J. EDUCATIONAL COMPUTING RESEARCH, Vol. 5(1) 89-114, 1989

BAYESIAN ADAPTATION DURING COMPUTER-BASED TESTS AND COMPUTER-GUIDED PRACTICE EXERCISES

THEODORE W. FRICK
Indiana University

ABSTRACT

One of the potential advantages of computer-based instruction (CBI) is individualization of instruction. However, this goal has not been fully realized in practice, due largely to limitations of natural language understanding and to combinatorial explosion. It is nonetheless possible to develop CBI programs which can adapt to students, depending on their performance, by adjusting the length of computer-guided practice exercises and computer-based tests. The validity of this approach is supported empirically. The number of questions can be significantly reduced for many individuals, while mastery and nonmastery decisions remain highly accurate.

In this article I demonstrate how Bayesian reasoning can be used to adjust the length of computer-guided practice exercises and computer-based tests, when the goal is to make mastery or nonmastery decisions. Next, results of an empirical study are presented which support the validity of this approach. Finally, an extension of this approach is considered when previous empirical information about the questions themselves is available.

THE PROBLEM OF INDIVIDUALIZATION IN COMPUTER-BASED INSTRUCTION

One of the potential advantages of computer-based instruction (CBI) is individualization of instruction. Individualization implies tailoring the events of instruction to fit the particular circumstances of a given student. Yet most extant CBI is only minimally adaptive, if at all [1]. For example, most computer-based tutorials are basically linear, where all students follow essentially the same

path of instruction, or they are menu driven, where students are free to choose which section to do next. In the latter case, once a section is chosen from a menu, the section itself will usually be linear in nature—though students may be able to page forwards and backwards within the section or escape to a menu. Alternatively, a section of questions or problems may be presented according to level of difficulty, or in a random order for the sake of variety.

It is true that minimal but useful adaptation can be achieved by anticipating specific incorrect responses to questions and providing contingent feedback or hints to help a student answer correctly. Remediation strategies may also be employed for anticipated misunderstandings, with additional practice provided for areas of particular difficulty. It is also true that many computer games and simulations will usually behave differently when user input is varied.

These latter two approaches to CBI tend to be more adaptive than computer-based tutorials, drills, and tests, since algorithms underlying games and simulations normally utilize a set of variables whose values depend on student performance and which affect the program's behavior.

However, two interrelated problems currently prevent CBI programs from fully adapting to students during instruction as do human teachers: 1) computer programs presently cannot be written which adequately understand natural language; and 2) the possible number of instructional paths necessary to individualize instruction for a wide variety of learners results in a combinatorial explosion of instructional frames to be developed.

Natural language understanding is currently a significant obstacle to intelligent CBI, though some success has been achieved in very narrow and well-defined content domains [2-5]. Thus, the CBI designer presently faces the practical constraint that a limited number of presentation, question-and-feedback, practice, and testing frames can be developed for each instructional objective. The developer is also constrained by the type of student responses that can be adequately judged by a CBI program—i.e., alternative-choice and brief constructed responses. It is currently not possible, due primarily to the problem of natural language understanding, for a CBI program to react intelligently to *unanticipated* responses and questions asked by students.

A CURRENTLY ACHIEVABLE FORM OF ADAPTATION IN CBI

In what other ways, then, can a CBI program realistically adapt to individual differences in students? One practical way a CBI program can adapt is by deciding, depending on student performance, *when* to terminate a mastery test, a set of embedded questions, or a drill/practice session. All students need not be presented with all questions in order to reach a conclusion about whether or not their learning is sufficient. Those who perform extremely well or extremely poorly *early* in the situation are likely to continue doing so, and thus the situation

may be ended without requiring them to do all items. Those who have done well can move on to the next objective. Those who have done poorly can repeat instructional sections or seek additional assistance outside the CBI program.

There are various ways that such mastery or nonmastery decisions can be made [6-10]. These methods use statistical decision algorithms, and result in shortened tests or practice situations for many individuals. After an item is presented to a student and his or her response is evaluated as either correct or not, the evidence collected thus far is considered in light of some criteria of acceptability. If these criteria are not met, the test or practice situation continues with the selection and administration of a further question. If the criteria are met, the situation is terminated with a mastery or nonmastery decision. Most of these decision methodologies are essentially Bayesian in nature, but differ in the specific ways that evidence is weighed and combined with prior information and in methods of item selection.

A straightforward and practical application of Bayesian reasoning is presented below. When combined with additional rules for choosing one outcome over another, it will be shown that Bayes' Theorem can be extended to become, in essence, a sequential probability ratio test (SPRT). The SPRT was developed by Wald [11], though not explicitly presented in a Bayesian framework at that time, nor developed for criterion-referenced testing *per se*.

AN EXAMPLE OF BAYESIAN REASONING DURING TESTING

Suppose that there is a pool of test items which match a particular instructional objective and that our goal is to decide whether or not a student has mastered that objective [12]. A test question matches an objective if the performance required to answer correctly is the same as that stated in the objective. This implies that one who has mastered the objective is very likely to answer such a question correctly, whereas one who has not mastered the objective is much less likely to answer correctly. Suppose further that past experience with the test item pool has revealed that students who are indeed masters of the objective tend to answer 85 percent of the questions correctly and that those who are nonmasters score an average of 40 percent.

We can express our knowledge by means of "If . . . , then . . ." rules:

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

   Rule 1A: Prob(Correct|Master) = .85
   Rule 1B: Prob(Incorrect|Master) = .15

2. If the student is a nonmaster, then the probability of *selecting* a question that will be answered correctly is .40.

Rule 2A: Prob(Correct|Nonmaster) = .40
Rule 2B: Prob(Incorrect|Nonmaster) = .60

The alternative choices here are mastery and nonmastery, and the aim is to collect information by administering test items to decide which of these alternatives is most likely to be true for a particular student at some point in time.

Suppose further that we have no prior information about a particular student when the test is begun. With no prior information, the alternatives are equally likely—i.e., Prior Prob(Mastery) = Prior Prob(Nonmastery) = 0.50. We now proceed by making observations in this simulated example.

Observation 1

We randomly select a test item for measuring mastery of the current instructional objective and give it to this student, who answers it incorrectly. Using Bayes' Theorem, the posterior probabilities of the alternatives for this student are derived as follows [13]:

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

It can be seen that the prior probability of each alternative is multiplied by the respective probability of the observation, given that the alternative is true.[1] Multiplying the prior probability by the conditional probability results in a joint probability for each alternative. Normalization of the joint probabilities yields the posterior probabilities of the alternatives. Normalization is accomplished by summing the joint probabilities and then dividing each by that sum. Thus, the sum of the posterior probabilities is always equal to one. At this point, following the observation of an incorrect response, the probability of nonmastery is .80, four times as likely as mastery.

Observation 2

Suppose that we continue observing by randomly selecting another question and administer it to this student, who correctly answers it. Bayes' formula is reapplied, only this time the above posterior probabilites of the alternatives become the new prior probabilities:

[1] Since the student missed the question, we use the conditional probabilities from rules *1B* and *2B*. If answered correctly, we would have used rules *1A* and *2A*—see Observation 2.

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

Note that in the third column the conditional probability of selecting a question that is answered *correctly* is used this time for each alternative (rules *1A* and *2A*—since we just observed a correct response). Further note that now the odds of nonmastery to mastery have dropped to about two to one, having observed one correct and one incorrect response thus far.

Observation 3

We continue observing by selecting and administering another item, which this student answers incorrectly. As before, we update by using the most recent posterior probabilities of the alternatives as the new priors:

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

At this point the odds of nonmastery to mastery are about eight to one, given the observation of two questions answered incorrectly and one correctly. One might wonder how many observations are necessary to confidently choose one of the alternatives. This problem will be addressed below. For the time being, let us continue making observations.

Observation 4

Another test question is sampled, and again the student for whom we are attempting to decide either mastery or nonmastery answers incorrectly. As before:

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

Now the nonmastery alternative is about thirty times as likely as the mastery alternative. Should we stop? Let us take one more observation, the reason for which will become evident subsequently.

Observation 5

A further test item is randomly selected and administered. The student in question also misses this one. We update using Bayes' Theorem as before:

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

At this point the nonmastery alternative is highly probable (.992), about 125 times as likely as the mastery alternative. Recall that we began with a flat prior distribution, meaning that each alternative was equally likely. After making five sequential observations, where four questions were answered incorrectly and one correctly, the posterior probabilities of the alternatives have changed markedly through successive applications of Bayes' Theorem–i.e., through Bayesian reasoning.

This Bayesian approach to updating probabilities of discrete alternatives on the basis of new evidence appears to be so simple and straightforward that one may immediately wonder what assumptions are necessary to use it appropriately.

BAYES' THEOREM

If two or more discrete alternatives (A_i's) are mutually exclusive and exhaustive, if we know or can estimate the prior probability of each alternative $[P_0(A_i)]$, and if we have made a new observation (X) and know the conditional probability of that kind of observation under each alternative $[P(X|A_i]$, then we can determine the posterior probability of each alternative $[P(A_i|(X)]$ according to Bayes' Theorem [13]:

$$P(A_i|X) = \frac{P_0(A_i)P(X|A_i)}{\Sigma_j P_0(A_j)P(X|A_j)} \qquad (1)$$

Formula (1) expresses more compactly what was illustrated in the above examples in a tabular format.

Bayes' Theorem can also be used when there are more than two discrete alternatives. It can further be used when the alternatives are continuous. The basic concept obtains but the mathematics are more complicated, and requires the combination of a continuous prior probability distribution with a current observation or set of observations to yield a continuous posterior probability distribution. And, instead of referring to the probability of a discrete alternative, we indicate the probability of a range of alternatives.

It also follows that the posterior probability of an alternative is *proportional to* its prior probability multiplied by the likelihood of the observation if the alternative is true. Thus, if we are taking a sequence of observations, it is not necessary to normalize to obtain the posterior probabilities until after the last observation is made. After the five observations in the above example, the posterior probability of the mastery alternative is *proportional to* .5 X .15 X .85 X .15 X .15 X .15, which is equivalent to $.5 \times .85^1 \times .15^4$, or .0002151. Similarly, the posterior probability of the nonmastery alternative in the above example is *proportional to* $.5 \times .40^1 \times .60^4$, or .02592. We can then normalize after the last observation by dividing .0002151 by the sum (.0002151 + .02592), yielding a posterior probability for the mastery alternative which is the same as after Observation 5 above (= .008). The normalized posterior probability for the nonmastery alternative is .02592/(.0002151 + .02592) = .992, as before. To summarize, using a tabular format:

Alternative	*Prior Probability of Alternative*		*Probability Sequence|Alternative*		*Joint Probability*		*Posterior Probability*
Mastery	.5	X	$(.85^1)(.15^4)$	=	.0002151	/Sum =	.008
Nonmastery	.5	X	$(.40^1)(.60^4)$	=	.0259200	/Sum =	.992
				Sum =	.0261351		

If the observations are independent, then the posterior probability of an alternative is proportional to the prior probability multiplied by the quantity, $[p^r(1-p)^w]$: where p is the probability of selecting a question that will be answered correctly under the alternative, r is the number of questions answered correctly (right), and w is the number answered incorrectly (wrong). In this context, observations are independent if the probability of *selecting* an item which will be answered correctly by a student, when an alternative is true, does not differ depending on which questions may have been answered previously. This is often referred to as the assumption of *local independence* and will be discussed in greater detail below.

THE SEQUENTIAL PROBABILITY RATIO TEST (SPRT)

Abraham Wald developed the SPRT as a decision methodology for choosing between two discrete hypotheses when observations are made sequentially [11]. This method has been widely applied in quality control of manufactured

products—e.g., to decide whether to accept or reject a batch of goods. The major advantage of the SPRT is a significant reduction in the average sample size necessary to reach a decision, compared to conventional statistical tests which are performed after a group of observations have been obtained. The SPRT is applied after each observation. Sampling ends in the SPRT when one of the alternatives can be chosen with a level of confidence that depends on *a priori* decision error rates. The SPRT decision rules are as follows:

Rule S1. If the ratio of the posterior probabilities of the alternatives is greater than or equal to $(1-\beta)/\alpha$, then choose the first alternative.

Rule S2. If the ratio of the posterior probabilities of the two alternatives is less than or equal to $\beta/(1-\alpha)$, then choose the second alternative.

Rule S3. If neither rule *S1* nor *S2* is true, then conduct another observation, update the posterior probabilities, and apply the three rules again.

The decision error α is the probability of choosing the first alternative when the second alternative is really true, and conversely for β.

In the context of trying to reach a mastery or nonmastery decision during a computer-based mastery test, a drill/practice exercise, or a set of embedded questions in a tutorial, the three decision rules can be stated more explicitly:

After randomly selecting a question and observing the student's response, we first calculate the probability ratio, *PR*:

$$PR = \frac{P_{om}P_m{}^r(1-P_m)^w}{P_{on}P_n{}^r(1-P_n)^w} \qquad (2)$$

P_{om} and P_{on} are the *initial* prior probabilities of mastery and nonmastery, respectively. Note that if the initial prior probabilities of the alternatives are equal, as Wald implicitly assumed, then they cancel each other out; thus, *PR* is simply the ratio of the probabilities of the sequence of observations under the two alternatives. P_m is the probability of selecting a test question that would be answered correctly if the mastery hypothesis is true, whereas P_n is the probability of selecting a question that would be answered correctly under the nonmastery hypothesis. The exponents r and w refer to the number of right and wrong answers observed thus far.

Rule S1'. If $PR \geqslant (1-\beta)/\alpha$, then discontinue observations and choose the mastery hypothesis.

Rule S2'. If $PR \leqslant \beta/(1-\alpha)$, then take no more observations and accept the nonmastery hypothesis.

Rule S3'. If $\beta/(1-\alpha) < PR < (1-\beta)/\alpha$, then randomly select another question, administer it to the student, increment r or w depending on whether it is answered correctly or not, re-calculate *PR*, and apply rules *S1'* to *S3'* again.

One might legitimately wonder if a test could continue indefinitely, given the iterative nature of the SPRT—i.e., rule *S3'*. Although Wald mathematically proved that the SPRT will terminate, in practice it is true that, with a relatively small number of test questions, the pool could possibly be depleted before a mastery or nonmastery decision is reached. The number of items necessary to reach a decision will depend on how a particular student performs and the above rules. The rules in turn depend on our willingness to make decision errors, the prior probabilities of the alternatives, and the probability of selecting a test question that would be answered correctly under each alternative. If there is a clear trend toward either mastery or nonmastery *early* in the sequence of observations, a decision can often be reached with a relatively small number of randomly selected questions. If the trend is less clear early in the sequence, then more test questions will be needed to reach a decision.

AN EXAMPLE OF THE SPRT

Before beginning the SPRT decision process, we need to specify our tolerances for decision errors. Suppose that we are willing to falsely decide mastery when someone is indeed a nonmaster 2.5 percent of the time (thus $\alpha = .025$). Similarly, suppose we are willing to falsely decide nonmastery when someone is in reality a master at the same error rate ($\beta = .025$). In the SPRT the lower bound for the mastery threshold, *LBM*, is $(1-\beta)/\alpha$, or $(1-.025)/.025 = 39$. When the posterior odds of the mastery vs. the nonmastery alternative become equal to or greater than 39 to 1, then we stop testing and choose mastery. Similarly, the upper bound for the nonmastery threshold, *UBN*, is $\beta/(1-\alpha)$, or $.025/(1-.025) = .025641$. When the posterior odds of mastery vs. nonmastery become less than or equal to .025641 to 1, then we terminate the test and choose nonmastery. Otherwise, if the probability ratio lies between .025641 and 39, then we randomly sample another test item, update the posterior probabilities and check the ratio against *LBM* and *UBN* once again.

SPRT Stage 1

Assuming equal prior probabilities, the ratio of the posterior probabilities after an incorrect answer to the first randomly selected question is .20/.80 = .25 (from Observation 1 above). Since .25 lies between .025641 and 39, we continue.

SPRT Stage 2

After the next question is given, which is answered correctly, the ratio of the posterior probabilities, *PR*, is now .347/.653 = .531 and still remains between the *LBM* and *UBN*.

SPRT Stage 3

The third randomly selected question is answered incorrectly, resulting in posterior odds of .117 to .883, or a *PR* of .1325, which is neither less than or equal to .025641 or greater than or equal to 39. Thus, we make no decision and continue testing.

SPRT Stage 4

The next question is also answered incorrectly, and now *PR* is .032/.968 = .033, approaching but still greater than the *UBN*.

SPRT Stage 5

The fifth randomly selected question is also missed, resulting in posterior probabilities of mastery and nonmastery of .008 and .992, respectively. *PR* is now .008/.992 or .0081, which is less than .025641, the threshold for nonmastery decisions. Thus, we stop sampling test items and conclude nonmastery for this particular examinee. We would expect in the long run to be wrong 2.5 percent of the time, since we set *a priori* our β level at .025 (i.e., falsely concluding nonmastery when the examinee was in fact a master).

In this example, only five randomly sampled test items were required to make a decision, given rules *1A, 1B, 2A, 2B, S1'*, *S2'* and *S3'* above, formula (2), $\alpha = \beta = .025$, and the examinee response pattern: wrong, right, wrong, wrong, wrong. The number of test items necessary to reach a conclusion with the SPRT will vary depending on: the probabilities of selecting a question a master or nonmaster would answer correctly, specified decision error rates, and the observed pattern of examinee answers during the test.

Probably the best way to understand how all these factors interact is to conduct a computer simulation and observe the number of items necessary to reach an SPRT decision while systematically varying the parameters. Generally, one will find in such a simulation that *fewer* test items are required to reach decisions when the gap between the probabilities in rules *1A* and *2A* is greater, or when decision error rates are higher. The converse obtains as well. These results should not be surprising given the formulation of the SPRT. Moreover, nonmastery decisions tend to be reached more quickly than mastery decisions when a pattern of mostly incorrect answers is observed, compared to a pattern of mostly correct ones–holding all other factors constant at what would be considered typical levels for mastery testing situations.

ASSUMPTIONS NECESSARY TO USE THE SPRT

Random Sampling without Replacement

It is assumed that, when testing a particular examinee, test questions are sampled at random from the available pool of questions matching the instructional objective and that no question is asked more than once. The reason for this assumption is that for any given student we want to make a generalization about how he or she would do if the entire universe of test questions were administered. That is, given a sample drawn from the universe of items, we make an inductive inference about the proportion of correct answers that would be given if the whole universe of test items were administered to this individual at this particular time. Random selection is one way to help guarantee that samples of test items given are representative of the total pool.

Local Independence

All major extant adaptive testing strategies, including the SPRT, assume independence of observations [14, 15]. This means that the probability of a correct response to any given item should not change depending on the order in which items are administered. This is required by probability theory in order to calculate a joint probability, as is done in Bayes' Theorem. In practice, if items are selected at random, no feedback is given during the test, and students are not permitted to return to previous items and change their answers, then this assumption should be generally met–though it could be empirically tested.

On the other hand, in a CBI drill/practice situation it is usually a good idea to give immediate feedback to a student after answering each question or group of questions. Moreover, item queuing strategies may be employed, such as interspersing a previously missed question several times subsequently, or moving from easier to harder questions–and vice-versa–as the situation dictates [16]. Under these conditions, the assumptions of local independence and random sampling without replacement could be violated to the extent that decisions rendered are not valid statistically. However, drill/practice is usually considered a stage of the instructional and learning process, in contradistinction to a criterion-referenced test which might be given before and after instruction. In a drill/practice situation it would appear that the SPRT could still be used to reach decisions, though it might err more often than would be expected by the specified α and β error rates. This may be of little consequence, compared to the consequences of decision errors in more formal testing situations.

CRITICISMS OF THE SPRT

Unaccounted Variation in Item Difficulties

The SPRT has been criticized as being inappropriate for test item pools where items vary in difficulty level, particularly when a rather precise estimate of a student's level of achievement is desired [8, 15]. For instance, the probability of a correct response to one item by a master might be different than the probability of answering correctly another item. However, *average* probabilities are used in rules 1 and 2 above. If one had item difficulty data on all test items for masters and nonmasters, respectively, then the known estimates of probabilities of correct responses could actually be used in the Bayesian updating process. In other words, we would have a separate rule pair for each test item for masters and nonmasters, respectively. Depending on which item was selected, the corresponding rule pair would be applied in estimating the new posterior probabilities. Though this method would be more complex to carry out in practice, it does not differ in principle from what was illustrated above.[2] In fact, Reckase has extended this idea of using item parameters with the SPRT and assumptions from item response theory (IRT) [8]. Instead of using a separate rule pair for masters and nonmasters for each item, an estimated item response function in the form of a logistic ogive is used to predict the probability of a correct response to the item in relation to an underlying achievement continuum.

If *expected* probabilities are used instead of specific item difficulty levels at each stage of Bayesian updating, then the basic issue is the *representativeness of the sampled items* with respect to the universe of items. We know from sampling theory that more precise estimates of measures can be made as sample size increases. Here the universe of generalization is the mastery status of an examinee at some point in time. Since the SPRT can terminate rather quickly when initially there is a clear trend towards one alternative or the other, sampling error could cause false decisions to be made more often than expected by the *a priori* decision error rates. Thus, if average probabilities of a correct response for masters vs. nonmasters are used with the SPRT when item parameters vary widely, then it should be used conservatively, so a test will not end too quickly. This can be accomplished—in theory—by choosing very low decision error rates.

[2] The author is currently investigating such an approach, termed EXSPRT. Preliminary results indicate that, with a database of at least fifty test administrations for rule construction, the EXSPRT is more efficient than both the basic SPRT discussed here and an item-response theory (IRT) approach. The EXSPRT appears to be more accurate than the SPRT and as accurate as IRT-based mastery testing.

AN EMPIRICAL STUDY OF THE VALIDITY OF THE SPRT

Although the SPRT has been used widely as a decision methodology in quality control settings, few references to the SPRT have been found in the educational and psychological testing literature. Ferguson used the SPRT in an individually prescribed instruction (IPI) framework [6]. Kingsbury and Weiss, Millman, McArthur and Chou, and Reckase have explored the use of the SPRT in computer-based mastery testing [7, 8, 15, 17].

The major criticism of the SPRT, as discussed above, is that it does not explicitly account for variability in test item parameters. It is true that the SPRT, in its original formulation, makes no provision for the fact that some items may provide more information than others. The question is raised: To what extent can the SPRT make correct mastery and nonmastery decisions when test items vary widely in terms of difficulty level, discriminatory power, and chances of guessing correctly? In other words, how robust is the SPRT decision model? Since no empirical studies addressing this question have been found, a study was undertaken.

Computer-based Tests

Two computer-based tests were developed for empirically investigating the robustness of the SPRT: 1) a test on the structure and syntax of the Digital Authoring Language, referred to as the DAL test, and 2) a test of knowledge of how computers functionally work, called the COM test. Test items relevant to these respective content domains were constructed so that difficulty levels would be expected to vary. About half the items on each test were multiple-choice, one-fourth binary-choice, and the remainder short-answer. Subsequent item analyses indicated that items did vary considerably in difficulty levels and discriminatory power (see Appendix).

The DAL test consisted of ninety-seven items and was found to be highly reliable (coefficient alpha = .98). The COM test contained eighty-five items and was also very reliable (coefficient alpha = .94). These reliability coefficients were based on results from the two groups described below.

Examinees

The examinees who took the DAL test were mostly either current or former graduate students in a course on computer-based instruction taught by the 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 the Digital Authoring Language (which they were required to learn in

order to develop CBI 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 known to be long and difficult, no one was asked to take the test who did not have some knowledge of DAL or other CBI authoring languages.

About two-thirds of the students who took the COM test were current or former graduate students in two sections of an introductory course on using computers in education, also taught by the author. Current students took the COM test as both a pre- and posttest. The remaining one-third were undergraduate students taking a beginning course in instructional computing and took the test once.

Though examinees were not chosen randomly, the timing of testing and other prior indications of their knowledge in these two content areas helped insure that there would be fairly wide ranges of scores on both tests. Almost all examinees had some first-hand experience with computers prior to testing and, with few exceptions, did not appear to be intimidated by using a computer terminal or appear to be especially anxious.

Procedure

Tests were individually administered by the Indiana Testing System [18]. 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.[3] Examinees were not allowed to go back and change previous answers to items, nor was feedback given during the test. When the test was finished, 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 (right or wrong). Examinees were informed of their total test scores at the end of the test. The COM test typically took about thirty to forty-five minutes to complete, whereas the DAL test usually took between sixty and ninety minutes.

There were 105 administrations of the COM test and fifty-three of the DAL test. The DAL test was generally perceived as a very difficult test, with a mean score of 63 percent (S.D. = 24.6). The COM test was easier on the whole, with a mean of 79 percent correct (S.D. = 13.6).

Method of Determining SPRT Decision Outcomes

The SPRT parameters were set *a priori* as follows: mastery level = .85, nonmastery level = .60, $\alpha = \beta = .025$. The SPRT was applied *retroactively*, since each student was originally given all the items in a pool. A computer program was written which retrieved test results for each examinee. Since the order of the randomly selected items for an examinee was stored in the database, it was possible to retroactively apply the SPRT after each item record was retrieved from the database for that individual. Items were retrieved in the order in which they were originally administered; the correctness of the student's response was noted; the number of right or wrong answers was incremented accordingly; and the SPRT was applied after each item record was examined. As soon as a mastery or nonmastery decision could be reached via the SPRT, it was recorded in a separate file, along with the number of items needed for the SPRT decision, the total test score, and administrative identification information. This process was repeated for all examinees for both tests.

Indeed, this *post hoc* process of determining and recording SPRT decisions is no different than if it were accomplished in real time during testing. It was more expedient to do this afterwards with a separate retrieval program than to modify the existing record keeping software in the testing system.

[3] Due to a minor oversight, only ninety-six items were administered on the DAL test.

Results

The mean numbers of items required for SPRT mastery and nonmastery decisions for both tests are reported in Table 1. It can be seen that about twenty items were required to reach mastery decisions and a few less for nonmastery decisions. Had the SPRT been actually used during testing, this would have resulted in savings of significant amounts of student test taking time, since about one-fourth to one-fifth of the total item pool was needed on the average for SPRT mastery decisions.

How well did the SPRT decisions predict the decisions reached on the basis of total test scores? This required classification of total test scores as either mastery or nonmastery. Several different methods were investigated for determining mastery status on the total item pool (see [19]). Based on a suggestion by Jason Millman (personal communication), it was decided that the simplest and most valid method was to choose the mid-point between the mastery and nonmastery level, and if a total test score was at or above the mid-point (72.5 percent correct), classify the student as a master. Otherwise, students whose total test scores were below the mid-point were classified as nonmasters. While it is true that some misclassifications would be expected to occur for examinees whose total scores were near the mid-point, it is also true that, in normal nonadaptive mastery testing, decisions are reached in just this manner—i.e., choosing a single cut-off and proceeding as above.

The agreement between SPRT mastery decisions and those decisions reached from total test scores was extremely high (see Table 2). Fifty-one out of fifty-three SPRT decisions for the DAL test were consistent with total test decisions based on all ninety-six items. In one case where the SPRT ended with a nonmastery decision, the total test decision was for mastery. In another case,

Table 1. Descriptive Statistics on the DAL and COM Test Decision Outcomes for Both the SPRT Condition and Conventional Total Test Results[a]

	SPRT		*Conventional*		
	Mastery	*Nonmastery*	*Mastery*	*Nonmastery*	*Total*
DAL Test					
N (Administrations)	24	29	25	28	53
Mean Percent Correct (S.D.)	92.5 (8.3)	38.5 (22.0)	89.0 (8.1)	46.3 (15.8)	63.2 (24.6)
Mean Number of Items (S.D.)	19.1 (12.9)	17.4 (16.3)	96 –	96 –	96 –
COM Test					
N (Administrations)	76	29	77	28	105
Mean Percent Correct (S.D.)	90.5 (6.9)	44.0 (23.1)	87.7 (6.4)	56.2 (10.8)	79.0 (13.6)
Mean Number of Items (S.D.)	21.6 (12.9)	18.6 (16.3)	85 –	85 –	85 –

[a] The DAL test consisted of ninety-seven items on the structure and syntax of the Digital Authoring Language (r_{xx} = .98). Due to a minor oversight, only ninety-six items were administered. The COM test consisted of eighty-five items on how computers functionally work (r_{xx} = .94). Items on both tests were presented in a different random order for each examinee.

the SPRT was unable to reach a decision before the DAL test item pool was exhausted.

For the COM test, 104 out of 105 SPRT decisions agreed with total test decisions. The single misclassification was the same type as with the DAL test, a mistaken SPRT nonmastery decision.

Across both tests, the SPRT accurately predicted total test decisions in 155 out of 158 cases–about 98 percent agreement. Expected agreement was 95 percent, since the overall rate was set *a priori* at 5 percent (the sum of α and β). Thus, in this study the SPRT made fewer classification errors than were expected.

Based on these results, it would appear that the SPRT is a fairly robust model for mastery decisions, even when test items vary considerably in difficulty level and discriminatory power. The criticism of the SPRT on theoretical grounds as being invalid for test item pools with varying parameters was not supported empirically in the present study. The SPRT does appear to predict well, if it is used conservatively–i.e., keeping α and β relatively small–and if test item pools are highly reliable as were the two pools used in this study. These results are reminiscent of those of the robustness of analysis of variance (ANOVA) when the homogeneity of variance assumption is violated.

Table 2. Agreement between SPRT Decision Outcomes and Those from Conventional Total Test Results

		Conventional	
		Mastery	*Nonmastery*
	DAL Test		
SPRT	Mastery	23	0
	Nonmastery	1	28
	No Decision	1	0
		Percent Agreement = .96	
	COM Test		
SPRT	Mastery	76	0
	Nonmastery	1	28
	No Decision	0	0
		Percent Agreement = .99	
	Both Tests		
SPRT	Mastery	99	0
	Nonmastery	2	56
	No Decision	1	0
		Percent Agreement = .98	

Moreover, the few times the SPRT did err in its predictions, it erred in the conservative direction–classifying someone as a nonmaster who turned out to be a master. Not once in this study did the SPRT mistakenly classify someone as a master who turned out to be a nonmaster.

DISCUSSION

Although other Bayesian procedures have been used for making mastery decisions during CBI and testing, they tend to be considerably more complicated than the straightforward Bayesian sequential probability ratio test (SPRT) [9, 15].

Use of the Beta Distribution

The procedure adopted by Tennyson and his associates utilizes *continuous* prior and posterior beta distributions and a single cut-off [9]. It does not take into account item parameter information, and can be criticized on the same theoretical grounds as the SPRT. Moreover, since continuous distributions are used, numerical integration methods are necessary to find the areas of the posterior

beta distributions (and hence probabilities) on one side of the cut-off or other. To obtain reasonable degrees of accuracy, a large number of iterations is necessary for the numerical integration—so many, in fact, that significant delays will occur on current microcomputers, making it impractical for real time testing. This necessitates choosing a cut-off and building in advance a table of probabilities of mastery for each numerical combination of right and wrong answers. This table can then be included in the CBI program or stored as a file to be accessed by the program. While this works for relatively small numbers of items, significant amounts of computer memory or disk storage are required for larger item pools.[4]

In short, the Bayesian approach adopted by Tennyson, et al., lacks flexibility and economy, since a different table must be accessed every time a different cut-off is chosen. On faster minicomputers and mainframes, probabilities can be calculated fairly accurately in real time without significant delays (e.g., less than a second), and the storage/flexibility issue becomes moot.

A further problem when using a posterior beta distribution is that for typical cut-offs, incorrectly answering the first few randomly selected items can immediately lead to a nonmastery decision—i.e., termination of the test. Thus, more Type II decision errors may occur when items vary considerably in difficulty levels.

Adaptive Mastery Testing

Another approach, adaptive mastery testing (AMT), holds considerable promise for computer-based tests for large numbers of people [10]. This AMT approach is based on the item response theory (IRT) initially developed by Lord and Novick [20]. Item parameter information is used in this approach, a distinct advantage over the SPRT. Moreover, when a test is being administered, items are selected to match as closely as possible the curent estimate of an examinee's achievement level, while at the same time maximizing item discrimination and minimizing guessing. In practice, this means that if an examinee misses an item, a slightly easier one is presented next, and conversely for harder items. An empirical study has shown that the AMT approach can yield fairly precise estimates of an examinee's achievement level with significant reductions in test length, compared to conventional fixed-length tests [10].

The major disadvantage of the AMT approach is the number of examinees necessary for good estimation of item parameters, which must be done *before* AMT is actually used. For the one-parameter model (item difficulty only), a minimum of 200 examinees are required. For the two- and three-parameter models, 500 and 1000 examinees are needed, respectively (Weiss, personal communication).[5] Therefore, the AMT approach is not practical for many CBI and testing applications, where it would be difficult or unrealistic to test such large numbers of persons in advance.

[4] For example, approximately 40,000 bytes would be required for 100 items for *each* prespecified cut-off, assuming four bytes per floating point entry in the table.

[5] See also [14, 21-23].

SPRT

On the other hand, the SPRT is reasonably simple and economical to implement on any computer system. Less than twenty-five lines of code in most high level programming and authoring languages is required (see formula [2] above, and rules $S1'$ to $S3'$). Moreover, since the computational overhead is small, SPRT decisions can be reached in a fraction of a second. Thus, the major advantage of the SPRT is its practicality for use in adapting CBI and testing.

The SPRT, like any statistic, can be used improperly. As discussed above, the danger in using the SPRT with item or question pools which vary widely in difficulty level and/or discriminating power is reaching a mastery or nonmastery decision before a *representative* sample of items has been administered to an examinee. It could happen, just by chance, that very difficult questions were sampled early, resulting in a premature and incorrect nonmastery decision (and vice-versa). To guard against this possibility, it was recommended that the SPRT be used with small α and β levels (Type I and II decision error rates). The present study indicated that SPRT decisions are highly accurate when *a priori* error rates are kept small (i.e., .025).

Furthermore, the nonmastery level should be set to a level higher than that which would be expected by guessing alone for pools with multiple-choice items. Preferably, both mastery and nonmastery levels should be established on the basis of past experience with an item pool, using the average proportions of correct responses for masters and nonmasters, respectively.

Two Cut-offs Versus One

The fact that the SPRT in effect requires two cut-off levels may cause problems for those accustomed to a single cut-off. However, when a single cut-off is used, it is known that misclassifications are likely to occur when examinees score near the cut-off [24]. In effect, there is a zone of uncertainty around the cut-off, where no decisions can be reached without significant risks of misclassification. Thus, the outer bounds of this zone of uncertainty effectively define two cut-off points, similar to the SPRT. The difference is that in the former situation, the zone of uncertainty will tend to shrink as the test length increases, all other things equal. The SPRT requires specification of the mastery and nonmastery levels *a priori*, which remain unchanged as an adaptive test progresses.

As was suggested in the above example, the mastery and nonmastery levels for the SPRT can be based on empirical results from prior use of the test item pool. Otherwise, following Wald's lead [11, p. 29], the levels can be established by answering two questions: 1) What is the highest proportion of correct responses

on the whole test *above* which we would *not* want to classify someone as a nonmaster? 2) What is the lowest proportion of correct responses on the whole test *below* which we would *not* want to classify someone as a master? The answer to the first question effectively determines the mastery level, and the second the nonmastery level. The area in between the levels was referred to by Wald as the "zone of indifference," comparable to the area of uncertainty surrounding a single cut-off.

The efficiency of the SPRT is also affected by the distribution of examinee achievement levels in relation to the zone of indifference. For example, if the average achievement level lies in the zone of indifference and the scores are normally distributed, then the SPRT will typically require more items to reach a decision than if they are distributed bimodally—as expected for pre- and post-testing occasions.

SUMMARY

It is not currently possible to develop computer-based instruction (CBI) that can adapt to individual differences in students as do human teachers. Given this limitation, it is still possible to adapt to students by adjusting the length of computer-guided practice exercises and computer-based tests. This was illustrated by a straightforward application of Bayes' Theorem. It was shown that the sequential probability ratio test (SPRT), originally devised by Abraham Wald, extends Bayes' Theorem by explicit consideration of decision error rates in choosing an alternative.

The SPRT has been largely ignored as a decision model for adapting computer-based instruction and mastery testing since it does not take into account variability in item difficulty, discrimination and guessing factors. More complex and presumably more accurate decision models have been developed which do account for item parameter variability. However, the utility of these models for many CBI contexts is questionable. The attractiveness of the SPRT is its relative simplicity and practicality for use in adapting CBI and testing. It was contended that, if the SPRT is used conservatively (i.e., with small decision error rates), it is a viable decision model for adapting CBI and tests.

An empirical study was undertaken to investigate the predictive validity of the SPRT for making mastery and nonmastery decisions from two pools of test items with varying parameters. One item pool contained ninety-seven items and the other eighty-five items. Test length was reduced considerably by use of the SPRT. Approximately twenty and eighteen items were required on the average to reach mastery and nonmastery decisions, respectively. Most importantly, decisions reached by the SPRT agreed very highly with those reached from administration of the entire item pools to examinees. The SPRT predicted correctly in 155 out of 158 cases (98% agreement, when expected to be 95 percent according to *a priori* decision error rates). In the few cases where the SPRT erred, it failed to correctly classify students who were determined to be masters from their total test scores. In *no* case did the SPRT classify a student as a master who was subsequently determined to be a nonmaster.

On the basis of these data, it would appear that the SPRT is a viable decision model for CBI and testing, if used conservatively when tests are suspected or known to contain items of varying difficulty and discriminatory power.

One could conclude, as did one reviewer of this paper, that if reliable mastery and nonmastery decisions could be made with about twenty items, then perhaps fewer test items were needed in the first place. That is, why not construct a 20-item test and simply use it, instead of randomly selecting from a much larger pool of items? While this solution is certainly more practical, it should be noted that there was considerable variation around the mean of twenty items. The standard deviations were between 14 and 25. For example, if the first four items were missed, a nonmastery decision was rendered. In other cases, forty to fifty questions were required to reach a decision with the SPRT parameters used in this study. In one case, the SPRT could not reach a mastery or nonmastery decision with a 96-item pool.

The clear advantage of the Bayesian approach is that a test or practice exercise is no longer than necessary for a given individual. At the same time accuracy of mastery decisions is not sacrificed, as would be more likely to occur with a fixed-length test of twenty items.

A second concern is that the Bayesian approach investigated here does not use information related to item characteristics such as difficulty and discrimination. If such information is available about test items or practice questions, then it certainly makes sense to use it—not only for selecting items but also in updating Bayesian posterior probabilities. This is precisely what is done by the adaptive mastery testing (AMT) procedure developed by Weiss and his associates. The disadvantage of this and other IRT-based approaches is that a large number of test administrations are necessary for adequate item parameter estimation (200-1000).

The author is currently investigating an extension of the SPRT (EXSPRT) when prior information is available about the proportion of masters and nonmasters, respectively, who have correctly answered each test item in the pool. Based on a sample of fifty test administrations, preliminary results indicate that the EXSPRT is even more efficient than the SPRT. Using random selection of items, the EXSPRT needed ten to twelve items on the average to reach a decision with no more errors than expected. If items were instead selected in terms of their difficulty, discriminatory power and compatibility with an examinee's estimated achievement level, then the EXSPRT became even more efficient. In other words, more difficult questions were chosen for examinees predicted to be masters, and easier questions were chosen for those expected to be nonmasters, while at the same time considering those items which maximally discriminated between masters and nonmasters. With this quasi-intelligent selection

procedure the EXSPRT required an average of six or seven items to reach a decision, while still remaining highly accurate. On the other hand, the one-parameter AMT model required between fifteen and twenty items on the average to reach a decision. The reader is cautioned that these EXSPRT results are tentative at this time.

It appears that the Bayesian method of adjusting the number of questions that is common to both the SPRT and EXSPRT has considerable practical merit in computer-based instruction. Computer-guided practice exercises and computer-based tests can be adapted to students depending on individual performances. It is also apparent that adaptation can be further enhanced if prior empirical information is available on the questions used.

Finally, the approach discussed here is comparable to the kind of reasoning used in some expert systems inference engines [c.f., 25]. Although the method of obtaining information from users clearly differs here from that in a typical expert system, the fundamental decision methodology is essentially the same—as those who have developed expert systems without the aid of shells may recognize. Whether we call this method a component of an expert systems approach or a decision-support system is less important than the Bayesian reasoning it entails and its demonstrated utility in computer-based instruction.

APPENDIX

Item analyses were performed on two tests: 1) the DAL test—on knowledge of the syntax and structure of the Digital Authoring Language ($n = 53$); and 2) the COM test—on knowledge of how computers functionally work ($n = 105$). Classical item analyses were first performed. A one-parameter (Rasch) model was also used to estimate item difficulty levels. Two- or three-parameter models were not used due to relatively small sample sizes. In the tables that follow the below notation is used:

p_{i+} = proportion of examinees who answered item i correctly.
r_{it} = correlation of scores on item i with total test scores.
b_i = difficulty level estimated by the Rasch model for item i.
$S.E._i$ = standard error of estimate of difficulty for item i.

DAL TEST

Item	p_{i+}	r_{it}	b_i	$S.E._i$	Item	p_{i+}	r_{it}	b_i	$S.E._i$
1	.89	.51	−1.89	.49	8	.64	.62	.16	.36
2	.77	.46	−.73	.39	9	.42	.61	1.65	.36
3	.66	.51	.03	.36	10	.70	.34	−.23	.37
4	.89	.33	−1.89	.49	11	.72	.43	−.36	.37
5	.77	.57	−.79	.39	12	.79	.65	−.95	.40
6	.57	.41	.65	.35	13	.91	.50	−2.15	.52
7	.53	.68	.90	.35	14	.60	.54	.41	.35

DAL Test (Cont'd)

Item	p_{i+}	r_{it}	b_i	$S.E._i$	Item	p_{i+}	r_{it}	b_i	$S.E._i$
15	.42	.72	1.65	.36	57	.83	.57	−1.28	.42
16	.23	.53	3.10	.41	58	.77	.47	−.79	.39
17	.55	.72	.78	.35	59	.62	.48	.29	.36
18	.87	.34	−1.67	.46	60	.91	.41	−2.15	.52
19	.55	.51	.78	.35	61	.68	.62	−.10	.36
20	.36	.60	2.05	.37	62	.72	.31	−.36	.37
21	.45	.73	1.40	.36	63	.68	.63	−.10	.36
22	.73	.51	−.50	.38	64	.66	.77	.03	.36
23	.68	.44	−.10	.36	65	.60	.72	.41	.35
24	.66	.75	.03	.36	66	.91	.49	−2.15	.52
25	.81	.35	−1.11	.41	67	.72	.61	−.36	.37
26	.68	.57	−.10	.36	68	.74	.50	−.50	.38
27	.57	.57	.66	.35	69	.94	.26	−2.81	.65
28	.91	.48	−2.15	.52	70	.58	.55	.53	.35
29	.81	.47	−1.11	.41	72	.47	.27	1.27	.35
30	.83	.43	−1.28	.42	73	.53	.80	.90	.35
31	.57	.28	.66	.35	74	.38	.74	1.91	.37
32	.89	.31	−1.89	.49	75	.55	.69	.78	.35
33	.81	.35	−1.11	.41	76	.51	.69	1.03	.35
34	.68	.32	−.10	.36	77	.68	.39	−.10	.36
35	.81	.44	−1.11	.41	78	.57	.71	.66	.35
36	.91	.41	−2.15	.52	79	.64	.45	.16	.36
37	.45	.65	1.40	.36	80	.79	.56	−.95	.40
38	.72	.49	−.36	.37	81	.81	.56	−1.11	.41
39	.45	.47	1.40	.36	82	.47	.50	1.27	.35
40	.85	.56	−1.47	.44	83	.62	.62	.29	.36
41	.89	.56	−1.90	.49	84	.79	.23	−.95	.40
42	.85	.51	−1.47	.44	85	.60	.62	.41	.35
43	.47	.67	1.27	.35	86	.53	.69	.90	.35
44	.57	.51	.66	.35	87	.40	.67	1.78	.36
45	.87	.36	−1.67	.46	88	.40	.70	1.78	.36
46	.64	.43	.16	.36	89	.51	.70	1.03	.35
47	.70	.73	−.23	.37	90	.49	.79	1.15	.35
48	.49	.69	1.15	.35	91	.52	.80	.90	.35
49	.81	.30	−1.11	.41	92	.57	.60	.66	.35
50	.60	.74	.74	.35	93	.43	.71	1.52	.36
51	.51	.72	1.02	.35	94	.55	.50	.78	.35
52	.60	.71	.41	.35	95	.66	.52	.03	.36
53	.58	.53	.53	.35	96	.64	.56	.16	.36
54	.85	.51	−1.47	.44	97	.55	.46	.78	.35
55	.79	.39	−.95	.40	98	.28	.53	2.62	.39
56	.60	.61	.41	.35					

COM Test

Item	p_{i+}	r_{it}	b_i	S.E.$_i$	Item	p_{i+}	r_{it}	b_i	S.E.$_i$
1	.65	.53	1.05	.24	44	.92	.31	−1.18	.39
2	.78	.49	.26	.26	45	.89	.43	−.78	.34
3	.98	.30	−2.71	.73	46	.72	.51	.71	.25
4	.87	.38	−.47	.31	47	.78	.35	.33	.26
5	.76	.64	.40	.26	48	.84	.41	−.11	.29
6	.87	.26	−.47	.31	49	.82	.49	−.03	.28
7	.77	.22	.33	.26	50	.69	.68	.88	.24
8	.91	.26	−.90	.36	51	.72	.49	.71	.25
9	.85	.44	−.28	.30	52	.81	.27	.12	.27
10	.74	.58	.52	.25	53	.97	.30	−2.28	.60
11	.89	.49	−.78	.34	54	.73	.47	.59	.25
12	.89	.35	−.78	.34	55	.85	.39	−.28	.30
13	.93	.26	−1.33	.41	56	.81	.33	.12	.27
14	.70	.22	.82	.24	57	.63	.41	1.21	.23
15	.89	.23	−.78	.34	58	.56	.45	1.57	.23
16	.88	.29	−.56	.32	59	.84	.45	−.20	.29
17	.88	.48	−.67	.33	60	.80	.42	.19	.27
18	.85	.52	−.20	.29	61	.91	.29	−.90	.36
19	.87	.59	−.37	.31	62	.94	.28	−1.33	.41
20	.65	.33	1.10	.23	63	.96	.23	−1.72	.48
21	.79	.10	.19	.27	64	.88	.28	−.56	.32
22	.77	.41	.40	.26	65	.64	.48	1.10	.23
23	.92	.26	−1.03	.37	66	.82	.62	−.03	.28
24	.86	.63	−.28	.30	67	.56	.41	1.62	.23
25	.88	.47	−.56	.32	68	.66	.33	.99	.24
26	.82	.59	−.03	.28	69	.63	.55	1.26	.23
27	.81	.51	.04	.28	70	.51	.56	1.81	.22
28	.93	.57	−1.33	.41	71	.74	.45	.52	.25
29	.50	.39	1.91	.22	72	.73	.31	.58	.25
30	.81	.53	.04	.28	73	.24	.29	3.31	.26
31	.90	.43	−.90	.36	74	.88	.29	−.67	.33
32	.80	.45	.12	.27	75	.91	−.18	−.90	.36
33	.67	.39	.99	.24	76	.79	.57	.26	.26
34	.83	.14	−.11	.29	77	.64	.04	1.10	.23
35	.43	.26	2.26	.23	78	.66	.48	.99	.24
36	.90	.48	−.90	.36	79	.83	.33	−.11	.29
37	.83	.63	−.11	.29	80	.82	.50	.04	.28
38	.81	.69	.12	.27	81	.84	.32	−.20	.29
39	.98	.10	−2.71	.73	82	.73	.28	.46	.26
40	.94	.43	−1.51	.44	83	.50	.39	1.91	.22
41	.89	.28	−.78	.34	84	.84	.21	−.11	.29
42	.92	.50	−1.18	.39	85	.72	.38	.71	.25
43	.87	.33	−.47	.31					

REFERENCES

1. O.-C. Park, R. Perez, and R. Seidel, Intelligent CAI: Old Wine in New Bottles, or a New Vintage?, in *Artificial Intelligence and Instruction: Applications and Methods*, G. Kearsley (ed.), Addison-Wesley, Reading, Massachusetts, pp. 11–45, 1987.
2. M. Arbib, *Computers and the Cybernetic Society*, Academic Press, Orlando, Florida, pp. 178–239, 1984.
3. 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.
4. M. Minksy, The Problems and the Promise, in *The AI Business: The Commercial Uses of Artificial Intelligence*, P. Winston and K. Prendergast (eds.), The MIT Press, Cambridge, Massachusetts, 1984.
5. D. Sleeman and J. Brown (eds.), *Intelligent Tutoring Systems*, Academic Press, New York, 1982.
6. R. Ferguson, *Computer-Assisted Criterion-Referenced Measurement*, University of Pittsburgh Learning Research and Development Center, Pittsburgh, 1969.
7. D. McArthur and C.-P. Chou, *Interpreting the Results of Diagnostic Testing: Some Statistics for Testing in Real Time*, University of California Center for the Study of Evaluation, Los Angeles, 1984.
8. M. Reckase, A Procedure for Decision Making Using Tailored Testing, in *New Horizons in Testing*, D. Weiss (ed.), Academic Press, New York, pp. 238–256, 1983.
9. R. Tennyson, D. Christensen, and S. Park, The Minnesota Adaptive Instructional System: An Intelligent CBI System, *Journal of Computer-Based Instruction*, *11*, pp. 2–13, 1984.
10. D. Weiss and G. Kingsbury, Application of Computerized Adaptive Testing to Educational Problems, *Journal of Educational Measurement*, *21*, pp. 361–375, 1984.
11. A. Wald, *Sequential Analysis*, Wiley, New York, 1947.
12. R. Mager, *Measuring Instructional Intent*, Fearon-Pittman, Belmont, California, 1973.
13. S. Schmitt, *Measuring Uncertainty*, Addison-Wesley, Reading, Massachusetts, 1969.
14. R. Hambleton and L. Cook, Latent Trait Models and Their Use in the Analysis of Educational Test Data, *Journal of Educational Measurement*, *14*:2, pp. 75–96, 1977.
15. G. Kingsbury and D. Weiss, A Comparison of IRT-Based Adaptive Mastery Testing and a Sequential Mastery Testing Procedure, in *New Horizons in Testing*, D. Weiss (ed.), Academic Press, New York, pp. 257–283, 1983.
16. S. Alessi and S. Trollip, *Computer-Based Instruction: Methods and Development*, Prentice-Hall, Englewood Cliffs, New Jersey, 1985.
17. J. Millman, Passing Scores and Test Lengths for Domain-Referenced Measures, *Review of Educational Research*, *43*:2, pp. 205–216, 1973.

18. T. Frick, The Indiana Testing System (ITS, Version 1.0), Department of Instructional Systems Technology, School of Education, Indiana University, Bloomington, 1986.
19. ____, An Investigation of the Validity of the Sequential Probability Ratio Test for Mastery Decisions during Criterion-Referenced Testing, paper presented to the American Educational Research Association, April 16, 1986.
20. F. Lord and M. Novick, *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, Massachusetts, 1968.
21. R. Hambleton and L. Cook, Robustness of Item Response Models and Effects of Test Length and Sample Size on the Precision of Ability Estimates, in *New Horizons in Testing*, D. Weiss (ed.), Academic Press, New York, pp. 31–50, 1983.
22. R. Hambleton, H. Swaminathan, L. Cook, D. Eignor, and J. Gifford, Developments in Latent Trait Theory: Models, Technical Issues, and Applications, *Review of Educational Research*, *48*:4, pp. 467–510, 1978.
23. F. Lord, Small *n* Justifies Rasch Model, in *New Horizons in Testing*, D. Weiss (ed.), Academic Press, New York, pp. 52–62, 1983.
24. M. Novick and C. Lewis, Prescribing Test Length for Criterion-Referenced Measurement, American College Testing Program, *Technical Bulletin No. 18*, Iowa City, 1974.
25. J. Heines, Basic Concepts in Knowledge-Based Systems, *Machine-Mediated Learning*, *1*:1, pp. 65–95, 1983.

Direct reprint requests to:

Dr. Theodore W. Frick
School of Education
W. W. Wright Education Building
3rd and Jordan
Bloomington, IN 47405