Improving Measurement Quality and Efficiency with Adaptive Testing

David J. Weiss
University of Minnesota

Approaches to adaptive (tailored) testing based on item response theory are described and research results summarized. Through appropriate combinations of item pool design and use of different test termination criteria, adaptive tests can be designed (1) to improve both measurement quality and measurement efficiency, resulting in measurements of equal precision at all trait levels; (2) to improve measurement efficiency for test batteries using item pools designed for conventional test administration; and (3) to improve the accuracy and efficiency of testing for classification (e.g., mastery testing). Research results show that adaptive tests based on item response theory (IRT) can achieve measurements of equal precision at all trait levels, given an adequately designed item pool; these results contrast with those of conventional tests which require a tradeoff of bandwidth for fidelity/precision of measurements. Data also show reductions in bias, inaccuracy, and root mean square error of ability estimates. Improvements in test fidelity observed in simulation studies are supported by live-testing data, which showed adaptive tests requiring half the number of items as that of conventional tests to achieve equal levels of reliability, and almost one-third the number to achieve equal levels of validity. When used with item pools from conventional tests, both simulation and live-testing results show reductions in test battery length from conventional tests, with no reductions in the quality of measurements. Adaptive tests designed for dichotomous classification also represent improvements over conventional tests designed for the same purpose. Simulation studies show reductions in test length and improvements in classification accuracy for adaptive vs. conventional tests; live-testing studies in which adaptive tests were compared with "optimal" conventional tests support these findings. Thus, the research data show that IRT-based adaptive testing takes advantage of the capabilities of IRT to improve the quality and/or efficiency of measurement for each examinee.

Since the inception of the field of psychological measurement in the early 1900s, virtually all tests of ability and achievement (as well as instruments designed for the measurement of personality characteristics, attitudes, and other psychological variables) with one major exception, have had one common characteristic. These instruments are similar in that they have all used a fixed set of items which are administered to all examinees, even though examinees may differ substantially on the trait or traits being measured. As a consequence, psychometric theory and its resultant techniques for constructing measuring instruments has also been concerned with the problems of constructing fixed length tests. In addition, the vast majority of developments in psychological scaling and personality

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 6, No. 4, Fall 1982, pp. 473–492
© Copyright 1982 Applied Psychological Measurement Inc.
0416-6216/82/040473-20$2.00

measurement focus on the construction of measuring instruments with a fixed set of items. Even the majority of applications of item response theory are concerned with tests constructed from a common set of items administered to all examinees (e.g., Hambleton, in press; Lord, 1980).

In the construction of a "conventional" test with a fixed set of items, however, the test constructor is faced with a "bandwidth-fidelity" dilemma (McBride, 1976), since the difficulty of a test relative to a specific individual may vary from examinee to examinee. A "peaked" conventional test will provide very precise (high fidelity) measurements at trait levels at which the test is peaked. However, for individuals for whom the test is too difficult or too easy, the items will not provide optimal measurement. A "rectangular" conventional test, with item difficulties equally represented at various levels throughout the difficulty continuum, will provide only a few items at the appropriate difficulty level for any examinee, and the rest of the items will be either too difficult or too easy. Consequently, the rectangular conventional test will provide relatively equal, but low, levels of precision/fidelity throughout the trait range, thereby providing a test with good bandwidth. Given the requirement of a conventional test that a fixed set of items be administered to all examinees, therefore, the constructor of a conventional test must trade bandwidth for fidelity or vice versa.

Purpose

This paper describes how adaptive (tailored) testing can provide a solution to this bandwidth-fidelity dilemma. Adaptive testing methods are described that permit measurements of equal precision throughout the range of the trait being measured while maintaining high levels of efficiency. In addition, variations of these methods are described that permit efficient testing using small item pools from a conventional test or test battery, as well as methods for improving the accuracy and efficiency of testing for classification (e.g., mastery testing). Relevant research results are summarized.

Early Adaptive Tests

The major exception to the predominant trend of the use of conventional tests in psychological measurement is the individually administered Binet intelligence test, which was the first adaptive test. It is adaptive because the difficulty level of the items administered to each individual adapts (or is tailored) to the individual's ability level as it is determined during the process of testing. Binet's test had all of the characteristics of an adaptive test:

1. It used a differential starting point for different individuals, based on prior estimates of the individual's ability.
2. Items were scored as they were administered and answered by the examinee.
3. Based on the examinee's responses to items already administered, an item selection rule was used to select subsequent items to be administered.
4. Testing was terminated according to a predetermined termination criterion based on the examinee's performance on the test.

Thus, in the Binet tests the test administered to each individual can be a different length for each examinee and the subset of items selected for administration to each examinee differs from individual to individual depending upon their responses to previous items administered. The procedure tends to identify a subset of items *for each individual* that constitutes a test in which the individual has answered half of the items correctly and half of the items incorrectly, i.e., the test will be of about .5 difficulty *for that individual,* and will provide highly precise measurements for that individual in comparison to other subsets of the same number of items that could be administered to that examinee.

A number of designs not based on item response theory (IRT) but using pre-
structured item p stigated (Weiss, 1974). These include two-stage (Betz & Weiss,
1973, 1974; Lord idal" tests (Larkin & Weiss, 1974; Lord, 1970), flexilevel tests
(Betz & Weiss, 19 . the stratified-adaptive (stradaptive) test (Vale & Weiss, 1975a,
1975b, 1978; We e Binet tests were designed to be administered by trained psy-
chologists, some aptive tests were administered by paper and pencil. Because in-
dividual adaptive is expensive, and paper-and-pencil adaptive test administration
is both cumberso nost current adaptive tests are administered by computers. In a
computerized ada ; are stored in the computer and administered to individuals on
a cathode-ray ter Examinees respond on the CRT keyboard, and the response is
processed immed ter. Based on branching or item selection procedures specified
by the test admin n to be administered is selected by the computer and presented
on the CRT scree response.

RT-Based Adaptive Testing

Advantages

Although it is sign and administer computerized adaptive tests without IRT,
IRT-based adapti ntial advantages over non-IRT-based adaptive testing. A major
problem with test s that the scores in which abilities are expressed are on a differ-
ent metric than th tems. This makes it difficult to select items in a meaningful way
and cumbersome ation from previously administered items for purposes of item
selection during t ministration. One major advantage of IRT-based adaptive test-
ing is that person ities, as they are usually called for simplicity but without loss of
generality) and ite ie same scale, since the difficulties of items (the *b* parameters)
and the trait (θ) es als in IRT are on the same metric. As a consequence, if an abil-
ity estimate for a mined in IRT terms, the appropriate level of difficulty of the
item to first be a ndividual—or, in later stages of the test, the next item to be
selected for admi ividual—is expressed on the same numerical scale. Thus, by
matching the θ est 1 of appropriate difficulty for an individual can be selected.

A related adv ty levels can be estimated based on any subset of items ad-
ministered to an sequence, different items can be deliberately selected for ad-
ministration to ea ie resulting ability estimates will be placed on the same metric
by the IRT scoring rast to the Binet scoring procedure, no "age norming" is neces-
sary in IRT scorin

A third majo -based adaptive testing is that the use of IRT parameters for
items (combined nistration) permits use of nonstructured branching models for
item selection. As cessary to predefine a branching structure based on item diffi-
culties. Rather, th T-based adaptive testing are designed to search an item pool
with known paran ify one item out of the pool that best meets some item selection
criterion. Further selected for administration on the basis of more than just their
difficulty levels, si an simultaneously take into account item difficulties as well as
their discriminati ing parameters, if these have been estimated in advance.

A fourth adva adaptive testing is that the termination of an adaptive test can
be based on the p surements obtained. IRT scoring procedures not only make it
possible to estima each item is administered and answered but also make it pos-

sible to determine the precision (or standard error) of each ability estimate. These standard errors can then be used as criteria for terminating the adaptive test. When test termination is based on precision data, it is possible to measure individuals to a predetermined level of precision, given an adequate supply of items from which to choose.

Maximum Information Item Selection

Maximum information item selection uses item information, a transformation of the item characteristic curve or item response function, to select items for an adaptive test. Equation 1 (Lord, 1970, p. 73) expresses item information in terms of the parameters of the item (assuming a three-parameter logistic item response function) and provides a convenient computing formula, given an estimate of θ:

$$I\{\theta\} = \frac{2.89a_i^2\ (1 - c_i)}{\left[(c_i + e^{1.7a_i(\theta - b_i)}\right]\left[1 + e^{-1.7a_i(\theta - b_i)}\right]^2} \qquad [1]$$

where

a_i = item discrimination,
b_i = item difficulty, and
c_i = the pseudoguessing parameter of the item.

If a two-parameter model is used, c_i is set to 0.0; for the one-parameter model, a_i = 1.00.

To implement IRT-based adaptive testing, the value of item information for each item in the pool can be determined from Equation 1, given its item parameter estimates and substituting an ability estimate, $\hat{\theta}$, for θ. Using the variable entry capability of adaptive testing, prior information on the examinee (derived from other test data in the examinee's file, the examinee's own estimate of his/her ability level, or the examiner's estimate of the examinee's ability level) can be expressed on the θ metric. Given this initial estimate of the individual's θ level (which could be as simple as the mean of some population, e.g., $\hat{\theta} = 0$), values of item information are determined for all items in the pool at that estimated θ level. The item is chosen that provides the maximum level of information at the current $\hat{\theta}$.

The effect of choosing items in this way is to maximize the sum of item information across a set of items administered to an individual (hence, the name maximum information item selection strategy). Because of the inverse relationship between information and the standard error of $\hat{\theta}$ (Lord, 1980), this item selection strategy will minimize the standard error of $\hat{\theta}$ for each examinee, yielding the most precise θ estimate for each examinee given the items available.

The usual procedure for scoring response vectors during the process of administering a maximum information adaptive test is maximum likelihood θ estimation (Lord, 1980), resulting in a likelihood function that gives the likelihood of the observed response pattern as a function of θ. The $\hat{\theta}$ associated with a given response pattern is the value of θ at which the likelihood function is observed to have its maximum.

Likelihood functions differ not only in the location of the maximum of the function but also in the value of the likelihood at its maximum, which is related to the height of the function and inversely to its spread. The spread or variance of the function is an indication of the precision of the θ estimate. The variance of the likelihood function is inversely related to the observed value of the second derivative of the log likelihood function evaluated at $\hat{\theta}$. This latter quantity is referred to as response pattern information (Bejar & Weiss, 1979; Samejima, 1973), indexing the precision of the θ estimate determined at the maximum of the likelihood function. Specifically, the standard error of $\hat{\theta}$ for a given response pattern resulting from administration of a specified set of items is evaluated as the reciprocal square root of response pattern information evaluated at $\hat{\theta}$. These standard errors are very useful in terminating adaptive test administration.

A problem with maximum likelihood scoring is that maximum likelihood θ estimates cannot be determined for response patterns in which an examinee answers all of the items correctly or all of the items incorrectly. Similarly, there are some unusual kinds of response patterns exhibited by examinees for which the maximum likelihood estimation procedure fails to converge. Solutions to this problem include assigning arbitrary θ estimates or using arbitrary predetermined branching sequences in the early phases of an adaptive test (e.g., Reckase, 1977), but these procedures are not entirely satisfactory.

Bayesian Item Selection

An alternative solution to the problems of maximum likelihood estimation is to use a Bayesian θ estimation procedure proposed by Owen (1969, 1975); the availability of this procedure also suggests an alternative means of adaptive item selection.

Bayesian scoring. Owen's Bayesian scoring method imposes upon the likelihood distribution a normal density distribution with a specified mean and variance that is used to modify the likelihood distribution according to the probabilities associated with the normal density distribution (Bejar & Weiss, 1979). The result is a posterior distribution with a specified mean, which is the θ estimate, and a posterior variance. This posterior variance is associated with the standard error of the θ estimate, or response pattern information, resulting from maximum likelihood scoring. The posterior variance of the Bayesian θ estimate indicates the lack of precision of the estimate; the square root of the posterior variance is a Bayesian estimate of the standard error of the θ estimate.

The effect of multiplying the likelihood function by the normal distribution is to eliminate the nonconvergence problem, since a maximum or a mode (or, in Owen's method, a mean) can always be determined for the modified likelihood function. A second effect is that the θ estimates tend to be regressed toward the mean of the prior distribution. As a consequence, Bayesian-scored θ estimates tend to be biased toward whatever prior mean is used in the estimation procedure (e.g., Gorman, 1980; McBride, 1977).

However, estimating θ by Bayesian methods during the early phases of an adaptive test for purposes of obtaining a nonarbitrary θ estimate for response vectors not scorable by maximum likelihood methods is very useful, since it is frequently only used with the first few items administered in the test. These θ estimates are then used only as a means of selecting the next item to be administered. Since few tests with very small numbers of items will be used to make decisions about individuals, there is little effect of the Bayesian scoring procedure for practical uses of adaptive tests. That is, the vast majority of the uses of test scores will be based upon the maximum likelihood scoring procedure, which results in unbiased θ estimates.

Item selection. Owen (1969, 1975) suggested that items be selected, not on the information that they provide at a given $\hat{\theta}$, but on the basis of minimizing the posterior variance of the θ estimate based on the administration of a given item. Owen's item selection strategy utilizes a current θ estimate and its Bayesian variance as the prior for the item selection process. In order to select the next item to be administered during the adaptive test, this method evaluates the posterior variance of the θ estimate for each item in the pool under two conditions: (1) if the item is answered correctly and (2) if the item is answered incorrectly. For each item these variances are averaged. The process continues for all unadministered items in the pool, and the next item to be administered is the item with the lowest value of the average posterior variance that would be obtained if that item were administered. The item is administered, the response of the examinee is recorded, and the value of the posterior variance and the mean of the posterior distribution (the θ estimate) are determined. This new θ estimate and its posterior variance are then used to determine the new posterior variances for all unadministered items. The next item that minimizes the posterior variance is selected, administered, and the actual posterior variance is computed.

Using this process, θ estimation and item administration can be continued until a prespecified termination criterion is obtained, based, for example, on the value of the observed posterior variance. Since the Bayesian posterior variance is related to the variance of the likelihood function, which in turn is inversely related to the information provided by a given response pattern, selecting items by maximum information is closely related to selecting items by minimizing the Bayesian posterior variance. A difference occurs in the actual items selected for administration. This is attributable to the fact that the Bayesian θ estimation procedure results in θ estimates biased towards the prior mean used. In some cases, particularly for individuals whose θ estimates are distant from the prior mean, this will result in different items being selected. The only study that has directly compared the two item selection procedures (Sympson, Weiss, & Ree, 1982) shows that in a live-testing implementation of the two adaptive testing procedures, an average of approximately 85% of the items selected by the two procedures were the same. However, because of the bias inherent in the Bayesian θ estimation procedure, the two procedures will result in measurements with somewhat different characteristics.

Applications of IRT-Based Adaptive Testing to Testing Problems

As indicated, an adaptive testing strategy consists basically of three components: (1) a means for selecting the first item to be administered to an individual, (2) a means for scoring items during the process of test administration and for selecting the next item to be administered, and (3) a means for terminating the adaptive test on an individual basis. These three characteristics of adaptive tests can be combined in a variety of ways in order to solve specific measurement problems. These problems include the following:

1. Simultaneously improving measurement efficiency and controlling the precision of the measurements obtained for all examinees (i.e., obtaining equiprecise measurements at all levels of the trait continuum), using an item pool specifically designed for adaptive testing.
2. Improving measurement efficiency using test item pools not designed for equiprecise measurement.
3. Improving the efficiency and accuracy of testing for mastery or other classification decisions.

Adaptive Testing for Equiprecise Measurements

The vast majority of research on adaptive testing has been concerned with the design and evaluation of adaptive testing strategies and item pools for equiprecise measurement. Equiprecise measurements are measurements of equal precision at all levels of the trait continuum being measured. Adaptive testing of this type is designed to eliminate the bandwidth-fidelity dilemma that results in the construction of conventional tests. Adaptive tests that have equiprecise measurement characteristics result from the design of an item pool which has highly discriminating items equally represented at the full range of difficulty associated with the range of trait levels anticipated in the entire population to be measured. The higher the discriminations of the items in the pool, the more rapidly will the adaptive testing strategy achieve desired levels of measurement precision; the more rectangularly distributed are the difficulty levels of the items, the more equiprecise will be the measurements resulting from the adaptive test. There must also be a sufficient number of items in the neighborhood of each θ level to attain the desired degree of precision in the θ estimates.

Adaptive testing for equiprecise measurement is designed to use differential entry points for starting the adaptive test. Obviously, the more accurate the initial $\hat{\theta}$ is for selecting the first item to be administered, the more quickly the adaptive tests will converge upon the correct θ estimate for the individual. However, experience indicates that most adaptive tests are shortened by only a few items with the use of accurate entry θ estimates. Nevertheless, the use of differential entry points, such as is used in the Binet tests, is an appropriate means of increasing adaptive test efficiency.

Equiprecise measurement is, of course, achieved by continuing the adaptive test until the level of information associated with each $\hat{\theta}$ equals a prespecified value. Given a perfect item pool constructed of items of equal and high discriminations and equally represented at all difficulty levels throughout the θ continuum, equiprecise measurement can be achieved with a fixed length adaptive test. However, since perfect item pools exist only in theory, variable length adaptive tests are necessary with real item pools in order to achieve equiprecise measurements. In this case, testing until a specified level of information/precision is achieved will result in longer tests for individuals whose abilities are in the range of difficulty of the item pool where fewer items are located, or where items of lower discrimination are located.

Data from Crichton (1981) illustrate the improved efficiency and other measurement characteristics of the two major IRT-based adaptive testing strategies in comparison to peaked and rectangular conventional tests. These data were derived from a monte carlo simulation study that was concerned primarily with the effects of errors in item parameters on the performance of adaptive testing strategies. The data used below, however, are from the baseline error-free case in which item parameters were assumed to be estimated without error. Crichton's simulation used a three-parameter logistic model with abilities ranging from -3.2 to $+3.2$ in intervals of .4. There were 100 simulated examinees (simulees) at each of the 17 levels of θ. To evaluate the effects of the independent variables using a realistic item pool, the item pool was modeled on the real, numerical reasoning item pool parameterized by Sympson et al. (1982). Thus, the item pool did not entirely have the characteristics that would result in equiprecise measurements for the adaptive tests but reflected an approximation to these desired characteristics that it was possible to obtain with real data.

Peaked and rectangular conventional tests ranging from 5 to 30 items in length were constructed. Adaptive tests, both Bayesian and maximum information, were administered to the 1,700 simulees. All adaptive tests used as the entry point $\hat{\theta} = 0.0$ with a variance of 1.0. Adaptive tests were scored by maximum likelihood at lengths of 5, 10, . . . to 30 items. All tests, conventional and adaptive, were scored by maximum likelihood to eliminate scoring method as a variable affecting the results.

Using a normally distributed random sample of 630 simulees from the 1,700, fidelity correlations—the correlation of true (generating) θ and observed $\hat{\theta}$—were computed at each test length for each testing strategy. These correlations are shown in Table 1. The fidelity correlations for the two conventional tests ranged from .64 to .94, while the correlations for the adaptive tests ranged from .81 to .97. The greater efficiency of the adaptive tests is illustrated by the fact that fidelity correlations for the 10-item Bayesian adaptive test (.94) were equal to the fidelities of the rectangular and peaked conventional test at 30 items. Thus, the Bayesian adaptive tests required one-third the number of items in the two conventional tests to achieve equal fidelity. The data also show that the fidelities of the two adaptive tests at 15 items (.95 and .96) were not achieved by either of the conventional tests at 30 items.

These simulation results were supported in live testing by McBride and Martin (in press). In that study, Bayesian adaptive tests were administered to a group of 263 Marine recruits using an item pool of 150 verbal ability items. A separate group of 267 recruits took rectangular conventional tests con-

Table 1
Fidelity Correlations of True and Estimated Ability Levels for Conventional and Adaptive Tests at 5- to 30-Item Lengths

Test	Number of Items					
	5	10	15	20	25	30
Conventional						
Peaked	.64	.73	.85	.89	.91	.93
Rectangular	.65	.83	.88	.92	.93	.94
Adaptive						
Bayesian	.85	.94	.96	.96	.97	.97
Maximum Information	.81	.90	.95	.96	.97	.97

structed from the same item pool; both groups completed a 50-item "criterion" conventional test. To compute alternate forms reliabilities, the conventional test group was administered two 30-item alternate forms conventional tests; the adaptive test group took two interleaved 30-item alternate forms of the same adaptive testing strategy. Bayesian ability estimates were determined at each test length from 1 to 30 items for the adaptive test, and the conventional test was scored by number correct at each test length. The results showed that the alternate forms reliability of the adaptive test at 9 items (.800) was equal to that of the conventional test at 17 items (.798). Thus, similar to the results shown in Table 1, the adaptive test obtained measurements of equal precision with about half the number of items. When the two testing strategies were compared in terms of validity based on correlations with the criterion test, the validity correlations for the adaptive tests at 11 items (r = .80 for both forms) were equal to those of the rectangular conventional tests at 29 items (r = .79 and .80), indicating an almost two-thirds reduction in the number of items required for the adaptive test to measure as well as the conventional test on the validity criterion.

The use of fidelity (or validity) correlations for evaluating the performance of testing strategies permits only limited comparisons among the strategies, because correlations are sensitive to the distribution of ability in the sample investigated and because they do not permit examination of the performance of the strategies at different levels of θ. IRT provides alternative means—particularly in simulation, but also in live-testing sudies (e.g., Bejar, Weiss, & Gialluca, 1977; Vale & Weiss, 1977)—for comparing the performance of adaptive testing strategies with each other or for comparing adaptive with conventional testing strategies. Crichton's (1981) study was designed to permit the comparison of testing strategies on IRT criteria. Although Crichton compared the testing strategies at various test lengths, the major effect of test length was to accentuate differences between strategies. The data reported below from Crichton's study are all based on the 30-item test length, since they provide the most conservative comparison between adaptive and conventional tests.

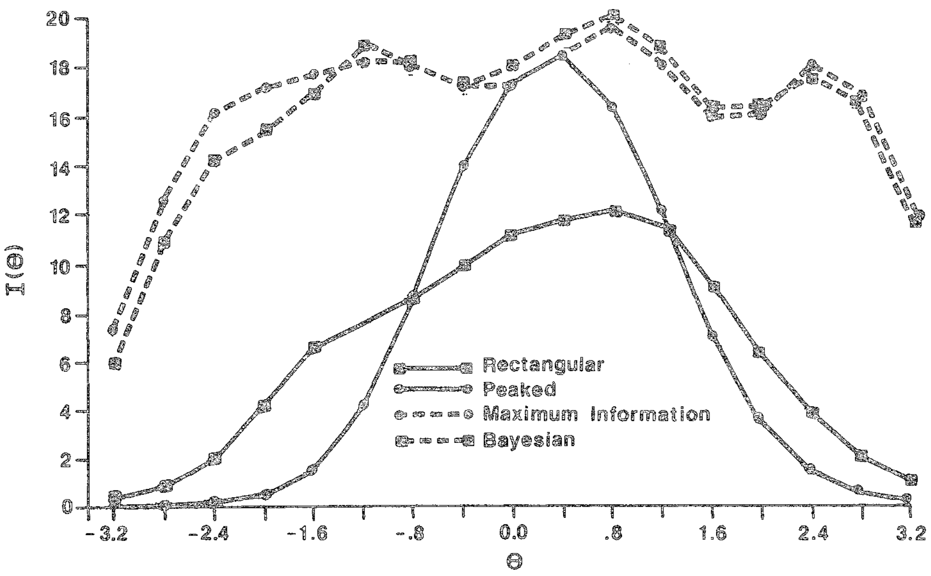
Figure 1 shows test information curves for the two 30-item adaptive tests and the two 30-item conventional tests; for the adaptive tests information was computed as the sum of item information values for the items administered to each simulee, averaged across the 100 simulees at each θ level. The two solid lines in Figure 1 for the conventional test illustrate the bandwidth-fidelity dilemma inherent in the use of conventional tests. As Figure 1 shows, test information for the peaked conventional test is very high for θ levels near zero (that the peak of information is not at zero is due to the nonzero guessing parameters used for the items), with information dropping off very rapidly for θ levels distant from zero. For the rectangular conventional test, test information is not as high at the center of the θ distribution but, similarly, does not drop off as rapidly as for the peaked conventional test.

The two adaptive tests shown in Figure 1 obtained relatively equiprecise measurement for θ levels from −2 to about +2.8. Outside these regions, information drops off due to the lack of items beyond the boundaries of the θ continuum above $b = 3.0$ and below $b = -3.0$. The variations in information near the center of the distribution are due to inadequacies in this item pool which, as indicated, was modeled after a live-testing item pool.

The greater efficiency of the adaptive test is shown by the ratio of the information functions, which can be interpreted as the increase in test length of the test with lower levels of information required for it to measure at the same level of information as the more informative test (Lord, 1980). For example, at $\theta = 0.0$, the peaked conventional test measured with approximately the same level of information as the maximum information and Bayesian adaptive tests. By contrast, the 30-item rectangular conventional test would need 46.5 items to measure as well as a maximum information adaptive test and would need 48 items to measure as well as the Bayesian adaptive test. At $\theta = -.8$ the average information of the rectangular and peaked conventional tests was 8.58 and that for the adaptive tests was 18.16; the conventional tests would need to be lengthened from 30 to 63.5 items to measure as well as the 30-item adaptive tests. Finally, at $\theta = +2.4$ the rectangular conventional test would need 131.7 items to measure as well as the 30-item adaptive tests, while the peaked conventional test would require 316.8 items.

These data support Crichton's (1981) fidelity correlation data and the live-testing data obtained by McBride and Martin (in press) but provide an interpretation of efficiency dependent upon θ levels.

Figure 1
Test Information for 30-Item Rectangular and Peaked Conventional Tests
and Maximum Information and Bayesian Adaptive Tests



As is obvious from Figure 1, the relative efficiency of adaptive compared to conventional tests increases as $\hat{\theta}$ deviates from the average level for the group being tested.

Crichton (1981) also compared adaptive and conventional tests in terms of bias, root mean square error, and inaccuracy of the θ estimates, conditional on θ. She defined bias as

$$B(\theta) = \frac{\Sigma(\hat{\theta}_i - \theta)}{N} \qquad [2]$$

root mean square error as

$$RMSE(\theta) = \sqrt{\frac{\Sigma(\hat{\theta}_i - \theta)^2}{N}} \qquad [3]$$

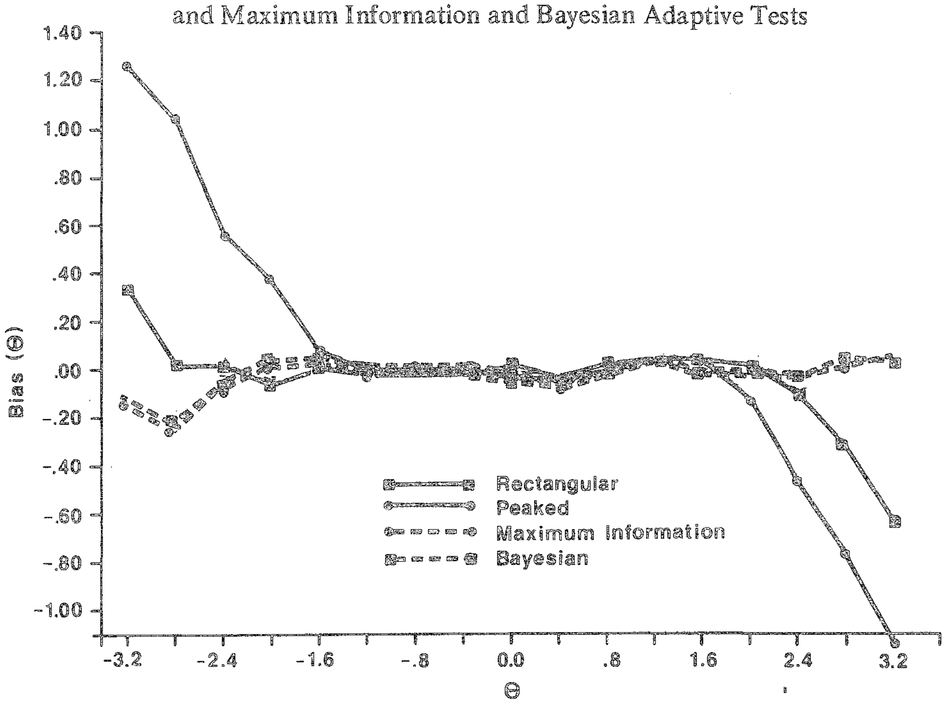
and inaccuracy as

$$IA(\theta) = \frac{\Sigma|\hat{\theta}_i - \theta|}{N} \qquad [4]$$

Each of these characteristics of the θ estimates provides additional information beyond that provided by fidelity and information.

Figure 2 shows bias for the adaptive and conventional tests as a function of θ for the 30-item test length. As can be seen, neither of the adaptive tests measures with substantial bias, except for θ values at the very lowest extreme of the θ distribution. By comparison, the two conventional tests re-

Figure 2
Bias for 30-Item Rectangular and Peaked Conventional Tests and Maximum Information and Bayesian Adaptive Tests



sult in differential bias at different θ levels. The peaked conventional test results in substantial bias for θ levels beyond $\theta = \pm 1.6$. The rectangular conventional test measures with less bias than the peaked conventional tests. In both cases the conventional tests tend to overestimate low θ levels and to underestimate high θ levels.

Figure 3 shows the root mean square error (RMSE), or standard deviation, of the θ estimates for the four tests. Again, both conventional tests had substantial variability of errors in their θ estimates, with the peaked conventional tests resulting in higher levels of RMSE for θ estimates than the rectangular conventional tests and both conventional tests providing higher RMSE throughout the θ range than either of the adaptive tests. There were virtually no differences between the two adaptive tests.

Finally, Figure 4 shows the inaccuracy of the four testing strategies. The patterns of inaccuracy were very similar to those for RMSE. Again, the conventional tests measured with higher levels of inaccuracy at all levels of θ than did the adaptive tests, with only minor differences around the mean of the θ distribution and extending to about 1.5 standard deviations above the mean. For low ability examinees, below $\theta = -.4$, both conventional tests resulted in measurements of higher inaccuracy than either of the adaptive tests. There was, again, very little difference between the two adaptive testing strategies.

Thus, the data from both live-testing studies and computer simulations indicate (1) that adaptive testing can result in higher levels of measurement efficiency than conventional tests and (2) that they can solve the bandwidth-fidelity dilemma inherent in conventional tests, resulting in equiprecise measurements throughout the θ range. In addition, other characteristics of the measurements resulting from the adaptive tests—such as bias, RMSE, and inaccuracy of the θ estimates—are more desirable for the adaptive tests than for the conventional tests.

Figure 3
Root Mean Square Error for 30-Item Rectangular and Peaked Conventional Tests
and Maximum Information and Bayesian Adaptive Tests

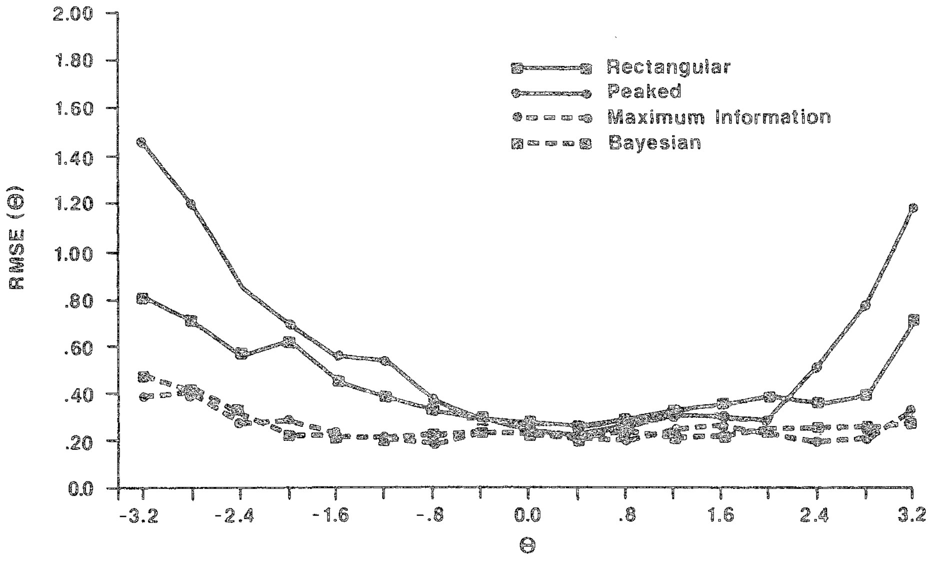
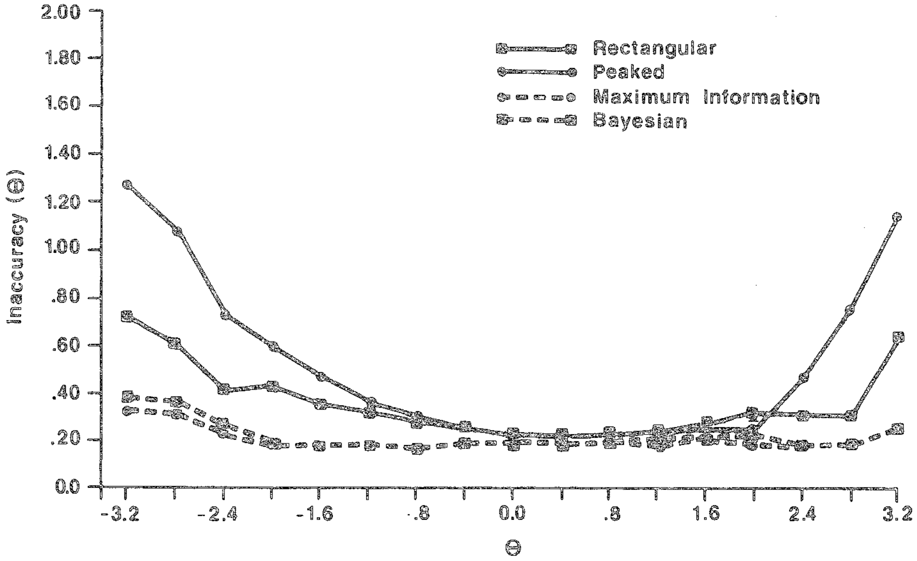


Figure 4
Inaccuracy for 30-Item Rectangular and Peaked Conventional Tests and Maximum Information and Bayesian Adaptive Tests



Adaptive Testing To Improve Measurement Efficiency

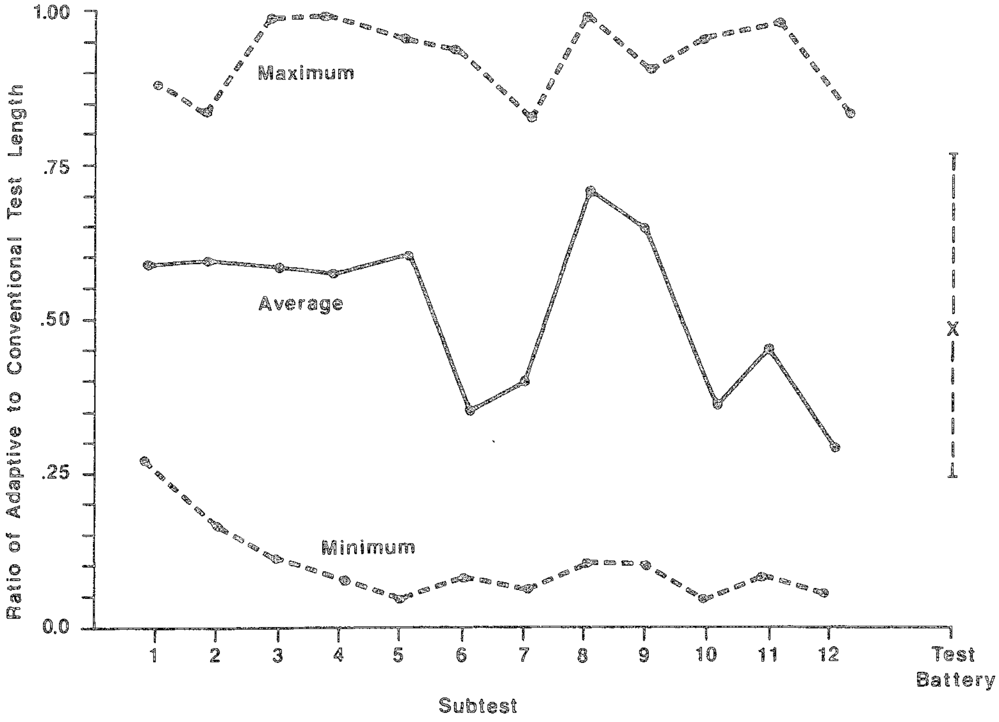
Although the use of item pools specifically constructed for adaptive testing can result in increases in measurement efficiency (by reducing test length), while at the same time achieving equiprecise measurements at all levels of the trait continuum, it is not always possible in the initial phases of implementing adaptive testing to develop item pools composed of items with high discriminations rectangularly distributed in difficulties. By varying the adaptive testing termination rule, however, it is possible to use adaptive testing strategies to improve measurement efficiency, i.e., to reduce test length, while maintaining (but not necessarily improving) the psychometric quality of the measurements. This approach is particularly useful as a first-stage implementation of adaptive testing using relatively small item pools originally designed for conventional tests. The reduction in test length possible from this implementation of adaptive testing will provide maximum efficiency in the administration of test batteries, since by reducing each subtest to its minimum required length for each individual, substantial time savings can be realized in the administration of a number of tests comprising a test battery.

To reduce test length using item pools originally developed for conventional tests, Brown and Weiss (1977) proposed the use of a maximum information item selection strategy with variable entry but with termination of the test based on a minimum value of item information for the item to be selected for administration at a given point in the test. Thus, testing continues for each individual until essentially there are no items available for administration at a given stage in the test administration that provide more than some trivial amount of information. In their initial implementation of this strategy, Brown and Weiss terminated test administration for each individual when the maximum information provided by any unadministered item was .01 or .001. In later studies of the procedure (Gialluca & Weiss, 1979; Maurelli & Weiss, 1981), termination criteria using information values of .01 and .05 were also studied.

Brown and Weiss (1977) combined this maximum information item selection procedure with Bayesian θ estimation for use within each subtest in a test battery. To further reduce test battery length, an intersubtest procedure based on multiple regression was used to obtain starting points for later subtests in the test battery. In the implementation of this procedure, Bayesian θ estimates from prior subtests were regressed on later subtests in order to obtain a predicted $\hat{\theta}$ for the next subtest to be administered. This $\hat{\theta}$ was then used by the intrasubtest maximum information item selection procedure to select the first item to be administered. The procedure was repeated with a new multiple regression equation at the end of each subtest.

Figure 5 summarizes the ratios of tests lengths achieved by this adaptive testing strategy to those of the conventional tests comprising a 12-test military achievement test battery totalling 201 items. These results were obtained by real-data simulation in which the adaptive administration was simulated based on the responses of 365 examinees to the conventional test battery. IRT item parameters were estimated by Urry's (1976, p. 99) three-parameter normal ogive method. As the data show, average adaptive test lengths for the 12 subtests ranged from 30% of conventional test length for the 18-item Subtest 12 to 71% for the 10-item Subtest 8. The shortest adaptive test required only 4% of the items in Subtest 10. Average test length for the entire adaptively administered test battery was 50.6% of the total battery length, indicating that only about 100 of the original 201 items were extracted, on the average, from the conventional test for administration by the adaptive testing strategy. Test battery lengths for the adaptive tests ranged from a minimum of 27 items (13% of the conventional test battery length) to 153 items (76% of battery length). None of the examinees required all 201 items.

Figure 5
Ratio of Mean, Minimum, and Maximum Adaptive Test Lengths
to Conventional Test Lengths for 12 Subtests and Total Test Battery



Brown and Weiss's (1977) data also showed that there was no reduction in test information for any of the subtests administered, as would be expected from the nature of the item selection strategy, since only items that provide nontrivial amounts of information are administered to any individual. In addition, their data showed high correlations between scores on the conventional subtests and ability estimates from the shortened adaptive tests. In a computer simulation of the same test battery, Maurelli and Weiss (1981) showed that this adaptive testing strategy resulted in equal fidelities (correlations with true θ) to those of the conventional tests, even with almost 50% average reduction in test battery lengths.

Maurelli and Weiss (1981) separated the effects of the intrasubtest and intersubtest adaptive branching strategies in their simulation, while Gialluca and Weiss (1979) studied the separate effects of the intersubtest and intrasubtest strategies using real-data simulation based on data from 1,600 students in a college biology test battery. The combined data showed that the intrasubtest item selection strategy accounted for 95% to 98% of the test length reduction, with intersubtest adaptive branching accounting for the remaining 2% to 5% of test length reduction. However, the magnitude of this reduction would be expected to be a function of the degree of subtest intercorrelations to some extent, since with highly intercorrelated subtests fewer items should be needed to locate the appropriate level of difficulty for an individual in a particular subtest. While intersubtest branching accounted for little test length reduction, Maurelli and Weiss's (1981) data showed that intrasubtest item selection with intersubtest branching helped to maintain mean test battery information levels closer to those of the conventional tests than did the use of intrasubtest item selection alone.

Thus, when a test battery is available on which IRT parameters can be estimated, this form of adaptive testing can be used to improve testing efficiency and to effect substantial reductions in test length without reductions in the psychometric quality of the measurements obtained. To improve measurement quality, however, it would be necessary to augment each subtest's item pool by increasing the number of items available for administration across the range of item difficulty so that individuals can be measured with equal precision across the range of the traits being measured. Adaptive testing designed solely for reducing test length (increasing measurement efficiency) will result in essentially the same information curves that exist in the conventional test item pools to which it is applied and, hence, cannot improve measurement precision.

Adaptive Testing for Classification or Mastery Decisions

By again varying the termination rule, maximum information item selection (combined with Bayesian scoring) can be used to increase the efficiency and accuracy of dichotomous classifications, such as those used in mastery testing. In this application (Kingsbury & Weiss, 1979) an adaptive test is terminated when the θ estimate for an individual is confidently above or below a prespecified cutting score, which has been converted to the θ metric from the proportion-correct metric (if the cutoff score for mastery or other classification has originally been expressed on that metric). This conversion is accomplished by the test characteristic curve or test charactertistic function (Lord, 1980, p. 49)—the regression of proportion correct on θ—based on the estimated *a*, *b*, and *c* parameters for the items that constitute the item pool for the mastery test. Given this function, the expected proportion of correct responses is then converted nonlinearly to a value on the θ metric, θ_m.

Testing for each individual begins at θ_m, administering the item at that level providing the maximum information. The item response is scored, and the next item is selected based on the individual's response to the first item and on the Bayesian θ estimate that results from the use of Owen's (1969, 1975) Bayesian scoring method. In this case, the Bayesian $\hat{\theta}$ is used rather than the maximum likelihood estimate, since it permits θ estimation after each item is administered and since it also provides Bayesian-based confidence intervals that can be used to implement the termination criterion. Although the Bayesian θ estimates have been shown to be biased for short tests (McBride, 1977), regressing θ estimates toward the mean, this use of the Bayesian θ estimate combined with the Bayesian prior for selecting items for administration by the maximum information item selection strategy will result in a conservative estimate of mastery or nonmastery, since individual θ estimates will tend to be regressed toward the cutoff score. The effect is to lengthen the adaptive test somewhat more than it would be if this regression effect did not exist.

Test administration continues for each individual with the Bayesian $\hat{\theta}$ and its posterior variance computed after each item. Based on a prespecified standard error confidence interval associated with the Bayesian $\hat{\theta}$ (e.g., 95%, 99%), testing is terminated when the Bayesian $\hat{\theta}$ and its confidence interval fall entirely to one side or the other of the IRT-based mastery criterion, θ_m. A major characteristic of this approach is that it results in classifications which have minimally equal confidence for all individuals tested, given an item pool with a sufficient number of items distributed along the achievement continuum. The method results in efficient classifications, requiring very few items for individuals whose achievement levels are distant from the cutoff and more items for individuals whose achievement levels are close to the mastery criterion.

Kingsbury and Weiss (1979, 1980) studied this adaptive mastery testing (AMT) strategy in comparison to conventional mastery tests by both simulation and live testing. Table 2 shows the observed test lengths for three conventional mastery tests of 10, 25, and 50 items and for three adaptive mastery tests with maximum test lengths of 10, 25, and 50 items in two item pool configurations. The uniform item pool was an unrealistic pool in which all items were equal in their *a, b,* and *c* parameters and functioned as random replacements for each other. The *a*-, *b*-, and *c*-variable pool was one in which items were allowed to vary in all three parameters, more realistically representative of a real item pool. As shown by the data, in both item pools the AMT strategy resulted in minor mean test length reductions for the 10-item maximum test length, reductions of about one-third for the 25-item conventional test, and test length reductions of over 50% for the 50-item test length.

Table 3 shows the fidelity phi correlations between observed mastery state and true mastery state for the two testing strategies in the two item pools. As the data indicate, the two testing strategies measured mastery states with equal accuracy in the unrealistic uniform item pool. However, for the

Table 2
Mean Number of Items Administered to Each Simulee
for Conventional and Adaptive Mastery Tests
In Two Item Pools at Three Maximum Test Lengths

Item Pool and Testing Strategy	Maximum Test Length 10	25	50
Uniform Pool			
Conventional	10.00	25.00	50.00
Adaptive	9.03	15.99	23.00
a-, *b*-, and *c*-Variable Pool			
Conventional	10.00	25.00	50.00
Adaptive	8.73	16.35	23.39

a-, *b*-, and *c*-variable item pool, the AMT strategy resulted in substantially higher fidelities, particularly at the shorter test lengths. When combined with the shorter test lengths of Table 2, the AMT strategy determined mastery status more efficiently and more accurately than did the conventional mastery test, particularly in the more realistic item pool.

Table 3
Phi Correlations Between Observed Mastery State and True Mastery State for Conventional and Adaptive Mastery Tests in Two Item Pools at Three Maximum Test Lengths

Item Pool and	Maximum Test Length		
Testing Strategy	10	25	50
Uniform Pool			
Conventional	.771	.837	.875
Adaptive	.775	.840	.871
a-, *b*-, and *c*-Variable Pool			
Conventional	.290	.670	.735
Adaptive	.470	.733	.787

Table 4 compares the kinds of classifications made by the two testing strategies. As the data show, there were essentially no differences in classification errors for the two testing strategies in the uniform item pool. In the *a*-, *b*-, and *c*-variable pool, however, the AMT tended to have a more equal distribution of false mastery and false nonmastery decisions than did the conventional tests, as well as lower overall misclassification rates.

These monte carlo simulation data were supported by live AMT administration to 463 students of an achievement test battery covering five content areas in a college biology course (Kingsbury & Weiss, 1981). The conventional mastery tests were "optimal" (Lord, 1980) in that they were constructed of sets of items that were selected to provide the most information at the cutting score, θ_m. The criterion was overall classroom mastery status based on all examinations administered in the course combined with the laboratory work done by the students. Table 5 shows the percentage of correct and incorrect mastery and nonmastery classifications made by the two testing strategies in all of the five content areas. As the data show, for all content areas except Content Area 5, the AMT had higher levels of total correct classifications, and in several cases a more balanced distribution between incorrect nonmastery and incorrect mastery classifications, than did the conventional mastery test. These mastery classifications made by the AMT strategy were achieved with an average 80% reduction in test length from the 20-item conventional tests; almost half of the AMT classifications in each content area required only three items or less. Across all five content areas the "optimal" conventional tests required 100 items to achieve the five mastery classifications, while only about 20 items were required for the adaptive tests to make the five mastery classifications for the average examinee. The adaptive testing strategy, therefore, arrived at classifications of higher quality, with considerably shorter test lengths, than did these "optimal" conventional mastery tests.

Table 4
Percentage of Incorrect Classifications by Type of Error Made by Conventional and Adaptive Mastery Tests in Two Item Pools at Three Maximum Test Lengths

Maximum Test Length and Classification	Item Pool and Testing Strategy			
	Uniform		a-, b-, and c-Variable	
	Conventional	Adaptive	Conventional	Adaptive
10 Items				
False Mastery	3.6	3.6	0.0	8.0
False Nonmastery	8.0	7.8	44.6	19.4
Total	11.6	11.4	44.6	27.4
25 Items				
False Mastery	2.6	3.0	2.6	5.2
False Nonmastery	5.6	5.0	15.2	8.2
Total	8.2	8.0	17.8	13.4
50 Items				
False Mastery	2.8	3.0	7.4	5.0
False Nonmastery	3.4	3.4	5.8	5.6
Total	6.2	6.4	13.2	10.6

Discussion and Conclusions

The data summarized above support the use of adaptive testing in a number of psychometric environments for improving both test efficiency and the quality of obtained measurements. They also show that some implementations of adaptive testing do not require specially constructed item pools in order to achieve desirable ends. Rather, variations of the basic adaptive testing paradigms can result in ability estimates that are obtained more efficiently than those of conventional tests, in some cases with improvements in measurement quality. Of course, when it is possible to construct an item pool specifically designed for adaptive testing, the desirable goals of increased measurement efficiency combined with equiprecise measurement (or classification) can be achieved. These characteristics are achieved in adaptive testing by dynamically selecting from an item pool a test of the appropriate level of difficulty (e.g., proportion correct of .50) separately *for each examinee.*

The desirable outcomes achieved by adaptive testing are best obtained by adaptive testing based on item response theory. The use of item information as an item selection mechanism results in an efficient and highly flexible means of adaptive item selection. Where computing time is a problem, such as in some microcomputer-based adaptive testing systems, the maximum information strategy can be modified somewhat to reduce computer processing times. In this case, the strategy is known as the stratified maximum information strategy (STMI; Sympson et al., 1982) and is based on a precalculated table for information values, as opposed to calculating information values for the items remaining in the pool at the current estimated level of θ after each item is administered.

To implement the STMI strategy, information values of all items in the pool are computed at some number of predefined levels of θ, such as 25 values of θ ranging from +3 to −3. These informa-

Table 5
Percentage of Correct and Incorrect Mastery and Nonmastery Classifications Made by Conventional and Adaptive Mastery Tests Within Each of Five Content Areas

Content Area and Classification	Testing Strategy	
	Conventional	Adaptive
Content Area 1		
Correct Nonmastery	45.3	50.7
Incorrect Nonmastery	41.1	30.5
Correct Mastery	12.6	16.0
Incorrect Mastery	.9	2.8
Total Correct	57.9	66.7
Total Incorrect	42.0	33.3
Content Area 2		
Correct Nonmastery	42.1	52.1
Incorrect Nonmastery	34.6	35.2
Correct Mastery	19.2	11.3
Incorrect Mastery	4.2	1.4
Total Correct	61.3	63.4
Total Incorrect	38.8	36.6
Content Area 3		
Correct Nonmastery	45.8	53.1
Incorrect Nonmastery	47.2	41.8
Correct Mastery	6.5	4.7
Incorrect Mastery	.5	.5
Total Correct	52.3	57.8
Total Incorrect	47.7	42.3
Content Area 4		
Correct Nonmastery	53.1	48.9
Incorrect Nonmastery	42.6	31.5
Correct Mastery	4.3	17.4
Incorrect Mastery	0	2.3
Total Correct	57.4	66.3
Total Incorrect	42.6	33.8
Content Area 5		
Correct Nonmastery	53.1	50.2
Incorrect Nonmastery	44.5	46.1
Correct Mastery	2.4	2.7
Incorrect Mastery	0	.9
Total Correct	55.5	52.9
Total Incorrect	44.5	47.0

tion values are then sorted, with the item providing highest information at each θ value listed first and the item providing the lowest amount of information at that θ value listed last. In the stored table, each item is listed at each θ level. Once these computations have been completed, the maximum test length is determined for the adaptive test to be administered. All items in the sorted list below that maximum test length are eliminated from the table. The table thus contains the n items at each θ level providing the maximum information at that θ level, where n is the maximum test length.

Items are administered and θ estimated as before, after each item is administered. Given a $\hat{\theta}$, the next item to be administered is chosen by determining the prestratified θ level closest to the current $\hat{\theta}$. The next unadministered item at that θ level is administered and that item number is eliminated from all remaining θ levels so that it will not again be administered. The item response is recorded, and θ is again re-estimated, with the next item administered based on the closest θ stratum to the observed $\hat{\theta}$. Unpublished data on the STMI strategy suggest that it closely approximates the performance of the full maximum information adaptive testing strategy, while minimizing computational requirements through the preprocessing of the data in the item pool.

The adaptive testing strategies described above are the major approaches that have thus far been implemented for improving the measurement process through the combined power of both item response theory and the use of adaptive item selection with variable entry and variable termination. Future developments in this area, as well as future research with these strategies, should show even more improvements in the quality and efficiency of psychological measurement through the combination of these two powerful psychometric methodologies.

References

Bejar, I. I., & Weiss, D. J. *Computer programs for scoring test data with item characteristic curve models* (Research Report 79-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1979.

Bejar, I. I., Weiss, D. J., & Gialluca, K. A. *An information comparison of conventional and adaptive tests in the measurement of classroom achievement.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

Betz, N. E., & Weiss, D. J. *An empirical study of computer-administered two-stage ability testing* (Research Report 73-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973.

Betz, N. E., & Weiss, D. J. *Simulation studies of two-stage ability testing* (Research Report 74-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1974.

Betz, N. E., & Weiss, D. J. *Empirical and simulation studies of flexilevel ability testing* (Research Report 75-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975.

Brown, J. M., & Weiss, D. J. *An adaptive testing strategy for achievement test batteries* (Research Report 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

Crichton, L. C. *Effect of error in item parameter estimates on adaptive testing.* Unpublished doctoral dissertation, University of Minnesota, 1981.

Gialluca, K. A., & Weiss, D. J. *Efficiency of an adaptive inter-subtest branching strategy in the measurement of classroom achievement* (Research Report 79-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1979.

Gorman, S. A comparison of the accuracy of Bayesian adaptive and static tests using a correction for regression. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Hambleton, R. K. (Ed.). *Applications of item response theory.* Vancouver BC: Educational Research Institute of British Columbia, in press.

Kingsbury, G. G., & Weiss, D. J. *An adaptive testing strategy for mastery decisions* (Research Report 79-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1979.

Kingsbury, G. G., & Weiss, D. J. *A comparison of adaptive, sequential, and conventional testing strategies for mastery decisions* (Research Report 80-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Kingsbury, G. G., & Weiss, D. J. *A validity comparison of adaptive and conventional strategies for mastery testing* (Research Report 81-3). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, 1981.

Larkin, K. C., & Weiss, D. J. *An empirical investigation of computer-administered pyramidal ability testing* (Research Report 74-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1974.

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

Lord, F. M. *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum, 1980.

McBride, J. R. Bandwidth, fidelity, and adaptive tests. In T. J. McConnell, Jr. (Ed.), *CAT/C 2 1975: The second conference on computer-assisted test construction.* Atlanta GA: Atlanta Public Schools, 1976.

McBride, J. R. Some properties of a Bayesian adaptive ability testing strategy. *Applied Psychological Measurement,* 1977, *1*, 121-140.

McBride, J. R., & Martin, J. T. Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press, in press.

Maurelli, V. A., & Weiss, D. J. *Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries* (Research Report 81-4). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, 1981.

Owen, R. J. *A Bayesian approach to tailored testing* (Research Report 69-92). Princeton NJ: Educational Testing Service, 1969.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association,* 1975, *70,* 351-356.

Reckase, M. D. Procedures for computerized testing. *Behavior Research Methods and Instrumentation,* 1977, *9*, 148-152.

Samejima, F. A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika,* 1973, *38*, 221-233.

Sympson, J. B., Weiss, D. J., & Ree, M. J. *Predictive validity of conventional and adaptive tests in an Air Force training environment* (AFHRL TR 81-40). Brooks Air Force Base TX: Manpower and Personnel Division, Air Force Human Resources Laboratory, 1982.

Vale, C. D., & Weiss, D. J. *A study of computer-administered stradaptive ability testing* (Research Report 75-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975. (a)

Vale, C. D., & Weiss, D. J. *A simulation study of stradaptive ability testing* (Research Report 75-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975. (b)

Vale, C. D., & Weiss, D. J. *A comparison of information functions of multiple-choice and free-response vocabulary items* (Research Report 77-2). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

Vale, C. D., & Weiss, D. J. The stratified adaptive ability test as a tool for personnel selection and placement. *TIMS Studies in the Management Sciences,* 1978, *8*, 135-151.

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

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

Acknowledgments

Preparation of this paper, and the research reported herein, were supported by Contract No. N00014-76-C-0627 NR 150-389 with the Office of Naval Research, with additional funding from the Army Research Institute, Air Force Human Resources Laboratory, Defense Advanced Research Projects Agency, and Navy Personnel Research and Development Center and by Contract No. N00014-79-C-0172 NR 150-433 from the Office of Naval Research, with additional funding from the Army Research Institute, Air Force Human Resources Laboratory, and the Air Force Office of Scientific Research.

Author's Address

Send requests for reprints or further information to David J. Weiss, Department of Psychology, N660 Elliott Hall, University of Minnesota, Minneapolis MN 55455 U.S.A.