JOURNAL OF
EDUCATIONAL
COMPUTING
RESEARCH

Edited by:
Dr. Robert H. Seidman

Volume 8, Number 2 — 1992

Computerized Adaptive Mastery Tests as Expert Systems

Theodore W. Frick

Baywood Publishing Company, Inc.
26 Austin Avenue, Box 337
Amityville, NY 11701
Call (516) 691-1270 • Fax (516) 691-1770
☎ Orders only —**call toll-free** (800) 638-7819

COMPUTERIZED ADAPTIVE MASTERY TESTS AS EXPERT SYSTEMS*

THEODORE W. FRICK
Indiana University

ABSTRACT

Expert systems can be used to aid decision making. A computerized adaptive test is one kind of expert system, though not commonly recognized as such. A new approach, termed EXSPRT, was devised that combines uncertain inference in expert systems with sequential probability ratio test stopping rules. Two versions of EXSPRT were developed, one with random selection of items (EXSPRT-R) and one with intelligent selection (EXSPRT-I). Two empirical studies were conducted in which these two new methods were compared to the traditional SPRT and to an adaptive mastery testing (AMT) approach based on item response theory (IRT). The EXSPRT-I tended to be more efficient than the AMT, EXSPRT-R and SPRT models in terms of average test lengths. Although further research is needed, the EXSPRT-I initially appears to be a strong alternative to both IRT- and SPRT-based adaptive tests for making categorical decisions about examinee mastery of single instructional objectives. The EXSPRT-I is clearly less complex than IRT, both conceptually and mathematically. It also 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

Second-Generation 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

*These studies were supported in part by a grant from the Proffitt Foundation, School of Education, Indiana University.

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 repair persons [1].

Early expert systems consisted of sets of production rules or frames, often called "knowledge bases." The name, "expert system," was coined because a knowledge base was typically constructed in early expert systems by interviewing one or more experts in some domain of knowledge. An attempt was 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 [2, p. 34].

This particular rule is one of over 400 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 categorical inferences by using both the rule set and specific answers to questions it asks a 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. It is also noteworthy that epidemiological data, in addition to expert physicians, were consulted in the formation and refinement of the MYCIN knowledge base.

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

In education and training, expert systems principles have been applied mostly in intelligent tutoring systems [3, 4]. As an example, GUIDON was later developed from MYCIN in an attempt to teach physicians how to identify different kinds of bacteria [5].

Evolution of Expert Systems

Expert systems have undergone three major phases of theoretical development According to Neapolitan, the first generation of expert systems of the 1960s utilized probability theory and simple Bayesian reasoning to perform uncertain inference [6, 7]. In the 1970s rule-based systems—originally called production

systems—were created that performed categorical inference (i.e., logical deductions with certainty). MYCIN is an example of such a second-generation expert system. Probability theory was questioned, however, by some researchers as a foundation for making *uncertain* inferences in rule-based systems which attempted to model human reasoning. Alternative approaches were proposed such as the use of certainty factors [8], the Dempster-Shafer theory of evidence [9], and fuzzy set theory and fuzzy logic [10]. In the 1980s researchers began to build normative expert systems [cf., 6]. These third-generation systems do *not* attempt to imitate actual human reasoning. Rather, Neapolitan takes the strong position that [6, p.2]:

> . . . the fundamental goal of an expert system is to make the best possible judgments, not to descriptively model human reasoning. If expert consensus and the results of an autopsy were different, I would want the system to agree with the results of the autopsy. I would not want the system to agree with the experts much less model the way the experts reason. If this view is taken, the normative approach of the 1980's is appropriate in expert systems which must perform uncertain inference.

Perhaps the term, "knowledge-based" or "data-based decision support systems" would be preferable to "expert" systems, since the name may connote that the expertise is modelled after human reasoning. Indeed, in the insurance industry actuarial tables are consulted by expert systems developed to aid underwriters. The "expertise" in the life expectancy tables is based on large statistical samples of persons who have characteristics and health profiles to similar to those of the individuals under consideration for policies. The knowledge is represented by a statistical database.

Computerized Adaptive Tests

In a computerized adaptive test (CAT), items are selected which are close to each particular person's 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. A CAT takes no longer than necessary to obtain a satisfactory estimate of an examinee's ability. Adaptive tests tend to be shorter than conventional fixed-length tests, and the results are as reliable if not more so. Most of the research on adaptive testing has been based on item response theory, building upon the seminal work of Lord, Novick and Birnbaum [11].

In an extensive review of the literature, Plew found that early notions of adaptive testing were emerging about the time computers were invented in the 1940s [12]. These early attempts employed relatively simple decision approaches

such as the sequential probability ratio test developed by Wald and hierarchical branching strategies. Test administration and computation had to be done by hand, since interactive computing was not to become available for several decades. Hence, adaptive tests were impractical at that time.

Frederick Lord invented item response theory in the early 1950s, but did little more with it until the middle 1960s when he found sufficient empirical evidence to justify the form of the item response function. Since the publication of *Statistical Theories of Mental Test Scores* in 1968, a large number of research studies on applications of IRT have been conducted [3, 4]. Item response theory has been applied to many areas of psychological measurement, including a relatively small but growing body of research on computerized adaptive testing [3, 5]. David Weiss, his colleagues and students conducted numerous studies of CATs in the 1970s and 1980s at the University of Minnesota Psychometric Methods Program. Item response theory, coupled with item selection strategies such as maximum information search and selection, proved to be a solid foundation for CAT [6].

The advent of powerful personal computers in the 1980s helped to bring CATs out of the research laboratories and into practical settings [12, 15-28]. For example, the Portland, Oregon public schools have implemented CATs [18]. CATs have also been approved by the Maryland State Board of Education for assessing mathematics and reading competencies required for high school graduation [20]. Bunderson, Inouye and Olson describe a number of applications of CATs, as well as speculate about future kinds of continuous and intelligent measurement systems [15].

It is noteworthy that most research on applications of CATs in based on an IRT or Rasch model where tests are taken by many thousands of individuals (e.g., by standardized testing agencies, U.S. armed forces, and state- or district-wide testing of student academic achievement). Relatively little effort has been directed toward practical applications of CATs for classroom tests constructed by teachers for evaluating student learning. There is also a paucity of research on use of CATs for determining student mastery of instructional objectives during computer-based instruction or in computer-managed instructional systems.

Third-Generation Expert Systems and Computerized Adaptive Tests

It might appear to many readers that expert systems and CATs are quite different kinds of entities. Indeed, developments in these two areas have been promulgated by different camps, one from the artificial intelligence and cognitive science area, and the other from psychometrics and education. Although a CAT may superficially seem quite different from an expert system such as MYCIN, both are examples of intelligent computer programs. Neapolitan considers an intelligent computer program to be ". . . one which makes judgments or gives

assistance in a complex area." He further states, "Such programs are often called expert systems." [6, p. 1].

If such a broad view of expert systems is taken, then a computerized adaptive test is clearly one type of an expert system:

1. The knowledge base for an IRT-based CAT is a set of item response functions estimated from prior test administrations. That is, each item response function (IRF) 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. CATs have inference engines that typically use Bayesian or maximum likelihood estimation methods.
3. Expert systems make judgments, typically by attempting to choose one alternative from a number of mutually exclusive and exhaustive conclusions. A typical goal of a CAT is to estimate an examinee's achievement or ability level with enough confidence to make a decision such as pass/fail or a grade classification.
4. Expert systems collect information in order to make judgments. A CAT does this by selecting questions on the basis of the amount of information they provide, depending on the ability of an examinee. His or her ability or achievement level is in turn estimated on the basis of which questions she or he previously answered correctly and incorrectly, and their respective IRFs. For example, Weiss and Kingsbury use a maximum information search and selection (MISS) procedure to choose each new question during a CAT. They also use Bayesian, or alternatively, maximum likelihood estimation methods of updating estimates of an examinee's achievement level [6].

Thus, although not widely recognized at this time, an adaptive testing system is one type of an expert system. I first realized this some years ago when developing computer code for an expert system, having already developed code for Bayesian decision methodologies and a computer-based testing system. Publications by computer scientists such as Neapolitan [6] and Heines [29] confirm this observation.

Nonetheless, in the research literature it appears that these two threads of development have been almost entirely independent. A 1990 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. Only one of these thirteen articles pertained to psychological testing; the remainder were in the field of engineering. The two camps not only use different language to describe their activities, but also tend to publish in different journals and attend different conferences. It also appears that neither camp understands the other's work very well. For example, one anonymous reviewer of an earlier version of this article found it difficult to accept that a CAT was an expert system. Based on his or her written comments,

this reviewer was apparently well-versed with the literature on CATs but took a narrow and somewhat naive view of expert systems (second generation only).

THE DEVELOPMENT OF *EXSPRT*

The Problem of Large Numbers

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 [cf., 30].

While the large number of examinees required for estimating item parameters may not be an issue for professional test developers (e.g., testing agencies, the armed forces, state-wide student assessment programs), it does pose a real problem for classroom teachers who make up their own tests and who want to administer them adaptively by computer. It also poses a problem for developers of computer-based instructional materials who also want to incorporate CATs. CATs appear to have considerable potential in mastery-based learning situations such as in computer-managed instructional systems, personalized systems of instruction, and so forth. For these kinds of situations, the IRT approach to adaptive testing can be likened to the use of a cannon to kill a mosquito.

For nearly ten years I have been searching for alternatives to IRT which would be appropriate for computer-based assessment of student mastery in classroom learning situations. I 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 [22]. 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 CATs, I 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.

The reader should note that the goal here is to determine student mastery of a *single* objective, where test items are developed to *match that objective*—following Robert Mager's philosophy [31]. This is referred to as criterion-referenced testing or mastery testing [22]. For example, if the objective is for students to be able to classify birds according to type, then the pool of test questions could consist of a number of pictures of different kinds of birds. Given a picture of a bird, the student is asked to name the type of bird (e.g., robin, cardinal, woodpecker, finch, etc.). If a student were to successfully identify most or all of the birds, then we would conclude that the student had mastered the instructional objective (i.e., is able to classify common birds in the United States according to type). On the other hand, questions such as, "In what regions of the United States are cardinals found?" would be inappropriate because they do not match the objective.

In contrast, the standardized achievement tests that are familiar to most teachers and students are designed to assess a wide range of instructional objectives. For example, on a mathematics subtest students might be asked to answer questions on simple arithmetic, fractions, decimals, algebraic equations, etc. Student performance on such tests can permit comparison of his or her achievement level to those of other students across a broad mathematics curriculum in elementary and secondary schools. The reader should note that this kind of norm-referenced test is *not* the kind of assessment for which the following methodologies were developed.

Development of the Rule Base for a Given Test Item Pool

Assume that we have developed a pool of test items which match a single instructional objective and that our goal is to choose between two alternatives for any given student: mastery or nonmastery of the objective. For each item i in the pool we create four rules:

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

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

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

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

Notice that these rules are essentially in the same if-then form as the sample rule from the MYCIN expert system at the beginning of this article. In MYCIN the rules were developed on the basis of expert knowledge and epidemiological data

on the co-occurrence of various kinds of gram stain, morphology and aerobicity and the incidence of each kind of bacterium. In our case, we will rely on empirical data collected on the test items with a representative sample of examinees who are characterized by the discrete categories from among which our expert system will later attempt to choose.

In the dichotomous case, the estimates of probabilities of correct responses to items by masters and nonmasters are determined as follows:[1]

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 32]:[2]

$$P(C_i \mid M) = (\#r_{im} + 1)/(\#r_{im} + \#w_{im} + 2) \quad [1.1]$$

$$P(\neg C_i \mid M) = 1 - P(C_i \mid M) \quad [1.2]$$

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:

$$P(C_i \mid N) = (\#r_{in} + 1)/(\#r_{in} + \#w_{in} + 2) \quad [1.3]$$

$$P(\neg C_i \mid N) = 1 - P(C_i \mid N) \quad [1.4]$$

Method of Making Inferences in the EXSPRT

The reasoning procedure employed in our expert systems approach is Bayesian, with the addition of stopping rules from the sequential probability ratio test (SPRT) [22, 32, 33]. A likelihood ratio is computed after each administration of a test item.

[1] Note that this reasoning can be extended to more than two categories, such as letter grade designations.

[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.

$$LR = \frac{P_{om}\prod_{i=1}^{n} P(C_i \mid M)^s[1-P(C_i \mid M)]^f}{P_{on}\prod_{i=1}^{n} P(C_i \mid N)^s[1-P(C_i \mid N)]^f} \quad [2]$$

where
P_{om} = prior probability that the examinee is a master,
P_{on} = prior probability that the examinee is a nonmaster, [3]
and $s = 1, f = 0$ if item i is answered correctly,
$s = 0, f = 1$ if item i is answered incorrectly,
$s = 0, f = 0$ if item i has not been administered.

The three stopping rules are:

If $LR \geq (1 - \beta) + \alpha$, then stop asking questions and choose mastery. [3.1]

If $LR \leq \beta + (1 - \alpha)$, then stop asking questions and choose nonmastery. [3.2]

Otherwise ask another question, update LR, and reiterate rules 3.1 to 3.3. [3.3]

Alpha and beta are Type I and II decision errors, respectively. Alpha is the probability of choosing mastery when the nonmastery alternative is actually true. Beta is the probability of choosing nonmastery when the mastery alternative is true. For numerical examples of this Bayesian reasoning process, the reader is referred to [34].

Item Selection

Random Item Selection: EXSPRT-R — When the EXSPRT was initially conceived, I viewed it as an extension of the Bayesian approach to the SPRT, as noted in [22], but used empirically derived data for estimating the probabilities of correct responses by masters and nonmasters to each test item rather than *average* probabilities across all items. In this initial approach to the EXSPRT, items were selected randomly without replacement, and it was assumed that observations were independent in order to multiply the conditional probabilities to form the likelihood ratio. This version of the EXSPRT is referred to as EXSPRT-R, in contrast to the intelligent item selection procedure discussed below.

Intelligent Item Selection: EXSPRT-I — Thomas Plew was not satisfied with EXSPRT-R, since it did not use information about test items in the selection process [12]. Plew and I jointly developed an item selection procedure that is modeled after basic principles used by Weiss and Kingsbury in the MISS

[3] Note that if the prior probabilities of mastery and non-mastery are equal, then they drop out of the formula for the likelihood ratio.

(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:

$$D_i = P(C_i \mid M) - P(C_i \mid N) \qquad [4]$$

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 fifty/fifty 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:

$$I_{ij} = abs\{(1 - P(C_i)) - E(\Phi_j)\} \qquad [5]$$

$$\text{where} \quad E(\Phi_j) = (\#r_j + 1)/(\#r_j + \#w_j + 2) \qquad [6]$$

$$\text{and} \quad P(C_i) = (\#r_i + 1)/(\#r_i + \#w_i + 2) \qquad [7]$$

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:

$$U_{ij} = D_i/(I_{ij} + \delta) \qquad [8]$$

where δ = some arbitrary small constant (e.g., .0000001), to prevent division by zero in case $I_{ij} = 0$.[4]

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) in that both item discrimination and item difficulty are considered in the item selection process [35].

Unanswered Questions

Since the EXSPRT-R and EXSPRT-I are new approaches to computerized adaptive testing, two empirical studies were conducted to compare these approaches to extant IRT-based adaptive mastery testing and SPRT approached [cf., 16, 22, 35]. 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 ninety-seven items, and referred to as the

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

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 sixty and ninety 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 I taught on computer-assisted instruction. 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 [36]. 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 fifty-three 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 [37]. For the EXSPRT-R and EXSPRT-I, the rule base for each parameter estimation sample was constructed using formulas (1.1), (1.2), (1.3) and (1.4). 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 [22]. 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 two doctoral assistants independently wrote computer programs in two different languages (Pascal and DAL) to construct the rule bases for the EXSPRT, and to carry out the four different adaptive testing methods on the same fifty-three 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.

It was discovered that on occasion the MISS procedure in the two programs would begin to select different items in the AMT model after fifteen to twenty 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 [35, 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 θ_c using the test characteristic curve [35, formula (24)] and the item parameter database constructed from the respective parameter estimation sample (either n = 25 or 50).

The value of θ_c was used as the initial prior θ and the prior variance was set to one, as recommended by Weiss and Kingsbury [16]. The MISS procedure was used to select the next test item for the re-enactment for each examinee [35, 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 θ and its variance was accomplished with Owen's method [38]. [See 35, formulas (13) to (22)]. After each item was "administered", the AMT stopping rules were applied using a .95 confidence interval [35, formulas (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 θ was greater than or equal to θ_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 β 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 rule bases constructed from the parameter estimation samples were used, of course, in the EXSPRT-R method of Bayesian updating (formula (2), with equal prior probabilities) and SPRT stopping rules (formulas (3.1) to (3.3). For a description of EXSPRT-R procedures, see [34] for 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 formulas (4) to (8)). 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 examinees 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 twenty-five and fifty examinees. EXSPRT rule bases 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 fifty there were twenty-three masters whose average test score was 87.3 percent, and twenty-seven nonmasters who scored 45.1 percent correct.

Mean test lengths of each of the four methods, variation in test lengths, and decision accuracies were compared. If the decision made by an adaptive method

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

Item Parameter Sample Description		Adaptive Testing Method			
	Mean Score (S.D.) *n*	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
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

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

*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.

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 twenty-eight nonmasters and twenty-five 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 twenty-five, compared to a mean of 18.3 for the sample of fifty. For the EXSPRT-I, 5.6 items were required for nonmastery decisions in the sample of twenty-five, compared to a mean of 5.9 for the parameter sample of fifty. Please note—and this is confusing—that the mean test lengths for each of the four methods are based on the same fifty-three test administrations, where all ninety-seven 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 rule bases were constructed for the EXSPRT-R and EXSPRT-I models.

Decision accuracies — For the fifty-three 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 [39]. 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.

When twenty-five 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 fifty 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 five and twenty items from the 97-item pool to reach those decisions—a very substantial reduction in test lengths (80 to 95% decrease).

Efficiency — A repeated measures ANOVA 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 fifty for the

fifty-three 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 for this kind of situation was conducted for all pair-wise contrasts of mean test lengths [40, pp. 373-381]. 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 sixty 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 eighty-five 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 $\alpha = .94$). 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.

Table 2. Efficiency and Accuracy of the Four Adaptive Testing Methods in the Second Study[a]

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

Results from the Second Study

Since there were more administrations of the COM test, parameter estimation samples of twenty-five, fifty, seventy-five and 100 were selected at random. Four sets of b_i coefficients were obtained for the AMT model and four rule bases were constructed for the EXSPRT models based on the same four parameter estimation samples. See the left sides of Table 2 for descriptive information about the parameter estimation samples.

Accuracy of predictions — When the parameter estimation sample was twenty-five, 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 rule base, so this is not surprising.

Table 2. (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
Masters	87.68	10.21	16.64	11.78	7.70
	(5.93)	(16.96)	(10.35)	(6.34)	(7.13)
	55	94.7	97.4	97.4	94.7
Nonmasters	56.00	28.93	16.29	14.75	7.82
	(10.17)	(28.58)	(16.96)	(15.54)	(4.85)
	20	82.1*	92.9	100.0	100.0
Total	79.23	15.25	16.55	12.58	7.73
	(15.84)	(22.21)	(12.39)	(9.71)	(6.57)
	75	91.3	96.2	98.1	96.2
Masters	87.47	13.75	16.97	13.58	7.64
	(6.33)	(20.80)	(10.75)	(9.51)	(6.12)
	75	93.4*	96.1	98.7	94.7
Nonmasters	56.00	31.50	13.04	12.32	8.93
	(11.34)	(29.77)	(10.66)	(10.78)	(7.41)
	25	78.6*	92.9	96.4	100.0
Total	79.60	18.53	15.91	13.24	7.99
	(15.77)	(24.70)	(10.82)	(9.83)	(6.48)
	100	89.4	95.2	98.1	96.2

[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.

*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 fifty, 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 seventy-five (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*.

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 was 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 quarter of Table 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 [40].

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 thirty-two 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.

DISCUSSION

Adaptive tests tended to be shorter with the DAL test than with the COM test (see Tables 1, and 2). Of the fifty-three administrations of the DAL test, there were twenty-eight nonmasters and twenty-five masters when the cut-off was set at 72.5 percent and when examinees answered all ninety-seven items. The overall average test score was 63.2 (S.D. = 24.6). In the second study with the COM test there were 104 administrations of this test, with seventy-six masters and twenty-eight nonmasters when the entire 85-item test was taken (grand mean = 79.0, S.D. = 13.6, mastery cut-off = 72.5 percent).

A similar study was conducted by Plew [12] with yet a different test on computer literacy (referred to as the LIT test here). In his sample of 183 examinees there were fifty-four masters and 129 nonmasters based on total test results from the fifty-five-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 Plew study, over two-thirds of the examinees were classified as nonmasters on the entire 55-item test [12]. The distribution of this group was close to normal, with the mean being about eight percentage points below the selected cut-off.

I 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 [35, 41]. 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 two empirical studies, as well as in Plew's.

The second factor that may affect 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 smaller item pools, as was the case with the LIT test in Plew's 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.

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 response functions (IRFs).

Normally an expert system consists of a set of questions and a knowledge base. An inference engine uses answers to the questions and the knowledge 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 IRFs. 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). Two 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 fifty 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.

Although further research is needed, the EXSPRT-I initially appears to be a strong alternative to both IRT- and SPRT-based adaptive testing when categorical decisions about examinee mastery of a single educational objective are desired. The EXSPRT-I is clearly less complex than IRT, both conceptually and mathematically. It also 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. It is important to note that 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, compared to IRT.

The reader is also reminded that EXSPRT was developed for criterion-referenced testing of single learning objectives. If assessment of multiple educational objectives is desired, then pools of test items can be developed to match each objective, respectively. The EXSPRT would then be applied to *each* objective individually. The overall outcome would not be a single test score but rather a list of objectives a student has mastered and not mastered.

The EXSPRT is expected to be most applicable to computer-based instruction, computer-managed instructional systems, personalized systems of instruction and other mastery learning contexts. The EXSPRT is not appropriate for norm-referenced testing or when it is desirable to obtain a precise estimate of achievement on some continuum.

Additional Research on the EXSPRT

While this article was under review additional studies of the EXSPRT were completed by several of my doctoral students. One of the limitations of the present study is that the sample for creating the item rule base was the same as that upon which the EXSPRT test re-enactments occurred. Cross-validation with a new sample would have provided greater assurance concerning the predictive validity of EXSPRT decisions. A second limitation of the present study is that because of the relatively small sample size it was not possible to compare EXSPRT to the two- and three- parameter IRT models. Hing-Kwan Luk recently completed a study in which these two limitations were addressed [28].

Luk compared the EXSPRT and the three-parameter IRT model by re-enacting two subtests of the College Entrance Examination Board Spanish Achievement Test: a 40-item listening subtest and a 60-item reading subtest. Results were available from 1672 students who took these tests to determine placement in Spanish courses at Indiana University. One thousand of these examinees were selected at random. Their data were used to estimate *a*, *b*, and *c* item parameters for the IRT model as well as for forming the rule quadruplets for the EXSPRT model. The sample used for cross-validation and test re-enactments consisted of the remaining 672 examinees.

A similar pattern of results obtained in Luk's study as in the present study. The EXSPRT-I tended to be much more efficient than the three-parameter IRT model when mastery decisions were reached. Surprisingly, the IRT model was significantly less accurate (73% and 83%) than the expected *a priori* theoretical rates (97.5%) for mastery decisions, whereas both the EXSPRT-I and -R models were within expected bounds on the two subtests. On the other hand, for nonmastery decisions average test lengths for all three approaches tended to be shorter than those for mastery decisions, with EXSPRT-I test lengths being shorter than AMT and EXSPRT-R. A further surprise was that both EXSPRT-I and -R resulted in nonmastery decision accuracies that were significantly below theoretical expectations (but still considerably higher than those for IRT-based *mastery* decisions). It is noteworthy that the EXSPRT decisions were correct 90 to 95 percent of the time, but less than the expected 97.5 percent accuracy. Luk concluded that the problems with decision accuracies in all three approaches may be due to the shape of the distribution of examinee scores in relation to the cutpoints chosen. Consistent with computer simulations by Frick [35] and a different college placement test studied by Plew [12], more decision errors were observed to occur

when examinees' scores were near the cutpoints on the two subtests in Luk's study [28].

In a different study Emily Powell was concerned that computerized adaptive tests might increase student test anxiety, which in turn might impair performance on CATs [24]. She measured student test anxiety and performance on students' *actual* exposure to EXSPRT-I, -R and -S (self-adaptive) methods. These were not test re-enactments from previously collected data as in Plew's, Frick's and Luk's studies. She found no significant differences in student performance levels across the three methods, nor in their anxiety levels. Mastery and nonmastery decisions reached by the three testing approaches and one-parameter IRT (Rasch) estimates of student achievement were highly correlated. She did find a significantly positive correlation between anxiety and performance for students experiencing the EXSPRT-I approach. That is, students who were more anxious tended to do better under EXSPRT-I conditions, and students who were less anxious did worse. Though not central to her study, Powell also found that test lengths for the EXSPRT-I were significantly shorter than those for EXSPRT-R.

One of the properties of rule-based (algorithmic) approaches such as the EXSPRT-I is that they are deterministic. That is, the item with the highest initial utility is first selected, and so forth as described above. What this means in practice is that a determinate pattern of items will be given for a particular examinee response pattern. The same problem also obtains in the IRT-based approach which uses the maximum information search and selection procedure, as well as any other algorithmic method. While the item sequences will differ for different examinee response patterns—as they should in an adaptive test—the door is open for possible student cheating. For example, if a pattern of all correct answers is given to the first five-ten questions, then a mastery decision is typically reached and that particular test is ended in the EXSPRT-I. While each question becomes successively more difficult as correct answers are given, it would not take long for students to discover the determinism inherent in both the EXSPRT-I and AMT with MISS. In effect, once the word was spread, students would only need to memorize the answers to a small number of test questions in order to achieve mastery. The validity of such results would be seriously jeopardized.[5]

One of my doctoral students, Susan Huang, has proposed a solution to the determinacy problem with rule-based item selection procedures. Her solution is the EXSPRT-RI method. When an examinee begins a test, items are selected at random (EXSPRT-R) until some predetermined minimum number have been chosen (e.g., ten questions). If a mastery or nonmastery decision cannot be confidently reached at that time, then the item selection procedure switches to the

[5] This problem does not occur in EXSPRT-R, since items are selected at random; hence the order is indeterminate.

intelligent selection method (EXSPRT-I) until a decision can be reached. In effect, her method is a compromise between EXSPRT-R and -I. With some items being chosen at random, the entire item pool is more likely to be better represented; and with some items being chosen intelligently, their difficulty levels are matched to estimates of student achievement while also maximizing item discrimination. In a preliminary study comparing EXSPRT-I, -R and -RI, Huang has found that average test lengths for the EXSPRT-RI method lie in between those for the -I and -R methods, with EXSPRT-I still resulting in the shortest adaptive tests. While these are only preliminary results with a relatively small sample of examinees, Huang's approach appears to be very promising.

Finally, since the EXSPRT is a relatively new approach to computerized adaptive testing, further research studies with a variety of test item pools and examinee groups are needed.

ACKNOWLEDGMENTS

Hing-Kwan Luk adapted and extended considerably the author's computer code for conducting the test re-enactments with the four methodologies. Luk also assisted with statistical analyses. Thomas Plew must be acknowledged for his significant contribution of an initial method of intelligent item selection for the EXSPRT, subsequently refined by the author. Had Tom not asked the questions he did in a doctoral seminar taught by the author in 1987, then EXSPRT might not exist today.

REFERENCES

1. P. Winston and K. Prendergast, *The AI Business: The Commercial Uses of Artificial Intelligence*, The MIT Press, Cambridge, 1984.
2. R. Davis, Amplifying Expertise with Expert Systems, in *The AI Business: The Commercial Uses of Artificial Intelligence*, P. Winston and K. Prendergast (eds.), The MIT Press, Cambridge, pp. 17-40, 1984.
3. G. Kearsley, *Artificial Intelligence and Instruction: Applications and Methods*, Addison-Wesley, Reading, Massachusetts, 1987.
4. D. Sleeman and J. Brown (eds.), *Intelligent Tutoring Systems*, Academic Press, New York, 1982.
5. W. Clancey, Methodology for Building an Intelligent Tutoring System, in *Artificial Intelligence and Instruction: Applications and Methods*, G. Kearsley (ed.), Addison-Wesley, Reading, Massachusetts, pp. 193-227, 1987.
6. R. Neapolitan, The Difference Between Uncertain and Approximate (Fuzzy) Inference, in *Fuzzy Logic for the Management of Uncertainty*, L. A. Zadeh and J. Kacprzyk (eds.), Wiley, New York, 1991, in press.
7. R. Neapolitan, *Probabilistic Reasoning in Expert Systems*, Wiley, New York, 1990.
8. B. Buchanan and E. Shortliffe, *Rule-based Expert Systems*, Addison-Wesley, Reading, Massachusetts, 1984.

9. G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, New Jersey, 1976.
10. R. Yager, S. Ovchinnikov, R. Yong and H. Nguyen (eds.), *Fuzzy Sets and Applications: Selected Papers by L. A. Zadeh*, Wiley, New York, 1987.
11. F. Lord and M. Novick, *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, Massachusetts, 1968.
12. G. T. Plew, *A Comparison of Major Adaptive Testing Strategies and an Expert Systems Approach*, doctoral dissertation, Indiana University Graduate School, Bloomington, Indiana, 1989.
13. R. Hambleton, Principles and Selected Applications of Item Response Theory, in *Educational Measurement*, R. L. Linn (ed.), Macmillan, New York, 1989.
14. F. Lord, *Applications of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum, Hillsdale, New Jersey, 1980.
15. V. Bunderson, D. Inouye and J. Olson, The Four Generations of Computerized Educational Measurement, in *Educational Measurement*, R. L. Linn (ed.), Macmillan, New York, 1989.
16. D. Weiss and G. Kingsbury, Application of Computerized Adaptive Testing to Education Problems, *Journal of Educational Measurement, 21*, pp. 361-375, 1984.
17. W. Sands and P. Gade, An Application of Computerized Adaptive Testing in U.S. Army Recruiting, *Journal of Computer-Based Instruction, 10*, pp. 87-89, 1983.
18. G. Kingsbury, E. Brzezinski and R. Houser, Computerized Adaptive Testing: A Four-Year-Old Pilot Shows that CAT Can Work, *Technological Horizons in Education Journal*, pp. 73-76, November, 1988.
19. J. Noonan and P. Sarvela, Implementation decisions in Computer-Based Testing Programs, *Performance and Instruction*, pp. 5-13, July, 1988.
20. H. Baghi, R. Gabrys and S. Ferrara, Applications of Computerized-Adaptive Testing in Maryland, paper presented at the annual meeting of the American Research Association, Chicago, April, 1991.
21. M. Reckase, Adaptive Testing: The Evolution of a Good Idea, *Education Measurement: Issues and Practice, 3*, pp. 11-15, 1989.
22. T. Frick, Bayesian Adaptation in Computer-based Tests and Computer-guided Practice Exercises, *Journal of Educational Computing Research, 5*:1, pp. 89-114, 1989.
23. T. Rocklin and A. O'Donnel, Self-adapted Testing: A Performance Improving Variant of Computerized Adaptive Testing, *Journal of Educational Psychology, 79*:3, pp. 315-319, 1987.
24. E. Powell, *Test Anxiety and Test Performance under Computerized Adaptive Testing Methods*, doctoral dissertation, Indiana University Graduate School, Bloomington, Indiana, 1991.
25. *ETS Developments*, Breakthrough Development in Computerized Testing Offers Shorter Tests, More Precise Pass-fail Decisions, Vol. *XXXIII*, pp. 3-4, Winter/Spring, 1988.
26. G. Dillon and L. Ross, The Effect of Targeting Test Questions at the Minimum Pass Point—Implications and Practical Considerations for a Large Scale Testing Program, paper presented at the annual meeting of the American Educational Research Association, Chicago, April, 1991.

27. S. Swinton and J. Hater, Block-Branching: Adaptive Mastery Testing for Multiple Cutpoints, paper presented at the annual meeting of the American Educational Research Association, Chicago, April, 1991.
28. H. K. Luk, An Empirical Comparison of an Expert Systems Approach and an IRT Approach to Computer-Based Mastery Testing, paper presented at the annual meeting of the American Educational Research Association, Chicago, April, 1991.
29. J. Heines, Basic Concepts in Knowledge-based Systems, *Machine-Mediated Learning, 1*:1, pp. 65-95, 1983.
30. B. Wright, Solving Measurement Problems with the Rasch Model, *Journal of Educational Measurement, 14*:2, pp. 97-116, 1977.
31. R. Mager, *Measuring Instructional Intent*, Fearon, Belmont, California, 1973.
32. S. Schmitt, *Measuring Uncertainty*, Addison-Wesley, Reading, Massachusetts, 1969.
33. A. Wald, *Sequential Analysis*, Wiley, New York, 1947.
34. T. Frick, Analysis of Patterns in Time: A Method of Recording and Quantifying Temporal Relations in Education, *American Educational Research Journal, 27*:1, pp. 180-204, 1990.
35. T. Frick, A Comparison of Three Decisions Models for Adapting the Length of Computer-based Mastery Tests, *Journal of Educational Computing Research, 6*:4, pp. 479-513, 1990.
36. T. Frick, The Indiana Testing System (ITS, Version 1.0), Department of Instructional Systems Technology, School of Education, Indiana University, Bloomington, Indiana, 1986.
37. R. Mead, B. Wright and S. Bell, BICAL (Version 3), Department of Education, University of Chicago, Chicago, 1979.
38. R. J. Owen, A Bayesian Sequential Procedure for Quantal Response in the Context of Adaptive Mental Testing, *Journal of the American Statistical Association, 70*, pp. 351-356, 1975.
39. G. Glass and K. Hopkins, *Statistical Methods in Education and Psychology*, Prentice-Hall, Englewood Cliffs, New Jersey, 1984.
40. L. Marascuilo and J. Levin, *Multivariate Statistics in the Social Sciences: A Researcher's Guide*, Brooks/Cole, Monterey, California, 1983.
41. T. Frick, H.-K. Luk and N.-C. Tyan, *A Comparison of Three Adaptive Decision-making Methodologies Used in Computer-based Instruction and Testing*, Final Report, Proffitt Foundation, Indiana University School of Education, Bloomington, Indiana, 1987.

Direct reprint requests to:

Dr. Theodore W. Frick
Indiana University
Department of Instructional Systems Technology
School of Education
W. W. Wright Education Bldg.
Bloomington, IN 47405