BANDWIDTH, FIDELITY, AND ADAPTIVE TESTS

James R. McBride
Department of Psychology
University of Minnesota
Minneapolis, Minnesota

Since this is a conference whose central topic is test construction, it seems fitting to direct my remarks to some extent towards that topic. Therefore I shall try to present a brief exposition of adaptive testing as a special case of computer-assisted test construction.

Obviously we do not -- or should not -- administer tests in a vacuum. There is, or ought to be, a purpose for administering any test. The purposes of testing may vary widely. In one setting a test may be used to select a small number of outstanding students for a scholarship. In another context tests are frequently used to rank large numbers of examinees as to their status on an important ability dimension. Yet another application of a test might be to diagnose reading disability -- say to select students from the bottom of the score scale for remedial treatment.

Each of these applications of testing is more or less commonplace. And it is well known in psychometric circles that different techniques of test construction are appropriate for each application. For the scholarship selection problem, for instance, the test items should be quite difficult and highly discriminating -- in order to permit accurate discriminations among the very ablest examinees. For the rank ordering problem the test items should be moderately discriminating and of about median difficulty, in order to permit maximal discrimination throughout the middle range of the construct -- where the people are concentrated most densely -- as well as some discrimination in the tails of the distribution. For selecting candidates in need of remedial treatment, the test items should be very easy, as well as highly discriminating, in order to discriminate the "low normals" from the ones who can benefit from differential treatment. Each of these different purposes of testing brings with it a technical mandate for a distinct sort of distribution of test scores, in order to make the kinds of discriminations among persons called for by the situation.

PSYCHOMETRIC INFORMATION

The concept of the "information" in a set of test items (Birnbaum, 1968) is a very useful one, both for clarifying the different technical problems involved in different purposes of testing, and for facilitating the selection of items to comprise any special-purpose test. Without giving a technical definition of information, let me just say that the "information value" in a set of test items can be estimated at any point along the continuum of what the test measures. The numerical magnitude of the information at any point tells us something about the error probabilities involved in making discriminations around that point on the basis of test scores. For any set of test items -- that is, for any test of fixed composition -- if we know certain psychometric characteristics of the items, we can evaluate the test information at any point on the continuum. If we construct an information curve from these values, we can readily determine the range of the construct in which the test discriminates best, as well as range(s) in which it might discriminate very poorly.

The test information curve is built up out of the information curves of the constituent items. In fact, the test information function is simply the sum of the item informations. The implication for test construction here is becoming well-known: by judicious item selection based on the notion of the item information, we can construct a test with a known information structure. For example, for the scholarship selection problem we know that we want maximal discrimination among the top one percent of the candidates. Thus we want a test whose information value is very high at and beyond two standard deviations above the mean. Our item selection rationale is to choose those items whose information values are highest in that same region. For the remedial treatment problem, we would select items with maximal information below, say one and one-half standard deviations below the mean. For the rank-ordering problem, we would want items providing best information in the middle range -- say between -1 and + 1 standard deviations about the mean.

If we have a large bank of items, all measuring the construct of interest, and all calibrated as to their psychometric characteristics, it is a fairly simple matter to write a computer program to select the best test items to construct a test for almost any purpose, provided we can specify the point or range of the construct at which we want maximal test information.

The results of using such a test should be quite favorable, provided the test is used in an appropriate examinee population,

the purpose of testing is formally congruent with the item selection rationale, and the construct range in which maximal information is needed has been accurately specified. If any of these provisions is violated, however, the test may fail to some extent to achieve its intended prupose. One of the simplest ways in which this can happen is to use a highly discriminating peaked test in a different group from that in which it was intended to be used. The utility of a test for its intended purpose may also be diminished if the critical point is badly in error, i.e., if the test is constructed so that maximal information is achieved in the wrong place or places on the continuum.

I won't pretend that the two sources of difficulty just mentioned occur frequently in large scale testing programs. They probably do not, because in such programs the characteristics of the group being tested are usually quite well known, and so are the characteristics of the test items used. I would suggest, however, that when tests are constructed ad hoc from a pool of pre-calibrated items, the smaller the examinee group, the greater the potential for mischief. It would be nice if there were some self-correcting method of test construction available to circumvent this potential problem. In a sense, computer-administered adaptive testing is just that.

To explain what I mean by that, let me revive (and modify somewhat) the notion of the "bandwidth-fidelity dilemma" posed by Cronbach some years ago (Cronbach, 1961, p. 602). "Fidelity" is a concept closely akin to information, as I have used the latter term. Thus the scores from a certain test have highest "fidelity" where the information function is highest. But if the test items are sufficiently discriminating, and about equal in difficulty, there will be a very narrow range of the construct over which high fidelity is maintained. This range, or distance along the trait continuum in which fidelity is highest, we might call the bandwidth. The dilemma is that, other things being equal, high fidelity is achieved at the expense of bandwidth, and vice versa. The "other things" which must be held constant in order to make the analogy true, are just those things we do hold constant in conventional testing: item discriminating power, and test length. Holding these things constant, the fidelity and bandwidth of the test are determined by the choice of item difficulty. Bandwidth is a direct function of the distribution of the item difficulties. The more they vary, the wider the bandwidth; the less they vary, the higher the fidelity.

An adaptive test has the potential of broadening the bandwidth with little loss of fidelity. To the extent that this potential is realized, an adaptive test is the self-correcting method of test construction I mentioned earlier. This is so

because the wider the bandwidth, the smaller the probability of losing utility by using a test in an inappropriate group, or by erroneous choice of the point of maximal information.

Adaptive tests have this potential by virtue of the fact that they can tailor the difficulty of the test items to the ability (or achievement level, or trait status) of the examinee, during the test. As all of you probably know, the simple rationale for this is to administer a more difficult item after a correct response, and an easier item after a wrong one. This procedure is greatly facilitated by administering the tests at a computer terminal, but for some strategies of adaptive testing less sophisticated devices will suffice.

Let me return briefly to the bandwidth-fidelity dilemma. The ideal outcome of any test construction endeavor would be an infinitely high, horizontal test information function, that is, a test whose scores had perfect fidelity and infinite bandwidth. The dilemma necessitates a compromise, however, and in a conventional test the compromise is usually struck in favor of high fidelity and low bandwidth. The compromise is necessitated by the fact that the constituent items of the conventional test are the same for all persons. Now, the adaptive tests give different sets of test items to different persons: can the adaptive tests therefore achieve perfect fidelity and infinite bandwidth? The answer is no. Fidelity -- whether the test is adaptive or conventional -- is limited by the item information functions, whose limits in turn are imposed by the items' discriminating powers. And in fact, the highest value of a test information function which can be achieved by administering a fixed number of items selected from a larger pool of items, is achieved by conventional test construction.

The unique contribution of any adaptive testing strategy is an increase in bandwidth. Different strategies of adaptive testing differ in the extent to which they achieve this. They differ in a number of respects as well, so perhaps it would be fruitful to differentiate several of these strategies and to evaluate them in terms of fidelity and bandwidth.

In order to have a basis for comparison, let us look at a typical information curve for a peaked test. The bell-shaped curve in Figure 1 was obtained by administering a 24-item test having identical item difficulties, to a large number of hypothetical examinees in a computer simulation study. (For clarity of presentation, item discriminations were all equal, and guessing was not allowed to influence the test scores. These same conditions will hold for the comparative studies discussed below.) Plainly, the test information curve is highest at the

mean (0) of the ability range, and diminishes rapidly in the outlying regions; the bandwidth, then, is fairly narrow. The low, flat curve in the same figure was obtained from simulated test data in which the item discriminating powers were the same as for the peaked test, but the item difficulties were uniformly distributed in the interval between -2 and + 2 units around the mean. I shall refer to this as a "rectangular" conventional test. The information value, and hence the fidelity, is low throughout the depicted range of ability -- but the bandwidth is quite broad. Bear in mind that both these conventional tests had the same number of items. The broad bandwidth test could achieve the same peak information level as the peaked test only if its length were increased.

The first adaptive test to be considered here is one called the pyramidal test. Its name reflects the conceptual structure of its item pool, and the simple algorithm it employs for tailoring item difficulty to the examinee's performance. Once the items are arranged in the pyramidal structure (e.g., Figure 2) item selection takes place by means of a mechanical branching algorithm based solely on item difficulty: a one unit increment in item difficulty follows a correct answer, and a one unit decrement in difficulty follows a wrong answer. The test continues, typically, until the examinee has answered a prespecified number of items.

Figure 3 shows the information function for the pyramidal adaptive test superimposed over those of the peaked and rectangular conventional tests. Its information value is everywhere higher than that of the rectangular test, and its bandwidth is broader than that of the peaked test. Its highest information value, is slightly lower than that of the peaked test scores. Except for discriminating in a very small range of ability level centered around 0, this pyramidal adaptive test appears to be clearly preferable to the peaked conventional test in terms of both fidelity and bandwidth. However, the information function of the pyramidal test is strongly influenced by a characteristic called step size. Step size is the magnitude of the difficulty increment between adjacent item difficulty levels in the pyramidal structure. The fidelity of the pyramidal test's function will increase as step size increases. In fact, its information function will approach that of the peaked test as step size approaches zero. Conversely, as step size increases, the peak information level decreases to that of the rectangular test.

Another strategy which uses a branching algorithm similar to that of the pyramidal test is the stratified adaptive, or "Stradaptive" test (Weiss, 1973). Its items are arranged into

levels or strata; item difficulty is homogeneous within each stratum. Each stratum's difficulty level differs from that of the adjacent strata by a constant step size. As in the pyramidal strategy, a correct answer to an item results in a more difficult subsequent item, and an incorrect answer leads to an easier item. Three major differences from the pyramidal strategy are 1) in the stradaptive test there are more items available at all but the middle difficulty level; 2) the first item administered may be from any one of the several strata, depending on what prior information is available about the examinee; and 3) test length may vary from one examinee to another. The differential starting points and the larger number of items at most difficulty levels should result in wider bandwidth for the stradaptive test than for the pyramidal one. Figure 4 shows the information function resulting from simulated administration of a stradaptive test. In this case, differential starting points were used, based on a prior estimate of each examinee's ability which correlated .50 with the simulated ability. A constant 24-item test length was used to permit direct comparison with the other simulated tests. The information function resulting was virtually flat throughout the range of abilities tested. It was lower than that of the peaked test in the interval (-1, + 1), and lower than the pyramidal strategy's information function from -1.5 to + 1.5 standard deviation units. It appears that the stratified test's bandwidth is extremely wide, however.

The third adaptive test strategy to be considered is a Bayesian sequential strategy proposed by Owen (1969). This strategy does not require the structured item pool typical of the other two. All it requires is a fairly large pool of items with known difficulty and discrimination parameters, preferably with a uniform and heterogeneous distribution of item difficulty, and items with moderate to high discriminating power.

This Bayesian strategy begins with an a priori estimate of each examinee's trait status. This may be the same for everyone, or may vary if differential prior knowledge is available. Associated with each initial prior estimate is a variance; the estimate and its variance are assumed to be the parameters of a normal distribution on trait level, characterizing the examinee. After the first test item response, the estimate and its variance are updated, contingent on the item score (right or wrong). The second item to be administered is (essentially) the one item in the pool which has the most information value at the trait level equal to the current estimate. That item is administered and scored; the estimate and its variance are updated again; and a third item is chosen — again the unused one with maximal information at the latest estimated trait level. This process continues until some test termination criterion is satisfied.

This may be a certain number of items administered, or an arbitrary small value of the Bayes posterior variance attained.

Several characteristics distinguish the Bayesian sequential procedure. Foremost is its mathematical elegance. It re-estimates the trait level parameter after every item, and explicitly searches the item pool for the most informative item to administer next. Thus, unlike the two strategies described earlier, the step size or difficulty increment is not constant from one item to the next. In practice, if the item pool contains sufficiently many and good items, the step size tends to be relatively large at first, and to diminish steadily, item by item. This usually results in less within-examinee variability of item difficulty than does, say, the stradaptive test. That fact, coupled with the conservative nature of the Bayesian ability estimation procedure, tends to regress item difficulty levels and ability estimates towards the initial ability estimate. The effect of this is a test information function like the one from our simulated testing, shown in Figure 5. It is not as flat or as broad-band as the one from the stradaptive tests, but is somewhat higher from -2.5 to + 2.5 units on the trait level scale. With only mild reservations, I would venture that under the conditions simulated the Bayesian sequential test appears to strike the best balance between bandwidth and fidelity of the three adaptive strategies discussed.

A number of other promising strategies were not considered here, including two which rival the Bayesian one in mathematical sophistication, and perhaps the stradaptive one in bandwidth. I refer here to strategies proposed by Lord (1975) and by Samejima (1975).

Something else not considered here was the effect of guessing on the shape of the test information functions. Typically, guessing destroys both the symmetry and the elevation of information functions for the conventional, pyramidal, and Bayesian tests (e.g., see Lord, 1970; McBride, 1975). It will certainly lower the elevation of the stradaptive test information function as well, but the data are not yet in reference to its effect on the bandwidth of that test.

Another aspect of the data summarized in my figures which some may find disquieting is the fact that no live people were tested to obtain them. People may not respond according to the neat model used to generate the simulation data reported here, thus invalidating some or all of the simulation results reported. These data do, however, portray accurately the characteristic ways in which different test strategies react to test item responses. I believe such data have strong implications for the

future of mental testing in light of the growing ubiquitousness of computers, and the consistent annual decline in the cost of terminals.

The notion of adaptive testing is an elementary logical extension of computer-assisted test construction. The data summarized here show that adaptive tests have the potential to equal conventional tests (which are never perfectly peaked) in what I have called fidelity, and to exceed the utility of conventional tests by virtue of their greater "bandwidth".

OTHER BENEFITS OF ADAPTIVE TESTING

Other potential gains from adaptive testing lie outside the realm of technical psychometrics, in the effects they may have on test-taking behavior. Take, for instance, the problem of guessing on multiple-choice items. The usual correction-for-guessing formulas penalize the non-guesser as well as the guesser, but in no way distinguish the two. An effective broad-band adaptive procedure would tailor item difficulty to all examinees such that the effects of guessing are equalized throughout the ability range, by minimizing the variability across persons of the proportion of items answered correctly.

Another potential advantage of adaptive testing lies in the motivational effect of self-administered feedback. The low ability subject taking a moderately difficult conventional test knows that he doesn't know the answer to certain test items. What effect this has on his test-taking motivation is uncertain, but there is some reason to believe that it may frustrate him, and perhaps cause debilitating anxiety and emotional "retreat" from the threatening situation. An effective adaptive strategy should result in low ability examinees averaging more than fifty percent correct on multiple choice items, just as high ability persons do. The test is bound to "feel" better to the person used to knowing, say, ten or twenty percent of the items on a standardized test. And feeling better about his test performance, perhaps he will be motivated to do his best rather than withdraw from the perceived threat.

That last point is largely speculative, but is not without some indirect support. In over three years of administering adaptive tests to live people at Minnesota, we have consistently seen a very favorable response from the examinees, including some minority students in an inner city high school. In one experiment, the effects of inter-item feedback on a conventional test (administered by computer) were observed to eliminate a significant race effect which occured on the same test without

feedback (Weiss, 1975). In another experiment explicitly comparing a low ability group with a high ability one, use of a stradaptive test appeared to have the same effect as did feedback in the earlier experiment. The low ability group performance was not significantly different from that of the other group on the adaptive test, despite a significant group difference on a conventional test (Betz, 1975). I have taken some liberties with the contexts of both experiments (and both require replication) but to me there is sufficient evidence in both to suggest that feedback has a facilitating effect on motivation, and that adaptive testing has an effect similar to that of feedback in low ability groups.

SUMMARY

In summation: computer-administered adaptive testing has an obvious intuitive appeal. More inportant, it has technical or psychometric advantages over conventional tests. I have pointed out some of these under the headings of "bandwidth" and "fidelity". Knowledge of these technical qualities of adaptive testing is not new, although they have only recently begun to be appreciated.

In comparison with the psychometric aspects of adaptive testing, very little is known about its psychological aspects. I have hinted at what some of these may be: a more favorable reception by examinees; a leveling of the guessing tendency across ability levels; and most important, a dramatic improvement in test-taking motivation among some of those who are now frustrated by our conventional tests.

All of these remarks are made from within the framework of adaptive ability testing. They should be applicable as well to achievement and performance testing, whether norm-referenced or criterion-referenced. We can expect to see wider applications of adaptive testing in the very near future.

FOOTNOTES

1. Research reported herein was supported by the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, under contract No. 00014-67-A-0113-0029, NR no. 150-343; David J. Weiss, Principal Investigator.

2. The figures presented here are based on data obtained by C. David Vale, and reported by him in Vale (1975).

REFERENCES

Betz, N. E. Prospects: New Types of Information and Psychological Implications. In D. J. Weiss (ed.), *COMPUTERIZED ADAPTIVE TRAIT MEASUREMENT -- PROBLEMS AND PROSPECTS*. Research Report 75-6, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, November, 1975.

Birnbaum, A. Some Latent Trait Models and Their Use In Inferring An Examinee's Ability. In Lord, F. M. and Novick, M. R., *STATISTICAL THEORIES OF MENTAL TEST SCORES*. Reading, Massachusetts: Addison-Wesley, 1968.

Cronbach, L. J. *ESSENTIALS OF PSYCHOLOGICAL TESTING*. New York: Harper and Row, 1961.

Lord, F. M. Some Test Theory For Tailored Testing. In Holtzman, W. H. (ed.), *COMPUTER-ASSISTED INSTRUCTION, TESTING, AND GUIDANCE*. New York: Harper and Row, 1970.

Lord, F. M. A Broad-Range Tailored Test of Verbal Ability. *PROCEEDINGS OF THE CONFERENCE ON COMPUTERIZED ADAPTIVE TESTING, JUNE* 1975. United States Civil Service Commission, in press (Also Research Bulletin 75-5. Princeton, N.J.: Educational Testing Service, 1975.)

McBride, J. R. Adaptive Testing Research At Minnesota: Some Properties of A Bayesian Sequential Adaptive Mental Testing Strategy. *PROCEEDINGS OF THE CONFERENCE ON COMPUTERIZED ADAPTIVE TESTING, JUNE* 1975. United States Civil Service Commission, in press.

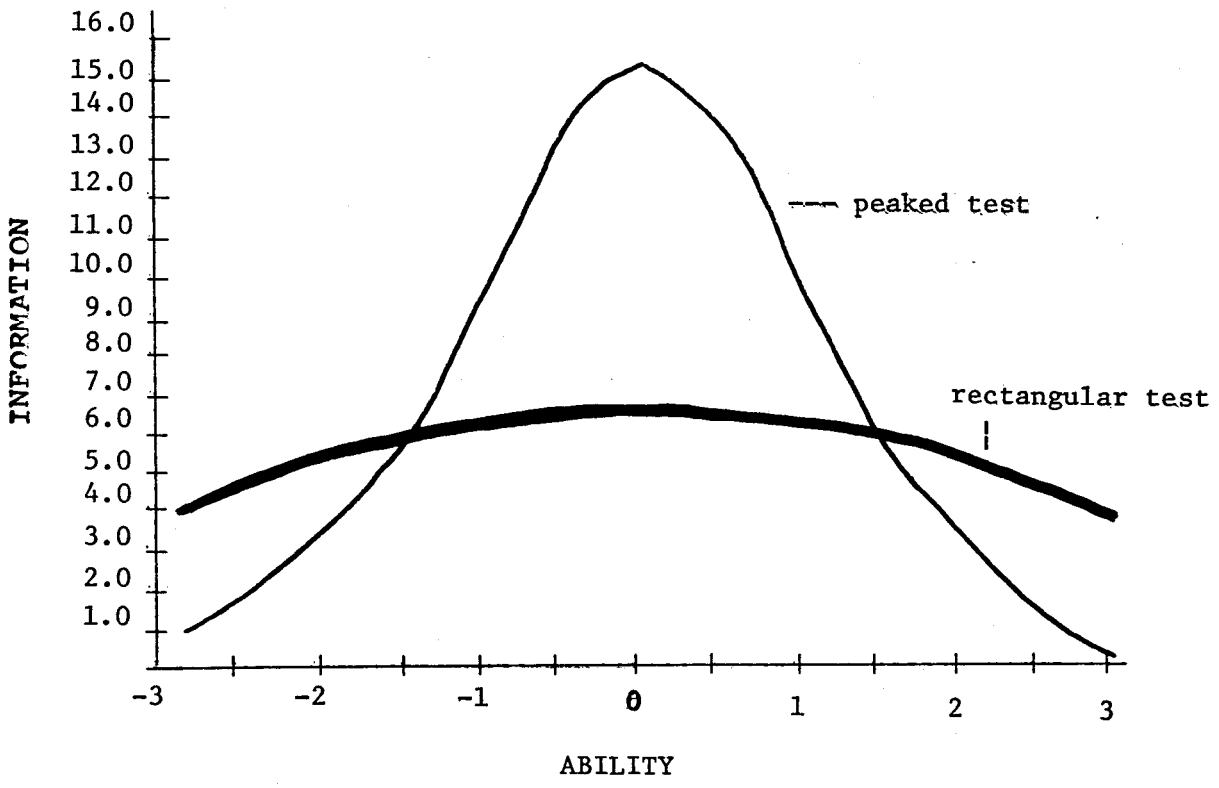
Owen, R. A. *A BAYESIAN APPROACH TO TAILORED TESTING*. RB-69-92. Princeton, New Jersey: Educational Testing Service, December, 1969.

Samejima, F. Behavior of the Maximum Likelihood Estimate In a Simulated Tailored Testing Situation. Paper presented at the meeting of the Psychometric Society, Iowa City, Iowa, April, 1975.

Vale, C. D. Problem: Strategies of Branching Through An Item Pool. In D. J. Weiss (ed.), *COMPUTERIZED ADAPTIVE TRAIT MEASUREMENT -- PROBLEMS AND PROSPECTS*. Research Report 75-6, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, November, 1975.

Weiss, D. J. The Stratified Adaptive Computerized Ability Test. Research Report 73-3, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1973.

FIGURE 1


16.0
15.0
14.0
13.0
12.0
11.0
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
INFORMATION
peaked test
rectangular test
-3
-2
-1
0
1
2
3
ABILITY

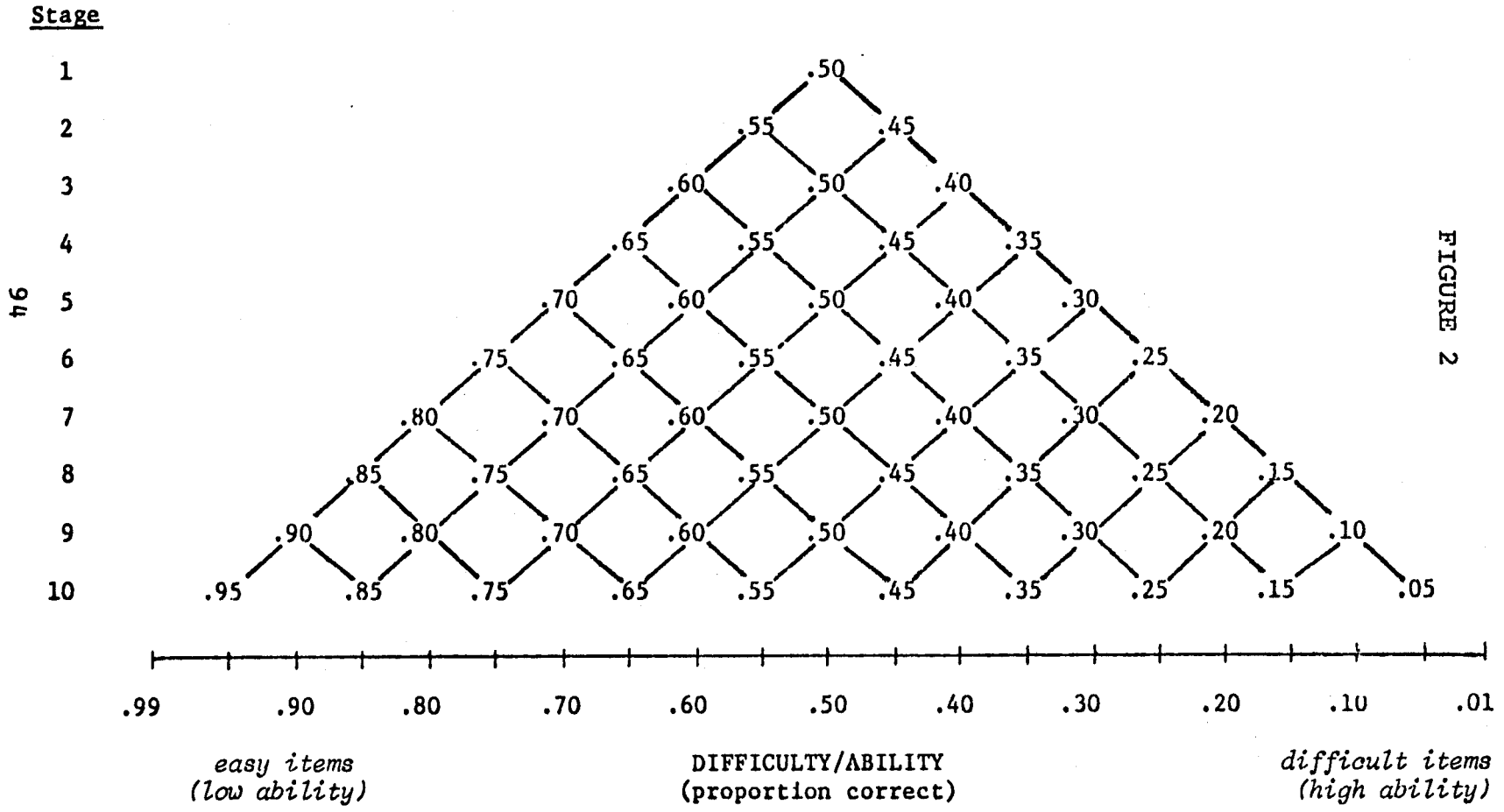

Stage
1
2
3
4
5
6
7
8
9
10
θ
.50
.55 .45
.60 .50 .40
.65 .55 .45 .35
.70 .60 .50 .40 .30
.75 .65 .55 .45 .35 .25
.80 .70 .60 .50 .40 .30 .20
.85 .75 .65 .55 .45 .35 .25 .15
.90 .80 .70 .60 .50 .40 .30 .20 .10
.95 .85 .75 .65 .55 .45 .35 .25 .15 .05
.99 .90 .80 .70 .60 .50 .40 .30 .20 .10 .01
easy items
(low ability)
DIFFICULTY/ABILITY
(proportion correct)
difficult items
(high ability)

FIGURE 2

FIGURE 3

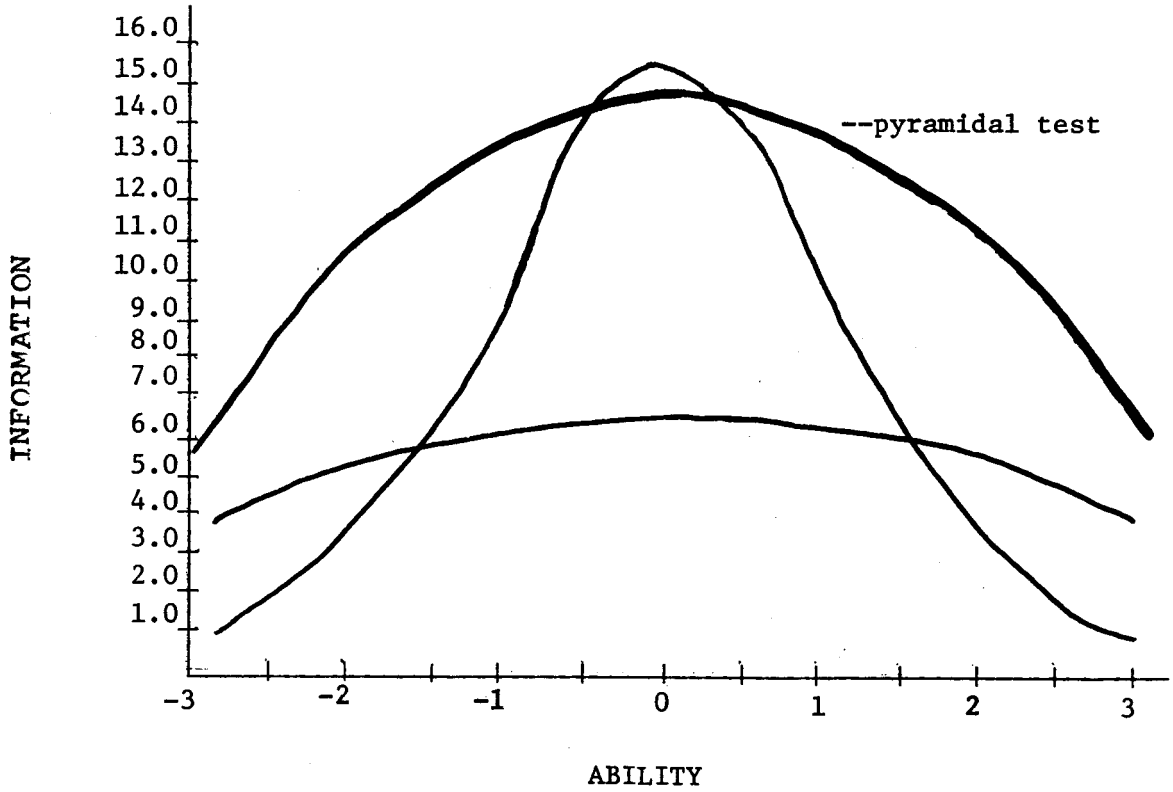

16.0
15.0
14.0
13.0
12.0
11.0
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
INFORMATION
--pyramidal test
-3
-2
-1
0
1
2
3
ABILITY

FIGURE 4

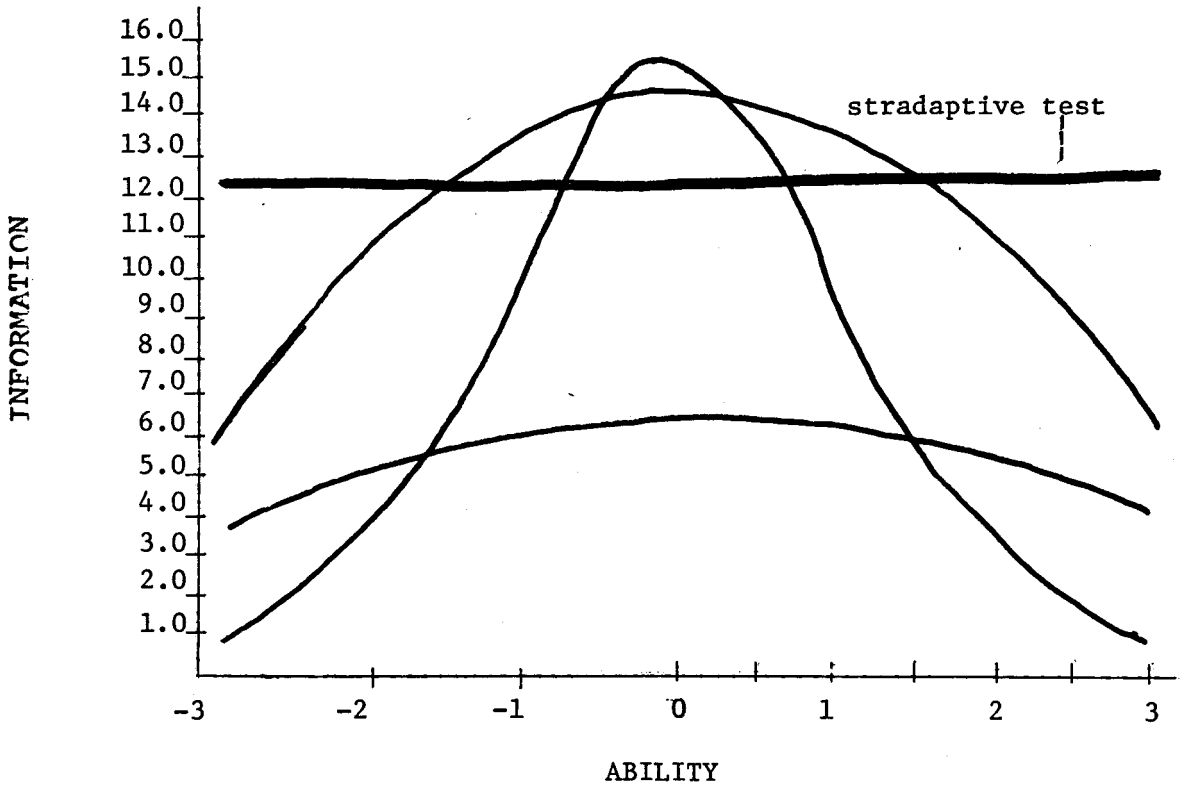

16.0
15.0
14.0
13.0
12.0
11.0
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
INFORMATION
stradaptive test
-3
-2
-1
0
1
2
3
ABILITY

FIGURE 5

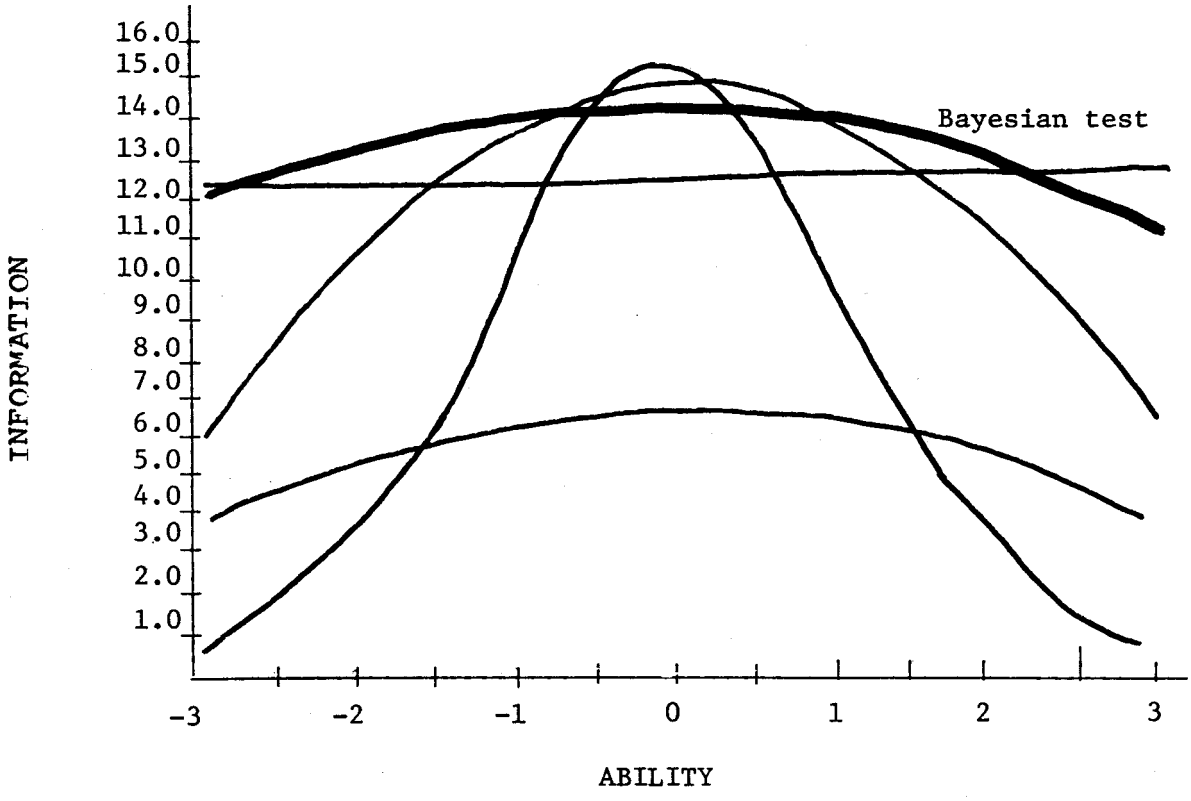

16.0
15.0
14.0
13.0
12.0
11.0
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
INFORMATION
Bayesian test
-3
-2
-1
0
1
2
3
ABILITY

FIGURE CAPTIONS

Figure 1. Information curves for two 24-item conventional tests: 1) a "peaked" test (bell-shaped curve) with homogeneous item difficulty values; 2) a "rectangular" test (flatter curve) with a wide range of item difficulty.

Figure 2. The schematic arrangement of items in a pyramidal adaptive test. The horizontal dimension represents item difficulty; the vertical dimension represents stage number, or sequential position of items in the test.

Figure 3. The information curve of a 24-item pyramidal adaptive test, superimposed on the conventional test information curves of Figure 1.

Figure 4. The information curve of a 24-item stradaptive test, superimposed over the information curves from Figure 3.

Figure 5. The information curve of a 24-item Bayesian sequential adaptive test, superimposed over the information curves from Figure 4.