J. EDUCATIONAL COMPUTING RESEARCH, Vol. 6(4) 479-513, 1990

A COMPARISON OF THREE DECISION MODELS FOR ADAPTING THE LENGTH OF COMPUTER-BASED MASTERY TESTS

THEODORE W. FRICK
Indiana University, Bloomington

ABSTRACT

Three extant methods of adapting the length of computer-based mastery tests are described and compared: 1) the sequential probability ratio test (SPRT), 2) Bayesian use of the beta distribution, and 3) adaptive mastery testing based on item response theory (IRT). The utility of the SPRT has been empirically demonstrated by Frick [1]. Research has also demonstrated the effectiveness of use of the beta function in the Minnesota Adaptive Instructional System by Tennyson et al. [2]. Considerably more empirical research has been conducted on IRT-based approaches [3]. No empirical studies were found in which these three approaches have been directly compared. As a first step, computer simulations were undertaken to compare the accuracy and efficiency of these approaches in making mastery and nonmastery decisions. Results indicated that the IRT-based approach was more accurate when simulated examinee ability levels were clustered near the cut-off. On the other hand, when ability levels were more widely dispersed—as would likely be the case in pre- and posttest situations in mastery learning—all three approaches were comparably accurate. While the IRT approach tended to be the most efficient, it is the least practical to implement in typical classroom testing situations.

OVERVIEW

Making the Future

Alan Kay advocates that "the best way to predict the future is to make it."[1] Concern has grown during the last decade about the quality of American education. Computers and related information technologies could play a significant role in how education is restructured.

[1] This was the focus of a speech to invited educators and other leaders at the Children's Museum in Indianapolis in April, 1988.

The present author envisions a future educational system that is carried out in a very individualized, self-paced and enjoyable manner. All students would have opportunities to learn, and learn successfully. They could exercise choice of learning activities in educational environments designed and continually modified by their teachers. There would be numerous self-contained learning activities that do not rely on teachers as the major disseminators of information and feedback. Instructional technologies such as computer-based tutorials, simulations, hypermedia and interactive video would be commonplace. Teachers would be very important, but their roles different—more as guides and confidants rather than governors. They could spend more time listening to and advising individual students and less time directing large groups.

To realize this dream, several things must fall into place. First, there must be a critical mass of effective, but primarily self-contained, learning activities mediated by appropriate technologies. Second, there must be educational systems structured or organized so that learning environments as described above can be realized. Third, there must be some system of educating educators to gain the competencies and confidence to work in these new environments.

If instruction is individualized in such a manner, then practical methods of assessing individual student progress will be needed. The educational practice of moving cohorts of students through the same learning activities and tests at the same time—the norm at present—can be discarded. Instead of grading students where a minority succeed (i.e., receive A's) and the majority do not, a mastery learning approach can be adopted. Mastery learning is not practical with teacher-led cohorts, but could become the norm in the individualized, technologically mediated education system envisioned.

This article addresses the issue of how student mastery can be assessed by computer software for those cognitive learning objectives amenable to such an approach. The approaches discussed are *not* appropriate for assessing student competence, for example, in playing a musical instrument, writing an essay, or riding a bicycle. The approaches are appropriate for many kinds of cognitive learning objectives traditionally assessed by paper-and-pencil tests.

Research on three extant methods of computerized adaptive mastery testing is first discussed. Each method is then examined in detail. Examples are provided to illustrate how each decision model actually works.[2] Next, the three methods are compared by two Monte Carlo simulation studies. Finally, recommendations are made on appropriate contexts for use of these methods of adaptive mastery testing.

[2] The author believes that this level of detail is necessary for two reasons. First, several researchers have made inadvertent mistakes in correctly carrying out two of the models as indicated in the next section. Second, these details will help other researchers to implement correctly and further study these adaptive methods. The author originally wrote this article for his doctoral seminar on methods of adapting computer-based instruction and tests. Students are required to write computer simulations to carry out these and other methodologies. The numerical examples are used to confirm initially the correctness of their computer code.

Introduction to Computerized Adaptive Mastery Testing

In a mastery learning context a teacher typically needs to verify that a student has indeed mastered an instructional objective, normally after the student has engaged in learning activities relevant to the objective. If mastery is evident, then the student is presumably ready to move on to the next unit of instruction. If not, then repetition of the current unit or some kind of remediation is pursued.

When the mastery assessment can be made by a conventional or computer-based test consisting of a set of questions appropriate for measuring achievement of the instructional objective, a cut-off score is typically chosen; and if the student's test score is at or above the cut-off, a mastery decision is rendered. Novick and Lewis [4] and Millman [5] have demonstrated, however, that this traditional method of deciding mastery is particularly error prone when obtained student scores are near the cut-off and tests are relatively short (e.g., 12 items or less).

Nonetheless, if mastery tests can be administered and adequately scored by a computer program, it is possible to adapt the *length* of such tests on an individual basis, depending on a particular student's ongoing performance during the test. If there is a clear trend towards mastery or nonmastery early in the test, it can be terminated without administering all items. When the trend is less clear, longer tests are required. The conclusion reached from the shortened test will tend to agree with that reached from a longer, fixed-length conventional test [1, 3]. Significant savings of test-taking time can be achieved with little or no loss of decision accuracy [6, 7].

The decision model developed by Weiss and Kingsbury [3], referred to as Adaptive Mastery Testing (AMT), is based on the item response theory (IRT) from Lord and Novick [8], and a Bayesian scoring algorithm developed by Owen [9]. While the AMT model has excellent predictive validity when compared to longer, fixed-length tests [3], it is not without its drawbacks. The AMT model requires empirically derived item parameter information, which is obtained by administration of the item pool to a large sample of examinees (approximately 1000 for the three-parameter model). Item parameters are then estimated from these data *before* the AMT model can actually be used for real-time testing decisions. For most teachers, this is an unrealistic demand. Consider, for example, a college professor who teaches 200 students a year and has developed a mastery test for one of the learning units in a course. She would need to collect test data for about five years before the three-parameter AMT model could be legitimately employed.

What is required, therefore, is a practical alternative decision methodology which—although not as accurate or as efficient as the AMT model—is nonetheless sufficient for instructional decisions concerning mastery. In an empirical study of the predictive validity of the sequential probability ratio test (SPRT), Frick found that an average of approximately twenty test items were required to reach mastery and nonmastery decisions on two different computer-based tests [1]. On each test, items were selected at random from a larger pool and varied considerably in difficulty levels and discriminatory power. Decision accuracy of the SPRT was 98 percent, when compared to decisions based on total test results (155 out of 158 SPRT decisions were in agreement with total test results). The expected accuracy was 95 percent given the *a priori* α and β error rates used in the study. The SPRT model was not compared with the AMT model in the Frick study [1], since the sample sizes were not adequate for good item parameter estimation, required by the AMT approach.

The practical advantages of the SPRT decision model are that it is relatively simple to implement in a computer-based testing system, and it requires no prior data collection on test item parameters—though some would be desirable. An apparent disadvantage of the model is that the probability of selecting an item that will be answered correctly by masters is assumed to be the same for all items in the pool; and similarly for nonmasters. Reckase [10], Kingsbury and Weiss [6], and others question these assumptions—particularly when a rather precise estimate of an examinee's achievement level is desired. On the other hand, Frick [1] demonstrated empirically that, even when items vary considerably in difficulty and discriminatory power, this does not seem to affect the predictive validity of the SPRT when the model is used conservatively (i.e., α and β levels = .025) with a typical mastery and nonmastery level (.85 vs .60). These results are analogous to those from studies of the robustness of ANOVA when the assumption of homogeneity of variances is violated.

Before a further attempt to compare the SPRT and AMT models empirically, Frick et al. [7] conducted Monte Carlo computer simulations with the aim of first replicating the Kingsbury and Weiss [6] simulations of the two models.[3] Average test lengths and decision accuracies in the Frick et al. [7] simulations of the AMT model closely matched the Kingsbury and Weiss [6] results, when both studies used the same experimental conditions and item parameter data.[4]

However, the results for the SPRT in the Frick et al. simulation were, surprisingly, quite *different* than those in the Kingsbury and Weiss study [7]. The differences were much larger than what could be attributed to sampling error. After reverifying the correct functioning of the SPRT in Frick's code, further investigation revealed that the Kingsbury and Weiss formulation of the SPRT decision model could not be algebraically transformed into Wald's original formulation [11]. Further checking revealed that Reckase's, Schmitt's, and Frick's interpretation of Wald's SPRT were consistent and could be algebraically transformed into Wald's original formula [10-13]. Frick et al. [7] *were* subsequently able to replicate Kingsbury and Weiss's [6] SPRT simulation results by using their incorrect SPRT formula. This gave impetus to redo the comparative computer simulations of the two models, since the SPRT results in the Kingsbury and Weiss study were clearly invalid.

At the same time, a third decision model was studied when conducting the simulations with the correct SPRT formulation. Tennyson, Christensen and Park have attempted to use Bayesian posterior *beta* distributions for decision making during instruction (e.g., for deciding how many interrogatory examples to provide in a concept learning task) [2]. Tennyson et al. [2] based their model on the work of Novick and Lewis [4] in which a ratio of posterior probabilities is compared to a loss ratio (disutilities associated with incorrect decisions). Another serendipitous discovery was made by Frick et al. [7] when verifying the computer code for calculating areas (probabilities) under various posterior beta distributions. Posterior probabilities calculated from Frick's code agreed exactly with those published by Novick and Lewis ([4], Table 3) and by Schmitt ([12], Appendix B). However, the values in Table 1 in Tennyson et al. could not be reproduced [2]. While Tennyson and his associates have empirically demonstrated the effectiveness of the Minnesota Adaptive Instructional System (MAIS), it is not clear whether they are indeed using a model based on legitimate Bayesian posterior beta probability ratios or some other model.

Moreover, Novick and Lewis acknowledge the problem of establishing a basis for choosing a loss ratio in an educational setting [4]. Frick et al. suggested that the ratio of posterior beta probabilities can be compared to ratios involving type I and II error rates, as did Wald with the SPRT [7, 11]. This would also make possible a direct comparison of the SPRT, AMT and beta models using the same theoretical decision error rates.

Each of the models will be described in detail with numerical examples. Then results of computer simulations will be discussed, comparing the three models in terms of average test lengths and decision accuracies.

EXPLICATION OF THREE ADAPTIVE MASTERY TESTING METHODS

The Sequential Probability Ratio Test

Abraham Wald originally developed the sequential probability ratio test (SPRT). He was finally permitted to publish his work after World War II, when the U.S. government had declassified it. In essence, Wald showed that if sampling is done sequentially and his decision rules are applied after each observation, then approximately *half* as many observations are required on the average to reach a decision—when compared to conventional methods in which the sample size is fixed in advance and a statistical test is applied after all observations are made [11]. Furthermore, Wald demonstrated that no more type I and II decision errors

[3] The major reason for repeating Kingsbury and Weiss's simulations was to further validate the accuracy of the present author's computer code for carrying out the rather complex AMT model.

[4] Some minor differences in results were expected, due to variation attributable to sampling error that is intentionally part of the simulation.

are expected with the SPRT than with conventional methods. The SPRT has been widely used as a decision model for quality control in manufacturing.

Ferguson may have been one of the first to apply the SPRT to criterion-referenced testing in individually prescribed instruction [14]. Millman, Kingsbury and Weiss, Reckase, and McArthur and Chow have explored the SPRT as a decision model for mastery testing [5, 6, 10, 15]. The use of the SPRT in computer-based testing is apparently not widespread, at least as inferred from the few references found in the educational and psychological testing literature.

The SPRT model is elegantly simple. Information is collected in order to choose between two alternatives. In the case of criterion-referenced testing, the choices normally are mastery and nonmastery, and the information is the observed sequence of a particular student's correct and incorrect answers to test questions. Assuming that each observation can be dichotomously characterized and that random sampling without replacement occurs, then a probability ratio is computed after each observation (i.e., administration of a test item):

$$PR = \frac{P_m{}^s(1-P_m)^f}{P_n{}^s(1-P_n)^f} \qquad (1)$$

where s successes and f failures have been observed up to this point, and where P_m is the probability of selecting an item that a master would answer correctly, and P_n is the probability of selecting an item that a nonmaster would answer correctly.

Wald's three decision rules—in the context of mastery testing—would be as follows:

Rule 1.1. If $PR \geq (1-\beta)/\alpha$, then stop the test and conclude that the present student is a master. (2)

Rule 1.2. If $PR \leq \beta/(1-\alpha)$, then stop the test and conclude that the present student is a nonmaster. (3)

Rule 1.3. If $\beta/(1-\alpha) < PR < (1-\beta)/\alpha$, then randomly choose another test item, give it to the present student, recalculate PR, and apply the three rules again. (4)

The decision errors α and β are type I and II error rates. Alpha is the probability of misclassifying a true nonmaster as a master. Beta is the probability of misclassifying a true master as a nonmaster.

As an example, suppose that we had previously given our test item pool to a sample of students, and those who scored 75 percent or higher were considered masters, and the remainder nonmasters. The average test scores for masters and nonmasters were found to be 89 and 46 percent, respectively. Thus, for this test, the probability of selecting an item that would be answered correctly by a master, P_m, is estimated to be .89. Similarly, the probability of selecting a question that would be answered correctly by a nonmaster, P_n, is estimated to be .46. Suppose further that we are willing to make false mastery and nonmastery decisions no more than a combined 5 percent of the time ($\alpha = \beta = .025$).

Now, suppose that we have a particular student whose mastery status is unknown.We randomly select an item and give it to the student, who answers it correctly ($s = 1, f = 0$ at this time). We compute PR:

$$PR = \frac{.89^1(1-.89)^0}{.46^1(1-.46)^0} = \frac{.89}{.46} = 1.93$$

We apply the three rules, and rule 1.3 applies, since $.025/(1-.025) < 1.93 < (1-.025)/.025$, or $.0256 < 1.93 < 39.0$. We continue the test by randomly selecting another question, which this student answers incorrectly (now $s = 1, f = 1$).

$$PR = \frac{.89^1(1-.89)^1}{.46^1(1-.46)^1} = .394$$

Again, rule 1.3 is still true, since .394 lies between .0256 and 39.0. Another question is selected at random, and the student also misses this one ($s = 1, f = 2$).

$$PR = \frac{.89^1(1-.89)^2}{.46^1(1-.46)^2} = .081$$

Since $.0256 < .081 < 39$, we continue. The next randomly selected question is also missed (now $s = 1, f = 3$).

$$PR = \frac{.89^1(1-.89)^3}{.46^1(1-.46)^3} = .016$$

Rule 1.2 is now true, since $.016 < .0256$. We stop the test and conclude that this student is a nonmaster, with the expectation that we would be mistaken 2.5 percent of the time.

A further way of understanding the SPRT is to consider the binomial distribution when the population proportion is .89 and the sample size is four, as illustrated in Figure 1. The probability of observing one success in four trials is approximately .0047, when the underlying proportion of successes is assumed to be .89. On the other hand, if the underlying proportion of successes is .46, as shown in Figure 2, then the probability of observing one success in four trials is .2897. The ratios of these two probabilities is .0047/.2897 = .016, which is the same ratio as above. In fact, the SPRT is based on the binomial model, but since the normalization factors in the binomial model are the same in both the numerator and denominator of the probability ratio for a given number of successes and failures, those factors cancel each other out in the SPRT.

The number of test items required to make a decision will, of course, depend on a particular student's performance. As discussed earlier, Frick observed that about twenty items were required on the average on two different tests [1]. Choice of α and β levels will also affect test length—higher values will tend to result in shorter tests, but also with the expectation of more decision errors. Furthermore, if the gap

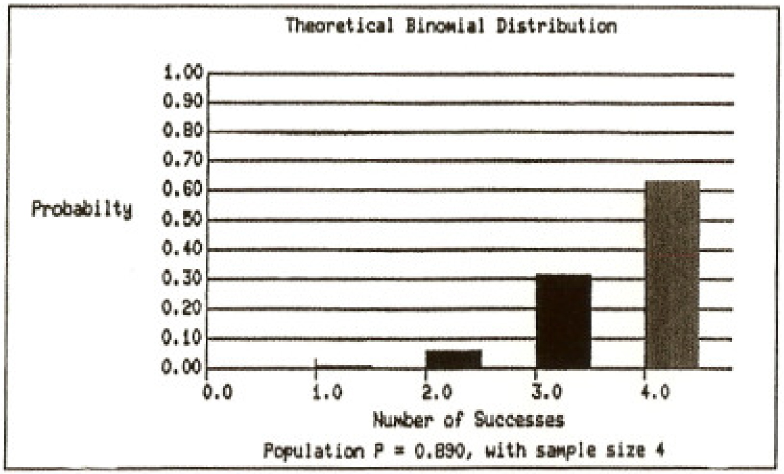


Figure 1. Binomial distribution with $p = .89$ and $n = 4$.

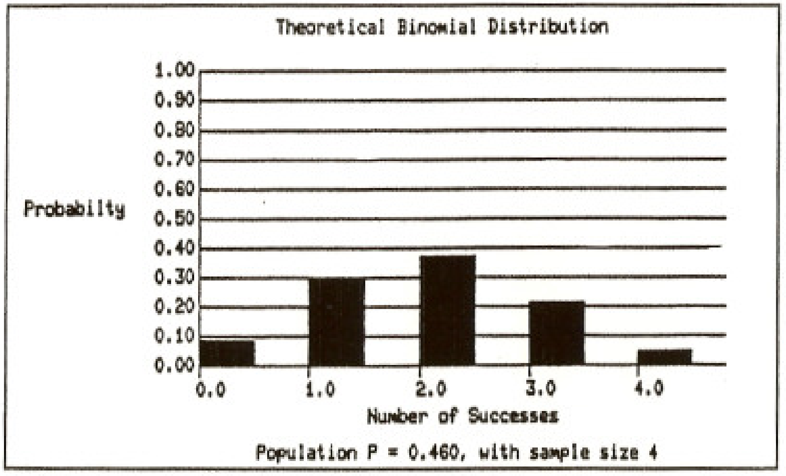


Figure 2. Binomial distribution with $p = .46$ and $n = 4$.

between the probability of selecting a test question that a master would answer correctly, compared to a nonmaster, is wider, then tests tend to be shorter—compared to narrower gaps.

Independence of observations is assumed in order to multiply the conditional probabilities in the numerator and denominator of *PR*. This means that the

probability of selecting an item that a master would answer correctly should not change depending on which items have been previously answered, and likewise for nonmastery. If no feedback is given to an examinee during a test and items are selected at random without replacement, then violation of this assumption is unlikely, though it could be empirically tested.

If the mastery and nonmastery levels are based on empirical results, as implied in the above example, then *average* item difficulties are used in updating *PR* for masters and nonmasters, respectively. This is where the SPRT has been criticized. If by chance a number of easy items were selected very early in a test, a premature and incorrect mastery decision might be rendered, and vice-versa for hard items. However, Frick contends that the basic issue is the *representativeness* of the sample of items chosen from the larger pool, and recommends that α and β be kept very small to prevent tests from being too short and decrease the likelihood that the sample will be unrepresentative. Empirical results support this contention [1].

Bayesian Posterior Beta Probabilities

Novick and Lewis explored the use of Bayesian posterior beta distributions in order to calculate the probability that Φ, an estimate of an examinee's true proportion correct, is greater than or equal to some prespecified cut-off (Φ_C), given an observed number of successes (s) and failures (f) [4]. Although Novick and Lewis were not concerned with interactively adapting the length of a test administered by a computer, Tennyson and his associates have attempted to do so with the beta distribution when assessing concept attainment during instruction (MAIS) [2].

The probability *density* function for the posterior beta distribution is defined [12]:

$$\text{beta}\,(\Phi \mid s,f) = \frac{(s+f+1)!}{s\,!f\,!}\,\Phi^{s}(1-\Phi)^{f} \tag{5}$$

This definition of the beta density assumes that current data are combined with a flat prior distribution [beta(Φ | 0,0)], and that s and f are positive integers greater than or equal to zero.[5] As Φ varies continuously from zero to one the beta distribution is defined. For example, the posterior beta density when $\Phi = .91$ and three successes and two failures have been observed, given a flat prior distribution, is:

$$\text{beta}(.91 \mid 3,2) = \frac{(3+2+1)!}{3!2!}\,.91^3(1-.91)^2 =$$

$$\frac{(6\times 5\times 4\times 3\times 2\times 1)}{(3\times 2\times 1)\,(2\times 1)}\,(.753571)\,(.0081) = .3662355$$

[5] Zero factorial (0!) is defined to be equal to one in this context.

However, we are not really interested in the probability density for some value of Φ but rather the posterior probability that Φ lies in some range [e.g., prob ($\Phi \geq .85$)]. This requires numerical integration methods in order to estimate such a probability. Simpson's rule will be used in the example below (cf., [12], Appendix A). In effect, the part of the distribution of interest is cut into many narrow intervals or slices and the areas of slices are added together to estimate the total area under that portion of the beta curve. The beta function is defined such that the entire area under a curve between zero and one is equal to one, which is the probability that [$0.0 \leq \Phi \leq 1.0$].

For example, suppose that we have observed three successes and two failures thus far in the test, and we are interested in the probability that Φ is greater than or equal to .85. We need to find the area under the beta curve between .85 and 1. According to Simpson's rule, we need to divide the range of interest into an *even* number of intervals and calculate the value of the beta function [beta (Φ) | 3,2] at the point where each interval begins. Then we weight the first and last probability density by one, and alternately weight by four and two the intermediate densities. For example, we will divide the area into ten equally spaced intervals beginning with .85 and ending with 1.0 (the width of each interval is (1 – .85)/10 = .015):

Φ	Beta Density	×	Simpson Weight	=	Product
.850	.829		1		.829
.865	.708		4		2.831
.880	.589		2		1.178
.895	.474		4		1.897
.910	.366		2		.732
.925	.267		4		1.068
.940	.179		2		.359
.955	.106		4		.423
.970	.049		2		.099
.985	.013		4		.052
1.000	.000		1		.000
			Sum of products =		9.468

[Area = Sum × Width/3 = Probability = (9.468) (.015)/3 = .047]

We then sum the products (9.468), multiply the sum by the width of the interval (.015), and always divide by three. Thus, the probability that [$\Phi \geq .85$] is approximately .047, which is derived from the numerical integration of [beta(Φ)| 3,2] from $\Phi = .85$ to $\Phi = 1.0$.

The entire beta distribution for three successes and two failures is plotted in Figure 3 (from $\Phi = 0$ to $\Phi = 1$ on the horizontal axis, and where the beta density is the vertical axis). It can be seen that a very small portion of the area under the curve lies between .85 and 1.0.

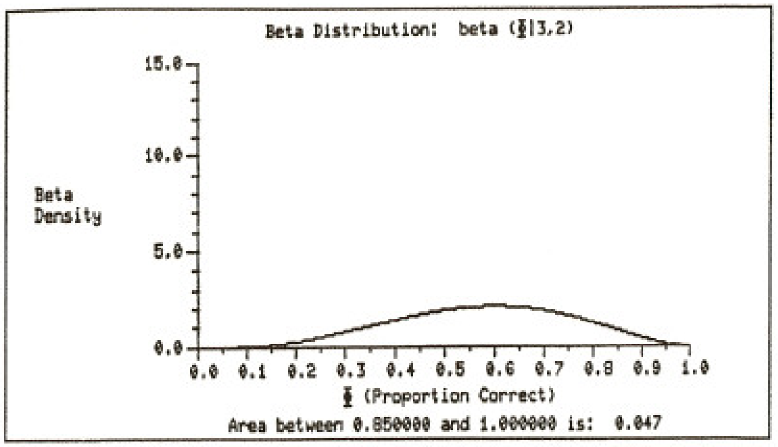


Figure 3. Posterior beta distribution of Φ given three successes and two failures.

The estimate of a probability when using Simpson's rule will become more accurate as the width of intervals is decreased (i.e., the number of intervals in some range is increased). In the author's experience, interval widths of .001 provide sufficient accuracy for most beta distributions, except extremely skewed and *J*-shaped beta distributions (the latter occurring when either *s* or *f* is zero). In the case of *J*-shaped distributions, the area on the high end of the distribution tends to be underestimated with Simpson's rule. A simple solution is to calculate the complementary area and then subtract from one. For example, suppose we want to know the probability that [$\Phi \geq .85$] when twelve success and zero failures have been observed. Instead of calculating the area between .85 and 1.0, we calculate the area between 0.0 and .8499 (the lower end of this distribution), which is the [prob ($\Phi < .85$) = .12]. Then subtract that result from one, to get the probability we really want [prob ($\Phi \geq .85$) = 1 – .12 = .88]. We would do the opposite when the number of successes is zero. With extremely skewed distributions (e.g., when *s* is quite large and *f* small, and vice-versa), a similar strategy can be employed in which the area under the longer tail is calculated. See Figures 4 and 5.

If we divide the range .85 to 1.0 into 150 intervals , when there are three successes and two failures , the Bayesian posterior probability that [$.85 \leq \Phi \leq 1.0$] is still .047. Thus, the posterior probability that [$0.0 \leq \Phi < .85$] is equal to (1.0 – .047) or .953. At this point, the odds are .953/.047 or about 20 to 1 in favor of nonmastery if .85 is our cut-off. There is still a small chance that

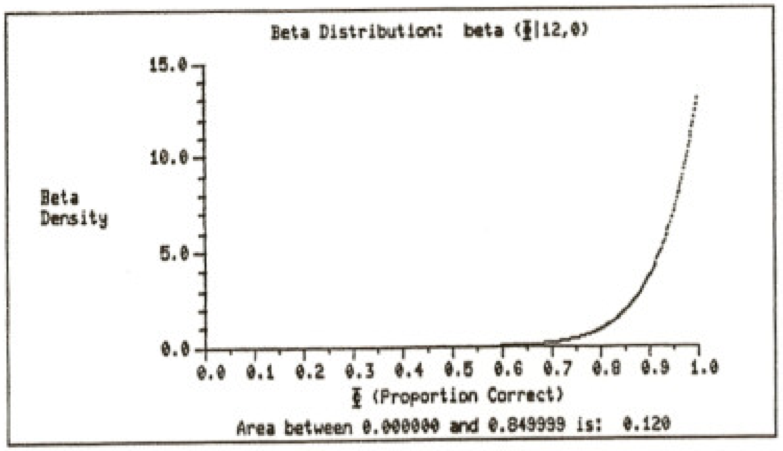


Figure 4. Posterior beta distribution of Φ given twelve successes and zero failures.

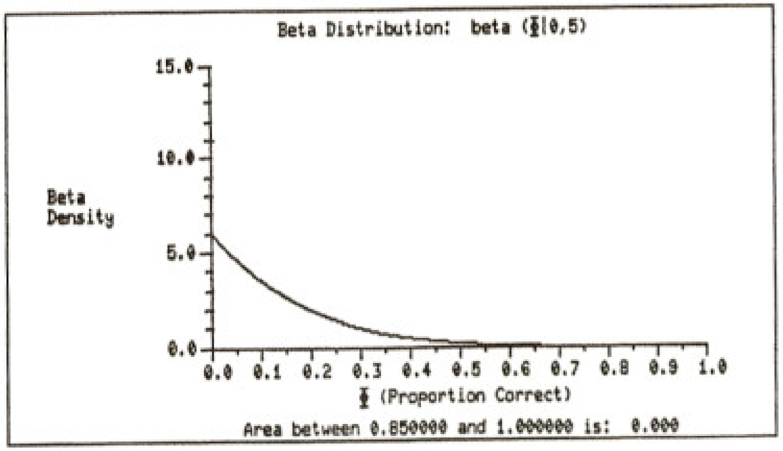


Figure 5. Posterior beta distribution of Φ given zero successes and five failures.

someone—whose true proportion correct for the universe of test items is .85 or higher—would answer three out of five randomly selected questions correctly.

The same assumptions required by the SPRT are also necessary for use of the beta model: independence of observations, random sampling without replacement, and treatment of items as if they each provide the same amount of

information. Thus, the beta model can be criticized on the same grounds as the SPRT. Nonetheless, the beta model is intuitively appealing, since it utilizes a single cut-off for mastery decisions.

One minor disadvantage of the beta model is that if numerical integration is done in real time to estimate probabilities, significant delays can occur on many microcomputers. A solution to this problem, if a cut-off is chosen in advance, is to calculate the areas above the cut-off for all combinations of successes and failures that could occur and store these results in a disk file, which can be loaded into memory when the program is run.

Novick and Lewis were interested in prescribing test lengths in order to make mastery and nonmastery decisions while minimizing losses due to false advancements (*a*) and false retentions (*b*) [4]. They used the rule:

$$\text{IF } \frac{\text{prob } (\Phi \geq \Phi_c \mid s,f)}{\text{prob } (\Phi < \Phi_c \mid s,f)} \geq \frac{a}{b} \text{ , then advance the student.} \tag{6}$$

Tennyson et al. apparently adopted this rule for deciding the amount of instruction to provide in terms of the number of interrogatory examples that are answered correctly in concept learning [2]. For example, in their Table 1 they want to know whether some student who has answered three out of four questions correctly should be advanced (i.e., a mastery decision) when the criterion level is .75 and the loss ratio (a/b) is .3. If they follow the Novick and Lewis rule, then the reasoning would be as follows:

$$\text{IF } \frac{\text{prob } (\Phi \geq .75 \mid 3,1)}{\text{prob } (\Phi < .75 \mid 3,1)} \geq \frac{.3}{1} \text{ , then advance the student.}$$

We need to perform a numerical integration between .75 and 1.0 for the beta distribution when three successes and one failure have been observed. The ratio of the posterior probabilities is .367/.633, or about .58. See Figure 6. Since .58 is greater than .3, the decision would be to advance the student. Tennyson, et al. report a beta value of .77 for three out of four correct in their Table 1 [2]. They do not indicate whether the beta value is a probability density or a probability of a range of values (i.e., prob[Φ ≥ .75]). In any case, their values reported in Table 1 do not agree with correct probabilities determined from posterior beta distributions given the numbers of successes and failures indicated in their table—assuming a flat prior distribution. Moreover, posterior beta probabilities do not depend on specification of a loss ratio, but only on the prior distribution of beta and the numbers of successes and failures currently observed. Perhaps Tennyson's posterior probabilities are based on a prior distribution that is not explicitly specified [2].

Another way of viewing the situation is that since the numerator and denominator on the left side of the inequality must sum to one, we can determine in advance the ratio of probabilities that is equal to the loss ratio. In this case,

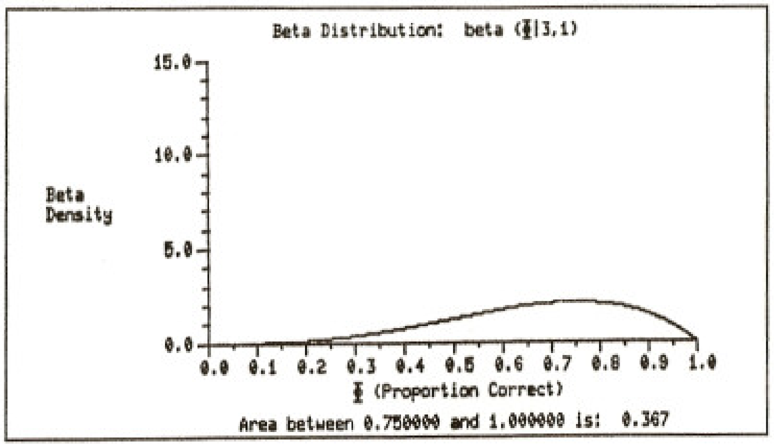


Figure 6. Posterior beta distribution of Φ given three successes and one failure.

.231/.769 is nearly equal to .3. Thus, whenever the posterior probability that [$\Phi \geq .75$] is greater than .231, we would advance the student.

It seems somewhat puzzling that Tennyson et al. would choose to *advance* a student when the odds could be as great as 3.3 to 1 (.769/.231) in favor of *nonmastery*! Nonetheless, whatever decision rule they do use does seem to improve overall learning achievement as measured by a posttest, when compared to a nonadaptive situation where students are free to control the number of interrogatory examples they undertake and to another nonadaptive situation where students are required to do all examples.

As an alternative to the Novick and Lewis use of loss ratios [4], Frick et al. suggested that instead of comparing the ratio of the posterior probabilities to a loss ratio, it appears sensible to compare the probability ratio to $(1 - \beta)/\alpha$ and to $\beta/(1 - \alpha)$ as did Wald with the discrete binomial case [7, 11]. In other words:

$$\text{IF } \frac{\text{prob}\ (\Phi \geq \Phi_c \mid \text{s,f})}{\text{prob}\ (\Phi < \Phi_c \mid \text{s,f})} \geq \frac{(1-\beta)}{\alpha}\text{, advance the student;} \tag{7.1}$$

$$\text{IF } \frac{\text{prob}\ (\Phi \geq \Phi_c \mid \text{s,f})}{\text{prob}\ (\Phi < \Phi_c \mid \text{s,f})} \leq \frac{\beta}{(1-\alpha)}\text{, retain the student;} \tag{7.2}$$

ELSE select another item at random and repeat this process. (7.3)

If mastery and nonmastery decisions are made by these rules, then it would be possible to compare the SPRT and use of the beta distribution with comparable type I and II error rates—and likewise for the IRT-based adaptive mastery testing model (AMT) developed by Weiss and Kingsbury.

Use of Item Response Theory and Bayesian Posterior Theta Distributions

One consideration not explicitly addressed by either the SPRT model or the beta model is the fact that not all test items designed to measure the same trait or instructional objective provide equivalent information about examinees. Some items appear to be harder than others, as indicated by a lower proportion of students who answer them correctly. Some items are good predictors of total test scores and some are not. Further items may turn out to be poor in that either no one answers them correctly, or everyone does; or even worse, those who answer correctly are those who overall are clearly nonmasters (and vice versa). With yet other items it may be easy to simply guess the correct answer.

These considerations are typically referred to in mental testing theory as item difficulty, discriminatory power, and chances of guessing (or lower asymptote). In classical item analysis, estimation of item difficulty and discrimination is heavily dependent on the sample of examinees who have taken the test. For example, if we administered the test only to persons who were masters of the instructional objective, then the item analysis would reveal that most of the items appear to be quite easy (low difficulty). On the other hand, if the sample consisted of only nonmasters, the analysis would reveal that items tended to be quite high in difficulty.

Such considerations are addressed in item response theory [8]. In essence, it is assumed that there is a relationship between the probability of a correct response to an item and an underlying (or latent) trait, and these item characteristics somehow enter into this relationship. The "trait" is what we are trying to indirectly measure by eliciting responses to test questions (e.g., mastery of a particular instructional objective). Persons who have more of this trait should be more likely to answer a question correctly than people who have less of this trait. Furthermore, some items may be useful for sorting out individuals who are high in this trait, but these same items would tell us practically nothing about people who have little of the trait we are trying to measure.

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

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

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

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

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

where:

$L = 1.7a_i\,(\theta - b_i)$,

a_i = discriminatory power of item i,

b_i = difficulty level of item i, and

c_i = lower asymptote of item i (chances of guessing).

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

The item discrimination parameter, a_i, normally varies between near zero and two. If a_i is zero, then the item does not discriminate at all across the spectrum of θ, and the ICC is a flat horizontal line. Such an item would be useless for trying to classify or rank individuals, since the probability of a correct response is the same for everyone. If a_i is fairly large, then the middle portion of the ICC is very steep. Such an item would be highly discriminating for a very narrow range of theta values. Everyone else would either be very likely to answer it correctly or very unlikely to answer it correctly.

The b_i parameter represents the difficulty of the item, and typically ranges between plus and minus three. Harder items have positive b_i values, easier items have negative b_i values, and average difficulty items have near-zero values of b_i. The effect of the b_i parameter on the ICC is to slide the whole curve to the right when b_i is positive, and to the left when negative.

The c_i parameter indicates the lower asymptote of the item, and is frequently referred to as the "guessing" factor. It can range between zero and one, though typical values would lie in the .1 to .3 range. The c_i parameter has the effect of compressing the bottom of the ICC vertically upwards.

Another important observation is that when θ is equal to b_i, then the probability of a correct response is half-way between the lower asymptote, c_i, and one.

In summary, the a_i parameter affects how steep the ICC is in the middle portion, the b_i parameter affects the horizontal displacement of the middle portion of the curve, and the c_i parameter affects the vertical displacement of the lower portion of the curve. This formulation of an item characteristic curve is known as the three-parameter model. In order to obtain fairly accurate estimates of the a_i, b_i, and c_i parameters, it is recommended that approximately 1000 individuals be tested with the item pool (cf. [16]).

As an example, suppose we have an item (#3) for which a_i is .5, b_i is −1, and c_i is .2. The expected probability of a correct response for someone whose theta level is −1.0 is:

$$\text{prob}\,(C_3 \mid -1.0) = .2 + (1 - .2)\frac{\exp(1.7(.5)((-1)-(-1))}{1 + \exp(1.7(.5)((-1)-(-1))} \approx .600$$

On the other hand, suppose another person's theta level is +2.0:

$$\text{prob}(C_3 \mid +2.0) = .2 + (1 - .2)\frac{\exp\,(1.7(.5)\,(2 - (-1))}{1 + \exp\,(1.7(.5)\,(2 - (-1))} \approx .942$$

As we should expect, a person with more of the trait being measured is more likely to answer this question correctly than a person with less. If we were to continue calculating probabilities of a correct response for a wide range of θ values, and we plotted these points on a graph where the probability value constitutes the vertical axis and the theta value the horizontal axis, then we would see the item characteristic curve for item #3, as shown in Figure 7.

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

If we consider all items to be equally discriminating and also do not consider chances of guessing, this is equivalent to setting a_i to a constant for all items and

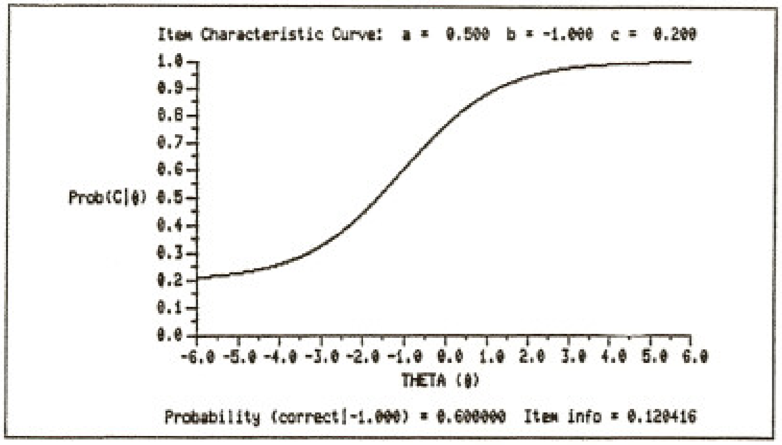


Figure 7. Theta distribution with $a = 0.5$, $b = -1.0$ and $c = 0.2$.

c_i to zero as above. This is known as a one-parameter model, equivalent to the Rasch model. All lower asymptotes are zero, and the middle portions of each ICC all have the same slope. The only thing that will differ is the horizontal displacement of the ICC's depending on the values of b_i's. A minimum of 200 examinees is needed to estimate the b_i parameters in the one-parameter model.

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

Brown and Weiss incorporate this concept of item information in selecting test items during adaptive mastery testing (AMT) [17]. Having some current estimate of an examinee's θ level, a computer program searches the pool of remaining items for the item which has the most information for this value of θ. This procedure is termed, 'maximum information search and selection' (MISS). The item which will have the most information is the one which has a difficulty level closely matching the current estimate of θ, and at the same time has the highest discriminatory power and lowest probability of being answered correctly by simply guessing.

Kingsbury and Weiss calculate information for item *i* for a given value of θ in the AMT model using Birnbaum's formula [6, 8]:

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

where

a_i, b_i and c_i, and L are defined as above,
$d = 1.7$ = constant scaling factor,

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

and

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

For example, let $a_i = 2$, $b_i = 1.5$, $c_i = .25$, and $\theta = 1.75$:

$$\text{L} = da_i\,(\theta - b_i) = 1.7(2)(1.75 - 1.5) = .85$$

$$\text{LPD} = \frac{\exp(.85)}{[1 + \exp(.85)]^2} = \frac{2.34}{[3.34]^2} \approx .210$$

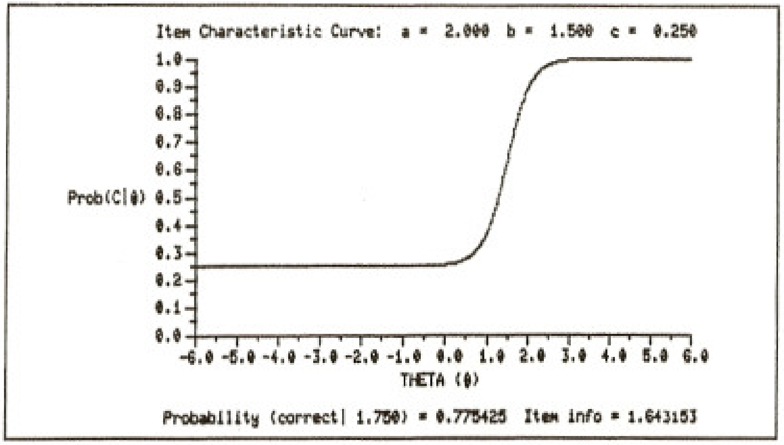
$$\text{LCP} = \frac{\exp(-.85)}{1 + \exp(-.85)} = \frac{.427}{1.427} \approx .299$$

Now we have all the pieces to calculate I_i ($\theta = 1.75$):

$$I_i(1.75) = \frac{(1 - .25)\,(1.7)^2(2)^2(.210)^2}{.210 + (.25)\,(.299)^2} \approx 1.64$$

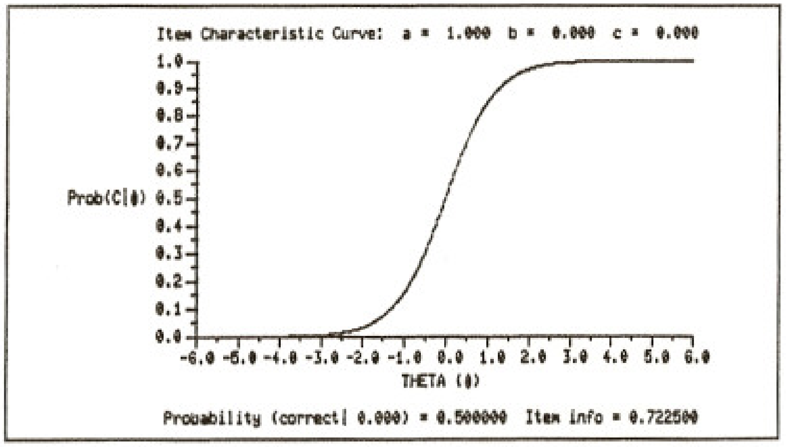
Information values for test items will typically range from near zero to three or so for various values of θ. The item in the example above would give us some information about a person whose estimated θ is 1.75. The probability of a correct response to this item is expected to be about .78, using the ICC with $a_i = 2$, $b_i = 1.5$, $c_i = .25$, and $\theta = 1.75$. See Figure 8.

This part of item response theory, while complex, is fairly straightforward, assuming we have trustworthy a_i, b_i, and c_i parameter estimates. The catch is that the probabilities of a correct response to an item and the information values of an



```
Item Characteristic Curve:  a =  2.000  b =  1.500  c =  0.250
Probability (correct| 1.750) = 0.775425  Item info = 1.643153
```

Figure 8. Theta distribution with $a = 2.0$, $b = 1.5$, and $c = 0.25$.

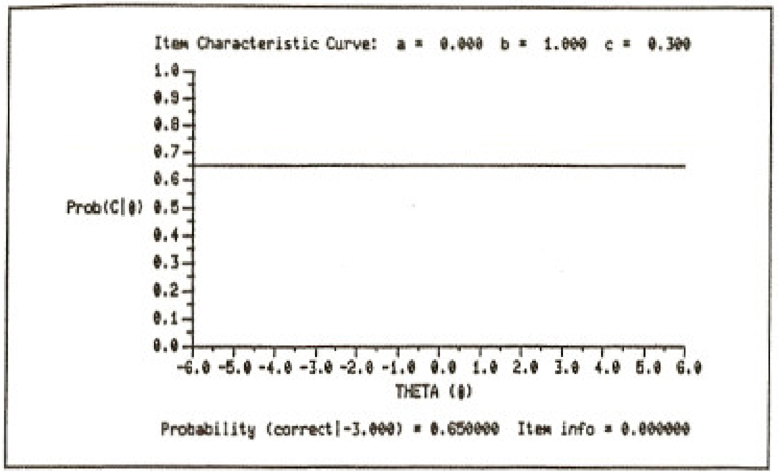


```
Item Characteristic Curve:  a =  1.000  b =  0.000  c =  0.000
Probability (correct| 0.000) = 0.500000  Item info = 0.722500
```

Figure 9. Theta distribution with $a = 1.0$, $b = 0.0$, and $c = 0.0$.

item vary as a function of θ, the underlying trait that we cannot directly measure or observe for some examinee. How can we estimate the value of θ for an individual during an adaptive test?

Bayesian posterior θ estimation — If we begin a test with a prior estimate of an examinee's θ level and its variance, and if we give an item to an examinee and know whether it was answered correctly or not, then we can determine the posterior distribution of θ and the variance of that distribution by using formulas developed by Owen [9]. The posterior estimate of theta given a *correct* response to the current item is:



```
Item Characteristic Curve:  a =  0.000  b =  1.000  c =  0.300
Probability (correct|-3.000) = 0.650000  Item info = 0.000000
```

Figure 10. Theta distribution with $a = 0.0$, $b = 1.0$, and $c = 0.3$.

$$E(\theta \mid C) = M_0 + \left\{\frac{[(1 - c_i)V_0 / W][\text{gau}(X)]}{Y}\right\}, \tag{13}$$

where

M_o = prior θ estimate,
V_o = prior variance of θ estimate,

$$W = [(1/a_i^{\,2}) + V_0]^{1/2}, \tag{14}$$

$$X = (b_i - M_0)/W \tag{15}$$

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

$$Y = c_i + (1 - c_i)\,[\text{logist}\,(-1.7X)], \tag{17}$$

and

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

The estimate of θ given an *incorrect* response to item i is defined:

$$E(\theta \mid -C) = M_0 - \left\{ \frac{[V_0/W]\,[\text{gau}(X)]}{\text{logist}(1.7X)} \right\}. \tag{19}$$

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

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

$$V(\theta \mid C) = V_0 \left[1 - \left\{ \frac{(1 - c_i)\,[\text{gau}(X)]}{U} \right\} \left\{ \frac{\dfrac{(1 - c_i)[\text{gau}(X)]}{Y} - X}{Y} \right\} \right], \tag{20}$$

where:

$$U = 1 + [1/(a_i{}^2 V_0)]. \tag{21}$$

The estimate of the θ variance, given an *incorrect* response is defined:

$$V(\theta \mid -C) = V_0 \left[1 - \left\{ \frac{\text{gau}(X)}{U} \right\} \left\{ \frac{\dfrac{\text{gau}(X)}{\text{logist}(1.7X)} + X}{\text{logist}(1.7X)} \right\} \right]. \tag{22}$$

Another observation is that the "guessing" factor, c_i, enters into the updating process of both θ and its variance when a question is answered correctly, but the c_i-related terms drop out if the question is answered incorrectly.

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

An example of the AMT in operation — Finally, we have all the pieces in order to show how the AMT works. We will assume that we have a 100-item test that has been used with 1000 examinees and we have good estimates of the a_i, b_i, and c_i parameters. Suppose initially we believe that a particular person has an average amount of the trait we are attempting to measure with our test items. In this case, we set our prior estimate of θ to zero (i.e., $M_o = 0$), and prior variance to one (i.e., $V_o = 1$). We next calculate the information that each item in the pool has when θ is zero. We find that I_{20} (0.0) ≈ 1.927 to be higher than any other item, so we administer this item for which $a_{20} = 2.0$, $b_{20} = 0.0$, and $c_{20} = 0.2$. The student in question answers item #20 correctly.

Since a correct response was given, we need to evaluate formulas (13) and (20) in order to estimate the posterior value of θ and its variance. To use these formulas we first need to compute the necessary pieces from (14) through (18):

$$W = [1/2^2 + 1]^{1/2} = [1.25]^{1/2} \approx 1.1180$$

$$X = (0 - 0)/1.118 = 0.0$$

$$\text{gau}\ (0.0) = (.3989)\ (\exp\ (-\ (0^2)/2)) \approx .3989$$

$$\text{logist}\ (-1.7(0.0)) = [\exp(-1.7(0))]/[1 + \exp(-1.7(0))] = 0.5$$

$$Y = .2 + .8(.5) = 0.6$$

$$U = 1 + [1/\ (2^2\ (1))] = 1.25$$

Now we can estimate the posterior θ and its variance:

$$E(\theta \mid C) = 0 + \left\{ \frac{[(.8)(1)/1.118][.3989]}{.6} \right\} \approx 0.4757$$

$$V(\theta \mid C) = (1) \left[1 - \left\{ \frac{(.8)(.3989)}{1.25} \right\} \left\{ \frac{\dfrac{(.8)(.3989)}{.6} - 0}{.6} \right\} \right] \approx .7737$$

We can now form a Bayesian confidence interval. For example, if we use a 95 percent confidence interval, then:

$$[E(\theta) - 1.96(V(\theta)^{1/2})] \le \theta \le [E(\theta) + 1.96(V(\theta)^{1/2})]\,, \tag{23}$$

or

$$[.4757 - 1.96(.7737^{1/2})] \le \theta \le [.4757 + 1.96(.7737^{1/2})]\,.$$

Thus, the probability is .95 that [$-1.248 \le \theta \le +2.20$].

If we decide to continue the test, we now set our prior estimate of θ to .4757 (i.e., $M_o = .4757$), and prior variance to .7737 (i.e., $V_o = .7737$). We next calculate the information that each item in the pool has when θ is .4757. We find that [$I_3(0.4757) \approx 1.523$] to be higher than any other item, so we administer this item for which $a_3 = 2.0$, $b_3 = 0.5$, and $c_3 = 0.3$. The student in question answers item #3 incorrectly.

Since an incorrect response was given, we need to evaluate formulas (19) and (22) in order to estimate the posterior value of θ and its variance. To use these formulas we first need to compute the necessary pieces from (14) through (18):

$$W = [1/\ 2^2 + .7737]^{1/2} = [1.0236]^{1/2} \approx 1.0118$$

$$X = (0.5 - 0.4757)/1.0118 \approx 0.0240$$

$$\text{gau}\ (0.0240) = (.3989)(\exp(-(0.0240^2)/\ 2)) \approx 0.3988$$

$$U = 1 + [1/(2^2(.7737))] = 1.323$$

$$\text{logist}(1.7(0.0240)) = \frac{[\exp(1.7(0.0240))]}{[1 + \exp(1.7(0.0240))]} \approx 0.5102$$

Now we can estimate the posterior θ and its variance:

$$E(\theta \mid -C) = 0.4757 - \frac{[.7737/\,1.0118][.3988]}{.5102} \approx -\,0.122$$

$$V(\theta \mid -C) = (.7737)\left[1 - \left\{\frac{(.3988)}{1.323}\right\}\left\{\frac{\frac{(.3988)}{.5102} + .0240}{.5102}\right\}\right] \approx 0.4049$$

We can now form a Bayesian confidence interval. If we use a 95 percent confidence interval, then:

$$[-\,0.122 - 1.96(.4049^{1/2})] \leq \theta \leq [-\,0.122 + 1.96(.4049^{1/2})]$$

Thus, the probability is .95 that $[-1.37 \leq \theta \leq +1.13]$. As can be seen, computation by hand is tedious and can be error prone. In particular, errors due to rounding can creep in and cause divergence from what is obtained if all computations are maintained to the highest degree of accuracy a computational device is capable of.[6]

Doing the AMT model by hand computation does reveal the importance of accurate item parameter estimates, since minor variations can dramatically effect these computations that frequently deal with relatively small numbers. And minor differences in new posterior θ estimates can make a difference in which item is selected next, according to its information value. In effect, small discrepancies can quickly magnify themselves into rather large differences, since the model contains many arithmetic multiplications and divisions.

The AMT stopping rule — We have still not addressed the basis for ending a mastery test under the AMT model and reaching a decision. Weiss and Kingsbury recommend using a test response function (more commonly referred to as a test characteristic curve, TCC) as follows [3]:

$$\text{prob}(C_t \mid \theta) = \left\{\Sigma_i \left[c_i + (1 - c_i)\frac{\exp(L)}{1 + \exp(L)}\ \right]\right\}/n \tag{24}$$

[6] I have also observed discrepancies when the AMT model is coded the same way in two different high-level languages on a VAX minicomputer, and then run on the same data set (i.e., same item pool, same item parameters, and identical correct or incorrect responses to items selected). Due to differences in the maximum precision of arithmetic in the two languages, on occasion after about 10 or 12 items a different item will be picked with the MISS procedure in one of the code versions. After that point the sequence of items selected by the MISS procedure will be different in the two code versions for that particular examinee's data set.

where n = number of items in the total pool, and $L = 1.7a_i(\theta - b_i)$, as before.

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

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

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

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

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

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

Thus, if the confidence interval does contain θ_c we continue the test by using the MISS technique to choose the next item.

Note that choosing a Bayesian confidence interval of .95 is the same as setting $\alpha = \beta = .025$ (see above discussion of the SPRT and posterior beta distribution).

Summary of the AMT model — The adaptive mastery testing (AMT) model is clearly the most complex of the three discussed here, and also has the requirement that a large number of examinees must take the test in advance in order to estimate item parameters. On the other hand, the SPRT and beta models treat items as if each provides the same amount of information about every examinee.

In the AMT model items are not selected at random as they are in the SPRT and beta models. Rather, in the AMT model the item selected next is the one which is predicted to provide the most amount of information about a particular examinee, given an estimate of that person's θ level. What this means in practice is that the next item chosen is one that has a difficulty level matching as closely as possible to a person's θ level, and which at the same time discriminates best in that area of θ, and which has the least chance of being answered correctly by guessing. The AMT model does more than adapt the length of a mastery test; it also adapts in terms of what items are chosen. These features would make this approach seem to be most desirable.

Three other features of the AMT are worth discussing:

1. In order to effectively take advantage of the MISS procedure, a rather large item pool is required which represents a wide range of difficulty levels. If a test continues to a point where there are no remaining items that closely match the current estimate of a person's θ level, then less than optimal items will be chosen subsequently. This means that Bayesian updates to θ and its variance will not be as dramatic as could be, and a test could become very long before a decision can be reached at the desired level of confidence.
2. Theta can alternatively be estimated by maximum likelihood methods as soon as an examinee has answered at least one question right and one wrong. Since the AMT model is being compared to two other Bayesian models here, only Bayesian estimation of θ is considered.
3. Computerized adaptive testing (CAT) includes the AMT procedures discussed here. Instead of just making mastery or nonmastery decisions, CAT can continue until θ estimates are precise enough for whatever decisions need to be made (e.g., grade classifications) [3].

MONTE CARLO STUDIES

Comparative Computer Simulations of the Three Models

Computer simulations of the sequential probability ratio test (SPRT), the beta model, and the adaptive mastery testing (AMT) model were conducted to compare efficiency and decision accuracy. Efficiency is measured by the average number of test items required to reach a mastery or nonmastery decision under a given set of experimental conditions. Accuracy is measured by the correctness of the decisions under those conditions—i.e., the percent of decisions in each model which agree with actual or " known" mastery status.

Since Kingsbury and Weiss did not use Wald's SPRT formulation [6], it was further desirable to replicate initially their simulation with the results for the correct SPRT model. No differences were expected in the AMT results, when comparing the present study with that of Kingsbury and Weiss [6], other than those attributable to sampling error. Furthermore, Kingsbury and Weiss did not compare the beta model to the SPRT and AMT [6]. In many respects, the beta model is more directly comparable to the AMT model than the SPRT, due to the basic problem of choosing both a mastery and nonmastery level *a priori* in the SPRT. The wider the gap between the two levels, the shorter tests tend to be, all other things equal. Ideally, these two levels would be chosen on the basis of prior empirical estimates. Kingsbury and Weiss choose .7 for P_m and .5 for P_n [6]. While it is obvious that .6 is the mid-point, .8 vs .4 could have been chosen as well, or .9 vs .3, etc. The average test length in the SPRT will be affected by the choice of mastery and nonmastery levels.

Methods — In the present study, the same experimental conditions were used as those by Kingsbury and Weiss in their a_i-, b_i-, and c_i-variable pool of 100 items. The same item parameters were also used in the present study (see [18]). The simulations were performed with three different maximum test lengths (10, 25, and 50 items). The test characteristic curve in the AMT model indicated that if the θ_c mastery level is zero, based on these 100 items, this corresponds to 60 percent correct on the whole test. Therefore, in both the beta model and the SPRT, if a simulee's "true" score was greater than or equal to .60, then he was determined to be an actual master, otherwise a nonmaster. In the AMT model, if the "true" θ level was greater than or equal to θ_c (=0), he was considered an actual master.

The "true" θ value for each of the 500 simulees was chosen from a normal distribution with a mean of zero and variance of one. In the AMT model, the prior estimate of θ was set to zero and variance to one at the beginning of each simulated test. For each simulee, a subset of the 100 items was chosen at random to create an item pool with *n* items, where *n* was either 10, 25 or 50 depending on the experimental condition. Test items actually administered to a simulee were drawn from his particular subset, not from the entire pool of 100 items.

The method of selecting a test item differed depending on the decision model. In the AMT model, the item selected next was the one in the pool with the most information for the current estimate of the simulee's θ level (MISS procedure). On the other hand, in the SPRT and beta models items were selected at random. No item in a given simulee's pool was administered twice within that model, but could have been "answered" up to three times (usually on different occasions), depending on if or when it was selected by one or more of the other decision models.

Simulation of examinee responses to test items — The method of determining whether or not a simulee answered a particular question correctly was, however, the *same* in all three models. The probability of a correct response to the item chosen in each model was calculated from the item characteristic curve, given the "true" θ value selected in advance for the particular simulee. A number between zero and one was then randomly chosen from a flat distribution. If this random number was less than or equal to the expected probability of a correct response to the item by the simulee, then the question was taken as being answered correctly. If the random number was greater than the ICC-derived probability value, then it was considered as being answered incorrectly.[7] This method of generating correct and incorrect answers is consistent with that used by Kingsbury and Weiss and is also consistent with the item response theoretic model that AMT is based upon

[7] This may surprise the reader, since initially it may seem backwards. The random number is drawn from a flat distribution between zero and one, where each value has an equal chance of being selected. For example, if the probability of a correct response is .75, then in the long run 75 percent of the numbers we randomly pick from the flat distribution should be between 0 and .75. Thus, this decision logic does simulate correctly the probability of a correct response.

[6]. Whether it is consistent with reality depends on the nature of the trait being measured, a question that can be answered empirically (e.g., see [3]).

Once it was determined that each selected question in each model was answered correctly or not, the total number of successes or failures was incremented accordingly for that model, and then the decision rules in that model were applied. For the SPRT, formula (1) and Wald's three rules were used ((2) to (4)). In the beta model, the probability that [$\Phi \geq .60$] was calculated using Simpson's rule and the beta density function (5), and decision rules (7.1) to (7.3). In the AMT model, formulas (13) and (20) were used if the question was answered correctly, and (19) and (22) if incorrect, to obtain posterior estimates of the simulee's θ value and variance. A .95 confidence interval was then generated using those new posterior values (formula (23)), and decision rules (25.1) through (25.3) were applied.

Decision error rates — Kingsbury and Weiss used a .95 confidence interval for mastery decisions in the AMT, but set α and β to .1 in the SPRT [6]. It is unclear why they chose to do this, since an .80 confidence interval is directly comparable ($1 - \alpha - \beta$). To make the theoretical decision error rates identical, α and β were set to .025 for both the SPRT and beta models in the present simulation, and a 95 percent confidence interval was used with the AMT model.

Determination of decision accuracies — If a mastery or nonmastery decision was reached by one of the models, then the number of items administered to that simulee was stored along with the accuracy of the decision. Decision accuracy was determined as follows: If the true θ for a simulee was greater than or equal to zero (θ_c) and a mastery decision was reached by that particular model, this was counted as a "hit." If the true θ was less than θ_c and a nonmastery decision was reached by the model, this was also counted as a hit. Otherwise, it was counted as a "miss."

Once a decision was reached by a model for a particular simulee, it dropped out of contention until the remaining models reached decisions or that particular simulee's item pool was exhausted. If it happened that a decision model could not reach a mastery or nonmastery decision when a pool was exhausted, then a decision was forced, realizing of course that this will cause more decision errors than expected by the theoretical *a priori* rates. This was apparently done by Kingsbury and Weiss to study the effect of maximum test length on decision accuracy [6]. In the case of pool exhaustion during both the SPRT and beta models, if the proportion of questions answered correctly was greater than or equal to .6, a mastery decision was rendered, otherwise nonmastery. Similarly, if the pool were exhausted during the AMT model and if the current estimate of θ was greater than or equal to zero (θ_c), mastery was concluded, otherwise nonmastery. These decisions were then compared to the "true" state of affairs, as described above, and the number of hits or misses was incremented accordingly for each model.

This entire process was repeated for a total of 500 simulated tests each for maximum test lengths of 10, 25, and 50 items. The process was carried out by a program run on a VAX minicomputer. At the end of the simulation for a given test length, the average number of items required in each model was calculated as well as the percent of hits.

In summary, the present simulation study parallels that done by Kingsbury and Weiss with the following differences [6]:

1. The correct SPRT formula was used here.
2. The beta model was also compared to the SPRT and AMT.
3. Alpha and beta decision error rates were set to .025 in both the SPRT and beta models.
4. An SPRT model with $P_m = .8$ and $P_n = .4$ was also included.

Results

First experiment — Results from the first experiment are reported in Table 1. It should be noted that results from Kingsbury and Weiss study [6] for the AMT model are also reported in the fifth column of Table 1 for purposes of comparison with the present AMT results. Due to random variation intentionally built into the simulations as described above, results do vary from experiment to experiment, but the pattern of results is consistent.

Table 1. Efficiency and Accuracy of the Sequential Probability Ratio Test (SPRT), Beta, and Adaptive Mastery Testing (AMT) Models Using Item Parameters from Kingsbury and Weiss (1983).

Distribution of True θ's: Standard Normal (Mean = 0, Variance = 1)

	SPRT[a]	*SPRT*[b]	*Beta*	*AMT*	*AMT (1983)*
Maximum Test Length = 10, n = 500					
Average test length (Efficiency)	8.59	9.98	9.40	8.96	8.73
Percent of correct decisions (Accuracy)	75.0	74.2	75.4	79.0	72.6
Maximum Test Length = 25, n = 500					
Average test length (Efficiency)	13.65	22.34	20.07	16.78	16.35
Percent of correct decisions (Accuracy)	79.6	81.8	83.6	89.4	86.6
Maximum Test Length = 50, n = 500					
Average test length (Efficiency)	16.53	33.74	33.07	22.38	23.39
Percent of correct decisions (Accuracy)	82.4	88.2	86.8	91.6	89.4

[a] $P_m = .8$, $P_n = .4$
[b] $P_m = .7$, $P_n = .5$

In each of the three maximum test length conditions, the SPRT (.8 vs .4) was the most efficient model. On the other hand, the SPRT (.7 vs .5) was the least efficient of the models. This apparently contradictory finding is a result of the choice of mastery and nonmastery levels in the SPRT, since wider differences in the two levels tend to reduce the number of items needed to reach a decision. It also illustrates the importance of the justification of the choice of levels, and whenever possible setting levels based on prior test results.

In addition, when comparing the two SPRT models, accuracy is relatively comparable, with a possible exception in the fifty-item tests where the accuracy was somewhat lower in the SPRT (.4 vs .8) model. A very important observation, however, is that the decision error rates for all models and all test lengths here were greater than the expected five percent. Both SPRT models had error rates of about 25 percent when the maximum test length was constrained to be ten items, about 20 percent when length was constrained to twenty-five items, and about 15 percent for the fifty-item tests. The reason for this was the relatively large proportion of simulees with true θ values near θ_c, which in the proportion correct metric is half-way between the mastery and nonmastery levels in the SPRT. Wald referred to this gap as the 'zone of indifference,' since the SPRT performs least optimally for samples with proportions of successes near the mid-point of the zone, both in terms of sample size and decision error rates [11].

The AMT model was more efficient than the beta and SPRT (.5 vs .7) models in all three test length conditions. The latter two models resulted in very comparable average test lengths. When the item pool subsets were constrained to fifty items, the AMT required about ten fewer items on the average to reach a decision. The AMT tended to be more accurate than any of the other models across all three test length conditions.

When the maximum test length was constrained to ten items, most of the time decisions were forced in all of the models, meaning that this would tend to push the decision error rate higher than that set *a priori*. Decision accuracies for all models tended to increase, respectively, in the twenty-five- and fifty-item tests. In the latter two conditions there were less forced decisions, with the highest decision accuracies in the fifty-item maximum condition. Still, however, the accuracies fell short of the expected 95 percent. Most likely this was due to the fact that there were a number of occasions where a simulee's item pool was exhausted and a decision was forced. The relatively high error rate (20% to 30%) in the ten-item maximum condition illustrates the problem with short tests that has been demonstrated theoretically by Novick and Lewis [4] and Millman [5].

The reader is again reminded that these results represent a severe test of the three models, since a large number of true θ levels were very near to the mastery cut-off point.

To further investigate the effects of the distribution of true θ values, the simulation was repeated where the mean was zero and the variance was 10—a much flatter normal distribution. In this case, about 95 percent of the θ values are expected to occur between –6.2 and +6.2, compared to the first simulation in which 95 percent of the values were in the –1.96 to +1.96 range. All other conditions were the same as in the first simulation.

Table 2. Efficiency and Accuracy of the Sequential Probability Ratio Test (SPRT), Beta, and Adaptive Mastery Testing (AMT) Models Using Item Parameters from Kingsbury and Weiss (1983).

Distribution of True θ's: Normal (Mean = 0, Variance = 10)

	SPRT [a]	*SPRT* [b]	*Beta*	*AMT*
Maximum Test Length = 10, n = 500				
Average test length (Efficiency)	7.36	9.89	8.23	6.90
Percent of correct decisions (Accuracy)	88.0	88.4	88.2	90.2
Maximum Test Length = 25, n = 500				
Average test length (Efficiency)	9.80	16.72	13.31	8.70
Percent of correct decisions (Accuracy)	90.0	93.8	92.8	95.8
Maximum Test Length = 50, n = 500				
Average test length (Efficiency)	10.16	21.93	19.11	11.07
Percent of correct decisions (Accuracy)	92.4	94.8	94.6	96.0

[a] $P_m = .8$, $P_n = .4$
[b] $P_m = .7$, $P_n = .5$

In Table 2 it can be seen that the overall effect of the greater spread of true θ values is to decrease average test lengths in all models and conditions, and to generally increase decision accuracy. In particular, in the fifty-item maximum test length condition the decision error rates are around 5 percent, the expected error rate specified in all models. One exception was the SPRT (.4 vs .8), which resulted in slightly higher than expected error rates.

The AMT model tended to be the most efficient and the most accurate over all conditions in the second simulation. A notable exception was in the fifty-item pools where the SPRT (.4 vs .8) resulted in slightly shorter tests. A further trend is that the beta model tended to be slightly more efficient than the SPRT (.5 vs .7) model with nearly equivalent accuracy levels.

Discussion

Perhaps the most important result is the observation of higher decision error rates when examinee θ levels are near the mastery cut-off, as evident from the first simulation. While this is not surprising, it does remind us of the severity of the problem when attempting to make mastery classifications under these conditions and when tests are relatively short.

On the other hand, when θ levels are further away from the cut-off, both decision accuracy and efficiency improve, as indicated in the second simulation. When the maximum test length was constrained to fifty-item pools, all models made fairly accurate decisions, though some were clearly more efficient than others.

It is clear that decision accuracy and efficiency are dependent on the shape and location of the distribution of true scores in the sample of examinees being tested. In a mastery learning context where students are pre- and posttested on instructional units, we would expect to find a flatter or a bi-modal distribution of true scores, as was the case, for example, in the Frick study [1].[8] Although the mastery and nonmastery levels were different in that study (.85 vs .6), decision accuracy was very high (about 98%) with average test lengths of about twenty items. There was no constraint on the maximum test length in the Frick study other than the actual sizes of the two test item pools, and only one forced decision was made in that study [1]. Those results are fairly comparable to the second simulation in the present study when maximum test length was constrained at fifty items, where the SPRT (.7 vs .5) and beta models resulted in test lengths of about twenty items, with decision accuracies at nearly 95 percent. Moreover, Weiss and Kingsbury reported that a validation study of the AMT with variable length tests of actual students was consistent with previous Monte Carlo simulation results with the AMT [3, 6].

Since there is consistency between validation studies such as Frick [1] and Weiss and Kingsbury [3] and the present simulation results, which replicate in part the Kingsbury and Weiss simulation [6], the credibility and generalizability of the results presented here are enhanced considerably.

Which Model Is Best?

It is noteworthy that all three models—the sequential probability ratio test (SPRT), the beta model, and the adaptive mastery testing model (AMT)—tend to provide fairly accurate predictions of mastery when used conservatively and when no constraints are placed on maximum test length. The differences lie in efficiency. Since the AMT model selects items according to the amount of information they provide, it is more adaptive than the SPRT and beta models, where random selection of items is assumed. Also, the AMT model is generally more efficient than the SPRT and beta models. The AMT model would appear, therefore, to be overall the best of the three.

On the other hand, the analogy of using a cannon to kill a mosquito is apt. In many classroom computer-based testing situations, the SPRT and beta models (fly swatters—to continue the analogy), would appear to be quite sufficient. The SPRT is the most practical of the three in that it is the least computationally complex, thus requiring less CPU time for rendering decisions. The SPRT is best used when test item pools have been previously used with the kinds of examinees for whom

[8] That study was not a simulation, but a validation of the SPRT with college students in classroom testing situations.

the tests are intended in order to empirically establish the mastery and nonmastery levels required by that model. Furthermore, if small decision error rates are used, the problem of uneven item representation is minimized.

The beta model is also attractive in that its computational overhead is relatively small, particularly if calculations of posterior probabilities are done in advance for a given cut-off with all combinations of numbers of successes and failures. Furthermore, the beta model can be more efficient than the SPRT in terms of average test length, depending on the distribution of examinee achievement levels and what SPRT mastery and nonmastery levels are used. The beta model also has the same drawback as the SPRT in that all items are treated as if each provides an equivalent amount of information about student mastery. The beta model can be particularly error prone if a student happens to get off to a poor start, since a test will end if he or she misses the first few items—using typical mastery levels, even when α and β are quite small. A simple solution is to wait until five or ten items have been answered before applying the beta model.

If a teacher is concerned that tests may be too short and/or students perceive an unfairness about the situation, a simple remedy is to establish a minimum test length that is satisfactory and then begin applying the SPRT or beta decision models from that point onwards (e.g., after 20 items have been answered).

It is clear that the SPRT and beta models are more practical than the AMT model for typical teacher-made mastery tests that are computer-administered. In addition, the consequences of occasional decision errors are probably not that severe for this kind of routine testing before and after student completion of a module or unit.

On the other hand, if a curriculum is standardized to the extent that a large number of students will be taking the same tests, then the AMT approach has more merit. First, in this kind of situation the problem of gathering data on 200, 500 or 1000 examinees to estimate item parameters for the one-, two-, or three-parameter IRT models becomes a less serious obstacle. Second, estimates of student achievement levels can be made more precise by choice of test items matching their ability levels. The AMT approach is definitely preferable if the goal of testing is to rank examinees along some continuum, rather than simply to classify as masters or nonmasters.

Finally, if the consequences of incorrect assessment decisions are severe, then it is advisable to collect as much good information as possible about an examinee. The goal of adapting the length of such a test is of little concern. Rather, the focus is on making a decision with the best information possible.

THE FUTURE

At present the notion of computerized adaptive tests may seem far-fetched and perhaps esoteric for the public schools. Clearly, many obstacles must be overcome to realize the kind of future educational system envisioned at the beginning of this article. Educators are not likely to use adaptive tests unless they understand how such tests work and that they trust the decisions reached by such methods. If this article has furthered that aim, then it has achieved its purpose.

REFERENCES

1. T. Frick, Bayesian Adaptation during Computer-Based Tests and Computer-Guided Practice Exercises, *Journal of Educational Computing Research, 5*:1, pp. 89-114, 1989.
2. 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.
3. D. Weiss and G. Kingsbury, Application of Computerized Adaptive Testing to Educational Problems, *Journal of Educational Measurement, 21,* pp. 361-375, 1984.
4. M. Novick and C. Lewis, *Prescribing Test Length for Criterion-Referenced Measurement*, American College Testing Program, Technical Bulletin No. 18, Iowa City, Iowa, 1974.
5. J. Millman, Passing Scores and Test Lengths for Domain-Referenced Measures, *Review of Educational Research, 43*:2, pp. 205-216, 1973.
6. 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.
7. 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.
8. F. Lord and M. Novick, *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, Massachusetts, 1968.
9. R. 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.
10. 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.
11. A. Wald, *Sequential Analysis*, Wiley, New York, 1947.
12. S. Schmitt, *Measuring Uncertainty*, Addison-Wesley, Reading, Massachusetts, 1969.
13. T. Frick, An Investigation of the Sequential Probability Ratio Test for Mastery Decisions during Criterion-Referenced Testing, paper presented to the *American Educational Research Association,* San Francisco, 1986.
14. R. Ferguson, *Computer-Assisted Criterion-Referenced Measurement*, University of Pittsburgh Learning Research and Development Center, Pittsburgh, 1969.
15. 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.
16. 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.

17. J. Brown and D. Weiss, *An Adaptive Testing Strategy for Achievement Test Batteries* (Research Report 77-6), University of Minnesota, Department of Psychology, Psychometric Methods Program, Minneapolis, 1977.
18. G. Kingsbury and D. Weiss, *A Comparison of Adaptive, Sequential, and Conventional Testing Strategies for Mastery Decisions* (Research Report 80-4), University of Minnesota, Department of Psychology, Psychometric Methods Program, Minneapolis, 1980.

Direct reprint requests to:

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