PROBLEM:

STRATEGIES OF BRANCHING THROUGH AN ITEM POOL

C. David Vale
University of Minnesota

The problem I am addressing has been the focus of much of the research in adaptive or tailored testing and provides, in fact, the major motivation for administering tests adaptively. The problem is: given a large pool of test items and a constraint to administer a relatively small number of them, what is the best way of selecting that small number of items? In this presentation, I am going to show 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, Dennis Dull, through a test administered under each strategy and note how his items are selected. We will further examine differences between strategies in terms of the amount of information about ability which each strategy provides.

Assumptions

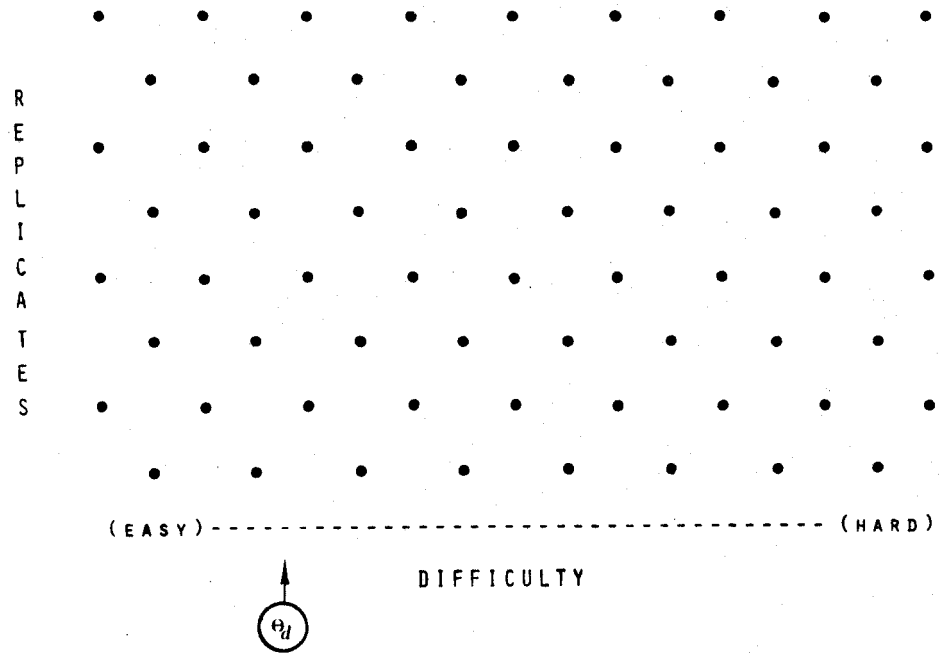
In order to make possible the analyses done for this presentation, some simplifying assumptions were made. First, it was assumed that a large pool of equally good items (i.e., items with equivalent discriminating power) was available to choose from. Second, it was assumed that these were free-response items and, hence, guessing was not possible. Third, it was assumed that all items 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 .5 with ability, was available.

Figure 1 shows schematically the item pool that will be used for testing with the various strategies. On the horizontal dimension are seventeen 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.

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 to allow some of the adaptive strategies to function well. Therefore, for evaluation of the strategies a 24-item test was used. Items for the 24-item tests were chosen in a manner analogous to the way items were chosen for the illustrated eight-item test. The results I'll present are from computer simulations (see Appendix for details of the simulation method; numerical results are in Appendix Table A-1).

Figure 1

Schematic Representation of the Item Pool Showing Dennis' Ability (Θ_d) in Relation to the Item Difficulties



Testing Strategies

Rectangular conventional test. 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. Dennis Dull, our low ability subject, produced the response record shown in Figure 2 with those items he answered correctly marked by a plus (+) and those he answered incorrectly marked by a minus (-). 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 Dennis encountered was beneath his ability level and, knowing the answer, he responded correctly. The second item was a bit more difficult for Dennis but he still answered it correctly. The third item, being a bit above his ability level, was too difficult for Dennis and he answered it incorrectly. Similarly, the fourth through eighth items were even more difficult and he answered all of them incorrectly.

Figure 2

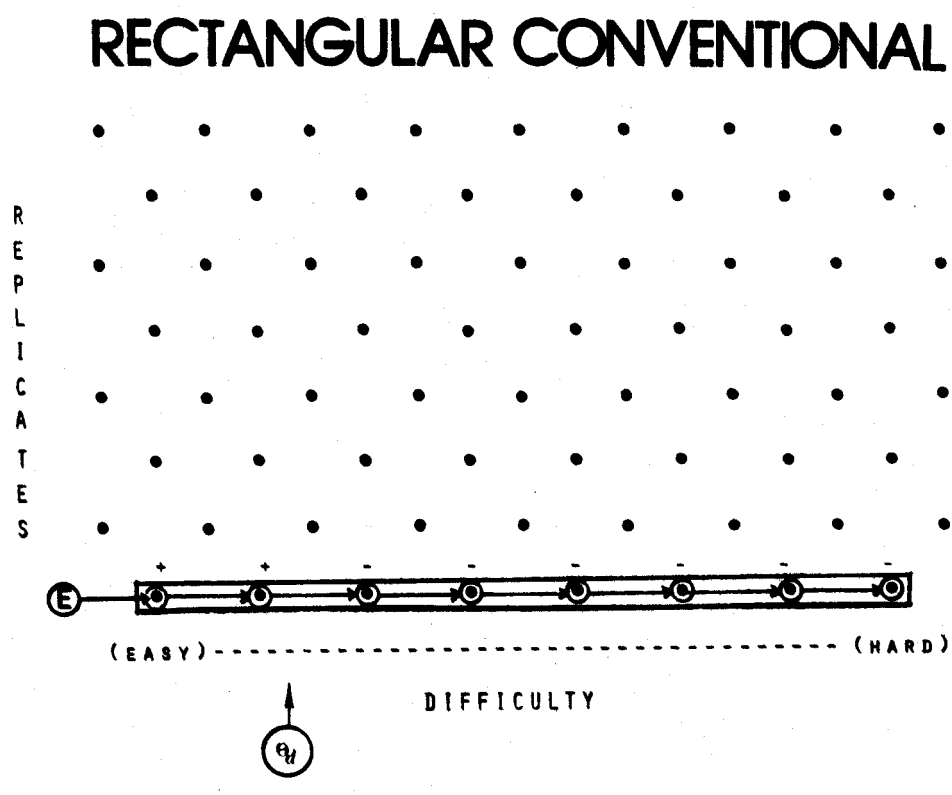


Figure 3

Information Curve for the Rectangular Conventional Test

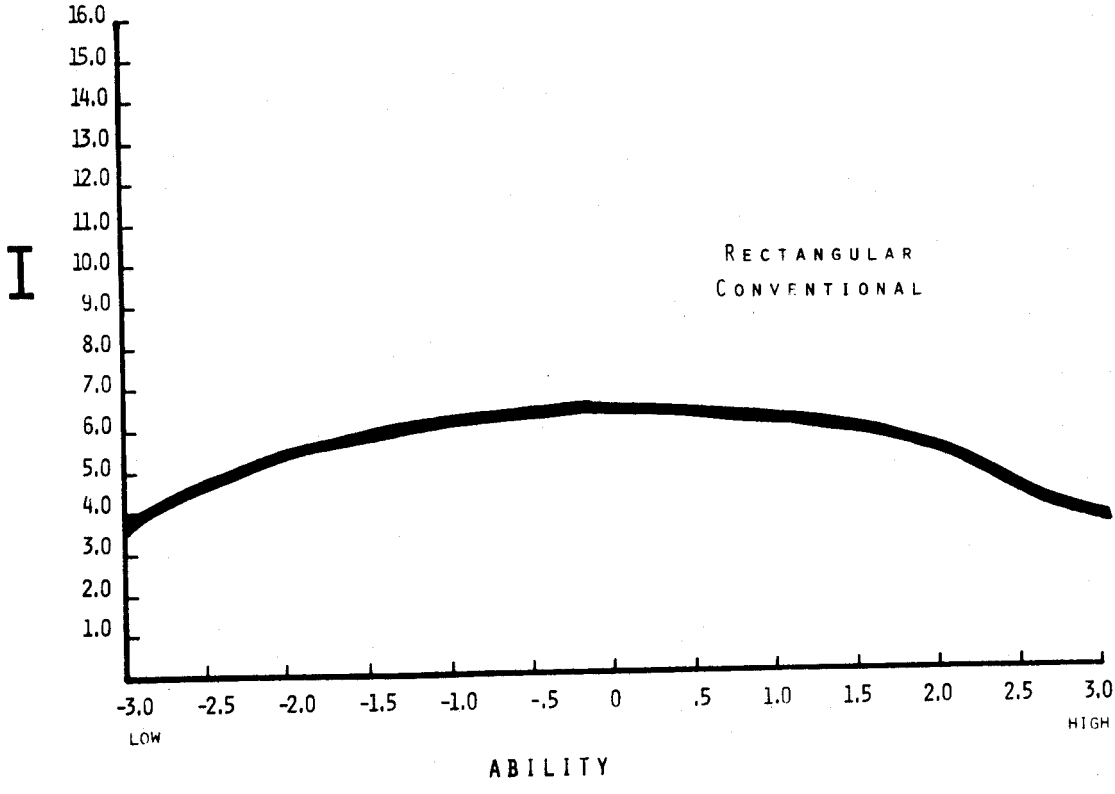
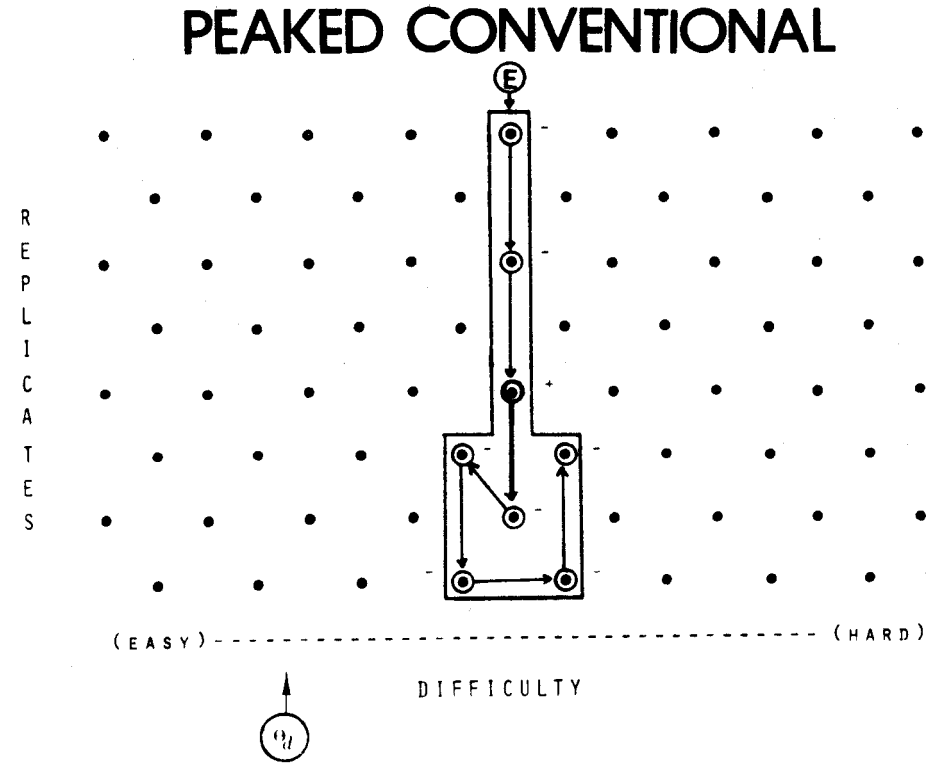


Figure 3 shows an information curve produced by the rectangular conventional test. Information can be thought of as related to the precision of measurement produced by a test at a given level of ability, or as how well a test can discriminate between two contiguous ability levels (see Lord, 1970, for a discussion of information curves). A good test produces an information curve that is high (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 information curves that the rectangular conventional test produces an information curve that is fairly flat but somewhat low. It can be seen, however, that even this information curve tapers off at the extremes indicating poorer measurement for testees where ability level is distant from the mean.

Peaked conventional test. 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 peaked conventional test. The four items from the median difficulty column and two from each of the adjacent columns were chosen for this test. Again, these items could have been administered in any order but we will begin at the top for clarity.

Figure 4

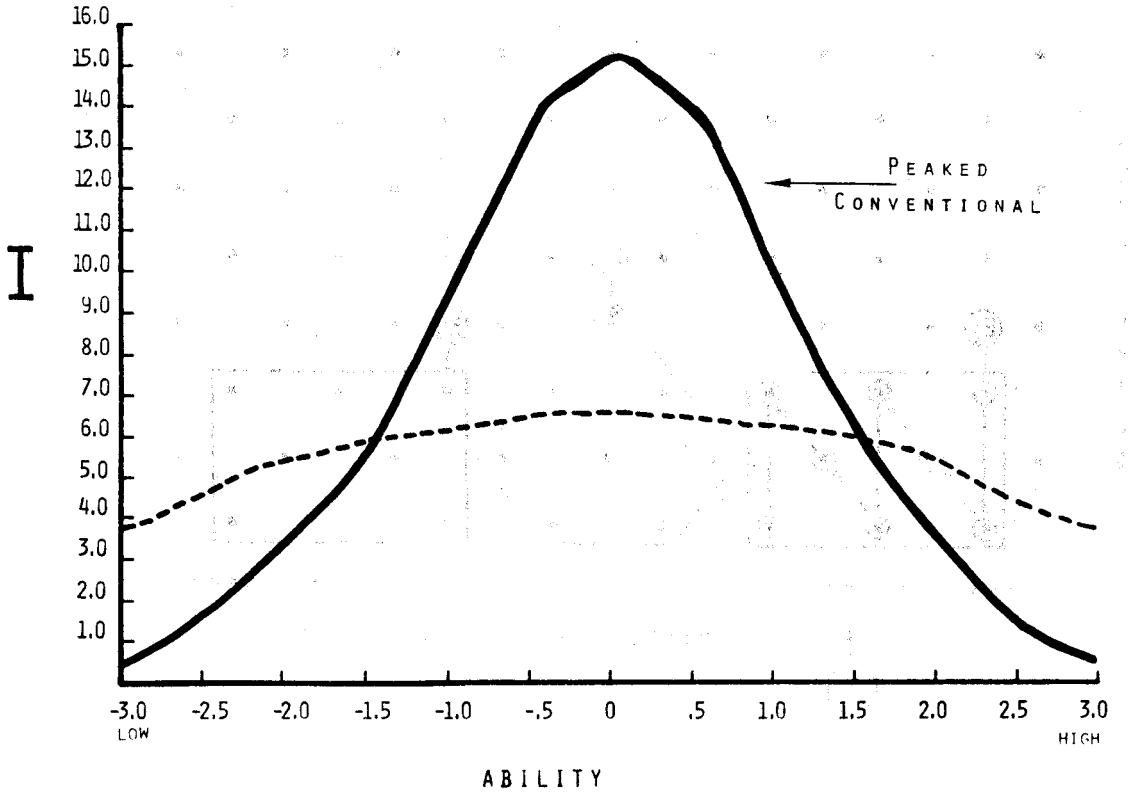


These items were intended for average ability testees and were all too difficult for Dennis. He answered incorrectly the first item, the second item, and most of the rest of the items. In fact, the only item he answered correctly asked for the definition of "Oedipal", a term he had picked up from his analyst.

The information curve for the peaked conventional test (Figure 5) shows graphically what Dennis 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 the ability range for which it was peaked but less outside of this range.

Figure 5

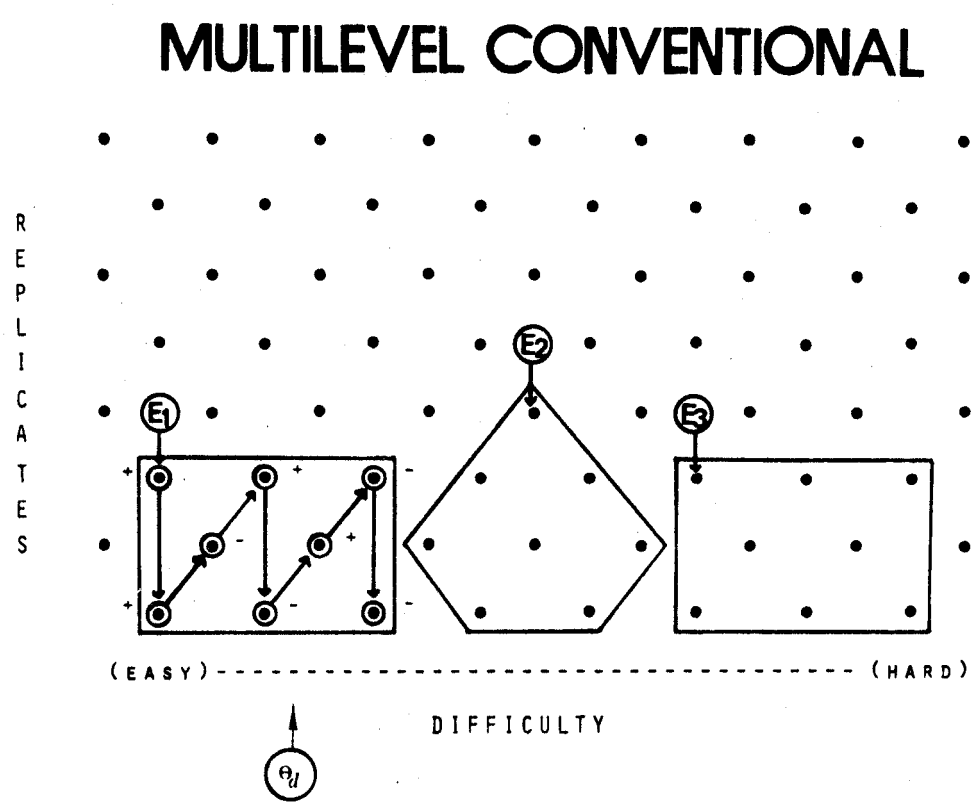
Information Curve for the Peaked Conventional Test



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 we 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 level 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 before testing, we would construct a test made up of those items with difficulties closest to his ability level (i.e., items which he/she 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.

Multi-level conventional tests. In practice we have, at best, a fallible prior estimate of the testee's ability level and may want to administer items more or less rectangularly distributed in a narrow range around that estimated ability level. 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 shows such a test. Knowing that Dennis ranked at the 27th percentile in his grade school graduating class, if this were a

Figure 6

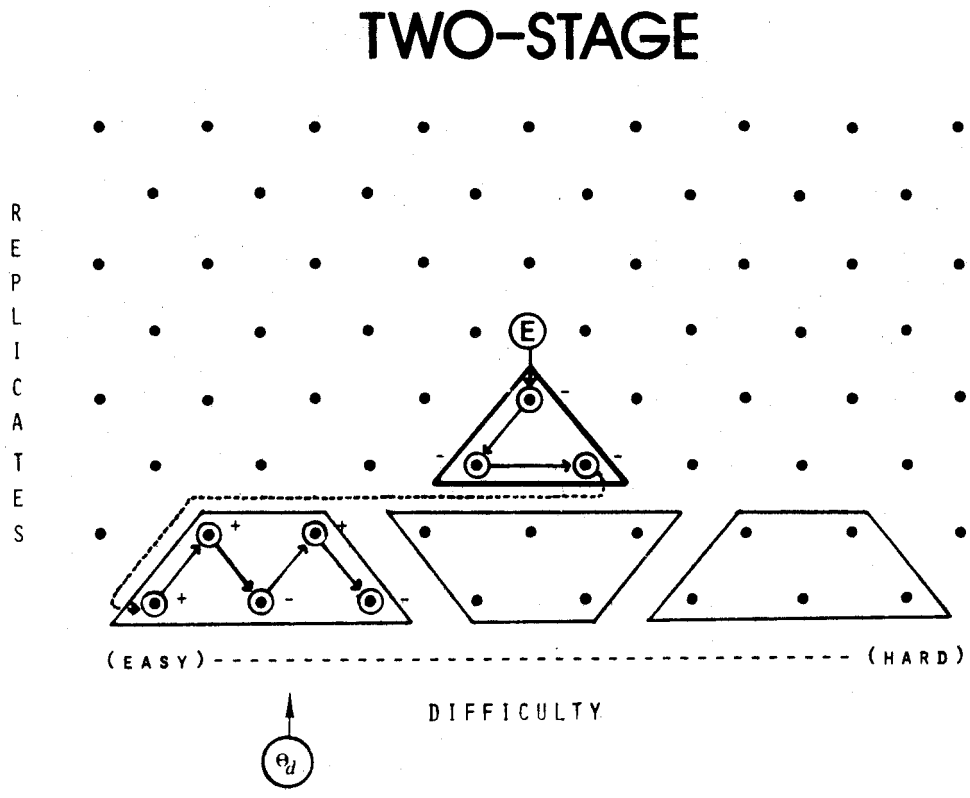


high school freshman achievement test, we might use this prior information to start Dennis at the easiest entry point (E_1). Or, if we had a testee with all A's in grade school, we might start him at the high entry point. Given a prior ability estimate, therefore, it is possible to adapt the test to the individual within the framework of a conventional test. But if prior information is not 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 (Angoff & Huddleston, 1958; Betz & Weiss, 1973, 1974) which is like the previous test but generates its own prior ability estimate.

Two-stage tests. 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. Dennis answered all three of the routing test items incorrectly as they were too difficult for him. Since this suggested that his

Figure 7



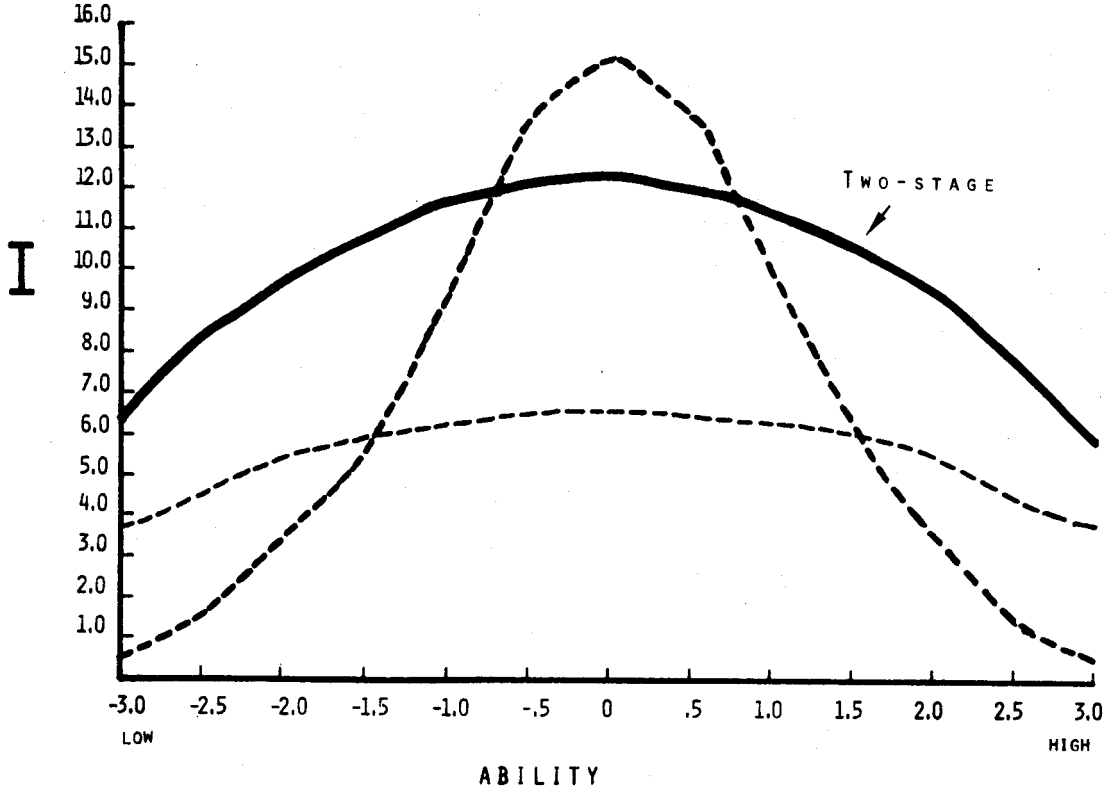
ability was low, he was branched to the easiest measurement test where he answered three out of the five items correctly.

As Figure 8 shows, this two-stage test yields an information curve that is at all points higher than that of the rectangular conventional test and higher than the information curve of the peaked conventional test except in the center. Thus, 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

Figure 8

Information Curve for the Two-Stage Test



is available in the form of the continuous second stage two-stage test (Sympson, 1975), a variant of the previous two-stage test, shown in Figure 9. As in the standard two-stage test, the testee is first administered the three-item routing test. Then, on the basis of the score on that test, he is branched to a five-item measurement test. But instead of using one of three pre-structured measurement tests, a measurement test consisting of the most appropriate item and two adjacent items on each side is individually composed for the testee. 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. For example, had the word "Oedipal" occurred in the two-stage routing test, Dennis would have answered one out of the three items correctly and might have been branched to the middle measurement test containing items that were too difficult for him.

Flexilevel test. 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 (Lord, 1971) shown in 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

CONTINUOUS TWO-STAGE

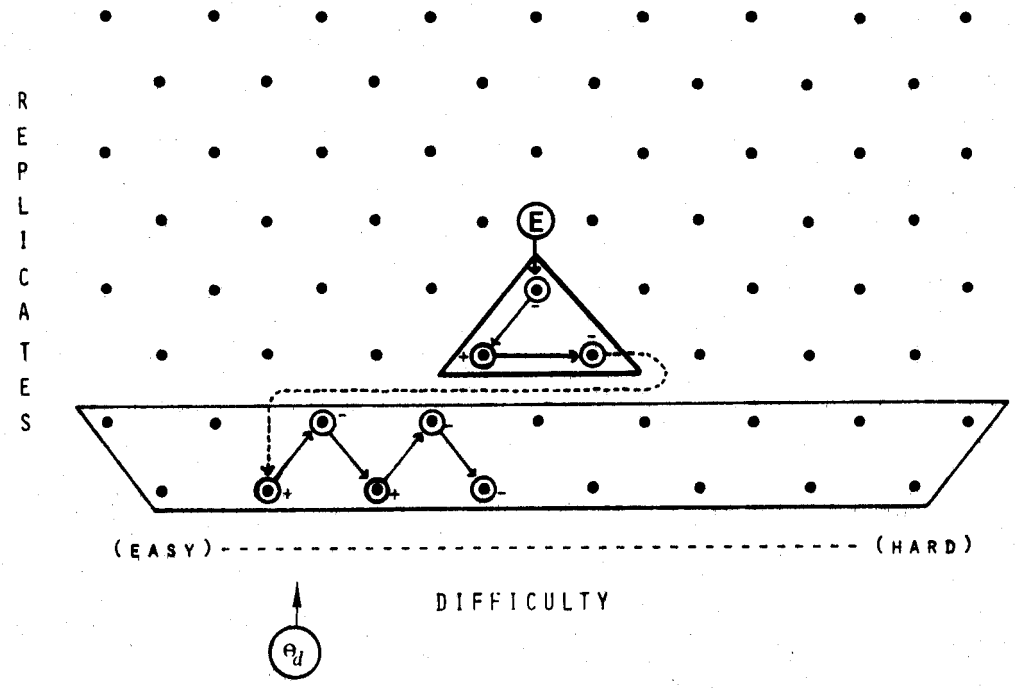
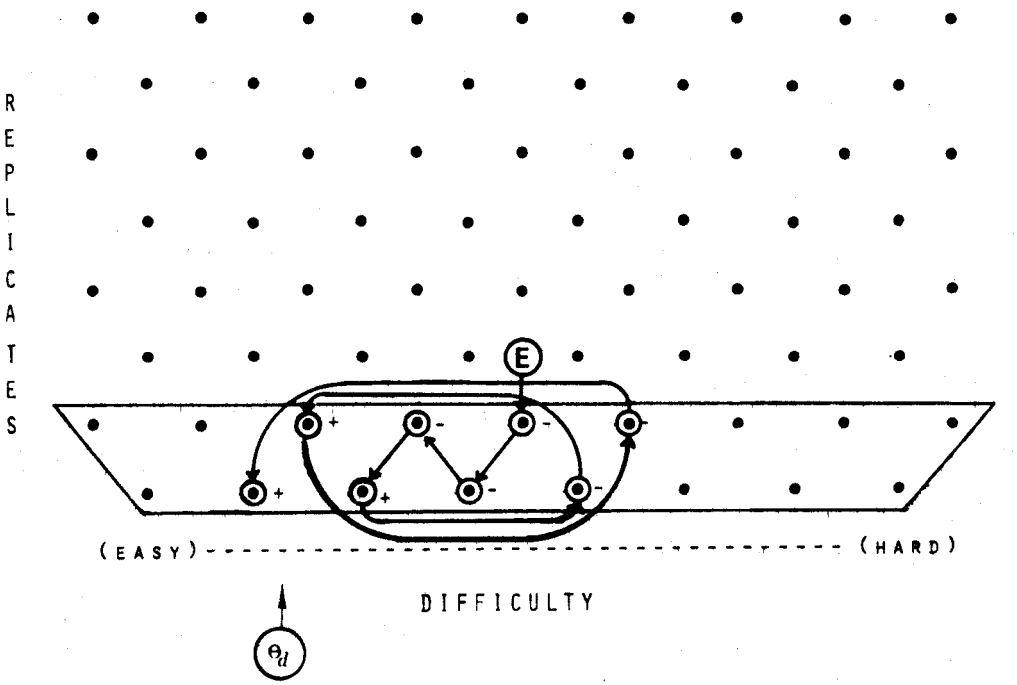


Figure 10

FLEXILEVEL

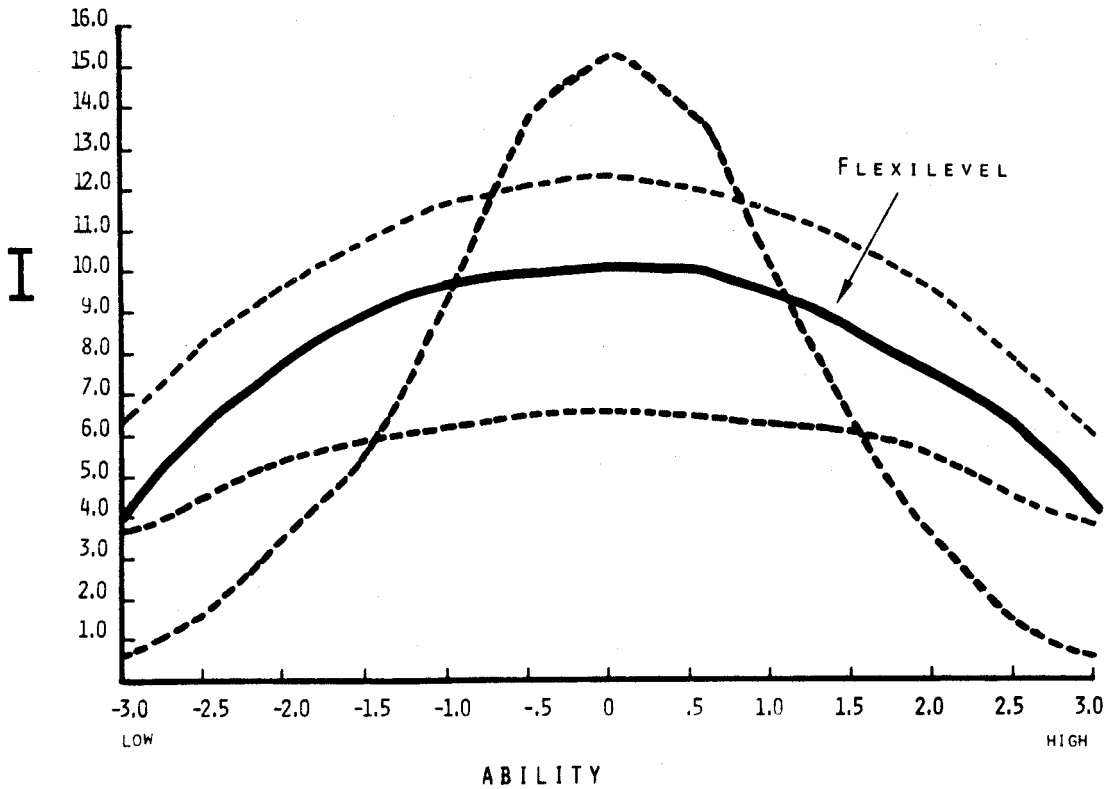


item set, and following each correct response is branched to the next more difficult unadministered item. Following an incorrect response, he is branched to the next less difficult unadministered item.

In Dennis' case, he answered incorrectly 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 that item 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 and continued oscillating between easy and difficult items until he had answered eight items.

Figure 11

Information Curve for the Flexilevel Test



The information curve for the flexilevel test is shown in Figure 11. Although the flexilevel test solves the problem of misrouting, the information it provides is always less than that provided by the two-stage test.

Three-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). 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

THREE-STAGE

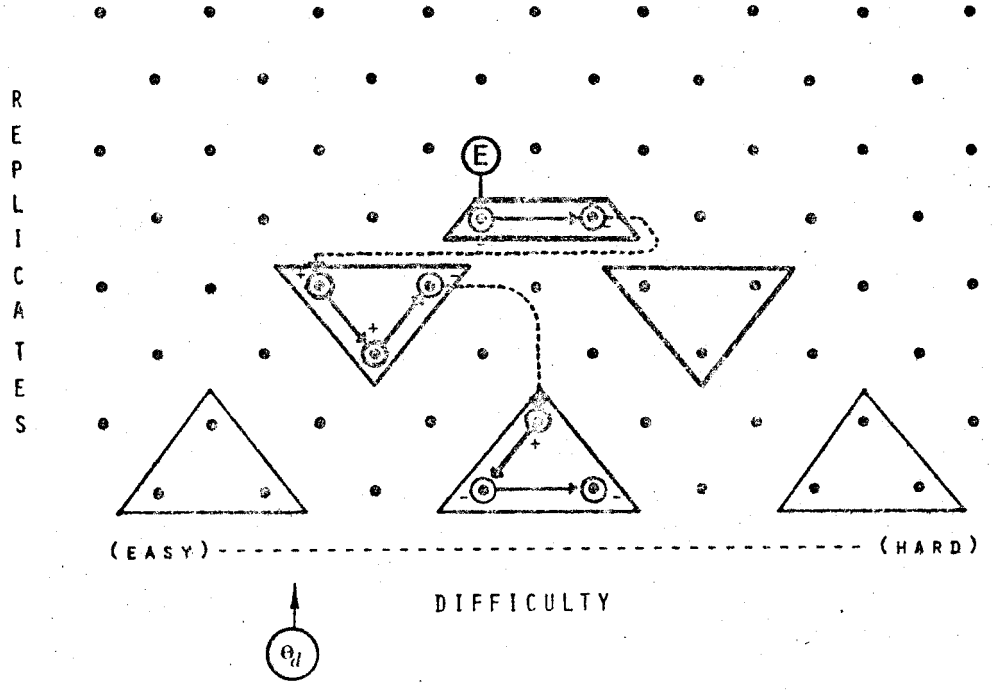
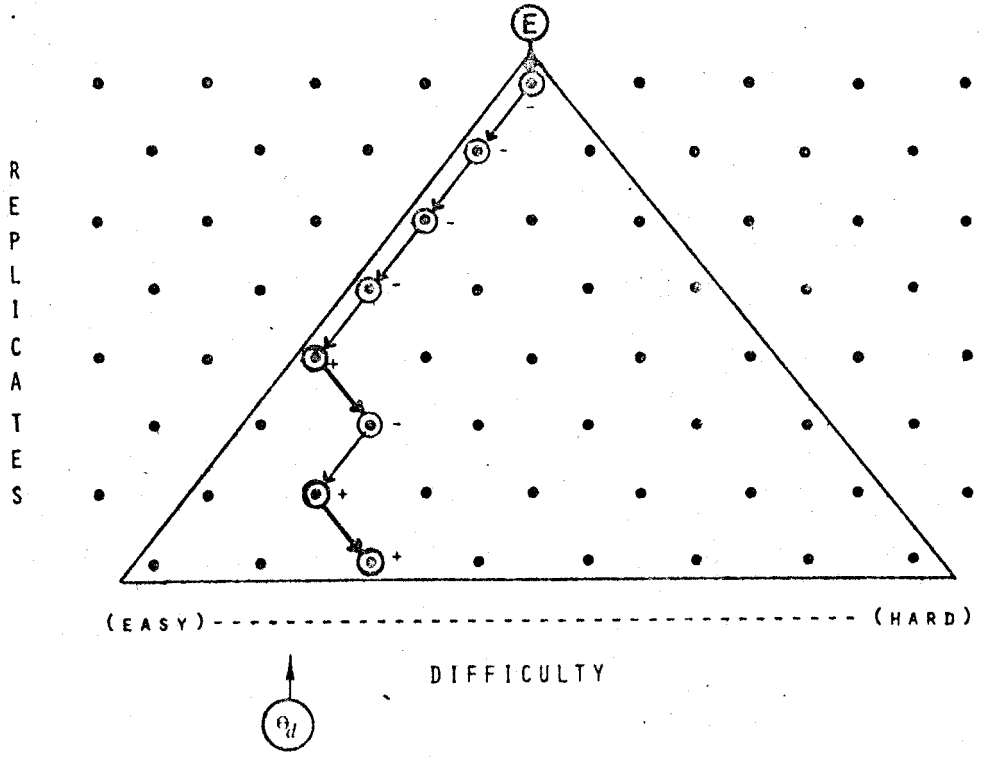


Figure 13

PYRAMIDAL

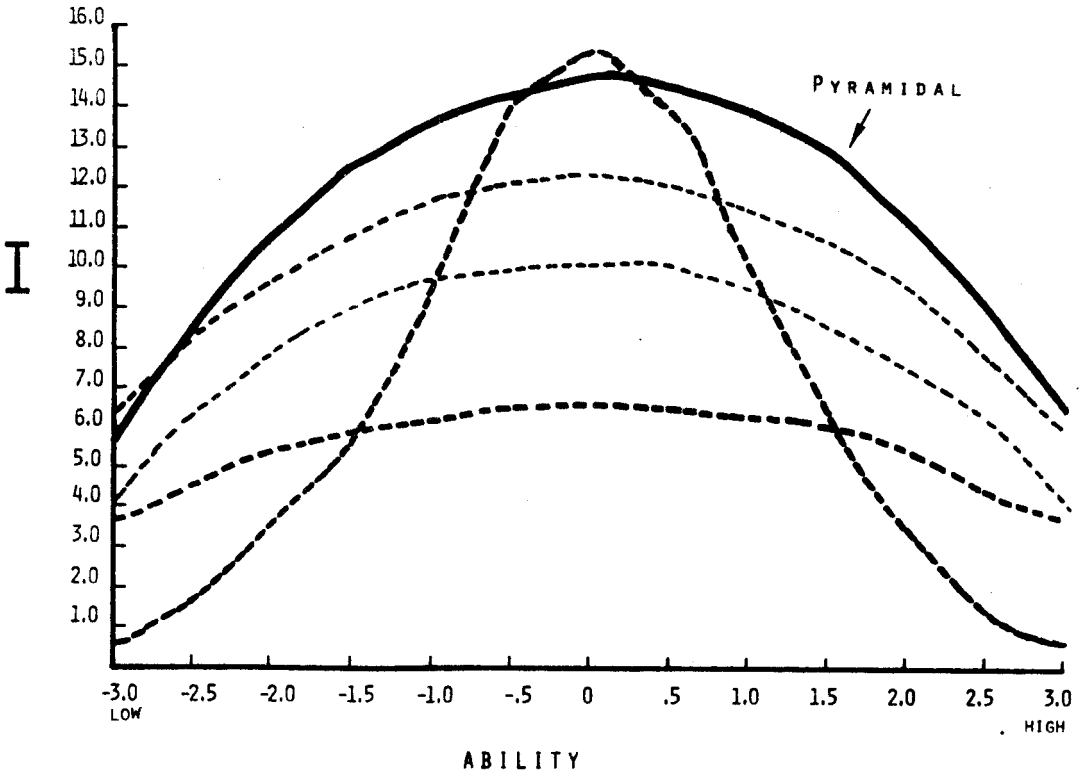


Pyramidal test. Carrying the idea of multiple routing to its logical extreme (i.e., using one item per stage) results, in this case, in the eight-stage test or, in the general case, the pyramidal test (Krathwohl & Huyser, 1956; Larkin & Weiss, 1974, 1975). As shown by Figure 13, in this strategy a testee starts with a median difficulty item and is branched after each item to a less difficult item following an incorrect response or to a more difficult item following a correct response.

The information curve for this test, shown in 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 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.

Figure 14

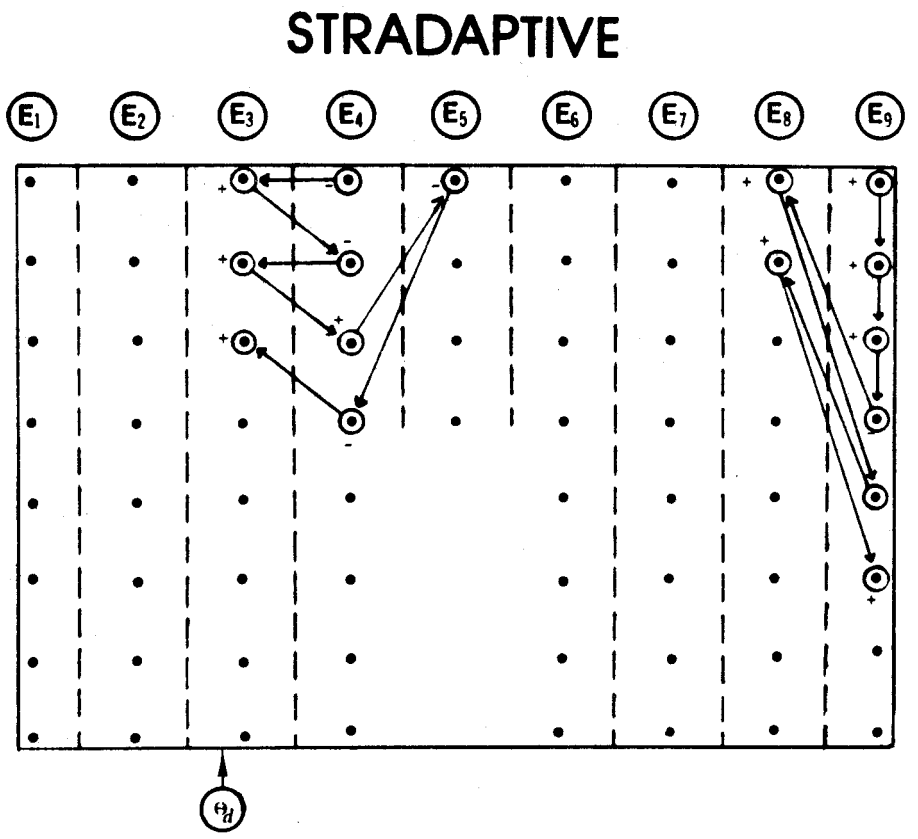
Information Curve for the Pyramidal Test



Stratified-adaptive test. 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 (Weiss, 1973) 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 difficulties; and 2) branching is between strata, with the item selected being the first unadministered item in a stratum.

Figure 15



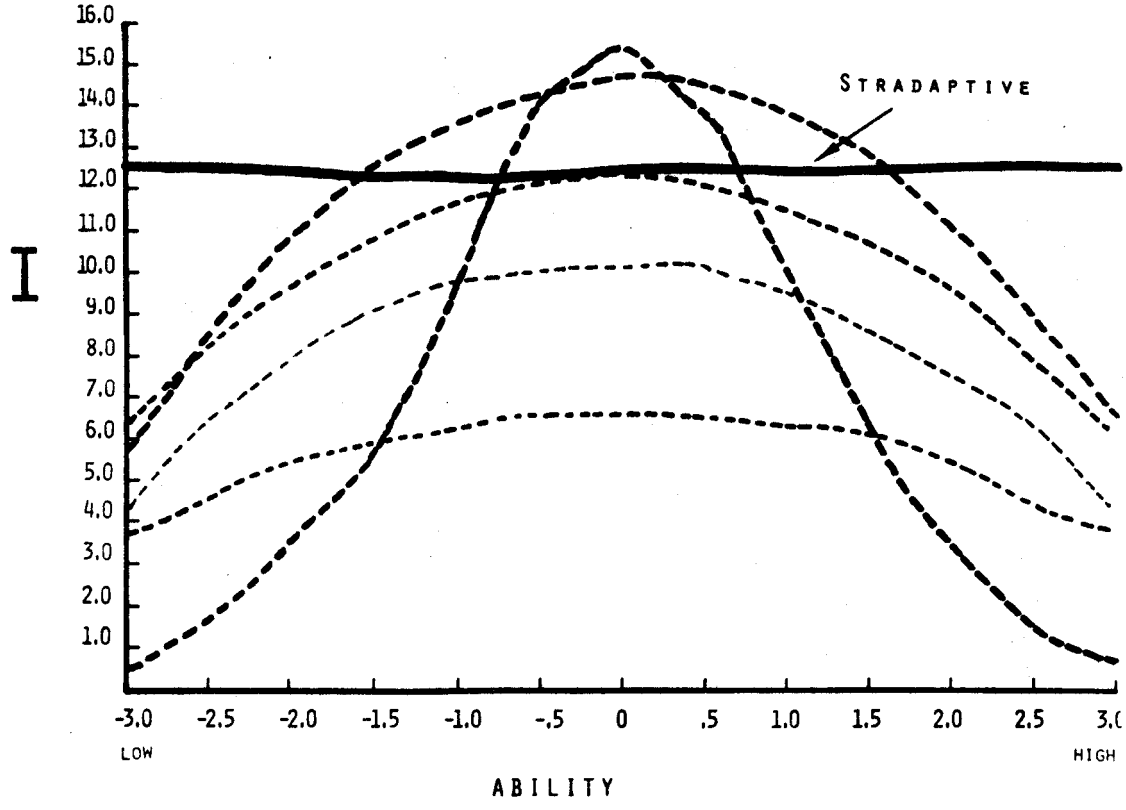
Dennis started at the fourth entry point. He did not correctly answer the first item in stratum four, was branched to the first item in stratum three, answered this item correctly, and alternated between these two strata until his fifth item. He correctly answered the fifth item, which was in the fourth stratum, and was branched to the first item in the fifth stratum. He did not know the correct answer to either this or 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; at the top of the stratum. The information curve for the stradaptive test, shown in Figure 16, is almost

flat indicating that the stradaptive test provides very equiprecise measurement. Its level is surpassed by several other strategies in the center, however.

Figure 16

Information Curve for the Stradaptive Test



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.

A Bayesian strategy. An example of the variable branching procedures is the Bayesian strategy (Owen, 1969), 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

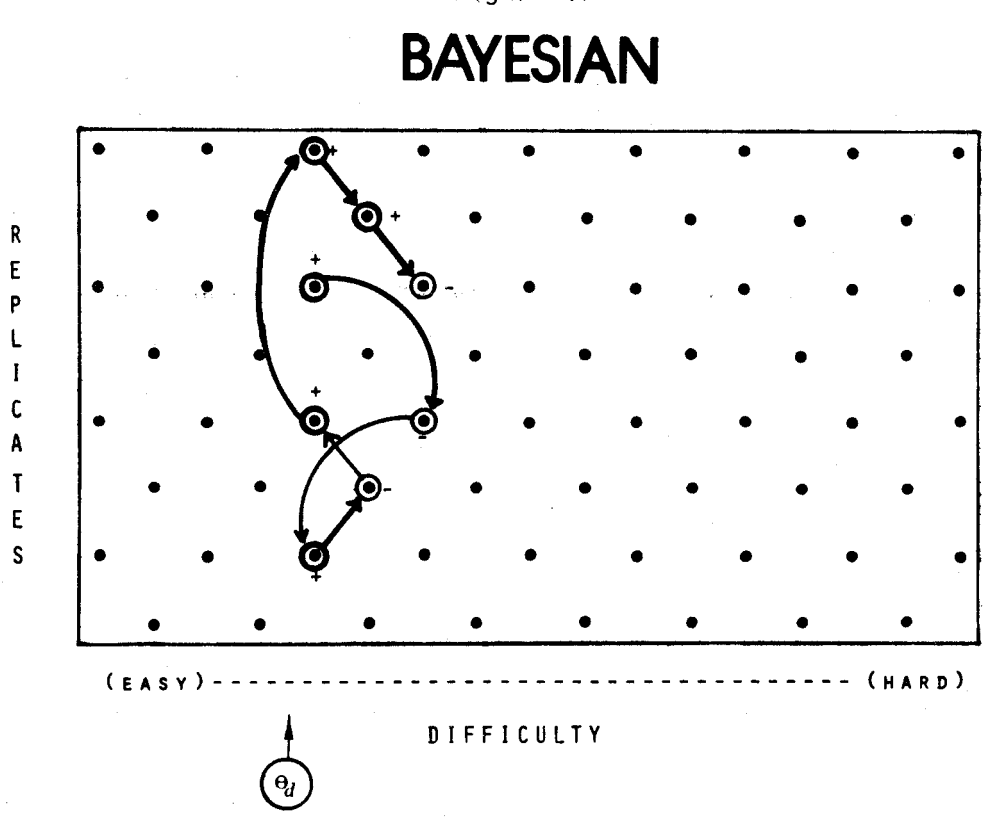
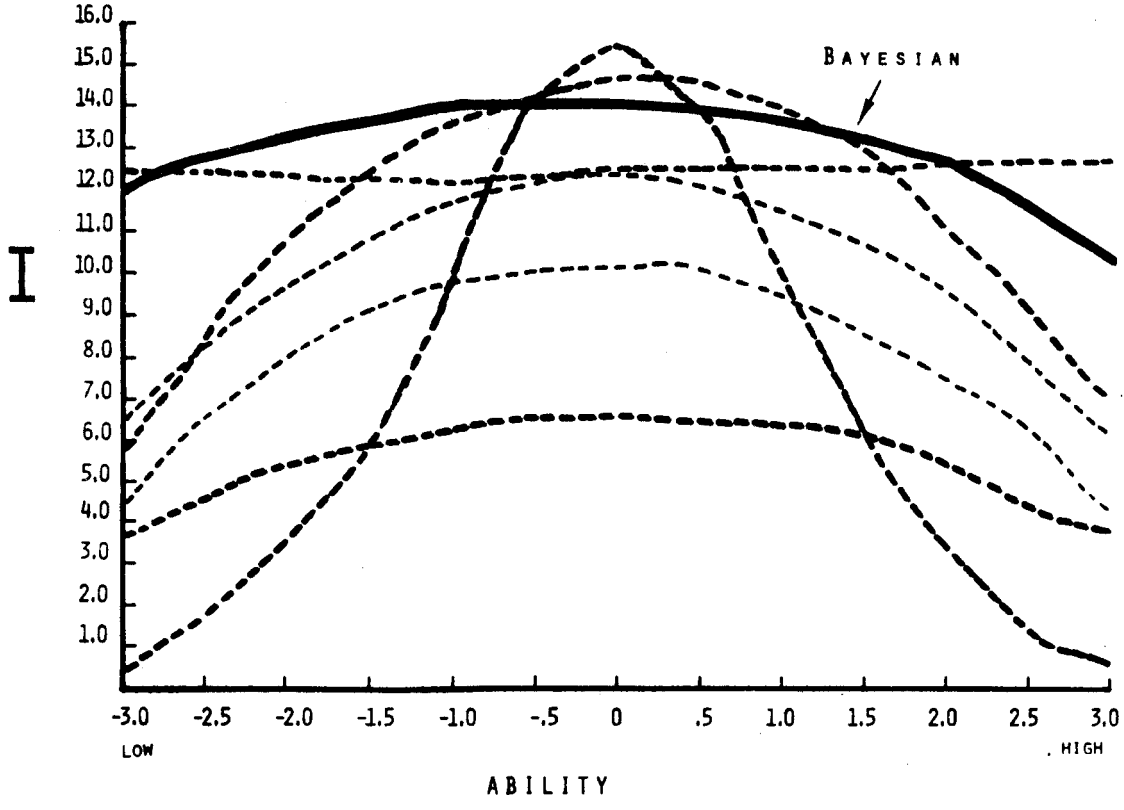


Figure 18

Information Curve for the Bayesian Test



The Bayesian test's information curve is shown in Figure 18. 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.

Limitations of the Results

In this presentation, I've attempted to give an idea why adaptive testing is needed, what some strategies of implementing adaptive testing are, and how these strategies compare in terms of the information they provide. If evaluation of adaptive testing were as simple as this presentation, however, our research would be unnecessary. This evaluation was very limited in a number of ways:

1. The information curves were calculated using a response model which may not accurately portray response tendencies of real subjects.

2. An unrealistic item pool containing equidiscriminating items was used. This would never be found in the real world.

3. Numbers of items per stage and peakedness of subtests were chosen arbitrarily and may not be optimal.

4. A common scoring technique was used which may not be optimal for all strategies. Mr. McBride will outline some of the alternative scoring procedures.

5. As you will see in Mr. Sympson's presentation, information curves are not the only way to evaluate the goodness of a testing strategy.

6. Strategies and scoring methods determined to be "best" in some situations may not be best in others.

The questions involved in adaptive testing are multifaceted and complex. The purpose of research in adaptive testing is to answer the questions necessary to decide when and how to use which kind of adaptive testing strategy. The illustrations provided here were designed simply to introduce the field and are, at best, limited in their generalizability.