DOCUMENT RESUME

ED 307 319 TM 013 459

AUTHOR Frick, Theodore W.; And Others
TITLE EXSPRT: An Expert Systems Approach to Computer-Based Adaptive Testing.
PUB DATE Mar 89
NOTE 37p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; College Students; *Computer Assisted Testing; *Expert Systems; Higher Education; *Latent Trait Theory; *Mastery Tests; Probability; Sequential Approach
IDENTIFIERS *EXSPRT

ABSTRACT

Expert systems can be used to aid decision making. A computerized adaptive test (CAT) is one kind of expert system, although it is not commonly recognized as such. A new approach, termed EXSPRT, was devised that combines expert systems reasoning and sequential probability ratio test stopping rules. EXSPRT-R uses random selection of test items, whereas EXSPRT-I incorporates an intelligent selection procedure based on item utility coefficients. These two new methods are compared to the traditional SPRT and to an adaptive mastery testing (AMT) approach based on item response theory (IRT). Three empirical studies using different tests and examinees were conducted. Study 1 included samples of 25 and 50 current or former graduate students who took the Digital Authoring Language Test; Study 2 included samples of 25, 50, 75, and 100 students in an introductory graduate-level course; and Study 3 included 333 college freshmen and sophomores. Results indicate that the EXSPRT-I is more efficient or as efficient as is the AMT model. When the distribution of examinees was not clustered near the mastery cutoff, all four methods made accurate mastery classifications. Although further research is needed, the EXSPRT initially appears to be a strong alternative to IRT-based adaptive testing when categorical decisions about examinees are desired. The EXSPRT is less complex conceptually and mathematically; and it appears to require many fewer examinees to empirically establish a rule base, when compared to the large numbers required to estimate parameters for item response functions in the IRT model. (TJH)

EXSPRT: AN EXPERT SYSTEMS APPROACH TO COMPUTER-BASED ADAPTIVE TESTING

Theodore W. Frick

Department of Instructional Systems Technology
School of Education, Indiana University

G. Thomas Plew

Computer Information Systems Department
Indiana Wesleyan University

Hing-Kwan Luk

Doctoral Candidate
School of Education, Indiana University

Paper Presented at the Annual Conference of the
American Educational Research Association
San Francisco

March, 1989

ABSTRACT

Expert systems can be used to aid decision making. A computerized adaptive test (CAT) is one kind of expert system, though not commonly recognized as such. A new approach, termed EXSPRT, was devised that combines expert systems reasoning and sequential probability ratio test stopping rules. EXSPRT-R uses random selection of test items, whereas EXSPRT-I incorporates an intelligent selection procedure based on item utility coefficients.

These two new methods are compared to the traditional SPRT and to an adaptive mastery testing (AMT) approach based on item response theory (IRT). Three empirical studies with different tests and examinees were carried out. Results indicated that the EXSPRT-I is more efficient or as efficient as the AMT model. When the distribution of examinees was not clustered near the mastery cut-off, all four methods made accurate mastery classifications.

Although further research is needed, the EXSPRT initially appears to be a strong alternative to IRT-based adaptive testing when categorical decisions about examinees are desired. The EXSPRT is less complex conceptually and mathematically, and it appears to require many fewer examinees to establish empirically a rule base, when compared to the large numbers required to estimate parameters for item response functions in the IRT model.

THEORETICAL ISSUES

An Overview of Expert Systems

One of the more practical results from extant research in artificial intelligence is the application of expert systems reasoning to aid in decision making or problem solving. Expert systems have been developed, for example, to help physicians identify types of bacterial infections, to aid investor decisions on buying and selling stock, for aid in assembling components of computer systems, for making decisions about where to drill for oil, for assisting underwriters in making insurance policies, and for diagnosing causes of equipment failures to help repairpersons (c.f., Winston & Prendergast, 1984).

An expert system consists of a set of production rules or frames, often called a 'knowledge base'. The name, 'expert system', was coined because a knowledge base is typically constructed by interviewing one or more experts in some domain of knowledge. An attempt is made to capture their reasoning processes, when they solve problems in that knowledge domain, in the form of "If..., then..." rules. For example, in MYCIN, a famous early expert system for diagnosing bacterial infections, one of the rules is:

> IF 1) the gram stain of the organism is negative, and
> 2) the morphology of the organism is rod, and
> 3) the aerobicity of the organism is anaerobic,
> THEN there is suggestive evidence (.7) that the identity of the organism is Bacteroides. (Davis, 1984, p. 34)

This particular rule is one of 400 to 500 such rules that comprise the MYCIN knowledge base. A computer program, called an 'inference engine', uses this rule set as data to help physicians identify unknown bacteria. The program makes inferences by using both the rule set and specific answers to questions it asks the physician about properties of the current situation (e.g., patient symptoms, white blood cell count, and other lab test results). MYCIN has been shown to be more accurate in its identifications of bacteria than typical practicing physicians, particularly in identifying those bacteria which are rarely observed.

Expert systems are not usually viewed as replacements for human decision makers, but as aids or tools for such persons. Expert systems obviously cannot perform in areas not covered by the knowledge base. Furthermore, decisions reached by expert systems can be no better than the accuracy of the knowledge or rules that comprise the database.

In education and training, expert systems principles have been applied mostly in intelligent tutoring systems (Kearsley, 1987; Sleeman & Brown, 1982). As an example, GUIDON was later developed from MYCIN in an attempt to teach physicians how to identify different kinds of bacteria (Clancey, 1987).

Similarities between Expert Systems and Adaptive Tests

One efficient and empirically validated approach to computerized adaptive testing (CAT) is based on item response theory (e.g., Weiss & Kingsbury, 1984). An adaptive test is no longer than necessary to

obtain a satisfactory estimate of an examinee's ability, and items are selected which are close to his or her estimated ability level. For example, if a person misses a question, a somewhat easier question is next asked. On the other hand, if a question is answered correctly, then a slightly more difficult question is subsequently selected. A computerized adaptive test does not waste time administering questions that are too hard or too easy for a particular individual. Adaptive tests tend to be shorter than conventional fixed-length tests and the results are as reliable if not more so (e.g., Weiss & Kingsbury, 1984).

Expert systems and adaptive computer-based tests have many properties in common:

1. The rule base for a CAT is a set of item characteristic curves estimated from prior test administrations. That is, each item characteristic curve is a compact way of saying, "If the examinee ability level is X, and item Y is asked, then the probability of a correct response is predicted to be Z."

2. Both expert systems and CATs use inference engines that are often Bayesian or Bayesian-like. Even if rules do not have probabilities (or confidence factors) associated with them, they can still be treated as a special Bayesian case where associated probabilities are either one or zero (c.f., Heines, 1983).

3. The goal of an expert system is to choose from a number of alternatives (e.g., causes of equipment failure) using the rule base and answers to questions it selects and asks of a particular user. The goal of a CAT is to estimate an examinee's achievement or ability level with enough precision to make a decision such as pass/fail or a grade classification using a rule base of item characteristic curves and answers to questions it selects and gives to examinees.

4. An expert system selects which questions it asks by using forward or backward chaining and the rule base. A CAT can select questions on the basis of the amount of information they provide, depending on the ability of an examinee. For example, Weiss and Kingsbury (1984) use a maximum information search and selection (MISS) procedure.

Thus, although not widely recognized at this time, an adaptive testing system is one type of an expert system. The first author realized this when developing computer code for an expert system, having already developed code for Bayesian decision methodologies and a computer-based testing system.

On hindsight, expert systems and adaptive tests have much in common. Yet in the research literature it appears that these two threads of development have been almost entirely independent. One camp has grown out of an artificial intelligence movement and the other from a psychological testing and measurement perspective. A recent computer search of numerous bibliographic databases only turned up thirteen articles where the terms, 'expert systems' or 'artificial intelligence' and 'adaptive' or 'computer' and 'testing' or 'test' were used as

descriptors. The two camps not only use different language to describe their activities, but also tend to publish in different journals and attend different conferences.

The Development of EXSPRT

A problem with the IRT-based approach to adaptive testing faced by many practitioners, however, is that a relatively large number of examinees must be tested in advance in order to estimate accurately item parameters of difficulty, discrimination, and lower asymptotes (200 to 1000 depending on the model used and the number of items in a pool). Furthermore, proponents of the Rasch model (one-parameter IRT model) have indicated that there is no valid way of estimating item discrimination and lower asymptotes for the two- and three-parameter models without imposing arbitrary constraints (c.f., Wright, 1977).

The first author has previously investigated the predictive validity of the sequential probability ratio test (SPRT) for making mastery decisions, where the lengths of tests were adapted according to student performance (Frick, 1989). He demonstrated that mastery decisions reached with the SPRT, when used conservatively, agreed highly with those based on total test results. Nonetheless, the SPRT does not explicitly take into account variability in item difficulty, discrimination or chances of guessing as does the three-parameter IRT model. Moreover, items are selected randomly in the SPRT, rather than on the basis of their characteristics and estimated examinee ability or achievement level as in the MISS procedure.

Is there some middle ground between the relatively simplistic SPRT decision model and the relatively sophisticated IRT-based approach? When considering the problem from an expert systems perspective, a solution became apparent. Instead of considering a continuum of alternatives, as is the case in IRT-based CAT, it was hypothesized that if the goal of an adaptive testing system is to choose between a few discrete alternatives (e.g., mastery or nonmastery; grades of A, B, C, etc.), then it should be possible to develop a satisfactory rule base from a smaller sample of examinee test data--compared to the IRT model.

An Example of Expert Systems Reasoning during Computer-Based Testing

Suppose that we have developed a pool of test items which match a particular instructional objective and that our goal is to decide whether or not a particular student has mastered that objective (e.g., Mager, 1973). Suppose further that our aim is to administer no more questions than are necessary to reach a mastery or nonmastery decision, and yet we want to be highly confident in our decision.

First, we need to construct a rule base. There are various ways that this could be done, but let us use a straightforward empirical approach. We obtain a sample of students representative of those who would be likely to be learning the instructional objective, who are learning, and who have learned (e.g., third grade students and

multiplication of two-digit numbers; college freshman taking a course in probability theory; graduate students in education learning how computers work).

Next, we give the whole test to this sample of students. We must then decide on a cut-off score for determining mastery and nonmastery. Suppose we are satisfied that anyone who scores 85 percent or higher on the test has minimally mastered the instructional objective being tested. This allows us to sort students into a mastery group and a nonmastery group. We now construct a rule set for each test item. For example (these are fictitious data, used for illustration only):

Rule 1.1. If the student is a master and item #1 is selected, then the estimated probability of a correct response is .92 (the proportion of masters in the sample who successfully answered the question).

Rule 1.2. If the student is a master and item #1 is selected, then the probability of an incorrect response is .08.

Rule 1.3. If the student is a nonmaster and item #1 is selected, then the probability of a correct response is .47.

Rule 1.4. If the student is a nonmaster and item #1 is selected, then the probability of an incorrect response is .53.

A quadruplet of such rules can be constructed for each item on the test, based on the proportions of masters and nonmasters, respectively, who answered the item correctly and incorrectly. We will assume that our student sample is large enough and representative enough of the population of those students of interest that we have sufficient confidence in the data used to derive the rules. The rules can be more conveniently summarized in tabular format. Some hypothetical data are provided below:

Item	$P(C\|M)$	$P(\neg C\|M)$	$P(C\|\neg M)$	$P(\neg C\|\neg M)$
1	.92	.08	.47	.53
.				
.				
23	.81	.19	.24	.76
.				
38	.98	.02	.86	.14
.				
47	.75	.25	.21	.79
.				
63	.89	.11	.65	.35
.				
76	.82	.18	.51	.49
.				
.				
etc.	.	.	.	.

where: $P(C|M)$ = Probability of a correct response, given a master,

$P(\neg C|M)$ = Probability of an incorrect response, given a master,

$P(C|\neg M)$ = Probability of a correct response, given a nonmaster,

and $P(\neg C|\neg M)$ = Probability of an incorrect response, given a nonmaster.

Now, we will use the above rule base to make a decision about the mastery status of a particular student about whom we presently know nothing with respect to her mastery or nonmastery of the instructional objective assessed by the test items. Therefore, our prior probabilities of mastery and nonmastery are equal to .50 for this student.

Observation 1. We randomly select an item from the pool (#63). We administer it to this student, who answers it incorrectly. Our expert systems inference engine will reason according to Bayes' Theorem as follows (c.f., Schmitt, 1969):

Alternative	Prior Prob. of Alternative		Prob. $\neg C$, Item 63 | Alt.		Joint Prob.	Posterior Prob.	
Mastery	.50	X	.11	=	.055	/Sum =	.239
Nonmastery	.50	X	.35	=	.175	/Sum =	.761
				Sum =	.230		

The prior probability of each alternative is multiplied by the probability of the observation, given that the alternative is true. The probability of an incorrect response by a master for item #63 is .11, which when multiplied by .50, yields a joint probability of .055. Similarly, the probability of an incorrect response by a nonmaster for item #63 is .35, and when multiplied by the prior probability of nonmastery (.50), results in a joint probability of .175. The joint probabilities are normalized by dividing each by the sum of the joint probabilities. After this observation, the posterior probability for mastery is now .055/.23 = .239. The posterior probability for the nonmastery alternative is .175/.23 = .761. At this point, the nonmastery alternative is about 3 times more likely than the mastery alternative.

Observation 2. We continue testing by selecting another item at random from the pool. We give item #23 to the student under consideration, who answers it correctly. We update as follows, only this time we use the most recent posterior probabilities as our new priors:

Alternative	Prior Prob. of Alternative		Prob. C, Item 23 | Alt.		Joint Prob.	Posterior Prob.	
Mastery	.239	X	.81	=	.194	/Sum =	.515
Nonmastery	.761	X	.24	=	.183	/Sum =	.485
				Sum =	.377		

This time in the third column we use the probability of a correct response to item #23, given each alternative. The odds of nonmastery to mastery have now become about equal, given the two observations made thus far.

Observation 3. This time we select at random item #1, which the student answers incorrectly. We update, as before, using the most recent posterior probabilities as our new priors.

Alternative	Prior Prob. of Alternative		Prob. ¬C, Item 1 | Alt.		Joint Prob.	Posterior Prob.	
Mastery	.515	X	.08	=	.041	/Sum =	.138
Nonmastery	.485	X	.53	=	.257	/Sum =	.862
				Sum =	.298		

The odds are a little over 6 to 1 in favor of nonmastery at this point.

Observation 4. We select another item, #38, at random, which our student also misses.

Alternative	Prior Prob. of Alternative		Prob. ¬C, Item 38 | Alt.		Joint Prob.	Posterior Prob.	
Mastery	.138	X	.02	=	.003	/Sum =	.024
Nonmastery	.862	X	.14	=	.121	/Sum =	.976
				Sum =	.124		

After the fourth observation, the posterior probability of the nonmastery alternative is about .98, roughly 40 times as great as the probability that the mastery alternative is true. Should we stop the test now? If so, on what basis should we do so? It appears that it is extremely likely that this particular student is a nonmaster, given just four test items, selected at random from the pool, given the response pattern [wrong, right, wrong, wrong], and given the Bayesian reasoning methods we have been employing with the item rulebase.

The decision as to when to terminate the test depends on how willing we are to make false mastery and false nonmastery decisions (type I and II errors). A type I error, α, is the probability of choosing mastery when the nonmastery alternative is really true. A type II error, β, is the probability of choosing the nonmastery alternative when the mastery alternative is really true. We will adopt

the rules developed by Abraham Wald (1947) for the Sequential Probability Ratio Test (SPRT):

> Stopping Rule 1. If the ratio of the posterior probabilities of the two alternatives (mastery vs. nonmastery) derived from Bayes' Theorem is greater than or equal to $(1 - \beta)/\alpha$, then stop making observations and choose the first alternative (mastery in this context).
>
> Stopping Rule 2. If the ratio of the posterior probabilities of the two alternatives (mastery vs. nonmastery) derived from Bayes' Theorem is less than or equal to $\beta/(1 - \alpha)$, then take no more observations and choose the second alternative (nonmastery).
>
> Continuation Rule. If the ratio of the posterior probabilities of the two alternatives is neither greater than or equal to $(1 - \beta)/\alpha$, nor less than or equal to $\beta/(1 - \alpha)$, then take a new observation, update the posterior probabilities using Bayes' Theorem, and apply the three rules once again.

Suppose that we set $\alpha = \beta = .05$. The threshold for the first rule is $(1 - .05)/.05 = .95/.05 = 19$. The threshold for the second rule is $.05/(1 - .05) = .053$. During the above observations, the first three result in posterior probability ratios which fall between the two thresholds. However, the ratio of the posterior probabilities after the fourth observation is .024/.976 = .025, which is less than .053, the threshold for stopping rule 2. Therefore, we would conclude that the present student is a nonmaster, knowing that we would tend to be wrong about 5 percent of the time, since we set β *a priori* at .05.

In summary, this example illustrates how data-based decision making could be made by a computer-based testing system, using expert systems reasoning--in particular, Bayesian reasoning--and rule quadruplets which were constructed from data derived from testing a representative sample of students who are masters and nonmasters. In effect, this approach combines Bayesian reasoning with empirically based rules and SPRT stopping rules. The new approach, which combines both expert systems and SPRT principles, is called EXSPRT-R, since items are selected randomly.

Intelligent Item Selection: EXSPRT-I

The second author was not satisfied with EXSPRT-R, since it did not use information about test items in the selection process. This stimulated the joint development of an item selection procedure that is modeled after basic principles used by Weiss and Kingsbury in the MISS (maximum information search and selection) procedure. Though the principles are comparable, the mathematical approaches are quite different.

In the EXSPRT-I (i.e., with "intelligent" item selection), the reasoning is as follows:

Item discrimination. If we are trying to choose between mastery or nonmastery alternatives, then an item is more discriminating when the difference between probabilities of correct responses by masters and nonmasters is greater. For example, if the probabilities of a correct response to item #5 are .90 for masters and .25 for nonmasters, then item #5 is very discriminating (difference = .65). On the other hand, if the probability of a correct response to item #53 is .85 for masters and .75 for nonmasters, then this item is much less discriminating (difference = .10). Or if the probability of a correct response to item #12 is .60 for masters and .80 for nonmasters, then such an item is negatively discriminating (difference = -.20).

Thus, the discrimination index for item i is defined:

$$D_i = P(C_i|M) - P(C_i|N) \qquad (1)$$

where $P(C_i|M)$ = estimate of the probability of a correct response to item i by a master;

and $P(C_i|N)$ = estimate of the probability of a correct response to item i by a nonmaster.

The estimates of probabilities of correct responses to items by masters and nonmasters are determined as follows:

1. Give the pool of test items to a representative group of examinees, about half of whom are expected to be masters and half nonmasters--i.e., for whom you expect a wide range of scores on the test.

2. Choose a mastery cut-off score (e.g., .85).

3. Divide the original group into a mastery group and nonmastery group based on their total test scores and the mastery cut-off.

4. For each item in the mastery group, estimate the probabilities of correct and incorrect responses by the following formulas (see Schmitt, 1969):

$$P(C_i|M) = (\#r_{im} + 1)/(\#r_{im} + \#w_{im} + 2) \qquad (2.1)$$

$$P(\neg C_i|M) = 1 - P(C_i|M) \qquad (2.2)$$

where $\#r_{im}$ = number of persons in the mastery group who answered the item correctly;

and $\#w_{im}$ = number of persons in the mastery group who missed the item.

5. Do likewise for the nonmastery group for each item:

$$P(C_i|N) = (\#r_{in} + 1)/(\#r_{in} + \#w_{in} + 2) \qquad (3.1)$$

$$P(\neg C_i|N) = 1 - P(C_i|N) \qquad (3.2)$$

Note that the estimates of these probabilities of correct responses to items by masters and nonmasters will never be one or zero. This means that, in the EXSPRT Bayesian updating process during the administration of a test to an examinee, the probabilities of the mastery and nonmastery alternatives will never be zero or one, though these extremes may be closely approached.

Item/examinee incompatibility. Not only do we want to select highly discriminating items, but also we want to select items that are matched to an examinee's estimated achievement or ability level. In theory, we gain little additional information by administering items which are very easy or very hard for a given individual. Better items would be those which a person has a 50/50 chance of answering correctly--i.e., which are very close to her or his achievement level. For example, if an examinee's achievement level is estimated to be .80 (on a scale from zero to one), then a good item would be one that was answered incorrectly by 80 percent of the examinees in the item parameter estimation sample [$P(C_i) = .20$ for masters and nonmasters combined].

Thus, the item/examinee incompatibility index is defined for each item:

$$I_{i,j} = \text{abs}((1 - P(C_i)) - E(\Phi_j)) \qquad (4)$$

where $$E(\Phi_j) = (\#r_j + 1)/(\#r_j + \#w_j + 2) \qquad (5)$$

and $$P(C_i) = (\#r_i + 1)/(\#r_i + \#w_i + 2) \qquad (6)$$

Note that $\#r_j$ and $\#w_j$ are the numbers of questions answered correctly and incorrectly, respectively, thus far in the test by the current examinee. Note also that the estimate of $P(C_i)$ is based on the total number of persons in the parameter estimation sample for item i, irrespective of mastery status. Thus, $\#r_i$ is the number of persons who answered item i correctly and $\#w_i$ is the number who answered it incorrectly. Finally, note that the item/examinee incompatibility index is based on the absolute value of the difference between the estimate of the probability of an incorrect response to the item and the estimate of the current examinee's achievement level (proportion correct metric).

Item utility. As a test proceeds, item utilities are re-calculated for all items remaining in the pool, in order to select and administer a new one that now has the most utility for an examinee:

$$U_{i,j} = D_i/(I_{i,j} + \delta) \qquad (7)$$

where δ = some arbitrary small constant (e.g., .0000001), to prevent division by zero in case $I_{i,j} = 0$.[1]

Thus, each utility value is simply the ratio of the discrimination of item i and its incompatibility with person j's achievement level. The item that is selected next in the EXSPRT-I (intelligent selection) is the remaining one with the greatest utility at that point for that particular examinee. This means that the item selected next is the one which discriminates best between masters and nonmasters and which is least incompatible with the current estimate of that examinee's achievement level. Note that item utilities change during a test, depending on an examinee's performance which affects the estimate of his/her achievement level in the item/examinee incompatibility index. In effect, the EXSPRT-I is comparable to the two-parameter item response theory model (IRT--see below) in that both item discrimination and item difficulty are considered in the item selection process.

IRT-Based Adaptive Mastery Testing

In classical item analysis, estimation of item difficulty and discrimination is heavily dependent on the sample of examinees who have taken the test. For example, if we administered the test only to persons who were masters of the instructional objective, then the item analysis would reveal that most of the items appear to be quite easy (low difficulty). On the other hand, if the sample consisted of only nonmasters, the analysis would reveal that items tended to be quite high in difficulty.

Such considerations are addressed in item response theory (Lord & Novick, 1968). In essence, it is assumed that there is a relationship between the probability of a correct response to an item and an underlying (or latent) trait, and these item characteristics somehow enter into this relationship. The 'trait' is what we are trying to indirectly measure by eliciting responses to test questions (e.g., mastery of a particular instructional objective). Persons who have more of this trait should be more likely to answer a question correctly than people who have less of this trait. Furthermore, some items may be useful for sorting out individuals who are high in this trait, but these same items would tell us practically nothing about people who have little of the trait we are trying to measure.

The relationship between the probability of a correct response to a test item and the underlying trait is assumed to follow a particular kind of mathematical function, called a logistic cumulative density function:

---

[1]Alternatively, δ_i could be considered as some kind of "guessing" factor for the item. However, this will not be considered in the present paper.

$$\frac{\exp(X)}{1 + \exp(X)} \qquad (8)$$

where exp(X) means raising the mathematical constant, e (=2.71828...), to the Xth power. This function is somewhat S-shaped in form (called an ogive). On a particular test item, there will usually be a range of examinees who are high in the trait and who all are very likely to answer it correctly (i.e., prob(C|High Range) ≈ 1.0.). This is the upper asymptote of the function. On the other hand, there will usually be a range of examinees who are low in the trait and whose probability of a correct response is at or near the chances of guessing (i.e., prob(C|Low Range) ≈ chances of guessing). This is referred to as the lower asymptote of the function. In between these two extreme ranges, there will be a middle range of examinees for whom the probability of a correct response will ideally vary linearly with the so-called amount of the latent trait (X) they possess--i.e., prob(C|Middle Range) = mX + B, where m is the slope of the line (Δprob / ΔX) and B is a constant. In other words, those who are at the higher end of the middle range should have a greater probability of a correct response than those who are at the lower end of the middle range.

This relationship between the probability of a correct response to a particular item, R_i, and an underlying trait, θ, is depicted by an item characteristic curve (ICC), later referred to as an item response function (IRF) by Lord. The formula for this function is:

$$\text{prob}(R_i|\theta) = c_i + (1 - c_i)\frac{\exp(L)}{1 + \exp(L)} \qquad (9)$$

where:

$L = 1.7a_i(\theta - b_i)$,
a_i = discriminatory power of item i,
b_i = difficulty level of item i,
and c_i = lower asymptote of item i (chances of guessing).

Theta, θ, can theoretically vary between zero and a very large value but it is typically scaled as a standardized variable with a mean of zero and a variance of one (i.e., z-scores). The parameters a_i, b_i, and c_i are fixed for a given item, i. These parameters are estimated from empirical data, having administered the item to a very large number of examinees. The scaling factor of 1.7 is used so that the logistic ogive will approximate a normal ogive.

The a_i parameter affects how steep the ICC is in the middle portion, the b_i parameter affects the horizontal displacement of the middle portion of the curve, and the c_i parameter affects the vertical displacement of the lower portion of the curve. This formulation of an item characteristic curve is known as the three-parameter model. In order to obtain fairly accurate estimates of the a_i, b_i, and c_i parameters, it is recommended that approximately 1000 individuals be tested with the item pool (c.f., Hambleton & Cook, 1983).

If we do not consider chances of guessing as part of an item characteristic, then c_i becomes zero for all items. The probability of a correct response for a given θ is then simply the ratio of [exp(L)/(1 + exp(L))]. This is known as the two-parameter model, involving difficulty level and discriminatory power only. All lower asymptotes of ICC's are zero in this model. A minimum of 500 examinees is recommended for estimating the a_i and b_i parameters for an item pool.

If we consider all items to be equally discriminating and also do not consider chances of guessing, this is equivalent to setting a_i to a constant for all items and c_i to zero as above. This is known as a one-parameter model, equivalent to the Rasch model. All lower asymptotes are zero, and the middle portions of each ICC all have the same slope. The only thing that will differ is the horizontal displacement of the ICC's depending on the values of b_i's. A minimum of 200 examinees is recommended for estimation of the b_i parameters in the one-parameter model (though see Lord, 1983).

Item information. As discussed earlier, not all items will provide us with useful information for all individuals. For example, if we are trying to discriminate between two or more examinees who have little of the trait being measured, then highly difficult test items will provide no useful information about these low-in-the-trait individuals, since they would be expected to answer correctly such items at a chance level only. It would be more desirable to choose items for these low-in-the-trait individuals which more closely match their ability--if our goal is to more precisely estimate the amount of the trait they possess. In other words, we want to find test items which have difficulty levels near the theta levels of the persons in question. Moreover, we want to find items which are highly discriminating and have a low probability of being answered correctly by chance for a range of difficulty levels that match the range of theta values of concern. These items will provide us with the most amount of information for the individuals in question--i.e., will allow us to sort out these individuals more precisely in terms of the amount of the trait being measured.

Brown and Weiss (1977) incorporated this concept of item information in selecting test items during achievement testing. Having some current estimate of an examinee's θ level, a computer program searches the pool of remaining items for the item which has the most information for this value of θ. This procedure is termed, 'maximum information search and selection' (MISS). The item which will have the most information is the one which has a difficulty level closely matching the current estimate of θ, and at the same time has the highest discriminatory power and lowest probability of being answered correctly by simply guessing.

Kingsbury and Weiss (1983) calculate information for item i for a given value of θ in their adaptive mastery testing (AMT) model using Birnbaum's formula:

$$I_i(\theta) = \frac{(1 - c_i)d^2a_i^2[LPD]^2}{[LPD] + c_i[LCP]^2}, \quad (10)$$

where

a_i, b_i and c_i, and L are defined as above,

$d = 1.7$ = constant scaling factor,

$$LPD = \frac{\exp(L)}{[\ 1 + \exp(L)\ ]^2} = \text{logistic probability density,} \quad (11)$$

and

$$LCP = \frac{\exp(-L)}{1 + \exp(-L)} = \text{logistic cumulative probability.} \quad (12)$$

This part of item response theory, while complex, is fairly straightforward, assuming we have trustworthy $\underline{a}_i$, $\underline{b}_i$, and $\underline{c}_i$ parameter estimates. The catch is that the probabilities of a correct response to an item and the information values of an item vary as a function of $\underline{\theta}$, the underlying trait that we cannot directly measure or observe for some examinee. How can we estimate the value of $\underline{\theta}$ for an individual during an adaptive test?

Bayesian posterior θ estimation. If we begin a test with a prior estimate of an examinee's $\underline{\theta}$ level and its variance, and if we give an item to an examinee and know whether it was answered correctly or not, then we can determine the posterior distribution of $\underline{\theta}$ and the variance of that distribution by using formulas developed by Owen (1975). The posterior estimate of theta given a correct response to the current item is:

$$E(\underline{\theta}|C) = M_o + \left(\frac{[\ (1 - \underline{c}_i)V_o\ /\ W\ ]\ [\ gau(X)\ ]}{Y} \right), \quad (13)$$

where M_o = prior $\underline{\theta}$ estimate,

V_o = prior variance of $\underline{\theta}$ estimate,

$$W = [\ (1/\underline{a}_i^2) + V_o\ ]^{1/2}, \quad (14)$$

$$X = (\ \underline{b}_i - M_o\ )\ /\ W \quad (15)$$

$$gau(X) = \left(\frac{1}{[2\pi]^{1/2}} \right) [\ \exp(\ -(X^2)/2\ )\ ], \quad (16)$$

$$Y = \underline{c}_i + (1 - \underline{c}_i)[\ logist(\ -1.7X\ )\ ], \quad (17)$$

$$\text{and} \qquad \text{logist}(Z) = \frac{\exp(Z)}{1 + \exp(Z)} . \tag{18}$$

The estimate of $\underline{\theta}$ given an incorrect response to item $\underline{i}$ is defined:

$$E(\underline{\theta}|\neg C) = M_{\theta} - \left(\frac{[\, V_{\theta} \,/\, W \,]\, [\, \text{gau}(X) \,]}{\text{logist}(\, 1.7X \,)} \right). \tag{19}$$

Although these formulas for Bayesian updating of the estimate of $\underline{\theta}$ are complicated, the principle is simple: If the examinee correctly answers a question, then the prior estimate of $\underline{\theta}$ is incremented by an amount that is related to characteristics of the item and to the prior variance of $\underline{\theta}$. If the examinee misses the question, then the prior estimate of $\underline{\theta}$ is decremented.

On the other hand, the Bayesian updating of the variance of $\underline{\theta}$ is multiplicative, not additive or subtractive. The variance of $\underline{\theta}$ will tend to decrease as more items are administered. The estimate of the variance of $\underline{\theta}$, given a correct response is defined:

$$V(\underline{\theta}|C) = V_{\theta}\, [1 - \left(\frac{(1 - \underline{c}_i)[\text{gau}(X)]}{U}\right)\left(\frac{\left(\frac{(1 - \underline{c}_i)[\text{gau}(X)]}{Y}\right) - X}{Y}\right)], \tag{20}$$

where:

$$U = 1 + [1/(\underline{a}_i^{\,2} V_{\theta})]. \tag{21}$$

The estimate of the $\underline{\theta}$ variance, given an incorrect response is defined:

$$V(\underline{\theta}|\neg C) = V_{\theta}\, [1 - \left(\frac{\text{gau}(X)}{U}\right)\left(\frac{\left(\frac{\text{gau}(X)}{\text{logist}(\, 1.7X \,)}\right) + X}{\text{logist}(\, 1.7X \,)}\right)]. \tag{22}$$

Another observation is that the "guessing" factor, $\underline{c}_i$, enters into to the updating process of both $\underline{\theta}$ and its variance when a question is answered correctly, but the $\underline{c}_i$-related terms drop out if the question is answered incorrectly.

Finally, the posterior estimate of $\underline{\theta}$ and its variance become the new priors after another test item is administered. Then new posterior estimates of $\underline{\theta}$ and its variance are estimated, and so on, until the posterior variance of $\underline{\theta}$ becomes small enough. How small that needs to be is discussed next.

The AMT stopping rule. We have still not addressed the basis for ending a mastery test under the AMT model and reaching a decision. Weiss and Kingsbury (1984) recommend using a test response function (more commonly referred to as a test characteristic curve, TCC) as follows:

$$\text{prob}(C_k|\theta) = \{ \Sigma_i [c_i + (1 - c_i) \frac{\exp(L)}{1 + \exp(L)}] \} / n \qquad (24)$$

where n = number of items in the total pool,

and $L = 1.7a_i(\theta - b_i)$, as before.

The TCC can be seen as an average of all the ICC's (see formula (9)). Normally, we think of a mastery level in terms of a proportion of correct answers (e.g., .85). However, in the AMT we are dealing with a θ metric. The problem is to convert a proportion correct as a mastery level to a corresponding theta cut-off, θ_c. This can be accomplished through use of the TCC by simply going up the TCC curve until a point is reached where the probability of a correct response is equal to the proportion correct wanted for the mastery level.

Once θ_c is determined, then after each test item is administered and a new posterior θ and variance estimate is calculated, we simply check to see whether or not the .95 confidence interval contains θ_c.

$$\text{If } [\, E(\theta) - 1.96(V(\theta)^{1/2}) \,] > \theta_c, \text{ then choose mastery.} \qquad (25.1)$$

$$\text{If } [\, E(\theta) + 1.96(V(\theta)^{1/2}) \,] < \theta_c, \text{ choose nonmastery.} \qquad (25.2)$$

That is, if the confidence interval does not contain θ_c, then we stop the test and choose mastery if the lower bound of the interval is above θ_c, or choose nonmastery if the upper bound is below θ_c.

$$\text{If } [\, E(\theta) - 1.96(V(\theta)^{1/2}) \,] \leq \theta_c \leq [\, E(\theta) + 1.96(V(\theta)^{1/2}) \,] \text{ then continue testing.} \qquad (25.3)$$

Thus, if confidence interval does contain θ_c we continue the test by using the MISS technique to choose the next item. Note that choosing a Bayesian confidence interval of .95 is the same as setting $\alpha = \beta = .025$ (see above discussion of the SPRT and EXSPRT).

An alternative to Owen's Bayesian method of estimating θ is maximum likelihood estimation, assuming that an examinee has answered at least one question correctly and one incorrectly. Since we will compare the AMT to other Bayesian approaches (SPRT, EXSPRT-R, EXSPRT-I), maximum likelihood estimation is not discussed here.

Basic Questions Addressed

Three empirical studies were conducted to compare IRT-based adaptive mastery testing, SPRT, EXSPRT-R and EXSPRT-I approaches. Of major concern was the accuracy with which each adaptive model could predict decisions based on total test scores. Does each adaptive

method make mastery and nonmastery decisions with no more errors than would be expected by a priori error rates? Second, how efficient is each adaptive method in terms of average test lengths for mastery and nonmastery decisions? Are any of the methods more efficient than others?

FIRST STUDY

Digital Authoring Language Test

A computer-based test on the structure and syntax of the Digital Authoring Language was constructed, consisting of 97 items, and referred to as the DAL test. This test was comprised of multiple-choice, binary-choice, and short-answer questions. The test was highly reliable (Cronbach α = .98). The DAL test was also very long, usually taking between 60 and 90 minutes to complete, and it was very difficult for most examinees (mean score = 63.2 percent correct, S.D. = 24.6).

Examinees

The persons who took the DAL test were mostly either current or former graduate students in a course on computer-assisted instruction taught by the first author. Those students who were currently enrolled at the time took the DAL test twice, once about mid-way through the course when they had some knowledge of DAL--which they were required to learn for developing CAI programs--and once near the end of the course when they were expected to be fairly proficient in DAL. The remainder of the examinees took the DAL test once. Since the test was long and difficult, no one was asked to take the test who did not have some knowledge of DAL or other authoring languages.

Test Administration

The DAL test was individually administered by the Indiana Testing System (Frick, 1986). As an examinee sat at a computer terminal, items were selected at random without replacement from the total item pool until all items were administered. Students were not allowed to change previous answers to questions, nor was feedback given during the test. Upon completion of test, complete data records were stored in a database, including the actual sequence in which items were randomly administered to a student, response time, literal response to each item, and the item scoring (correct or incorrect). Examinees were informed of their total test scores at the end of the test. There were a total of 53 administrations of the DAL test in the first study.

Experimental Methods

The basic procedure was to re-enact each test, using actual examinee responses in the database, for each of the four adaptive

methodologies: 1) IRT-based adaptive mastery testing (AMT--with maximum information search and selection [MISS]), 2) sequential probability ratio test (SPRT), 3) EXSPRT-R (random selection of items), and 4) EXSPRT-I (intelligent selection of items--see above descriptions).

Item parameter estimation. Two random samples of examinees were used to estimate item parameters (n = 25 and n = 50), the latter containing the former. This was done to see if increasing the sample size used for parameter estimation would result in fewer decision errors in the four methods. Due to the relatively small sample sizes, the one-parameter AMT model was used--i.e., only b_i estimates were obtained for the two samples using program BICAL (Mead, Wright & Bell, 1979). For the EXSPRT-R and EXSPRT-I, the rule base for each parameter estimation sample was constructed using formulas (2.1), (2.2), (3.1) and (3.2). The mastery cut-off was set at 72.5 percent, half way between the established .85 mastery level and .60 nonmastery level used in an earlier study of the SPRT only (Frick, 1989). In the current study, however, the mastery and nonmastery levels for the SPRT were established empirically from the .725 cut-off and the two parameter estimation samples. The mean proportion correct for masters was used as the mastery level and the mean proportion correct for nonmasters was used as the nonmastery level in each sample. In effect, the SPRT was treated just like the EXSPRT-R, except that the rule quadruplets for all items were the same in the SPRT, based on the sample means for masters and nonmasters, respectively.

Test re-enactments. Once the parameter estimation samples were chosen, then the latter two authors independently wrote computer programs in two different languages (Pascal and DAL) to construct the rulebases for the EXSPRT, and to carry out the four different adaptive testing methods on the same 53 sets of test administrations. This was done to reduce the possibility of error in coding these rather complex methodologies, especially the AMT model. When results did not agree, as was occasionally the case, this helped to identify and ameliorate errors in coding. The one difference that was not correctable was traced to the precision of arithmetic in DAL and Pascal on a VAX minicomputer.

We discovered that on occasion the MISS procedure in the two programs would begin to select different items in the AMT model after 15 to 20 items had been retroactively "administered" to an examinee. This occurred because the updating of the estimate of θ and its variance, and in turn the item information estimates for that θ estimate, would tend to differ very slightly in the two code versions as a test progressed. Consequently, the MISS procedure would occasionally pick a different item in the two different versions when estimates of item information were very close for two or more items remaining in the pool. From that point on in a test, different item sequences were observed. The average AMT test length in the DAL version tended to be about one item shorter, compared to the Pascal version, but the decisions reached were the same with one exception.

These discrepancies do point out a problem inherent in the IRT-based approach, which contains numerous multiplications, divisions, and exponentials (see formulas (9) to (25)). Very small errors due to rounding or differences in precision of arithmetic can magnify themselves rather quickly. This problem was not observed with the EXSPRT-I, EXSPRT-R, or SPRT--other than differences in the millionth's decimal place when computing probability ratios.

1. AMT re-enactment. The mastery cut-off was converted to $\underline{\theta}_c$ using the test characteristic curve (see (24)) and the item parameter database constructed from the respective parameter estimation sample (either $\underline{n}$ = 25 or 50). The value of $\underline{\theta}_c$ was used as the initial prior $\underline{\theta}$ and the prior variance was set to one, as recommended by Weiss and Kingsbury (1984). The MISS procedure was used to select the next test item for the re-enactment for each examinee (see formulas (10) to (12)). The correctness of the examinee's response to that item was determined by retrieving it from the database. Bayesian updating of $\underline{\theta}$ and its variance was accomplished with Owen's method (1975). See formulas (13) to (22). After each item was "administered", the AMT stopping rules were applied using a .95 confidence interval (see (25.1) to (25.3)). If a decision could be reached, the re-enactment was ended at that point. The number of questions answered correctly and incorrectly in the AMT and the decision reached for that examinee were written to a computer data file. Also stored in that file were the total test score for that examinee and the agreement between the AMT decision and the total test decision. If no decision could be reached by the AMT model before exhausting the test item pool, then a decision was forced at the end of the test: If the current estimate of $\underline{\theta}$ was greater than or equal to $\underline{\theta}_c$, the examinee was considered to be a master; otherwise a nonmaster.

2. SPRT. The mastery and nonmastery levels required by the SPRT were empirically established from the parameter estimation samples, as described above. Since the SPRT requires random selection of items, test items were "administered" in a random order. Alpha and $\underline{\beta}$ levels were set at 0.025, to make the overall decision error rate (.05) equivalent to the .95 confidence interval method used in the AMT approach. When the SPRT reached a mastery or nonmastery decision, results were stored in a separate data file in the same manner as described above for the AMT.

3. EXSPRT-R. As in the SPRT, items were "administered" in a random order. However, the rulebases constructed from the parameter estimation samples were used, of course, in the EXSPRT-R method of Bayesian updating. For a description of EXSPRT-R procedures, see the above section on an example of expert systems reasoning during computer-based testing. When the EXSPRT-R reached a decision, the test re-enactment was ended and results written to a data file as before.

4. EXSPRT-I. This method was the same as the EXSPRT-R, except that items were selected intelligently, based on their utility indices (see (1) to (7)). Thus, like the AMT, items were not "administered" randomly for each re-enactment. Since no feedback was given during the test it is unlikely that decisions reached by both AMT and EXSPRT-I

methods would be systematically affected by factors other than differences in the adaptive methods themselves. One mitigating factor might be examinee fatigue, where they were more likely to answer questions incorrectly at the end of the long and difficult test. However, since all test items were originally administered in a different random order for each individual, it is very unlikely that fatigue would systematically bias any findings.

Results from the First Study

For the DAL test, IRT item parameters (b_i's) were estimated from samples of 25 and 50 examinees. EXSPRT rulebases were also derived from the same samples. Descriptive information is given about the two samples in the left side of Table 1. It can be seen that there were about the same proportions of masters and nonmasters in each sample. In the sample of 50 there were 23 masters whose average test score was 87.3 percent, and 27 nonmasters who scored 45.1 percent correct.

We were interested in comparing the mean test lengths of each of the four methods, variation in test lengths, and decision accuracies. If the decision made by an adaptive method was the same as that reached on the basis of the entire test item pool, this was considered to be a "hit". Thus, the accuracy measures are the percent of correct predictions made by each method. There were 28 nonmasters and 25 masters identified by the entire 97-item test, when the cut-off score was set at 72.5 percent correct.

First, note that the parameter sample size seems to make little difference in the mean test length within each method. For example, within the AMT model 20.6 items were required for nonmastery decisions when item parameters were based on a sample of 25, compared to a mean of 18.3 for the sample of 50. For the EXSPRT-I, 5.6 items were required for nonmastery decisions in the sample of 25, compared to a mean of 5.9 for the parameter sample of 50. Please note--and this is confusing--that the mean test lengths for each of the four methods are based on the same 53 test administrations, where all 97 items were originally given, and which were re-enacted under each adaptive method. The size of the parameter estimation sample refers to the number of examinees randomly selected on whom the item difficulties were estimated for the AMT model and on whom the item rulebases were constructed for the EXSPRT-R and EXSPRT-I models.

Decision accuracies. For the 53 administrations of this DAL test there does seem to be some difference in decision accuracies within each model for the two parameter estimation sample sizes. The decision accuracies tended to be high for all methods. Decision accuracies were compared to expected values of .975 correct mastery decisions and .975 correct nonmastery decisions, using Chi-square goodness of fit tests (e.g., see Glass & Hopkins, 1984). A significant Chi-square ($p < .05$) means that the observed decision accuracies departed from what was expected according the a priori decision error rates that were established for each of the four adaptive testing methods.

Table 1. Efficiency and Accuracy of the Four Adaptive Testing Methods in the First Study.[1]

Item Parameter Sample Description		ADAPTIVE TESTING METHOD			
		AMT	SPRT	EXSPRT-R	EXSPRT-I
	Mean Score (S.D.) n	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy
Masters	87.46 (7.90) 12	8.40 (9.62) 100.0	8.72 (5.16) 92.0	7.56 (3.22) 100.0	5.44 (1.23) 100.0
Nonmasters	42.66 (15.83) 13	20.57 (24.45) 96.4	10.54 (7.14) 85.7*	12.71 (15.46) 96.4	5.64 (2.02) 85.7*
Total	64.16 (25.99) 25	14.83 (19.77) 98.1	9.68 (6.29) 88.7	10.28 (11.65) 98.1	5.55 (1.68) 92.5
Masters	87.27 (7.89) 23	8.28 (8.19) 100.0	10.36 (6.92) 96.0	8.44 (5.74) 96.0	6.84 (2.64) 100.0
Nonmasters	45.06 (16.25) 27	18.29 (24.43) 96.4	10.11 (10.97) 89.3*	9.39 (9.15) 92.9	5.93 (2.28) 92.9
Total	64.47 (24.89) 50	13.57 (19.14) 98.1	10.23 (9.20) 92.5	8.94 (8.94) 94.3	6.36 (2.47) 96.2

*Percent accuracies were tested by goodness of fit, where .975 accuracy was expected according to the a priori error rates for masters and nonmasters. Only those percent accuracies which differed significantly from the expected accuracies, according to a chi-square test (d.f. = 1, $p < .05$) are marked with an asterisk.

[1]Alpha = β = 0.025 for the SPRT, EXSPRT and EXSPRT-I; a .95 confidence interval was used with the AMT. There were 53 administrations of the DAL test which were re-enacted for each of the four adaptive methods.

When 25 examinees were used for parameter estimation, there were two significant departures from expected accuracy. The EXSPRT-I was 85.7 percent accurate in nonmastery decisions, which significantly differed from the expected 97.5 percent accuracy. At the same time, however, the EXSPRT-I was reaching decisions when the other models were requiring two to four times as many items. The SPRT accuracy for nonmastery decisions was also significantly lower than expected.

When 50 examinees were used for parameter estimation, the AMT, EXSPRT-R, and EXSPRT-I models were within the expected range of accuracy. The SPRT failed to make as many correct nonmastery decisions as were expected. What is notable is how well all of the adaptive methods predicted total test decisions, while using between 5 and 20 items from the 97-item pool to reach those decisions--a very substantial reduction in test lengths (95 to 80 percent decrease).

Efficiency. A repeated measures ANOVA (MANOVA) was conducted to see if there were significant differences among the mean test lengths for the four adaptive methods. This was done for the results based on the parameter sample of 50 for the 53 test administrations. Hotelling's T^2 was significant at the .05 level. However, the sphericity assumption was violated, due to the large differences in variances among the four methods. A post hoc comparison procedure suggested by Marascuilo and Levin (1983, pp. 373-381) for this kind of situation was conducted for all pair-wise contrasts of mean test lengths. One statistically significant difference was found. The mean test length for the SPRT was significantly greater than that for the EXSPRT-I. Even though some of the other contrasts have greater magnitudes of difference, the within-method variances are very different themselves. It can be noted that, overall, the AMT model required about twice as many items to reach decisions (13.6) as did the EXSPRT-I (6.4), though it was not statistically significant at the .05 level.

The variances in average test lengths within each adaptive method were significantly different, as noted above in violation of the sphericity assumption. The variance in test lengths for the AMT model was approximately 60 times larger than that for the EXSPRT-I model (19.14^2 vs. 2.47^2). In the AMT model, tests tended to be longer before nonmastery decisions were reached, and there was much more variation in test lengths compared to the remaining models. The variation in lengths of tests with EXSPRT-I method was relatively small compared to variation in the remaining models.

SECOND STUDY

Computer Functions Test

A computer-based test on how computers work, consisting of 85 items, was constructed. The COM test, as it is referred to here, was comprised of about half multiple-choice, one-fourth binary choice, and one-fourth fill-in type questions (Cronbach α = .94). Compared to the

DAL test, the COM test was much easier for most examinees (mean score = 79.0 percent, S.D. = 13.6).

Examinees

About half of those who took the COM test were from two sections of an introductory graduate-level course on use of computers in education. The remainder were mostly volunteers from an undergraduate-level course for non-education majors who were learning to use computers. A small number of students were volunteers recruited at the main library on campus.

Test Administration and Experimental Methods

The COM test was individually administered by the Indiana Testing System in the same manner as the DAL test. There were a total of 104 administrations of the COM test in the second study. The same four adaptive testing methods were re-enacted from actual examinee test data in the very same manner as described above for the DAL test.

Results from the Second Study

Since there were more administrations of the COM test, parameter estimation samples of 25, 50, 75 and 100 were selected at random. Four sets of b_i coefficients were obtained for the AMT model and four rulebases were constructed for the EXSPRT models based on the same four parameter estimation samples. See the left sides of Tables 2.1 and 2.2 for descriptive information about the parameter estimation samples.

Accuracy of predictions. When the parameter estimation sample was 25, all four adaptive methods did not perform as well as expected in correctly predicting nonmasters in the 104 administrations of the COM test. Chi-square goodness of fit tests showed that all four methods significantly departed from the expected accuracy rates. EXSPRT-I had the worst accuracy, but it should be noted that there were only seven nonmasters in the estimation sample for creating the rulebase, so this is not surprising.

When the parameter estimation sample was 50, the AMT and EXSPRT-I models still made significantly fewer correct nonmastery decisions than expected a priori. On the other hand, the SPRT and EXSPRT--both of which use random selection of items vs. intelligent selection in the AMT and EXSPRT-I--predicted masters and nonmasters correctly within the bounds of expected error rates.

When the parameter estimation sample was 75 (55 masters and 20 nonmasters when the cut-off was 72.5 percent correct), all models predicted well except the AMT, which made significantly fewer correct nonmastery decisions than were expected a priori.

Table 2.1. Efficiency and Accuracy of the Four Adaptive Testing Methods in the Second Study.[a]

Item Parameter Sample Description		ADAPTIVE TESTING METHOD			
		AMT	SPRT	EXSPRT-R	EXSPRT-I
	Mean Score (S.D.) n	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy
Masters	86.21 (6.35) 18	8.37 (13.59) 97.4	11.71 (7.80) 98.7	10.05 (5.42) 98.7	4.57 (3.60) 98.7
Nonmasters	48.07 (9.10) 7	33.93 (31.83) 82.1*	14.39 (15.81) 85.7*	15.00 (10.41) 82.1*	7.07 (2.36) 67.9*
Total	75.53 (18.84) 25	15.25 (23.02) 93.3	12.43 (10.55) 95.2	11.38 (7.39) 94.2	5.24 (3.49) 90.4
Masters	87.16 (5.68) 35	11.83 (18.23) 94.7	15.08 (9.06) 96.1	11.71 (8.28) 98.7	5.72 (3.92) 96.1
Nonmasters	53.65 (10.44) 15	31.89 (29.97) 78.6*	17.39 (14.50) 92.9	15.82 (12.70) 96.4	7.93 (6.21) 89.3*
Total	77.11 (17.15) 50	17.23 (23.61) 90.4	15.70 (10.77) 95.2	12.82 (9.78) 98.1	6.32 (4.72) 94.2

*Percent accuracies were tested by goodness of fit, where .975 accuracy was expected according to the a priori error rates for masters and nonmasters. Only those percent accuracies which differed significantly from the expected accuracies, according to a chi-square test (d.f. = 1, $p < .05$) are marked with an asterisk.

[a]Alpha = β = 0.025 for the SPRT, EXSPRT-R, and EXSPRT-I; a .95 confidence interval was used with the AMT. There were 104 administrations of the COM test which were re-enacted for each of the four adaptive methods.

Table 2.2. Efficiency and Accuracy of the Four Adaptive Testing Methods in the Second Study (cont'd).

Item Parameter Sample Description	Mean Score (S.D.) n	ADAPTIVE TESTING METHOD			
		AMT	SPRT	EXSPRT-R	EXSPRT-I
		Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy
Masters	87.68 (5.93) 55	10.21 (16.96) 94.7	16.64 (10.35) 97.4	11.78 (6.34) 97.4	7.70 (7.13) 94.7
Nonmasters	56 0 (10.17) 20	28.93 (28.58) 82.1*	16.29 (16.96) 92.9	14.75 (15.54) 100.0	7.82 (4.85) 100.0
Total	79.23 (15.84) 75	15.25 (22.21) 91.3	16.55 (12.39) 96.2	12.58 (9.71) 98.1	7.73 (6.57) 96.2
Masters	87.47 (6.33) 75	13.75 (20.80) 93.4*	16.97 (10.75) 96.1	13.58 (9.51) 98.7	7.64 (6.12) 94.7
Nonmasters	56.00 (11.34) 25	31.50 (29.77) 78.6*	13.04 (10.66) 92.9	12 32 (10.78) 96.4	8.93 (7.41) 100.0
Total	79.60 (15.77) 100	13.53 (24.70) 89.4	15.91 (10.82) 75.2	13.24 (9.83) 98 1	7.99 (6.48) 96.2

*Percent accuracies were tested by goodness of fit, where .975 accuracy was expected according to the a priori error rates for masters and nonmasters. Only those percent accuracies which differed significantly from the expected accuracies, according to a chi-square test (d.f. = 1, $p < .05$) are marked with an asterisk.

When the parameter estimation sample was 100, the AMT model still had problems with accuracy of nonmastery classifications. And strangely enough, the AMT model also made significantly fewer correct mastery decisions than were expected. The SPRT, EXSPRT-R, and EXSPRT-I all correctly predicted masters and nonmasters within the bounds of expected accuracies. It should be noted that it is generally recommended that a minimum of 200 examinees be used for estimating b_i parameters in the IRT-based, one-parameter AMT model. Only half that number were available in this study. Thus, it is not surprising that the AMT model performed less well than it should, since estimation of the item difficulty parameters was not as precise as desired.

Efficiency. Average test lengths of the four adaptive methods were compared for the 100 examinee parameter estimation situation only. See the bottom half of Table 2.2. A MANOVA again revealed that the sphericity assumption was violated, and so the same procedure as described above for the DAL test was used in post hoc comparisons of the adaptive COM test length means (Marascuilo & Levin, 1983).

When nonmastery decisions were made, the AMT model required significantly longer tests than either the SPRT, EXSPRT-R or EXSPRT-I. The AMT model required about 32 items to reach nonmastery decisions, compared to the EXSPRT-I, which required about nine items. Moreover, the AMT made significantly fewer correct nonmastery decisions than expected, as noted above. When mastery decisions were reached, test lengths for the SPRT and EXSPRT-R methods (15 and 12) were significantly longer than the EXSPRT-I (6 items). Mean test lengths for mastery decisions in the AMT and EXSPRT-I models were not significantly different at the .05 level.

When looking at decisions overall, the following contrasts were significantly different: the AMT, SPRT, and EXSPRT-R methods each required significantly longer tests than did the EXSPRT-I model. The AMT model required over twice as many items as did the EXSPRT-I (19 vs. 8).

Summary. It would appear from the COM test data that the EXSPRT-I is significantly more efficient than the other adaptive methods. Indeed, it is rather remarkable that the EXSPRT-I can make such highly accurate mastery and nonmastery decisions with relatively few test questions. It is also notable that the EXSPRT-R and SPRT also made highly accurate predictions, but were less efficient than the EXSPRT-I. The AMT performed worst of all, not only resulting in longer adaptive tests but also in making significantly more prediction errors than theoretically expected.

One limitation of the first and second studies is that both respectively used the same sets of test administrations for not only estimating item parameters and rulebases, but also for re-enacting the four adaptive testing methods. It would have been preferable to use one set of test administration data for parameter estimation and rulebase construction, and then to use an independent sample of examinees to compare the four adaptive methods for efficiency and accuracy. This latter strategy was followed in the third study.

THIRD STUDY

The Computer Literacy Test

A 55-item test was constructed for purposes of screening undergraduate students at Indiana Wesleyan University. The eventual goal is to use the test for deciding whether or not undergraduate students must take a course in general computer literacy. A paper-and-pencil version of the test was initially given to 40 students. Cronbach's α was .84, and the average test score was about 50 percent with examinees distributed normally. This test is herein referred to as the LIT test.

Examinees

A new sample of roughly half freshman and half sophomores was obtained (n = 333). Fifteen different majors and a wide range of academic ability levels were represented by this sample.

Test Administration

It was not feasible to test the 333 examinees by computer at the time the data were collected. Instead, individual test booklets were constructed for each examinee with one item per page. Each test booklet contained a different random order of the 55 items. Examinees were instructed to answer items in the order they appeared in the booklet, and they were not permitted to flip back to previous pages during the test. Most students completed the test in 30 to 45 minutes.

Tests were hand-scored and item-by-item results were entered into a computer database so that it would be possible to conduct re-enactments of the LIT test under the four adaptive conditions, as described above for the DAL and COM tests.

Results of the Third Study

Four samples were selected at random from the 333 examinees for estimating item parameters in the AMT model and constructing rulebases in the EXSPRT models (n = 25, 50, 75, 150). A cut-off of 59.5 percent correct was chosen for sorting students into mastery and nonmastery categories required for constructing the EXSPRT rulebases. Examinees who were selected for each parameter estimation sample were subsequently excluded from further analyses of results from the four adaptive testing methods. For example, when the parameter sample was 50, then 333 minus 50, or 283 administrations were re-enacted under each adaptive method. S Tables 3.1 and 3.2.

Table 3.1. Efficiency and Accuracy of the Four Adaptive Testing Methods in the Third Study.[a]

Item Parameter Sample Description		ADAPTIVE TESTING METHOD			
		AMT	SPRT	EXSPRT-R	EXSPRT-I
	Mean Score (S.D.) n	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy	Mean Length (S.D.) Accuracy
			n = 308		
Masters	62.7	21.72	22.34	25.59	21.49
	(1.3)	(26.62)	(10.82)	(15.60)	(16.28)
	2	95.3	100.0	89.5*	97.7
Nonmasters	39.9	20.80	38.23	33.02	27.10
	(10.0)	(18.18)	(15.53)	(15.44)	(19.49)
	23	89.2*	68.7*	77.5*	86.9*
Total	41.7	21.06	33.79	30.95	25.54
	(11.5)	(19.17)	(16.03)	(15.81)	(18.80)
	25	90.9	77.6	80.8	89.9
			n = 283		
Masters	65.6	25.40	23.40	25.26	24.07
	(4.7)	(20.80)	(12.45)	(15.78)	(14.87)
	6	79.3*	100.0	89.0*	96.5
Nonmasters	41.4	16.18	34.92	28.68	19.37
	(9.8)	(17.15)	(16.43)	(15.59)	(18.11)
	44	91.5*	76.6*	81.1*	90.5*
Total	44.4	18.85	31.58	27.70	20.73
	(12.2)	(18.72)	(16.23)	(15.69)	(17.34)
	50	88.0	83.4	83.4	92.2

*Percent accuracies were tested by goodness of fit, where .975 accuracy was expected according to the a priori error rates for masters and nonmasters. Only those percent accuracies which differed significantly from the expected accuracies, according to a chi-square test (d.f. = 1, $p < .05$) are marked with an asterisk.

[a]Alpha = β = 0.025 for the SPRT, EXSPRT-R, and EXSPRT-I; a .95 confidence interval was used with the AMT. There were (333 - parameter sample size) administrations of the LIT test which were re-enacted for each of the four adaptive methods.

Table 3.2. Efficiency and Accuracy of the Four Adaptive Testing Methods in the Third Study (cont'd).

Item Parameter Sample Description	Mean Score (S.D.) n	ADAPTIVE TESTING METHOD AMT Mean Length (S.D.) Accuracy	SPRT Mean Length (S.D.) Accuracy	EXSPRT-R Mean Length (S.D.) Accuracy	EXSPRT-I Mean Length (S.D.) Accuracy
			n = 258		
Masters	65.0	23.22	26.37	28.27	25.52
	(4.3)	(19.20)	(12.57)	(16.22)	(16.36)
	11	90.9*	100.0	92.2*	93.5*
Nonmasters	43.8	18.72	44.69	31.11	19.72
	(9.8)	(17.04)	(13.22)	(16.54)	(18.61)
	64	91.2*	85.1*	82.9*	91.7*
Total	46.9	20.07	34.36	30.26	21.45
	(11.9)	(17.80)	(15.97)	(16.46)	(18.13)
	75	91.1	89.6	85.7	92.2
			n = 183		
Masters	66.8	25.35	23.07	21.33	22.72
	(6.0)	(21.03)	(12.90)	(13.69)	(15.44)
	34	96.3	100.0	98.1	98.1
Nonmasters	43.2	20.85	34.85	30.73	22.35
	(10.4)	(19.37)	(15.69)	(16.18)	(18.58)
	116	90.7*	85.3*	77.5*	92.2*
Total	48.6	22.18	31.38	27.96	22.46
	(13.7)	(19.92)	(15.83)	(16.04)	(17.67)
	150	92.4	89.6	83.6	94.0

*Percent accuracies were tested by goodness of fit, where .975 accuracy was expected according to the a priori error rates for masters and nonmasters. Only those percent accuracies which differed significantly from the expected accuracies, according to a chi-square test (d.f. = 1, $p < .05$) are marked with an asterisk.

Note the characteristics of parameter samples in relation to the chosen cut-off of 59.5 percent correct on the left sides of Tables 3.1 and 3.2. For example, in the sample of 25, there were only two masters and 23 nonmasters. In the sample of 150, there were 34 masters and 116 nonmasters. Also note that the average total test scores were well below the chosen cut-off in the third study, as would be expected with such a disproportionate number of nonmasters.

Decision accuracies. The most notable result is that all four adaptive methods were unable to predict nonmastery decisions as accurately as expected across all four parameter estimation sample sizes (25, 50, 75 and 150). When the parameter estimation sample size was 50 or greater, both the AMT and EXSPRT-I correctly predicted nonmastery decisions slightly more than 90 percent of the time. Nonetheless, these were significant departures from the expected 97.5 percent accuracy rate.

When the parameter estimation sample size was 150, all four adaptive methods predicted mastery decisions within the bounds of the a priori error rates. Interestingly, the SPRT was 100 percent accurate in its mastery decisions when using mastery levels determined from each of the four different parameter estimation samples.

It would appear from these data and the LIT test that, when a 59.5 percent cut-off is used, the tendency of each of the methods is to fail to make as many correct nonmastery decisions as expected.

Efficiency. A MANOVA was run on the 183 examinees whose tests were re-enacted using parameter information derived from a sample of 150 other examinees. See the bottom half of Table 3.2. Hotelling's T^2 was highly significant. The mean test length for the SPRT (31.38) was significantly greater than each of the other three adaptive methods. The mean test length for the EXSPRT-R (27.96) was significantly larger than either the EXSPRT-I (22.46) or AMT (22.18). There was no significant difference between the EXSPRT-I and AMT mean test lengths.

It is noteworthy that average adaptive LIT test lengths tended to be much larger than those for the DAL and COM tests.

DISCUSSION

Why Do the Adaptive Methods Appear to Behave Inconsistently?

Adaptive tests tended to be shortest in the first study with the DAL test. See Table 1. Of the 53 administrations of the DAL test, there were 28 nonmasters and 25 masters when the cut-off was set at 72.5 percent and when examinees answered all 97 items. The overall average test score was 63.2 (S.D. = 24.6).

Next shortest were the adaptive tests in the second study with the COM test. See Tables 2.1 and 2.2. There were 104 administrations of this test, with 76 masters and 28 nonmasters when the entire 85-item

test was taken (grand mean = 79.0, S.D. = 13.6, mastery cut-off = 72.5 percent).

The LIT test resulted in the longest adaptive tests overall (Tables 3.1 and 3.2). In the sample of 183 examinees there were 54 masters and 129 nonmasters based on total test results from the 55-item pool. The cut-off for this test was 59.5 percent, and the overall average score was 51.5 percent (S.D. = 14.2).

One thing that appears to affect the average test lengths is the location and shape of the distribution of examinee achievement levels in relation to the cut-off selected. In the first study, the distribution was somewhat bimodal and relatively flat, with about half the examinees scoring above and below the cut-off. In the second study, the distribution was positively skewed, with about three-fourths of the examinees scoring above the 72.5 percent cut-off on the entire test item pool. In the third study, over two-thirds of the examinees were classified as nonmasters on the entire 55-item test. The distribution of this group was close to normal, with the mean being about 8 percentage points below the selected cut-off.

We have previously conducted a number of computer simulations comparing the three-parameter AMT model with the SPRT and a third adaptive method based on Bayesian posterior beta distributions (Frick, 1988; and Frick, Luk & Tyan, 1987). One important finding in those studies was that none of the adaptive methods performed as well as expected--and average test lengths tended to be longer--when the distribution of examinees was mostly clustered around the cut-off. Adaptive tests were shorter and accuracies agreed with theoretical expectations when the distributions of examinee achievement levels were much flatter. The same phenomenon appears to have occurred in the present three empirical studies.

The second factor that apparently affected the results is the number of test items in each pool and their properties. When there are more test items, and there are more items available at each ability or achievement level, then both the AMT and EXSPRT-I tend to be more efficient and more accurate. In both adaptive methods which rely on "intelligent" selection of items, Bayesian posterior estimates are affected more dramatically when there are highly discriminating items available whose difficulty levels are close to the current estimate of an examinee's achievement level. A real problem occurs with relatively small item pools (as was the case with the LIT test in the third study): After the best items have been administered early in a test, the remaining items tend to provide little additional information. That is, there are diminishing returns after some point because there are no really appropriate items left.

In the third study we observed a kind of "yo-yo" effect. Since most of the examinees tended to be average in achievement, those items of average difficulty were first chosen. After 10 to 15 items had been administered, there were very few items left of average difficulty. The remainder were either easier or harder. If a harder question was next picked, it was answered--not to our surprise--incorrectly. If an

easier question was chosen next--guess what? In other words, the remaining items tended to provide progressively less additional information about an examinee, belaboring and elongating a test until a decision could be reached. As a result, adaptive tests tended to be longer in both the AMT and EXSPRT-I models in the third study (compared to the first two studies), and the number of correct nonmastery decisions was not as high as expected a priori.

Summary

Expert systems can be used to aid decision makers. A computerized adaptive test (CAT) is one kind of expert system, though not commonly recognized as such. When item response theory is used in a CAT, then the knowledge or rule base is a set of item characteristic curves (ICC's).

Normally an expert system consists of a set of questions and a rule base. An inference engine uses answers to the questions and the rule base to choose from a set of discrete alternatives. If an adaptive test is viewed this way, then it is possible to construct "If ..., then ..." rules about test items that are not functions, as are ICC's. A new approach, termed EXSPRT, was devised that combines expert systems reasoning and sequential probability ratio test stopping rules. EXSPRT-R uses random selection of test items, whereas EXSPRT-I incorporates an intelligent selection procedure based on item utility coefficients.

These two new methods were compared to the traditional SPRT and to an IRT-based approach to adaptive mastery testing (AMT). Three empirical studies with different tests and types of examinees were carried out.

In the first study the EXSPRT-I model required about half as many items as did the AMT approach (6 vs. 14), though the difference was not statistically significant. When 50 examinees were used for item parameter estimation and rule base construction, all four methods (AMT, SPRT, EXSPRT-R and EXSPRT-I) made highly accurate mastery and nonmastery decisions.

In the second study the EXSPRT-I method again required about half as many items as did the AMT model (8 vs. 19), and this time the difference was statistically significant. When 100 examinees were used for estimation purposes, the SPRT, EXSPRT-R, and EXSPRT-I correctly predicted masters and nonmasters within the bounds of the expected theoretical error rates. The AMT model, however, made significantly more prediction errors than expected.

In the third study, the EXSPRT-I and AMT models each required about 22 items to reach decisions, whereas the SPRT and EXSPRT-R required significantly more test questions. None of the models was able to predict nonmasters as well as expected by theoretical error rates--even when the parameter estimation sample was as large as 150 examinees--though all models satisfactorily predicted masters at that

point. Two factors appeared to affect these outcomes. First, examinees tended to be mostly clustered near the mastery cut-off. Second, the item pool was much smaller in this study, and it appears that the adaptive methods tended to run out of test items which were of appropriate difficulty levels and also highly discriminating.

Overall, results indicated that the EXSPRT-I is more efficient or as efficient as the AMT model. When the distribution of examinees was not clustered near the mastery cut-off, all four methods were usually able to make highly accurate mastery and nonmastery classifications.

Although further research is needed, the EXSPRT initially appears to be a strong alternative to IRT-based adaptive testing when categorical decisions about examinees are desired. The EXSPRT is less complex conceptually and mathematically, and it appears to require many fewer examinees to establish empirically a rule base--compared to the large numbers required to adequately estimate parameters for ICC's in the IRT model. On the other hand, the EXSPRT is vulnerable, as is classical test theory, in that a representative sample of examinees must be selected for constructing rule quadruplets. This seems to be a small price to pay for the advantages of theoretical parsimony and operational efficiency.

REFERENCES

Birnbaum, A. (1968) In Lord, F. and Novick, M., Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

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

Clancey, W. (1987). Methodology for building an intelligent tutoring system. In G. Kearsley (Ed.), Artificial intelligence and instruction: Applications and methods. Reading, MA: Addison-Wesley, 193-227.

Davis, R. (1984). Amplifying expertise with expert systems. In P. Winston and K. Prendergast (Eds.), The AI business: The commercial uses of artificial intelligence. Cambridge, MA: The MIT Press, 17-40.

Frick, T. (1986). The Indiana Testing System (ITS, Version 1.0). Bloomington: Department of Instructional Systems Technology, School of Education, Indiana University.

Frick, T. (1988). A comparison of three decision models for adapting the length of computer-based mastery tests. Paper submitted to the Journal of Educational Computing Research, October.

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

Frick, T., Luk, H.-K., and Tyan, N.-C. (1987). A comparison of three adaptive decision-making methodologies used in computer-based instruction and testing. Bloomington, IN: Final Report, Proffitt Foundation. Indiana University School of Education.

Glass, G. and Hopkins, K. (1984). Statistical methods in education and psychology. Englewood Cliffs, NT: Prentice-Hall.

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

Heines, J. (1983). Basic concepts in knowledge-based systems. Machine-Mediated Learning, 1(1), 65-95.

Kingsbury, G. G., and Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), New horizons in testing. New York: Academic Press.

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

Lord, F. and Novick, M. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Mager, R. (1973). Measuring instructional intent. Belmont, CA: Fearon-Pittman.

Marascuilo L. and Levin, J. (1983). Multivariate statistics in the social sciences: A researcher's guide. Monterey, CA: Brooks/Cole.

Mead, R., Wright, B., and Bell, S. (1979). BICAL (Version 3). Chicago: Department of Education, University of Chicago.

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.

Schmitt, S. (1969). Measuring uncertainty. Reading, MA: Addison-Wesley.

Sleeman D. and Brown, J. (Eds.) (1982). Intelligent tutoring systems. New York, NY: Academic Press.

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

Weiss, D. and Kingsbury, G. (1984). Application of computerized adaptive testing to education problems. Journal of Educational Measurement, 21, 361-375.

Winston, P. and Prendergast, K. (1984). The AI business: The commercial uses of artificial intelligence. Cambridge, MA: The MIT Press, 17-40.

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