves for the stradaptive test (and a conventional test) with guessing. With items of low discrimination (a=.5) the stradaptive information curve is still quite horizontal. For items of higher discrimination, the information curve drops somewhat for the low ability testees, but remains reasonably horizontal through most of the ability range.

Figure 21



The results I've just presented are limited in yet another way. That is, all comparisons are in terms of information curves. Although information curves are a very valuable way of studying the relative utility of testing strategies, they don't tell the whole story. Testing strategies can also be compared in terms of the statistical bias in the scores they provide. Holding ability constant, bias can be defined as the difference between ability level and the average

ability estimate for all testees at that ability level. If the average ability estimate is equal to the ability level, the scores are unbiased for that ability level. The greater the difference between ability and ability estimate, the greater the bias. Bias is particularly important if it differs at different ability levels or, in other words, if ability and ability estimate are curvilinearly related.

Figure 22

BIAS CURVES FOR BAYESIAN AND MAXIMUM LIKELIHOOD SCORING

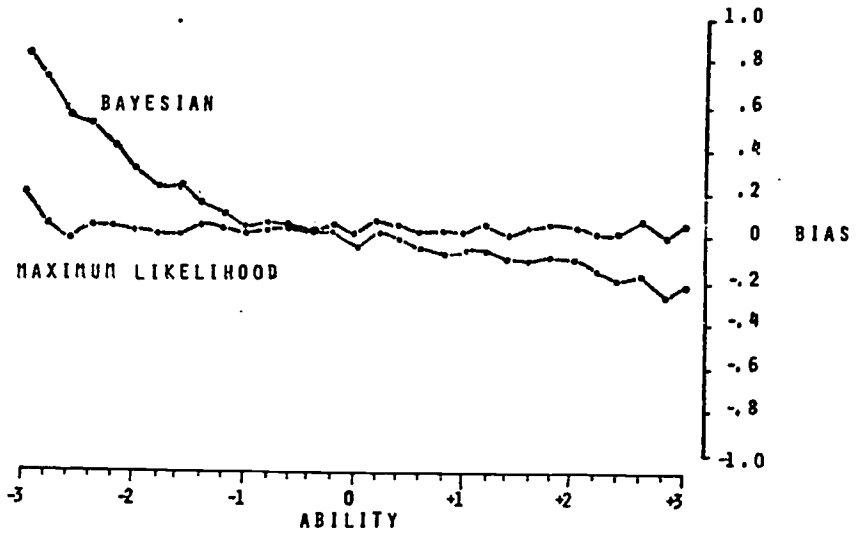
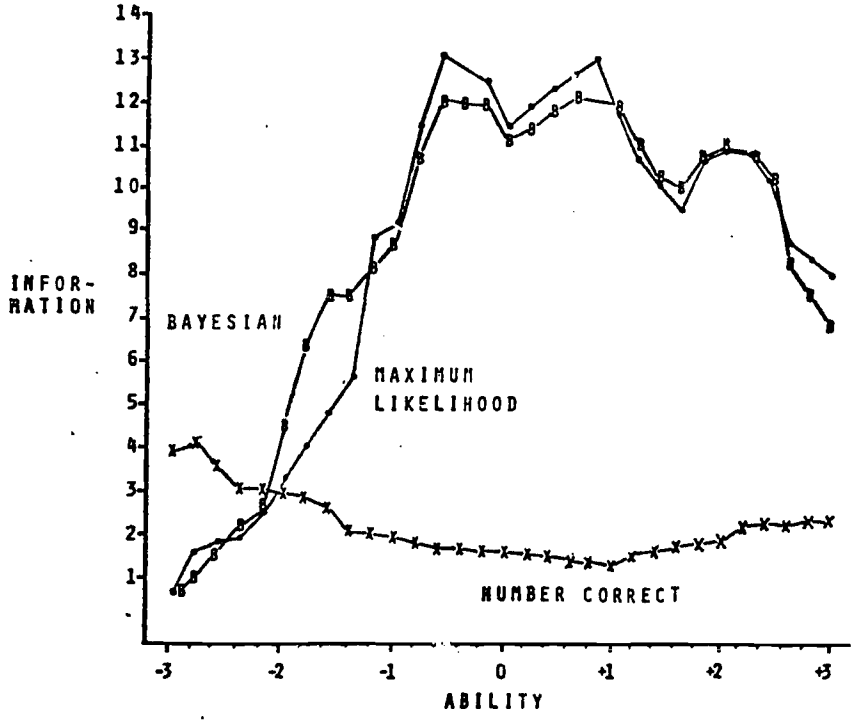


Figure 22 illustrates the bias characteristics of two methods of scoring the same adaptive test. From this we can extrapolate to the kinds of bias curves (which we haven't yet studied) which might result from two different adaptive strategies using the same scoring method. As Figure 22 shows, a Bayesian scoring technique applied to a set of data results in scores which are increasingly biased as ability deviates from the mean. Maximum likelihood scoring, applied to the same item response data, results in scores which are essentially unbiased estimators of true ability. But the information curves for the two scoring methods, shown in Figure 23, reflect very little difference between the information characteristics of the two scoring methods. Thus, different evaluative criteria (e.g., bias vs. information) can lead to different conclusions about scoring methods, in this case, or adaptive testing strategies, in general. Incidentally, Figure 23 also shows the information curve derived

from number correct scoring on a Bayesian adaptive test. As can be seen, number correct score provides a very low level of information in most of the ability range. However, for very low ability testees, when random guessing is in effect, number correct score is more useful than either the Bayesian or maximum likelihood scoring methods.

Figure 23

INFORMATION CURVES FOR THREE SCORING METHODS



Some researchers in adaptive testing evaluate the "goodness" of a testing strategy in terms of the correlation between ability level and ability estimate (e.g., test score), within simulation studies. However, using this correlation as the sole evaluative criterion for comparing strategies is inappropriate, since it conceals a substantial amount of information. Figure 21 shows information functions for conventional and stradaptive tests, when both consist of items with the same discriminations. A comparison of the upper two curves in that figure show that the stradaptive test yields measurement of almost

constant precision throughout the ability range, although the conventional test measures more accurately for average ability testees. The correlations of test score with ability for these data, however, are .97 for stradaptive and .95 for the conventional test. From the correlations alone, we would conclude that the stradaptive test is slightly better than the conventional test, but not dramatically so. But the information functions show considerable differences in the measurement accuracy of the two testing strategies for testees of different ability levels.

Similarly, the product-moment correlation will not reflect bias in ability estimates, as illustrated in Figure 22. Since the bias in the Bayesian score is non-linear, the correlation of ability and ability estimate will not include that non-linearity. Thus, the two scoring methods shown in Figure 22 would be evaluated similarly by the use of correlation indices, but they provide scores with quite different characteristics. And different scores will result in different decisions about people.

To summarize, we do not yet know which are the best strategies of adaptive testing. We do know, however, that adaptive tests in general have much better measurement characteristics than conventional tests, in which the same items are administered to all testees. The evaluation of adaptive testing strategies to identify those which are best will depend in part on the complex interaction of such variables as evaluative criteria, scoring methods, item pool characteristics and branching methods. We have considerable research to do on this topic, but should have some firmer answers within the next year or so.

Intra-Individual Dimensionality

As I indicated earlier, deviations from unidimensionality can result in errors in test scores. This is particularly true when summative scores are used, as they almost always are, since summation of any kind assumes one dimension underlying the responses that are summed to yield a total score. Thus, to the extent that two individuals obtain the same total score in two different ways, it can be assumed that they are not operating within the same unidimensional scale.

Some adaptive testing models permit us to begin to study the dimensionality of a particular individual's response record on an ability test. Given this capability, if we can identify different levels of intra-individual unidimensionality, resulting from the interaction of different individuals with the same "unidimensional" item pool, we can then study the consequences of deviations from unidimensionality in terms of both psychometric criteria and practical utility. One such hypothesis we can make is that scores which are unidimensionally determined should be more error-free than scores which are non-unidimensionally determined.

In ability measurement, we would expect that an individual should, in general, respond correctly to items below, or easier than, his ability level, and incorrectly to items above, or more difficult, than his ability level. If a person answers most easy items correctly and most difficult items incorrectly, we would say that he is responding consistently--that is, his response pattern seems to be influenced primarily by his position on the underlying trait continuum. However, if a person gets many easy items wrong and many difficult items correct, he is responding inconsistently, indicating that something besides the trait of interest is influencing his responses. Thus, inconsistency of this type reflects lack of unidimensionality.

In an ability test, response inconsistency may be caused by such extraneous variables (i.e., other response dimensions) as guessing, partial knowledge, or adverse psychological conditions such as test anxiety or lack of motivation to do one's best on the test. Whatever its cause, response inconsistency may reduce the reliability and/or validity of a given test score, and knowing the degree of consistency of an individual's response pattern may be important when we intend to use that score in making practical decisions.

We have operationalized the notion of response consistency in the stradaptive testing strategy. As you may recall (see Figure 15), in the stradaptive test items are organized into a series of levels or strata according to their difficulty. A correct response to an item in one stratum leads to the administration of the most discriminating item remaining in the next more difficult stratum. An incorrect response leads to the administration of the most discriminating item remaining in the next less difficult stratum.

Figure 24 shows a relatively consistent response pattern on the stradaptive test along with 10 ability scores and five consistency scores. This person entered the stradaptive test at stratum 5, based on some prior information. Stratum 5 items were too easy for him and he answered items correctly until, at item 4, he had been branched to stratum 8, which contained very difficult items. Notice that he consistently responded incorrectly to the stratum 8 items, which were too difficult for him, and correctly to the stratum 6 items, which were too easy for him. The items in stratum 7 seem most appropriate in difficulty for him, and he answered about half of them correctly and the other half incorrectly.

The consistency of this individual's response pattern was reflected in his relatively low consistency scores. Score 11, the standard deviation of the difficulties of the items encountered by this person, was .59. Further, in the stradaptive test, items are administered until a termination criterion is reached. Similar to the Stanford-Binet, the stradaptive test terminates when a stratum is identified at which the testee answers no items correctly, or only a chance number. The consistency of this individual's response pattern

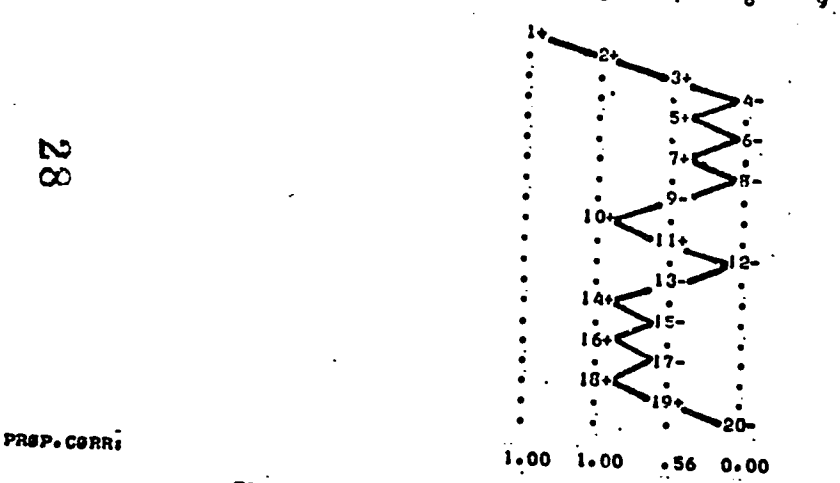
Figure 24

Report on a Stradaptive Test for a Consistent Testee

REPORT ON STRADAPTIVE TEST

ID NUMBER:

DATE TESTED: 73/07/12



TOTAL PROPORTION CORRECT= .550

SCORES ON STRADAPTIVE TEST

Ability Level Scores

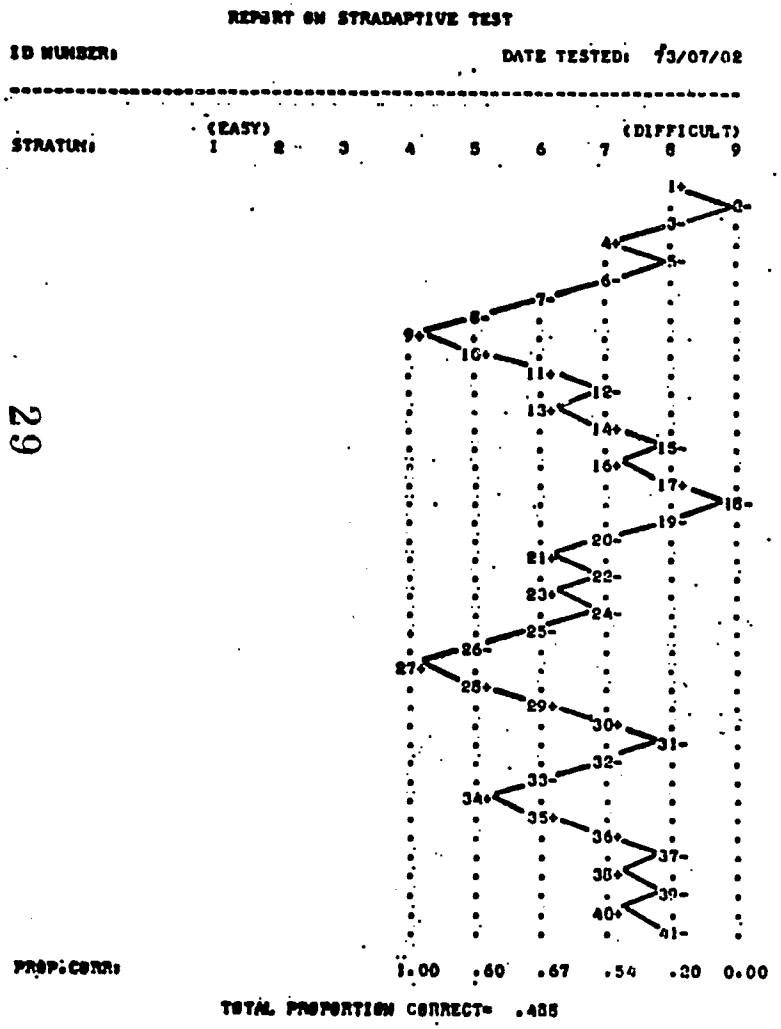
1. DIFFICULTY OF MOST DIFFICULT ITEM CORRECT= 1.49
2. DIFFICULTY OF THE N+1 TH ITEM= 1.44
3. DIFFICULTY OF HIGHEST NON-CHANCE ITEM CORRECT= 1.49
4. DIFFICULTY OF HIGHEST STRATUM WITH A CORRECT ANSWER= 1.33
5. DIFFICULTY OF THE N+1 TH STRATUM= 1.33
6. DIFFICULTY OF HIGHEST NON-CHANCE STRATUM= 1.33
7. INTERPOLATED STRATUM DIFFICULTY= 1.37
8. MEAN DIFFICULTY OF ALL CORRECT ITEMS= .88
9. MEAN DIFFICULTY OF CORRECT ITEMS BETWEEN CEILING AND BASAL STRATA= 1.28
10. MEAN DIFFICULTY OF ITEMS CORRECT AT HIGHEST NON-CHANCE STRATUM= 1.28

Consistency Scores

11. SD OF ITEM DIFFICULTIES ENCOUNTERED= .59
12. SD OF DIFFICULTIES OF ITEMS ANSWERED CORRECTLY= .46
13. SD OF DIFFICULTIES OF ITEMS ANSWERED CORRECTLY BETWEEN CEILING AND BASAL STRATA= .18
14. DIFFERENCE IN DIFFICULTIES BETWEEN CEILING AND BASAL STRATA= 1.36
15. NUMBER OF STRATA BETWEEN CEILING AND BASAL STRATA= 1

Figure 25

Report on a Stradaptive Test for an Inconsistent Testee



SCORES ON STRADAPTIVE TEST

Ability Level Scores

1. DIFFICULTY OF MOST DIFFICULT ITEM CORRECT= 1.89
2. DIFFICULTY OF THE N+1 TH ITEM= 1.01
3. DIFFICULTY OF HIGHEST NON-CHANCE ITEM CORRECT= 1.83
4. DIFFICULTY OF HIGHEST STRATUM WITH A CORRECT ANSWER= 2.01
5. DIFFICULTY OF THE N+1 TH STRATUM= 1.33
6. DIFFICULTY OF HIGHEST NON-CHANCE STRATUM= 1.33
7. INTERPOLATED STRATUM DIFFICULTY= 1.36
8. MEAN DIFFICULTY OF ALL CORRECT ITEMS= .72
9. MEAN DIFFICULTY OF CORRECT ITEMS BETWEEN CEILING AND BASAL STRATA= .76
10. MEAN DIFFICULTY OF ITEMS CORRECT AT HIGHEST NON-CHANCE STRATUM= 1.24

Consistency Scores

11. SD OF ITEM DIFFICULTIES ENCOUNTERED= .86
12. SD OF DIFFICULTIES OF ITEMS ANSWERED CORRECTLY= .74
13. SD OF DIFFICULTIES OF ITEMS ANSWERED CORRECTLY BETWEEN CEILING AND BASAL STRATA= .50
14. DIFFERENCE IN DIFFICULTIES BETWEEN CEILING AND BASAL STRATA= 2.64
15. NUMBER OF STRATA BETWEEN CEILING AND BASAL STRATA= 3

enabled him to meet the termination criterion after only 20 items had been administered.

Contrast this testee with the one shown in Figure 25. This person's response pattern was far less consistent and ranged over a larger number of strata. For example, this person answered some relatively easy items at stratum 5 incorrectly (note items 8 and 26) and answered some difficult items at stratum 8 correctly (items 1 and 17). By responding inconsistently, it took many more items before the termination criterion was reached, and the individual's consistency scores are higher, reflecting less consistency.

Figure 26

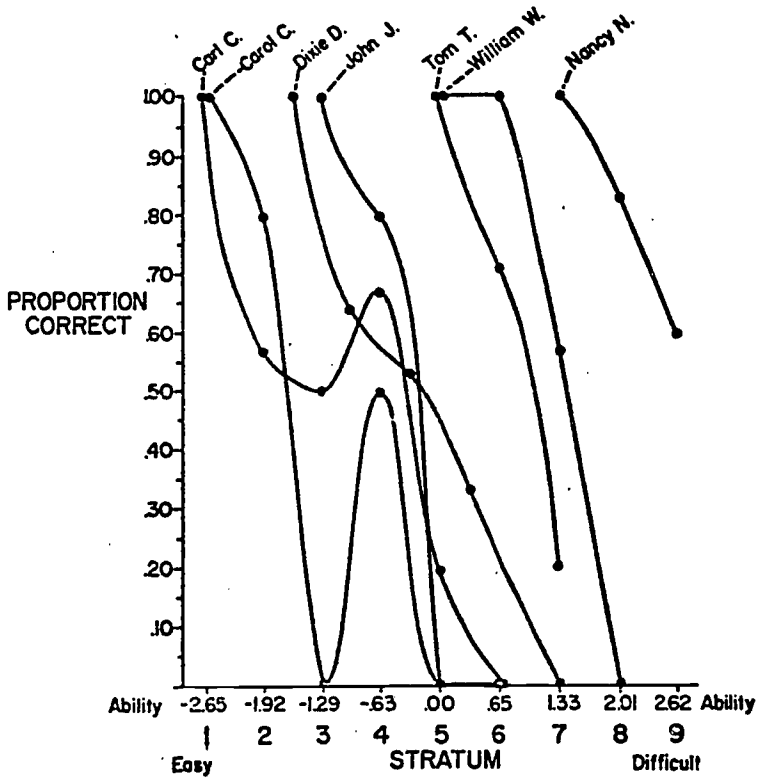


Figure 11. Proportion correct at each stratum, by individual

That consistency is related to dimensionality is illustrated in Figure 26. That figure shows a plot of proportion correct by stratum in the stradaptive test or what I have called "subject characteristic

curves", for eight different testees. If the testee is responding unidimensionally, or consistently, his subject characteristic curve should show a regular decrease with increasing item difficulty--such is the case for Nancy N., William W. and Tom T., in Figure 26. For these testees, all consistency scores would be low. For the inconsistent testee, or one who is responding non-unidimensionally, proportion correct does not decrease regularly with increasing item difficulty (as is the case for Carl C. and Carol C.) or it decreases more slowly (as for Dixie D.). For these testees, inconsistency scores will be considerably higher.

We used data from live administration of the stradaptive test to study the hypothesis that the scores of individuals who are responding unidimensionally should be more error-free than those of individuals who are responding non-unidimensionally. To study this hypothesis, we used test-retest stability as an indication of score reliability, and divided a group of 200 subjects into 5 groups, according to their consistency scores on the first stradaptive test administration, in a test-retest design. Within each group, we calculated the test-retest stability of the obtained scores. Table 1 shows the results obtained for consistency score 11, the standard deviation of the difficulties of all items encountered.

As Table 1 shows, the highest test-retest stability was found in the most consistent group of examinees for all 10 ability scores. The clearest pattern is that for ability score 1, where the scores in the most consistent group had a test-retest stability of .94, and the scores in the least consistent group had a stability of .65. The stabilities in the intermediate groups decreased with decreasing consistency. Note also that the stability for the most consistent examinees on scores 8 and 9 was .98, an extremely high five-week test-retest correlation.

The possible utility of consistency scores as a moderator variable is that they might permit us to make more stable predictions for some groups of individuals (consistent testees). If these results can be replicated over longer periods of time, the consistency score might prove to be a very useful and powerful moderator variable derivable from a stradaptive testing response record. It appears to be powerful because it also moderates the test-retest reliability, but not as systematically, on the conventional test administered at the same time. Table 1 shows a test-retest reliability of .979 on the conventional test for the highly consistent group using the consistency scores derived from the stradaptive test. But consistency scores are not derivable from a conventional test so it is necessary to implement this finding within the framework of the stradaptive testing strategy.

Thus by studying the consistency or dimensionality of a set of item responses we might be able to identify individuals whose scores on a given test are more error-free. For these individuals we will

Table 1

STRADAPTIVE AND CONVENTIONAL TEST
TEST-RETEST CORRELATIONS AS A
FUNCTION OF CONSISTENCY SCORE 11
ON INITIAL TESTING

		STATUS ON CONSISTENCY SCORE 11				
		VERY HIGH	HIGH	AVERAGE	LOW	VERY LOW
MEAN CONSISTENCY SCORE		.517	.625	.706	.815	1.038
NUMBER OF TESTEES IN INTERVAL		27	30	41	43	29
STRADAPTIVE ABILITY SCORE:	1	.940	.849	.847	.768	.652
	2	.875	.721	.799	.778	.751
	3	.956	.813	.879	.826	.708
	4	.934	.840	.869	.731	.664
	5	.896	.722	.791	.756	.741
	6	.950	.798	.886	.820	.704
	7	.970	.844	.902	.851	.758
	8	.981	.927	.915	.853	.869
	9	.983	.939	.907	.899	.889
	10	.951	.792	.882	.822	.718
CONVENTIONAL TEST		.979	.890	.918	.826	.878

be able to have more confidence and greater accuracy in making long-term predictions, and consequently increase our validity in the prediction of occupational criteria.

Psychological Effects

In the past, psychometricians have paid considerable attention to characteristics of tests administered to groups, for example, their reliability and validity. But we have ignored the fact that it is an individual who takes a test, not a group. Highly valid and reliable tests can be rendered useless for an individual if we do not have the cooperation of each individual or if that individual, for one reason or another, is not performing to his or her fullest capacity. For example, substantial amounts of error in the test score of an individual may result if that person's performance is hindered by high levels of test anxiety or if examinees are not motivated to do their best on each test item.

Ability tests are typically geared to the ability level of the average member of a group. Such tests will be a rather different experience for examinees of differing ability levels. The low ability individual receives a series of items which are far too difficult for

him or her and may react by becoming threatened, anxious, or frustrated--the test may seem hopeless and he may simply stop trying. The high ability individual, on the other hand, receives items which are too easy for him--this person may find the task boring and unchallenging and, in a fashion similar to that of the low ability examinee, may simply stop trying to do his best. It is only for the average ability examinee that the items are likely to be sufficiently difficult to be challenging and yet not so difficult as to seem hopeless.

Adaptive testing procedures, however, tend to maintain an appropriate level of item difficulty for each individual. As a result they should keep motivation at high levels and anxiety and frustration at low levels. Or, at least, adaptive tests should equate these variables across individuals instead of, as in conventional testing procedures, allowing them to covary with ability level, which is what we are trying to measure.

Computerized test administration also allows us the capability of providing the examinee with feedback immediately after each test response as to the correctness or incorrectness of that response. Immediate knowledge of results, or feedback, may have positive motivating effects on some examinees and, therefore, they may perform at higher levels. Knowledge of results has long been considered important in the area of learning and instruction and has been built into methods of programmed and computer-assisted instruction. Further, the constructors of individually-administered intelligence tests, for example, Binet, Terman and Wechsler, stressed that some form of encouragement by the examiner was essential in keeping the examinee motivated and performing to his fullest capacity, although this encouragement was not to include knowledge of results on each test item.

Since the effects of immediate feedback on performance on objective tests of ability has been only rarely studied, we have incorporated immediate feedback into some of our research designs.

In one study[2], both a conventional test and a pyramidal adaptive test were administered by computer to a group of inner-city high school students. The group was racially mixed, consisting of both black and white students. Tests were administered such that half the group received the conventional test first, while the other half received the pyramidal test first. Within each order of test presentation, half the group received feedback and the other half did not.

[2]These data were analyzed by Ms. Clara DeLeon.

We analyzed the data for the conventional test only--thus, the dependent variable in this analysis was number correct on the conventional test. The design was a 2x2x2 analysis of variance. The independent variables were 1) race--black and white; 2) feedback--immediate or none; and 3) order--conventional test administered first or second in the pair.

In order to make the feedback relevant to the high school group, we had previously asked a subgroup of students from the same school to generate a set of statements which would, to them, indicate that they answered an item correctly. We used six such statements, in pseudo-random order, including "right on", "that's cool, now try this one", and "all right, how about this one". This was done on the hypothesis that feedback can have an effect only if it is meaningful or relevant to the testee.

Table 2

Mean Test Scores for Blacks and Whites on the 40-Item Test in Two Orders and With and Without Feedback

Group	Feedback N	Feedback Mean	No Feedback N	No Feedback Mean	Total Group N	Total Group Mean
Blacks--First	8	26.38	6	13.83	14	21.00
Second	7	13.86	6	14.67	13	14.23
Whites--First	15	26.07	14	30.93	29	28.41
Second	15	30.00	19	25.53	34	27.50
Blacks	15	20.53	12	14.25	27	17.74
Whites	30	28.03	33	27.82	63	27.92
First	23	26.17	20	25.80	43	26.00
Second	22	24.86	25	22.92	47	23.83
Total	45	25.53	45	24.20	90	24.87

3-Way Anova

Source of Variation	DF	Mean Square	F	Est. P
Order	1	105.76	1.36	.25
Race	1	2,013.26	15.84	<.00
Feedback	1	81.74	1.05	.31
Race x Order	1	161.54	2.07	.15
Order x Feeback	1	28.74	.37	.55
Race x Feedback	1	170.40	2.19	.14
Order x Race x Feedback	1	599.46	7.69	<.01
Error	82	77.92		

The results for the three-way analysis of variance are shown in Table 2. The only significant main effect was for race. Mean score for the blacks was 17.74 and that for the whites was 27.92, on the

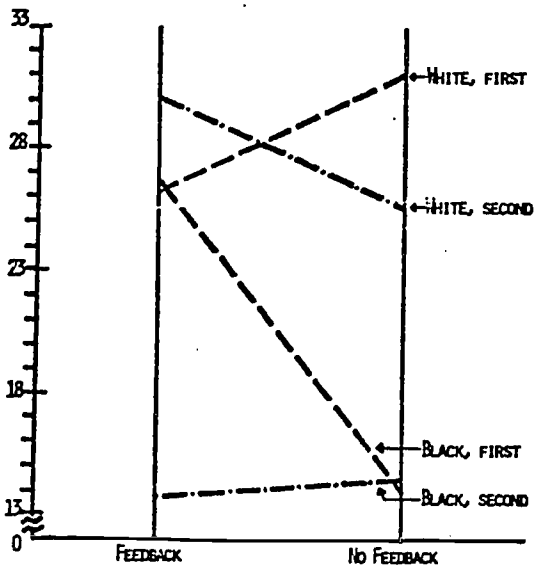
40-item test. Neither order nor feedback effects were significant, nor were any of the two-way interactions. However, the three-way order x race x feedback interaction was significant at $p<.01$.

Figure 27 shows the means for the three-way interaction. Under conditions of immediate feedback, when a conventional test was administered first, the mean of the black students (26.38) was not significantly different from the mean of the white students (26.0) who completed the conventional test under the same set of conditions.

This result implies, if it can be replicated, that race differences observed in test scores may be a function not of differences in ability but of differences in the psychological effects of the conditions of administration.

Figure 27

MEAN SCORES FOR BLACKS AND WHITES COMPLETING THE 40-ITEM CONVENTIONAL TEST FIRST AND SECOND, BY FEEDBACK CONDITION



There are some data in our results which suggest that the three-way interaction results might be due to motivational effects. In addition to analyzing test scores, we also analyzed the proportion of items skipped on the conventional test under the two experimental conditions and for the two racial groups. These results showed that blacks skipped more items than whites, in general, but when the

conventional test was administered first to the black students and they received feedback, they skipped almost no items. This is also the same set of conditions under which the test scores for the blacks were not significantly different than those of the whites. This appears to be a motivational effect since when the blacks are given feedback the test becomes relevant to them; and when it becomes relevant they can answer the questions just as well as the whites.

In a second study, either a conventional test or a stradaptive test was administered with or without feedback to two groups of subjects. One group consisted of students from the College of Liberal Arts at the University of Minnesota while the other consisted of students from the University's General College. The General College group is a much less select group and has significantly lower scores on conventional ability tests. Since the tests were constructed for the higher ability Liberal Arts group, it was expected that the conventional test would be particularly inappropriate, specifically too difficult, for the General College sample.

Table 3

MAXIMUM LIKELIHOOD ABILITY ESTIMATES
FOR 40-ITEM CONVENTIONAL TEST

	FEEDBACK		NO FEEDBACK		TOTAL	
GROUP	N	MEAN	N	MEAN	N	MEAN
COLLEGE OF LIBERAL ARTS	60	-.19	57	- .52	117	- .35
GENERAL COLLEGE	28	-.88	28	-1.26	56	-1.07
TOTAL	88	-.41	85	-.76	173	- .58

TWO-WAY ANALYSIS OF VARIANCE

SOURCE OF VARIATION	DF	MEAN SQUARE	F	EST. P
GROUP	1	19.37	19.66	.001
FEEDBACK	1	5.18	5.26	.022
GROUP X FEEDBACK	1	.02	.02	.999
ERROR	169	.99		

Table 3 shows the mean maximum likelihood scores for the two groups on the conventional test according to whether feedback was or was not given. The maximum likelihood scores are in standardized units, with mean = 0.0, and s. d. = 1.0. The analysis of variance indicated a significant main effect for feedback; in both subject groups the provision

of feedback resulted in significantly higher test scores. For example, in the College of Liberal Arts sample, the mean score under feedback conditions was -.19, while that under no-feedback conditions was only -.52. This difference of one-third of a standard deviation (about 3.5 raw score points) could be highly influential in a practical decision about an individual.

These results for the conventional test showed that feedback had a positive effect on test performance. But the results for the stradaptive test were quite different. Table 4 shows maximum likelihood scores on the stradaptive test under feedback and no feedback conditions. Note that in Table 4 not only is there no significant effect for feedback, but that the difference in group ability was also not statistically significant.

Table 4

ABILITY ESTIMATES FOR STRADAPTIVE TEST FOR TWO SUBJECT GROUPS WITH AND WITHOUT FEEDBACK

	FEEDBACK		NO FEEDBACK		TOTAL	
GROUP	N	MEAN	N	MEAN	N	MEAN
COLLEGE OF LIBERAL ARTS	60	-.66	62	-.62	122	-.64
GENERAL COLLEGE	28	-.96	27	-.81	55	-.89
TOTAL	88	-.76	89	-.68	177	-.72

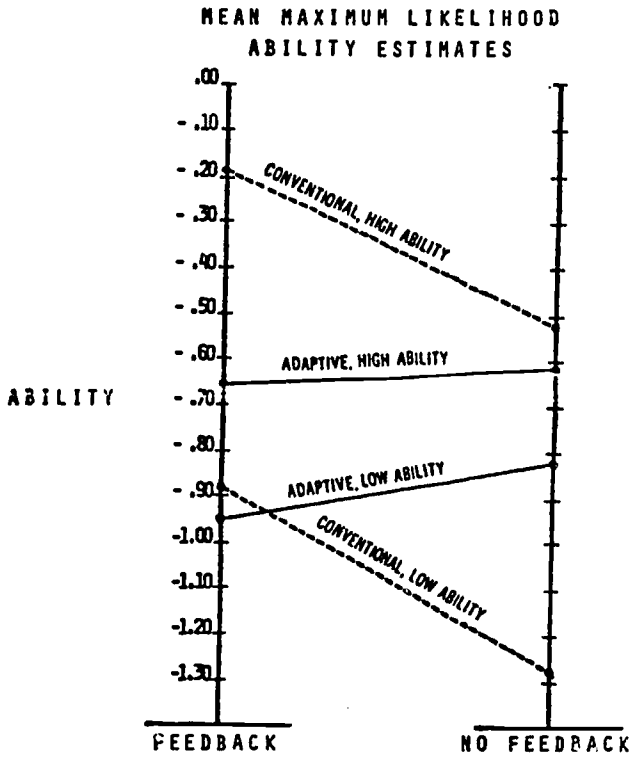
TWO-WAY ANALYSIS OF VARIANCE

SOURCE OF VARIATION	DF	MEAN SQUARE	F	EST. P
GROUP	1	2.27	1.75	.184
FEEDBACK	1	.24	.19	.999
GROUP X FEEDBACK	1	.10	.07	.999
ERROR	173	1.29		

On the surface, these results for the conventional and the stradaptive test appear to be contradictory. In the conventional test, feedback had a positive effect on test scores, and the groups differed significantly on mean ability level. On the stradaptive test, neither group differences nor feedback were significant. Figure 28 shows the means for the adaptive and conventional tests, for both groups, by feedback conditions.

If the feedback condition is interpreted as the "motivated" condition, and no feedback as "unmotivated" the apparently conflicting results can be explained. In the "low ability" (General College) group the mean

Figure 28



for the conventional test administered under feedback conditions is not significantly different from the means for the adaptive test under either condition. Or, in other words, the adaptive test itself yields scores which are intrinsically motivating to the "lower ability" testee. For the "higher ability" testee, the adaptive test scores are not significantly different from those obtained on the conventional test under unmotivated (no feedback) conditions.

The key to explaining this difference lies in the nature of the adaptive test itself. On an adaptive test--specifically, on the stradaptive test--each testee answers about 50% of the items correctly. Apparently, because of the subjective feedback the testee gets during testing, the "low ability" testee finds this "reinforcement ratio" better than what he has experienced in the past (since he is used to doing poorly on tests), and performs better even without formal feedback. The "high ability" testee, on the other hand, is used to getting a large proportion of items correct on a conventional test. But he finds the

adaptive test much more difficult than he is used to, and may experience some of the frustration that the typical low ability testee usually encounters. The fact that the mean ability estimates for the high ability group were not significantly different from those of the low ability group on the adaptive test, suggests that the adaptive test reduces error variance for the low ability testees which artificially depresses their test scores.

These results are obviously not conclusive and replications and further studies are certainly necessary. But given the current furor over test fairness and bias, it seems that we should pursue further the effects of various conditions of test administration upon performance, particularly for "low ability" testees, whose abilities might not be so low after all. Adaptive testing and immediate knowledge of results may be able to provide testing conditions more conducive to each individual's capability to demonstrate his/her fullest capacities in test performance. And, since computerized adaptive trait measurement, can provide us with important additional information of a variety of types, as well as providing more precise measurement throughout the ability range, it has promise of supplanting the paper and pencil tests which have dominated psychological testing for the last 50 years.