Computerized Adaptive Testing in Instructional Settings

☐ R. Edwin Welch
Theodore W. Frick

Item response theory (IRT) has most often been used in research on computerized adaptive testing (CAT). Depending on the model used, IRT requires between 200 and 1,000 examinees for estimating item parameters. Thus, it is not practical for instructional designers to develop their own CAT based on the IRT model. Frick improved Wald's sequential probability ratio test (SPRT) by combining it with normative expert systems reasoning, referred to as an EXSPRT-based CAT. While previous studies were based on re-enactments from historical test data, the present study is the first to examine how well these adaptive methods function in a real-time testing situation. Results indicate that the EXSPRT-I significantly reduced test lengths and was highly accurate in predicting mastery. EXSPRT is apparently a viable and practical alternative to IRT for assessing mastery of instructional objectives.

☐ For many years, instructional developers and classroom teachers have used fixed-length, paper-and-pencil tests to assess student achievement of cognitive learning objectives. With the advent of interactive mainframe and minicomputers in the 1970s, it became possible to implement item response theory in the testing environment (Weiss & Kingsbury, 1984). Questions arose as how to best use this technology for computerized adaptive testing (CAT) and which decision algorithms to use.

In order for readers who are not familiar with CATs to better understand the concept, the various forms of CAT are compared and contrasted here. Two examples are provided to illustrate how a CAT functions. Described in the last section of the article is a study that examined the efficiency and accuracy of the various CATs discussed.

The goal of a CAT is to use the least amount of questions necessary to determine the level of performance of the examinee. Moreover, many CATs attempt to tailor a test to an individual's achievement level by avoiding questions that are extremely easy or difficult for a particular examinee and instead choosing questions that are closer to his or her ability level.

It should be noted that several assumptions are made in order to use CATs appropriately. First, only one learning objective can be tested at a time. This objective can be as narrow or wide as desired, but only a unidimensional

trait or ability can be assessed by a CAT. In other words, the pool of test questions used in a CAT should all be measuring the same thing. This can be confirmed empirically by a factor analysis of the item pool which results in a single strong factor. Second, local independence of items is assumed in a CAT. In other words, the probability that examinees will answer any given question correctly should not be affected by the order in which the questions are asked. To avoid violating this assumption, feedback should not be given to students during an adaptive test, nor should students be allowed to skip questions. Lastly, the questions are selected without replacement. Once a question is presented to a learner, it is not used again during that particular administration of the test to that individual.

Item response theory (IRT) has most often been used in research on computerized adaptive testing (Bunderson, Inouye, & Olson, 1989). IRT-based CATs have been shown to significantly reduce testing time without sacrificing reliability of measurement (Weiss & Kingsbury, 1984). IRT requires from 1 to 3 item parameters in order to function. These item parameters are referred to as *a*, *b*, and *c*. Parameter *a* is the discrimination index. It tells us how well item *i* discriminates among examinees at various levels of ability or achievement. Parameter *b* is the difficulty level of item *i*. Parameter *c* is the lower asymptote for an item, sometimes referred to as the "guessing" factor. The probability of a correct response to a test question is represented by an item response function. The value of the probability varies according to the ability of the examinee, the difficulty of the question, its discriminating capacity, and the lower asymptote.

Item response theory is a complex subject which often is not well understood even by measurement experts themselves. Debate continues among proponents of competing models, which include the Rasch model and classical test theory. Only a brief overview of IRT is presented here. Interested readers are referred to introductions by Hambleton, Swaminathan, and Rogers (1991), Frick (1990), and Wright (1977).

Aside from the mathematical and conceptual complexity of item response theory, a practical problem with IRT is that it requires that a lengthy history of test items be established. Before computer adaptive tests can be implemented, item parameters in IRT should be estimated based on a minimum of 200 to 1,000 students, depending on the model used (cf., Hambleton & Cook, 1983; Hambleton et al., 1991; Lord, 1983; Weiss & Kingsbury, 1984). This requirement is seldom practical for instructor-made tests, although IRT-based CATs show considerable promise for large-scale testing such as military and state-wide tests of student learning achievement. Besides the large amount of historical data needed for IRT, the complex mathematical formulas involved could easily deter its use by instructional designers and classroom instructors.

PRACTICAL ALTERNATIVES TO IRT FOR ADAPTIVE TESTING

Now that many classroom instructors and students have access to powerful desktop computers, it is possible to do adaptive testing in school computer laboratories and in corporate training settings. As an alternative to the IRT-based approach, Frick (1989, 1990) suggested the sequential probability ratio test (SPRT) developed by Wald (1947). However, a potential limitation of SPRT is that it does not explicitly take item difficulty or discrimination into account. An improvement on the SPRT was made by combining it with expert systems reasoning. This method, jointly developed by Frick (1992) and Plew (1989), became known as EXSPRT (see Frick, 1992, pp. 192–197). EXSPRT assigns a weight to each question in the item pool, thus allowing the item difficulty and discrimination to be used in decision making. A further enhancement was made to the EXSPRT method by employing an "intelligent" item selection method. This method became known as EXSPRT-I (Plew, 1989; Frick, 1992), in contrast to EXSPRT-R, in which items are chosen randomly. With EXSPRT-I, the next item for presentation is chosen based on its utility. In other words, the item selected next is the one remaining in the item pool which

best discriminates between masters and nonmasters and which is *least incompatible* with the current estimate of the examinee's achievement level.

Common to SPRT, EXSPRT-I, and EXSPRT-R is the formation of discrete likelihood ratios. Instead of assuming a continuum for measurement of achievement as does IRT, these methods instead classify examinees into discrete categories, i.e., the achievement metric is nominal, not ordinal, interval, or ratio.

In contrast, in item response theory, maximum likelihood estimation (MLE) is one method for making point estimates of an examinee's ability or achievement level, which is measured on a theta (θ) scale. Theta typically varies from -3 to $+3$. A θ value of zero indicates an average examinee ability or achievement level on the trait or objective being measured. In MLE, the pattern of right and wrong answers to specific test questions made by a particular examinee is used in conjunction with empirically derived *a priori* test item response functions in order to estimate the likelihood of that examinee's response pattern for each value of θ (i.e., at each point along the achievement continuum between -3 and $+3$). The value of θ is chosen where that likelihood is highest, and hence becomes the estimate of that examinee's ability or achievement level. There is also an error of measurement associated with that estimate of θ, which is in turn dependent on the amount of information provided by the examinee's response pattern and the respective test item response functions.

On the other hand, in EXSPRT and SPRT the likelihood of each discrete category is estimated based on an examinee's response vector and an expert system rule base (described below). The *category* of achievement is chosen where that likelihood is highest. This is a very important distinction: A category is chosen rather than a point on some continuum (i.e., percentage correct).

In IRT, a CAT ends when the variance of θ—and hence the standard error of measurement at that level of θ—becomes small enough to satisfy the decision maker. In EXSPRT and SPRT, a CAT ends when the likelihood of a category of achievement is sufficiently high to satisfy the decision-maker, following the logic of Wald (1947) for terminating sequential observations.

The apparent advantages of EXSPRT and SPRT for classroom CATs are parsimony and efficiency. Since only categorical decisions are made (compared to point estimates in IRT), an item rule base can be generated from a smaller but necessarily representative sample of examinees (compared to the large numbers required for estimation of *a*, *b*, and *c* parameters in IRT). Moreover, the logic of the EXSPRT and SPRT is relatively straightforward and simple compared to the complexity of item response theory. We contend that instructors and instructional designers will be more likely to use a form of CAT that they can understand, particularly if it can be shown to be dependable and to reduce overall time spent on student testing. Table 1 compares and contrasts the characteristics of each type of CAT examined in this article.

TABLE 1 ☐ Characteristics of the Various CAT Methods

CAT Method	*Item Selection Method*	*Amount of Historical Data Required*	*Accounts for Item Difficulty and Discrimination*	*Easy to Implement*
SPRT	Random	None	No	Easy
EXSPRT-R	Random	Approx. 50 cases	Yes	More difficult than SPRT
EXSPRT-I	Intelligent	Approx. 50 cases	Yes	More difficult than SPRT
IRT	Intelligent	200 to 1,000 cases depending on model used	Yes	Very difficult

FORMATION OF EXPERT SYSTEMS RULES IN THE SPRT

A basic logic underlies the discrete likelihood estimation procedure that is the foundation of SPRT, EXSPRT-I, and EXSPRT-R. Before discussing EXSPRT, we address the simpler approach: the sequential probability ratio test.

Suppose we have an item pool for assessing mastery of a *single* learning objective. For purposes of illustration, let us assume that past experience has shown that students who have mastered the objective (masters) score .85 on the average, and those who have not (nonmasters) score .40 (Frick, 1989). From an expert systems perspective, these conditions are expressed through the use of "If . . . then" rules (conditional probabilities from a mathematical/statistical perspective).

1. If the student is a master, then the probability of selecting a question that will be answered correctly is .85. Restated as conditional probabilities:

Rule 1A: Prob(Correct|Master) = .85
or P(C|M) = .85

Rule 1B: Prob(Incorrect|Master) = .15
or P(~C|M) = .15

Note that the mathematical notation is standard here. P(C|M) literally means the probability (P) of a correct response (C) given (|) that the student is a master (M).

2. If the student is a nonmaster, then the probability of selecting a question that will be answered correctly is .40 (Frick, 1989, pp. 96–97):

Rule 2A: Prob(Correct|Nonmaster) = .40
or P(C|N) = .40

Rule 2B: Prob(Incorrect|Nonmaster) = .60
or P(~C|N) = .60

These are the four basic types of rules needed for the execution of SPRT and its variants.

During an SPRT-based adaptive test, a randomly selected item is chosen from the item pool and presented to the student. After observing and evaluating the student's response, a probability ratio is calculated.

$$PR = \frac{P_{om}P_m^r(1-P_m)^w}{P_{on}P_n^r(1-P_n)^w}$$

where:

PR = probability ratio
P_{om} = prior probability of mastery[1]
P_{on} = prior probability of nonmastery
P_m = probability of a correct response for a master
P_n = probability of a correct response for a nonmaster
r = number of correct answers so far
w = number of wrong answers so far

The probability ratio derived is then compared to three decision rules.

SPRT Decision Rule 1.

If $PR \geq (1-\beta)/\alpha$, then choose the mastery hypothesis and discontinue observations.

SPRT Decision Rule 2.

If $PR \leq \beta/(1-\alpha)$, then choose the nonmastery hypothesis and discontinue observations.

SPRT Decision Rule 3.

If $\beta/(1-\alpha) < PR < (1-\beta)/\alpha$, then randomly select another question and continue observations (Frick, 1989).

The value of α depends on the decision-maker's willingness to erroneously call someone a master who is actually a nonmaster—the probability of making a false mastery decision—whereas β is the probability of making a false nonmastery decision (cf. Wald, 1947). For example, if α is set to .05 by the decision-maker, this means that whenever the SPRT reaches a mastery decision, that decision would be expected to be wrong in 5 percent of the cases in the long run. Or, if β is set to .001, this would mean that whenever the SPRT reaches a nonmastery decision, that decision would be expected to be wrong in 1 out of 1,000 cases.

[1]Prior probabilities of mastery and nonmastery are set to 0.5 if one has no dependable prior information about a particular student which can be used, or if one does not adopt a Bayesian perspective.

In contrast, during conventional criterion-referenced testing, the decision-maker typically sets a cut-off score, e.g., a student who answers at least 21 of 25 questions correctly is considered to be a master. What most people do not realize or make explicit is the error with which such a decision is rendered, particularly when a given student scores at or near the cut-off score. When tests are relatively short, errors in decisions about students whose test scores are near the cut-off score may occur 40 to 60 percent of the time (cf. Frick, 1990).

On the other hand, the SPRT requires the decision-maker to specify explicitly *in advance* his or her tolerance for misclassification of masters and nonmasters. What this means is that as α and β are decreased, adaptive tests become increasingly longer, all other things being equal.

In two empirical studies, Frick (1989) found that if the SPRT is used conservatively ($\alpha = \beta = .025$, when $P_m = .85$ and $P_n = .60$), approximately 20 test items were required to reach a mastery or nonmastery decision. When SPRT decisions were compared to decisions reached from much longer tests (97 and 85 items each), the error rates observed were actually lower than those expected theoretically, even when test items varied widely in their difficulty and discriminating power.

An Example of the SPRT

Before implementing SPRT, we must specify how confident we want to be with the decisions made ($1 - \alpha$ and $1 - \beta$). If we wished to be 95% confident with the overall results of SPRT and were willing to make misclassifications of masters and nonmasters equally often, we would set $\alpha = \beta = .025$.

We must also set the probability of a correct response for masters (Rule 1A) and nonmasters (Rule 2A). Rules 1A and 2A can be empirically derived or specifically chosen. Let us assume that experience with the item pool has revealed that students who are truly masters answer on the average 85% of the questions correctly (Rule 1A), and that those who are truly nonmasters answer on the average 40% of the questions correctly (Rule 2A). Thus, we have the following parameters:

$\alpha = .025$

$\beta = .025$

Rule 1A: Prob(Correct|Master) = .85
or P(C|M) = .85

Rule 1B: Prob(Incorrect|Master) = .15
or P(~C|M) = .15

Rule 2A: Prob(Correct|Nonmaster) = .40
or P(C|N) = .40

Rule 2B: Prob(Incorrect|Nonmaster) = .60
or P(~C|N) = .60

Question 1. A question is randomly selected from the item pool and presented to the student, who answers it *incorrectly.* The probabilities for each alternative using Bayes' Theorem would be derived as shown below. Note that Rules 1B and 2B are used in the second column of probabilities, since the question was answered incorrectly.

Alternative	*Prior Probability of Alternative*		*Probability Incorrect|Alternative*		*Joint Probability*		*Posterior Probability*
Mastery	.5	×	.15	=	.075/Sum	=	.20
Nonmastery	.5	×	.60	=	.300/Sum	=	.80
				Sum =	.375		

As can be seen, the probability of nonmastery after one question is .80. Using Wald's SPRT formula, the probability ratio would be figured and the decision rules applied. In actuality, Rules 1 and 2 give us an upper and lower boundary for Rule 3 to be true. For this example, as long as the probability ratio falls between 39 $[(1 - \beta)/\alpha)]$ and .025641 $[\beta/(1 - \alpha)]$,

we continue testing. If the probability ratio becomes less than or equal to .025641, we choose nonmastery, whereas if the probability ratio becomes greater than or equal to 39, we choose mastery and stop testing. Here is observation 1 using the SPRT formula:

$$PR = \frac{(.5)\ .85^0\ (1 - .85)^1}{(.5)\ .40^0\ (1 - .40)^1} = \frac{.075}{.300} = .25$$

Since .25 is between .025641 and 39, we continue testing. Note that in the Bayesian representation above, the ratio of the posterior probabilities (.20/.80) is equal the SPRT probability ratio (.25). These two forms of representation are numerically equivalent, but the Bayesian representation may help some readers better understand the technique. Thus, both modes will be used in the examples provided.

Question 2. Another question is randomly chosen from the item pool and the student answers it *correctly*. The posterior probabilities for Question 1 now become the prior probabilities for Question 2. It is very important to note that Rules 1A and 2A are used in the second column of probabilities this time, since the question was answered correctly.

Alternative	*Prior Probability of Alternative*		*Probability Correct\Alternative*		*Joint Probability*		*Posterior Probability*
Mastery	.20	×	.85	=	.170/Sum	=	.347
Nonmastery	.80	×	.40	=	.320/Sum	=	.653
				Sum =	.490		

$$PR = \frac{(.5)\ .85^1\ (1 - .85)^1}{(.5)\ .40^1\ (1 - .40)^1} = \frac{.06375}{.120} = .53125$$

Since the probability ratio is still between .025641 and 39, we continue testing. Again, note above that .347/.653 = .531. (Also note that the probability ratios may not be exactly the same due to rounding errors, since to save space the probabilities in the Bayesian representation are rounded to the nearest one-thousandth.)

Question 3. Another item is randomly selected from the item bank and the student answers it *incorrectly*. Again, the posterior probabilities of the last question become the prior probabilities for this question. Note that Rules 1B and 2B are used in the second column of probabilities.

Alternative	*Prior Probability of Alternative*		*Probability Incorrect\Alternative*		*Joint Probability*		*Posterior Probability*
Mastery	.347	×	.15	=	.052/Sum	=	.117
Nonmastery	.653	×	.60	=	.392/Sum	=	.883
				Sum =	.444		

At this point in the test, the odds are about 8 to 1 in favor of nonmastery. Nevertheless, as can be seen from the probability ratio, a decision cannot yet be confidently made.

$$PR = \frac{(.5)\ .85^1\ (1 - .85)^2}{(.5)\ .40^1\ (1 - .40)^2}$$

$$= \frac{.0095625}{.072} = .1328125$$

Question 4. Another question is randomly chosen from the remaining items in the item bank. The student whom we are trying to classify again answers *incorrectly*. Updating occurs as follows:

Alternative	*Prior Probability of Alternative*		*Probability Incorrect|Alternative*		*Joint Probability*		*Posterior Probability*
Mastery	.117	×	.15	=	.018/Sum	=	.032
Nonmastery	.883	×	.60	=	.530/Sum	=	.968
				Sum =	.548		

It appears that we can choose nonmastery and stop here, since there appears to be almost a 97% probability this is true. However, the probability ratio reveals that we still cannot decide at the *a priori* confidence level ($\alpha = \beta = .025$).

$$PR = \frac{(.5)\ .85^1\ (1 - .85)^3}{(.5)\ .40^1\ (1 - .40)^3}$$

$$= \frac{.001434375}{.0432} = .033203125$$

An alternative way of viewing the SPRT decision rules is to examine the posterior probabilities in the Bayesian representation. Whenever the posterior probability of mastery or nonmastery becomes equal to or greater than $1 - \alpha$ or $1 - \beta$ (.975 here), we can stop the test.

Question 5. Another question is randomly selected from the item pool, and the student again answers *incorrectly*. Updating occurs as follows:

Alternative	*Prior Probability of Alternative*		*Probability Incorrect|Alternative*		*Joint Probability*		*Posterior Probability*
Mastery	.032	×	.15	=	.005/Sum	=	.008
Nonmastery	.968	×	.60	=	.581/Sum	=	.992
				Sum =	.586		

$$PR = \frac{(.5)\ .85^1\ (1 - .85)^4}{(.5)\ .40^1\ (1 - .40)^4}$$

$$= \frac{.000215156}{.02592} = .008300781$$

The probability ratio is now less than .025641; thus, we choose nonmastery and stop testing. Alternatively, note that, at .992, the Bayesian posterior probability of the nonmastery alternative being true is now greater than .975.

The above example demonstrates the underlying logic for all three SPRT methods examined by this study.

FORMATION OF EXPERT SYSTEMS RULES IN THE EXSPRT

The difference between EXSPRT and SPRT is that a different set of weights is created for each question in the item pool (or different rules if a normative expert systems perspective is taken; Frick, 1992). Alternatively, all questions are treated equally by SPRT (e.g., a master has an 85% chance at answering any question correctly). In EXSPRT, that probability varies from item to item. Weights are assigned using a set of four rules for each item *i* in the pool.

Rule i.1: If the examinee is a *master* and item *i* is selected, then the probability of a correct response is $P(C_i|M)$.

Rule i.2: If the examinee is a *master* and item *i* is selected, then the probability of an incorrect response is $P(\sim C_i|M)$.

Rule i.3: If the examinee is a *nonmaster* and item *i* is selected, then the probability of a correct response is $P(C_i|N)$.

Rule i.4: If the examinee is a *nonmaster* and item *i* is selected, then the probability of an incorrect response is $P(\sim C_i|N)$.

The probabilities for each of the items are created by using historical data collected by administering the entire item pool to a representative group of approximately 50 or more examinees. Note that it is very important that sufficient numbers of masters and nonmasters—roughly 25 in each category—are represented in this sample, and that the examinees are typical of those who are expected to take the CATs in the future. Although more research needs to be done in this area, Frick (1992) observed that when there were fewer than 25 examinees in each category, the accuracy of classification in that category was less than expected from the *a priori* error rates.

A decision is now made as to the cut-off score for separating masters and nonmasters (e.g., .85). Using this cut-off score, we divide the examinees into two groups, masters and nonmasters, based on their total test scores. For each item in the mastery group, the following formulas are applied to determine the probabilities of correct and incorrect responses:

$$P(C_i|M) = (\#r_{im} + 1)/(\#r_{im} + \#w_{im} + 2) \quad [1]$$

$$P(\sim C_i|M) = 1 - P(C_i|M) \quad [2]$$

where:

$\#r_{im}$ = Number of persons in the mastery group who answered the item correctly.

$\#w_{im}$ = Number of persons in the mastery group who answered the item incorrectly.

The same is done for the nonmastery group:

$$P(C_i|N) = (\#r_{in} + 1)/(\#r_{in} + \#w_{in} + 2) \quad [3]$$

$$P(\sim C_i|N) = 1 - P(C_i|N) \quad [4]$$

The decision formula for EXSPRT is as follows:

$$LR = \frac{P_{om} \prod_{i=1}^{K} P(C_i|M)^s [1 - P(C_i|M)]^f}{P_{on} \prod_{i=1}^{K} P(C_i|N)^s [1 - P(C_i|N)]^f} \quad [5]$$

where:

LR = likelihood ratio

P_{om} = prior probability that the examinee is a master

P_{on} = prior probability that the examinee is a nonmaster

and:

s = 1, f = 0 if item *i* is answered correctly,

or:

s = 0, f = 1 if item *i* is answered incorrectly,

s = 0, f = 0 if item *i* has not been administered.

The rules for termination of the testing situation are the same as above for Wald's SPRT.

An Example of the EXSPRT

To save space, in this example we use empirical data from four test items: 1, 23, 38, and 63. After giving these items to a representative group of examinees, assume that the rule quadruplets in Table 2 are formed when .85 is used as a cut-off score for separating masters and nonmasters (see formulas 1–4 above).

TABLE 2 ☐ Rule Quadruplets for Items 1, 23, 38, and 63

	MASTER		NONMASTER	
	Correct *Rule i.1*	*Incorrect* *Rule i.2*	*Correct* *Rule i.3*	*Incorrect* *Rule i.4*
Item 1	.92	.08	.47	.53
Item 23	.81	.19	.24	.76
Item 38	.98	.02	.86	.14
Item 63	.89	.11	.65	.35

Observation 1. Suppose that we are using EXSPRT-R and we want to be 95 percent confident in our classification of a student as a master or nonmaster. We begin by randomly selecting an item from the pool of items for measuring mastery of the instructional objective. Say that we administer randomly selected item 63 to this student, about whom we know nothing, and that the student answers it *incorrectly.* Thus, rules 63.2 and 63.4 are relevant, as shown below:

Alternative	*Prior Probability of Alternative*		*Probability #63 Incorrect|Alternative*		*Joint Probability*		*Posterior Probability*
Mastery	.500	×	.11	=	.055/Sum	=	.239
Nonmastery	.500	×	.35	=	.175/Sum	=	.761
				Sum =	.230		

Note that we are using the same Bayesian representation as we did above for the SPRT. What is different are the numbers that we insert in the second column of probabilities. These numbers depend on the rule quadruplet associated with the item administered *and* upon whether the student answers this question correctly or incorrectly. At this time the odds that the nonmastery alternative is true, compared to mastery, are about 3 to 1 (.761/.239). If we want to be 97.5 percent confident in choosing one of the alternatives, then we must continue the test.

Observation 2. We next randomly select item 23 from the item pool and administer it to the same student, who answers it *correctly.* Thus, rules 23.1 and 23.3 are relevant, as shown below. Note also that the above posterior probabilitites become our new prior probabilities.

Alternative	*Prior Probability of Alternative*		*Probability #23 Correct|Alternative*		*Joint Probability*		*Posterior Probability*
Mastery	.239	×	.81	=	.194/Sum	=	.515
Nonmastery	.761	×	.24	=	.183/Sum	=	.485
				Sum =	.377		

At this point, the two alternatives are about equally likely, given that item 63 was missed and item 23 was answered correctly.

Observation 3. We next randomly select item 1 from the item pool and administer it to the student, who answers it *incorrectly.* Thus, rules 1.2 and 1.4 are relevant, as shown below:

Alternative	*Prior Probability of Alternative*		*Probability #01 Incorrect|Alternative*		*Joint Probability*		*Posterior Probability*
Mastery	.515	×	.08	=	.041/Sum	=	.138
Nonmastery	.485	×	.53	=	.257/Sum	=	.862
				Sum =	.298		

At this point, the odds are slightly more than 6 to 1 in favor of nonmastery.

Observation 4. We next randomly select item 38 from the item pool and administer it to the student, who again answers *incorrectly.* Thus, rules 38.2 and 38.4 are relevant, as shown below:

Alternative	Prior Probability of Alternative		Probability #38 Incorrect|Alternative		Joint Probability		Posterior Probability
Mastery	.138	×	.02	=	.003/Sum	=	.024
Nonmastery	.862	×	.14	=	.121/Sum	=	.976
				Sum =	.124		

After the fourth observation, the posterior probability of nonmastery has exceeded 0.975, given that item 63 was answered incorrectly, item 23 correctly, item 1 incorrectly, and item 38 incorrectly. If we had set $\alpha = \beta = .025$ in advance of the test, then it would end at this time with a nonmastery decision, since the nonmastery alternative has a probability greater than $1 - \beta$.

The reader should note that if instead of the Bayesian representation we had used formula 5 presented earlier and Wald's three decision rules, the result would have been computationally equivalent and we would have reached the same decision. Formula 5 expresses in a succinct mathematical way the products of the conditional probabilities based on the observed response pattern. Given the response pattern observed in this example, the likelihood ratio would be:

$$LR = \frac{\underset{\downarrow}{\text{Prior Prob.}}\ \underset{\downarrow}{\text{Item \#63 wrong}}\ \underset{\downarrow}{\text{Item \#23 right}}\ \underset{\downarrow}{\text{Item \#01 wrong}}\ \underset{\downarrow}{\text{Item \#38 wrong}}}{.5(.89^0 \times .11^1)(.81^1 \times .19^0)(.92^0 \times .08^1)(.98^0 \times .02^1)}$$

$$.5(.65^0 \times .35^1)(.24^1 \times .76^0)(.47^0 \times .53^1)(.86^0 \times .14^1)$$

$$= \frac{.00007128}{.00311640} = .0228725$$

Since 0.0228725 is less than .025641 ($\beta/(1 - \alpha)$), we would conclude that this student is a nonmaster and stop the test.

RESEARCH ON SPRT AND EXSPRT

The above formulas and decision rules are based on Frick's (1991, 1992) work. There are two versions of EXSPRT: EXSPRT-R, which randomly selects the items to present to the examinee, and EXSPRT-I, which intelligently selects the items based on the examinee's responses and the properties of the items that maximize discrimination between masters and nonmasters and compatibility with the examinee's achievement level (see Frick, 1990, 1992, for further details).

These methods (SPRT, EXSPRT-R, and EXSPRT-I) can be used in mastery testing situations as well as in instructional situations. By using them in embedded testing situations, one can reduce the time needed for testing and allow more time for instructional activities. If these methods are shown to be effective and efficient, the way in which classroom testing is conducted could be radically changed. They could also offer the teacher more time for instruction by reducing the amount of time required for testing.

Frick (1992) conducted empirical studies of the SPRT, EXSPRT-R, and EXSPRT-I which indicated that the three methods were quite accurate in predicting students' mastery or nonmastery of a learning objective. A limitation of these studies is that they were done retroactively through re-enactments, rather than by controlling the test in real time (Plew, 1989). The test re-enactments were carried out by using test responses from former administrations in which the examinees answered all of the items on the test.[2] The algorithm being examined selected an item, and then the evaluation of the student's response to that item

[2]It is important to note that the original tests were administered by computer, that items were randomly selected without replacement until the pools were exhausted, and that students were not permitted to skip items or change prior answers later in the test. Re-enactments with the adaptive methods were performed on those test data, not data from conventional paper-and-pencil tests in which students can change their answers and determine the order in which they answer questions. If the latter kind of data had been used, the assumption of local independence required by the adaptive testing methods would be more likely to be violated, since students could search for clues to answer a question from other questions on the test.

(correct or incorrect) was retrieved from the computer file and entered into the algorithm as if it had been done in real time. The reenactment continued until a decision could be made by the CAT. The decision was then compared to the decision reached when the test was originally given by administering all items. The disadvantage of this procedure is that test conditions did not truly reflect a real adaptive testing situation.

As yet, no studies have tested the accuracy and efficiency of the EXSPRT-R and EXSPRT-I in real-time testing, i.e., when actually controlling the test. Hence, there are a number of issues that call for further study: How would these methods hold up in a "real world" situation? How accurate are the methods when the student knows that the length of the test will depend on how well he or she does? Are some methods more efficient than others?

METHODOLOGY OF THE PRESENT STUDY

In the present study, the item bank was one developed for a graduate course on computers in education. Subjects were volunteers from graduate and undergraduate courses at Indiana University. The 38 subjects were randomly assigned to the two groups. Twenty subjects were given an adaptive test governed by the EXSPRT-R method (random item selection) and 18 subjects were given an adaptive test governed by the EXSPRT-I method (intelligent item selection). The 85-item pool contained a variety of true/false, multiple-choice, and fill-in-the-blank questions. Subjects in the EXSPRT-I group had a mean score of 60.78 on the test, with a standard deviation of 17.80. The EXSPRT-R group had a mean score of 61.6, with a standard deviation of 13.92.

Both tests operated in the following manner. The student was informed that she or he would be taking two tests, an adaptive test, where the length of the test would depend on his or her performance, and a traditional test. The student was then asked to complete both tests. Actually only one test was given, but it appeared to the student that two tests were administered.

The first test continued until a mastery or nonmastery decision could be made using EXSPRT. When a decision was made as to mastery or nonmastery, the student was informed. The computer also recorded in a file the decision reached and when it was made. The second test consisted of the remaining items, so that the student would answer all 85 questions. The item selection strategy used for the first test (i.e., random or intelligent selection) continued throughout the second test. Remaining items were given so that a comparison could be made between the adaptive decision and the conventional test decision.

A second feature of the test was that the SPRT algorithm was executed by the computer parallel to the EXSPRT-R (without informing the student). When the SPRT method arrived at a decision, the computer recorded when the decision occurred. The SPRT decision did not affect the administration of the test. This was done so that information could be collected on how well the SPRT performed compared to EXSPRT and the traditional test. This design was used so that it would not be necessary to administer the test to a third group. The SPRT algorithm could not run parallel to EXSPRT-I because the fact that the items were chosen intelligently would violate the assumption of random selection for SPRT.

Because we wanted to be 98% confident about the adaptive decisions, the alpha and beta levels for all algorithms were set at .01. The cut-off levels for the SPRT method were derived from the historical data collected (n = 185).[3] Rather than using the traditional levels of .60 for nonmasters and .85 for masters, the empirically derived levels used were .63 for nonmasters and .90 for masters. The prior probability of being a master was set at .50 for all methods upon entrance into the test. EXSPRT rules were also based on these past 185 cases, as well as Rasch estimates of item diffculty (cf. Wright, 1977).

After data were collected on the 38 students in the present study, comparisons were made of EXSPRT-I, EXSPRT-R, and SPRT with reenactments using the IRT-based procedure developed by Weiss and Kingsbury (1984), called

[3]These historical data were obtained from earlier studies using the same test with graduates and undergraduates in beginning computer courses in education (Frick, 1989; Powell, 1991).

Adaptive Mastery Testing (AMT). During the re-enactments on the previously recorded item responses, the order in which the items were chosen was based on their maximum information search procedure, and θ estimates were Bayesian (following Owen, 1975). AMT test re-enactments ended when mastery or nonmastery could be decided at the .98 confidence interval, which is comparable to setting α and β to .01 in the EXSPRT and SPRT discrete likelihood procedure.

Finally, θ and its variance were estimated from responses to *all* items on the test for each examinee using both maximum likelihood and Bayesian approaches. The maximum likelihood estimate and Bayesian procedures resulted in virtually identical θs and standard errors for each examinee on the whole 85-item test. Also, IRT-based total test decisions (mastery, nonmastery, and no decision) were in 100% agreement with those reached from a conventional proportion correct metric with a .85 cut-off score. Both methods of making decisions from the total test item pool used a .98 confidence interval. If that interval did not contain the cut-off (1.422 on the θ scale, .85 on the proportion correct metric), then a mastery or nonmastery decision was made accordingly. If the confidence interval *did* include the respective cut-off, then no decision could be reached with the .98 confidence interval.

RESULTS

CAT Test Lengths

Average test lengths and standard deviations for the various CATs are reported in Table 3.

Independent *t* tests were conducted for the pairwise comparisons of adaptive methods in Group 1 versus Group 2 (e.g., EXSPRT-I in Group 1 vs. EXSPRT-R in Group 2), whereas correlated *t* tests were conducted for pairwise comparisons of means from different adaptive methods *within* groups (e.g., EXSPRT-I vs. AMT in Group 1, and EXSPRT-R vs. SPRT in Group 2). To be conservative, the type I error rate was set to .005 for individual pairwise contrasts, since these were nonorthogonal, *a posteriori* comparisons. Thus, the overall Type I error rate for the 9 contrasts performed was equal to or less than $1 - (1 - .005)^9$, which is slightly less than .05 for the overall experiment (see Kirk, 1982). The AMT method in Group 1 was compared to EXSPRT-I only, whereas all other pairwise contrasts were tested.

TABLE 3 ☐ Average Numbers of Items Required to Make Mastery and Nonmastery Decisions by the Various CAT Methods

		Mean	*SD*
Group 1 (n = 18)	EXSPRT-I	13.4	15.1
	AMT	17.7	25.5
Group 2 (n = 20)	EXSPRT-R	26.9	20.0
	SPRT	23.4	13.0
	AMT	20.0	28.0

EXSPRT-I tests were about half as long as EXSPRT-R tests, and the difference was statistically significant. There were no significant differences found among any of the remaining pairwise comparisons of mean test lengths in the various adaptive testing methods.

The reader is reminded that the AMT was done through re-enactments, not in real time as were EXSPRT-I, EXSPRT-R, and SPRT. This is admittedly a somewhat messy design, but the reader should also remember that the main purpose was to see how well EXSPRT would work in actual classroom testing situations, rather than through re-enactments as had been studied in the past by Frick (1989, 1992), Luk (1991), Plew (1989), and Powell (1991). EXSPRT-I test lengths tended to be significantly shorter than EXSPRT-R tests in most of the re-enactment studies. EXSPRT-R and AMT CATs have tended to be similar in length. With some tests, EXSPRT-I has been significantly shorter than AMT, and with other tests it has not. Lengths of adaptive tests depend on a number of factors in addition to the CAT method used (see Frick, 1990, 1992).

The fact that EXSPRT-I was significantly shorter than EXSPRT-R in the present study is consistent with most past studies. Most importantly, the pattern observed in the present study, in which these adaptive methods were in actual classroom settings and in real time, is consistent with patterns observed in past studies which used re-enactments of adaptive methods. This consistency is reas-

suring in that it adds credence to the predictive validity of the re-enactment studies. Those past studies are analogous to wind-tunnel experiments with model airplanes conducted to predict actual flight behavior of real airplanes.

CAT Decision Accuracies

Questions that arise in adaptive testing are: How trustworthy are the decisions reached by an adaptive test? How do those decisions compare with decisions that would be arrived at by a more lengthy (and presumably more dependable) measurement procedure? In the present situation, we can compare the decisions reached by the CATs with those reached by the total test item pool, since the latter is presumably more dependable.

Another question is: How should we make decisions on the basis of the total test results? There are three different ways we could make decisions based on the *entire* 85-item pool:

1. discrete likelihood estimation with Type I and II errors (α and β errors, which are used in EXSPRT);
2. conventional proportion correct with a confidence interval based on a standard error of measurement; and
3. IRT θ estimation with a standard error of measurement based on test information at that θ level.

We can then compare the decisions reached by the different CAT approaches to see how well they agree with these three different methods of classification based on total test results. Since in this study the IRT and conventional total test decisions agreed perfectly, they are combined in Table 4 (3 columns at right) for purposes of comparison with the CAT decisions.

In Table 4, the agreement between total test decisions and the various CAT methods is reported. This is a set of eight 3 × 3 contingency tables, where the 3 main diagonal elements in each represent agreement, and the 6 off-diagonal elements indicate disagreement.

The 3 columns of frequencies at the left of Table 4 indicate agreement between those test decisions reached by the discrete likelihood method on the 85-item test, and the test decisions reached by each of the CATs. For example, in Group 1 there were 6 total-test mastery decisions out of 18 cases. In all 6 of these cases, the EXSPRT-I had also reached a mastery decision—perfect agreement. The total test indicated 8 nonmasters. The EXSPRT-I also concluded nonmastery in all 8 of those

TABLE 4 ☐ Agreement between Computer Adaptive Test Decisions and Decisions Based on Total Test Results

		Decisions Based on Total Test Results					
		DISCRETE LIKELIHOOD ($\alpha = \beta = .01$)			IRT AND CONVENTIONAL (.98 C. I.)		
CAT DECISIONS		M	NM	ND	M	NM	ND
EXSPRT-I	M	6	0	1	3	0	4
(n = 18)	NM	0	8	3	0	8	3
	ND	0	0	0	0	0	0
EXSPRT-R	M	9	0	0	1	0	8
(n = 20)	NM	0	10	0	0	10	0
	ND	0	0	1	0	1	0
SPRT	M	8	0	0	1	0	7
(Same *S's*	NM	1	10	1	0	11	1
as EXSPRT-R)	ND	0	0	0	0	0	0
AMT							
(Same *S's*	M	11	0	0	4	0	7
as EXSPRT-I & -R	NM	2	18	3	0	19	4
(n = 38)	ND	2	0	2	0	0	4

Key: M = Mastery, NM = Nonmastery, ND = No Decision

cases—also perfect agreement. On the total test under the discrete method, the remaining 4 cases in Group 1 could not be classified clearly as masters or nonmasters (no decision). The EXSPRT-I classified one of those persons incorrectly as a master, and 3 as nonmasters, for a total of 4 disagreements.

The IRT/Conventional methods of making total test decisions are compared to the CAT decisions at the right of table 4.[4] The same 18 cases in Group 1 are compared to EXSPRT-I when this total test decision method is used. Notice that the IRT/Conventional total test could only make clear mastery decisions in 3 cases, compared to 6 when the discrete likelihood approach was used on the whole test. These 3 cases were also classified as masters by EXSPRT-I. The IRT/Conventional method made 8 nonmastery decisions, as did the discrete likelihood method, and these 8 were also classified as nonmasters by the EXSPRT-I method. The remaining 7 cases could not be classified clearly as masters or nonmasters by the IRT/Conventional method on the whole test. The EXSPRT-I classified 4 of these no-decision cases as masters and 3 as nonmasters. Notice that there were a total of 7 disagreements between IRT/Conventional total test decisions and EXSPRT-I CAT decisions, in contrast to a total of 4 disagreements when the discrete likelihood method was used for total test decisions.

The remainder of Table 4 can be similarly interpreted. For example, when EXSPRT-R is compared to the discrete likelihood total test decision method, there was only one disagreement in the 20 cases in Group 2. However, when compared to the IRT/Conventional approach, there were a total of 9 disagreements.

What may be confusing in Table 4 is that the same 20 cases in Group 2 that were exposed to the EXSPRT-R CAT method could also be categorized by the SPRT method. Finally, all cases in both groups ($n = 18 + 20 = 38$) were retroactively categorized by the IRT-based AMT CAT method by means of re-enactments based on the recorded right and wrong answers to test questions when originally taken under either EXSPRT-I or EXSPRT-R conditions. For example, under the discrete likelihood total test approach, the 5 no-decisions ($0 + 3 + 2$), when compared to the AMT CAT, are the same 5 no-decisions when compared to the EXSPRT-I and EXSPRT-R ($1 + 3 + 0 + 0 + 0 + 1$).

As can be seen from Table 4, it was not possible under the IRT/Conventional total test approach to make clear mastery and nonmastery decisions at the .98 confidence interval in many of the cases ($4 + 3 + 8 = 7 + 4 + 4 = 15$ no-decisions out of 38 cases). On the other hand, there were one-third as many ambiguous outcomes in the discrete case (5 out of 38) using EXSPRT likelihood estimation procedures (see formula 5). How can this be? If we think in terms of a *continuum* of achievement, a cut-off and a confidence interval, it would appear that we can be less certain more often (given the same test results), compared to making decisions when using *discrete categories* as in EXSPRT. The trade-off appears to be this: While the EXSPRT is less "precise" in that the choices are mastery versus nonmastery as opposed to estimating some point on a continuum, most of the time we can be *more* certain about our conclusions when given the same test results.

It can be seen that, when clear mastery or nonmastery decisions could be reached by the total test, all of the CAT methods tended to reach the same decisions. As mentioned above, both the IRT and conventional 85-item tests were unable to reach clear decisions with a .98 confidence interval in 15 out of 38 cases, whereas this occurred only 5 times when discrete likelihoods were based on the entire 85-item pool. AMT adaptive test decisions tended to agree less well with total test decisions based on discrete likelihoods ($2 + 2 + 3 + 2 = 9$ disagreements out of 38).

DISCUSSION

As can be seen from examining the data, all computerized adaptive methods studied—SPRT, EXSPRT-I, EXSPRT-R, and AMT—reduced the amount of time needed for the

[4] In the conventional proportion correct metric, the cut-off score is .85. This corresponds to a $\theta_c = 1.422$, as determined from the test response function for the one-parameter IRT or Rasch model.

test. Between one-third and one-fifth of the item pool was needed to reach a CAT decision. EXSPRT-I test lengths were significantly shorter than EXSPRT-R. None of the other CAT methods resulted in significantly different test lengths.

It should also be noted that SPRT, EXSPRT-I, and EXSPRT-R methods resulted in categorizations that agreed almost perfectly with total test results when decisions were not forced (i.e., could be made with α and β set to .01). The SPRT method made one false decision when compared to the total test discrete likelihood, but it should be noted that it was made in the conservative direction: calling a master a nonmaster. In the EXSPRT-I group there were 4 no-decisions on the total test, whereas the CAT reached mastery and nonmastery decisions on those 4 cases. Why did this happen? The most obvious explanation is that these were borderline cases, that is, students who were not clearly classifiable as masters or nonmasters on the total test. Frick (1990) demonstrated that more decision errors will occur when the distribution of examinees is clustered near the cut-off, regardless of the adaptive testing methodology used.

It is noteworthy that the SPRT performed about as well as both AMT and EXSPRT approaches. This is significant in that the SPRT requires no prior test administrations in order to estimate item parameters or a statistical rule base, as do the AMT and EXSPRT. Of course, this observation must be tempered by the facts that it occurred with a highly reliable test (Cronbach α = .94); that the decision error rates were set very low (.01 for false mastery and nonmastery decisions); and that the mastery and nonmastery levels were not too far apart (.90 versus .63).

WHICH CAT METHOD IS BEST?

Since SPRT, EXSPRT, and AMT all appear to be viable computerized adaptive testing methodologies, on what basis can one choose among them?

Perhaps the first consideration is whether categorical classifications are desired, or whether it is important to estimate a student's achievement level at a precise point on some continuum. Since SPRT and EXSPRT are not intended for precise estimates of achievement levels, it would appear that the IRT-based adaptive mastery testing method (AMT) would be the best choice for the latter situation. However, there is a very steep price to pay for using the AMT. If the three-parameter model is used, then a minimum of 1,000 or more examinees need to be tested *in advance* for item parameter estimation. This requirement is likely to be viewed as highly impractical by classroom instructors, trainers, and developers of mastery tests used in computer-based instruction, and by instructional designers in general.

If categorical classifications such as mastery and nonmastery are sufficient for the decision-maker, then the EXSPRT appears to be the best choice, since it requires a much smaller *a priori* sample for developing an item rule base which is used for discrete likelihood estimation during an adaptive test. The price to pay here is that the initial sample must be *representative* of students for whom the CAT is eventually intended, which means roughly 25 or more examinees per discrete category. If efficiency is important as well, then the EXSPRT-I would appear to be the best choice for categorical classifications, since it results in shorter tests than EXSPRT-R. On the other hand, if it is important that the items selected be representative of the *content* in the total pool, then EXSPRT-R would be the best choice, since items are chosen at random. EXSPRT-I may result in inadequate content coverage (as does the AMT) because items are selected based on their statistical properties and not on their content.

A compromise between content coverage and efficiency may be achieved by combining EXSPRT-R and EXSPRT-I. For example, a minimum test length can be specified (e.g., 15 questions) and the EXSPRT-R method used. Then, if no decision can be reached with EXSPRT-R, the CAT can switch to the EXSPRT-I method. This mixed approach also has the advantage that it is more difficult for students to "cheat" by memorizing answers to a small set of highly discriminating test items that are likely to be chosen early by the EXSPRT-I approach. It is possible for students to memo-

rize the first few items since the system always begins the exam with the same question, given a prior probability of mastery set at .50. If the student answers the first question correctly, the system always selects the same second question. If the second question is answered correctly, the system always selects the same third question. This pattern continues, so students could discover the pattern and learn the answers to only the first 5 to 8 questions and be classified a master. The validity of EXSPRT-I would be compromised in this "cheating" situation.

Finally, if no advance data are available on item properties and categorical decisions will suffice, then the SPRT is a good choice if used conservatively. That is, mastery and nonmastery levels should not be set too far apart (.85 and .60 work reasonably well), and *a priori* error rates must be kept small (.025 or less).

It is important to remember that EXSPRT and SPRT are appropriate for testing mastery of a single learning objective. If multiple objectives are of concern, then separate CATs should be conducted for each objective. The overall outcome is not a single score but a list of objectives that have and have not been mastered by a given student. This can be made transparent to students so that they perceive one test rather than a series of subtests, and feedback on test results can be suspended until decisions are reached on all objectives being assessed.

In conclusion, computerized adaptive tests can significantly reduce testing time in corporate education and training settings. Reduced testing time with no loss of reliability means decreased training costs, all other things being equal. In public school classrooms, CATs have the potential to support self-paced mastery learning and individualized instruction, in contrast to traditional group-paced instruction in lock-step grade levels. □

R. Edwin Welch is with the Learning Support Center, Taylor University, in Upland, Indiana. Theodore W. Frick is with the Department of Instructional Systems Technology of the School of Education at Indiana University, Bloomington.

REFERENCES

Bunderson, V., Inouye, D., & Olson, J. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational measurement*. New York: Macmillan.

Frick, T. W. (1989). Bayesian adaptation during computer-based tests and computer-guided practice exercises. *Journal of Educational Computing Research, 5*(1), 89–114.

Frick, T. W. (1990). A comparison of three decision models for adapting the length of computer-based mastery tests. *Journal of Educational Computing Research, 6*(4), 479–513.

Frick, T. W. (1991). *A comparison of an expert systems approach to computerized adaptive testing and an item response theory model*. Paper presented at the annual conference of the Association for Educational Communications and Technology, Orlando, Florida.

Frick, T. W. (1992). Computerized adaptive mastery tests as expert systems. *Journal of Educational Computing Research, 8*(2), 187–213.

Hambleton, R., & Cook, L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. Weiss (Ed.), *New horizons in testing* (pp. 31–50). New York: Academic Press.

Hambleton, R., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Kirk, R. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed, pp. 101–105). Belmont, CA: Brooks/Cole.

Lord, F. (1983). Small *n* justifies Rasch model. In D. Weiss (Ed.), *New horizons in testing* (pp. 52–62). New York: Academic Press.

Luk, H.-K. (1991). *An empirical comparison of an expert systems approach and an IRT approach to computer-based adaptive mastery testing*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois.

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

Plew, G. T. (1989). *A comparison of major adaptive testing strategies and an expert systems approach*. Unpublished doctoral dissertation, Indiana University, Bloomington.

Powell, E. (1991). *Test anxiety and test performance under computerized adaptive testing methods*. Unpublished doctoral dissertation, Indiana University, Bloomington.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Weiss, D., & Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361–375.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*(2), 97–116.