THE RELIABILITY OF SEQUENTIAL TESTING

by

Jesse M. Heines, Ed.D.
The CBT Artisan
University of Lowell
Dept. of Computer Science

This paper was first presented at the
Third Canadian Symposium on Instructional Technology
in Vancouver, British Columbia, in February 1980

ABSTRACT

A study was conducted to assess the feasibility of using computer-managed instruction (CMI) to control the quality of self-paced training in a customer environment. The study centered on a self-paced course on BASIC language programming and its complementary interactive CMI system.

The CMI system employed a mastery algorithm based on a sequential probability test ratio. The purpose of this algorithm was to reduce testing time while retaining a high level of criterion-referenced test reliability. These factors were assessed by comparing results on tests that were terminated by the sequential probability test ratio to those on test that were extended to 30 items in length. Average test time differences were computed, and four reliability indices were reported that compared mastery classifications on the shorter tests with those on the extended tests.

The study found that the sequential probability test ratio reduced testing time by an average of 29.8% over the extended tests and that criterion-referenced reliability was not significantly effected.

STATEMENT OF THE PROBLEM

Many factors have influenced the evolution of training offered by computer manufacturers in recent years. Two of the strongest factors have been:

- drastic decreases in the cost of computer systems due to new technologies
- significant increases in the costs of customer training due to inflation

Even with the regional training centers, these factors often make it financially unfeasible for some customers to send their employees to the manufacturer's facilities for training. For example, a company that buys a complete computer system for under $10,000 cannot be expected to spend several hundred dollars to cover tuition, travel, and per diem expenses for each employee that they wish to train to use that system.

Many computer manufacturers have begun developing self-paced instruction (SPI) courses to deal with these evolutionary factors. These courses are typically written in a modularized, criterion-referenced format. Many use audiovisual media for delivering instruction, and all are designed for use in an on-the-job customer environment without requiring the presence of an instructor.

The introduction of self-paced training has created numerous quality control problems. For example, learners often misuse the tests included in SPI courses in their haste to move on to new subject matter. They may skip the tests entirely, read the tests and look at the answers simultaneously, or take the tests once, check their answers, and move on without really trying to understand why they answered certain items incorrectly.

Working on the assumption that these behaviors occur because learners like to get through tests as quickly as possible, the major problem of this study was to develop a testing system whose use required a minimum amount of time. However, when test lengths are reduced to minimize testing time, reliability is usually sacrificed. The secondary problem was therefore to implement an algorithm that maintained reliability with short tests, and to devise and implement a scheme for assessing the validity of this algorithm.

MODELS OF CRITERION-REFERENCED MASTERY AND RELIABILITY

Mastery Models

The ideal criterion-referenced test (CRT) is one which yields a single, unambiguous answer to the question: "does the learner possess the skill being tested?" This ideal is described by Adams (1974) as the "Dichotomous Outcomes Model." In this model, a learner may be either in the mastery state or the non-mastery state, exclusively. On an ideal, valid test item, all learners in the mastery state will always give correct responses, and all learners in the non-mastery state will always give incorrect responses.

The Dichotomous Outcomes Model implicitly demands 100% correct performance, but this goal is unattainable in an imperfect world with imperfect measuring instruments. Meskauskas (1976) states that "considerations of measurement error essentially always result in the adoption of standards that demand less than the model seeks." Adams acknowledges this limitation by remarking that an "error of testing occurs whenever learner performance on an item does not reflect his true competence in the trait in question."

Thus, Adams points out that two types of errors can occur. One type occurs when a learner who is in the mastery state gives an incorrect response on a valid item. The other occurs when a learner who is in the non-mastery state gives a correct response on a valid item. Ferguson and Novick (1973) define these errors as follows (see also Figure 1):

> A Type I error occurs when an examinee has sufficient proficiency in a skill but the outcome of the testing suggests that he does not. A Type II error occurs whenever the examinee, in fact, lacks proficiency in a skill but on the basis of test results is said to have sufficient proficiency.

		LEARNER'S TESTED STATE: Master	LEARNER'S TESTED STATE: Non-Master
LEARNER'S ACTUAL STATE	Master	No Error	Type I Error
	Non-Master	Type II Error	No Error

Figure 1. **Types of test error.**

The goal of the test designer is to minimize the probabilities of these errors by requiring learners to respond to a large enough number of test items to assure reliability, yet to maximize the cost effectiveness of the testing procedure by keeping the number of items as small as possible. To do this, Millman (1974) proposed that allowance be made for the error of testing by computing a test score "Uncertainty Band" as follows:

$$UB = 2 \times \sqrt{\frac{N - n}{N - 1} \times \frac{P\emptyset \times (1 - P\emptyset)}{n}}$$

where UB is the size of the raw score uncertainty band
N is the number of items in the domain
n is the number of items in the test
Pø is the passing standard in percent correct

Millman claimed that "when scores fall outside of the Uncertainty Band, correct decisions are made [on the learner's mastery state] over 95% of the time."

Emrick (1971) approached the problem from the other side, i.e., given the error probabilities and test length, what is the optimal passing standard? This model includes a factor called the "Ratio of Regret," which is computed by summing quantitative expressions of the Bayes risks associated with each of the two types of decision errors. Emrick's formula is:

$$K = \frac{\log\frac{a}{1-b} + \frac{1}{n} \times \log(RR)}{\log\frac{ab}{(1-a)(1-b)}}$$

where K is the passing standard in percent correct
a is the probability that a Type I error will occur
b is the probability that a Type II error will occur
RR is the Ratio of Regret of Type I errors to Type II errors
n is the test length in number of items

In evaluating Emrick's work, Meskauskas (1976) concluded that:

> Emrick's model ... seems worthwhile to pursue. However, empirical quantification of the variables is likely to be a difficult and time-consuming matter.

Ferguson (1971) developed a Bayesian decision analysis model for computing two criterion scores, PØ and P1, each of which is a percentage of correct responses. A learner is said to have "sufficient proficiency" (mastery) on the skill being tested if his or her score is greater than PØ, and "insufficient proficiency" (non-mastery) if the score is less than P1. The area between PØ and P1 is similar to Millman's Uncertainty Band. The probabilities of Type I and Type II errors in this model are respectively expressed as a and b as in Emrick's model. This model is based on the principles of a sequential probability test ratio (Wald, 1947).

The beauty of Ferguson's model is that it allows the test administrator or developer to assign values to PØ, P1, a, and b to determine the learner's proficiency level to any desired degree of accuracy. This is done as follows. After each test item is administered, the student's score, S, is computed using the formula:

$$S = c \times \log\frac{P1}{P\text{Ø}} + w \times \log\frac{1-P1}{1-P\text{Ø}}$$

where c is the number of items answered correctly
w is the number answered incorrectly

The learner is said to have "sufficient proficiency" if:

$$S < \log\frac{b}{1-a}$$

and "insufficient proficiency" if:

$$S > \log\frac{1-b}{a}$$

If neither of the above inequalities is true, that is, if:

$$\log\frac{b}{1-a} < S < \log\frac{1-b}{a}$$

another test item is presented.

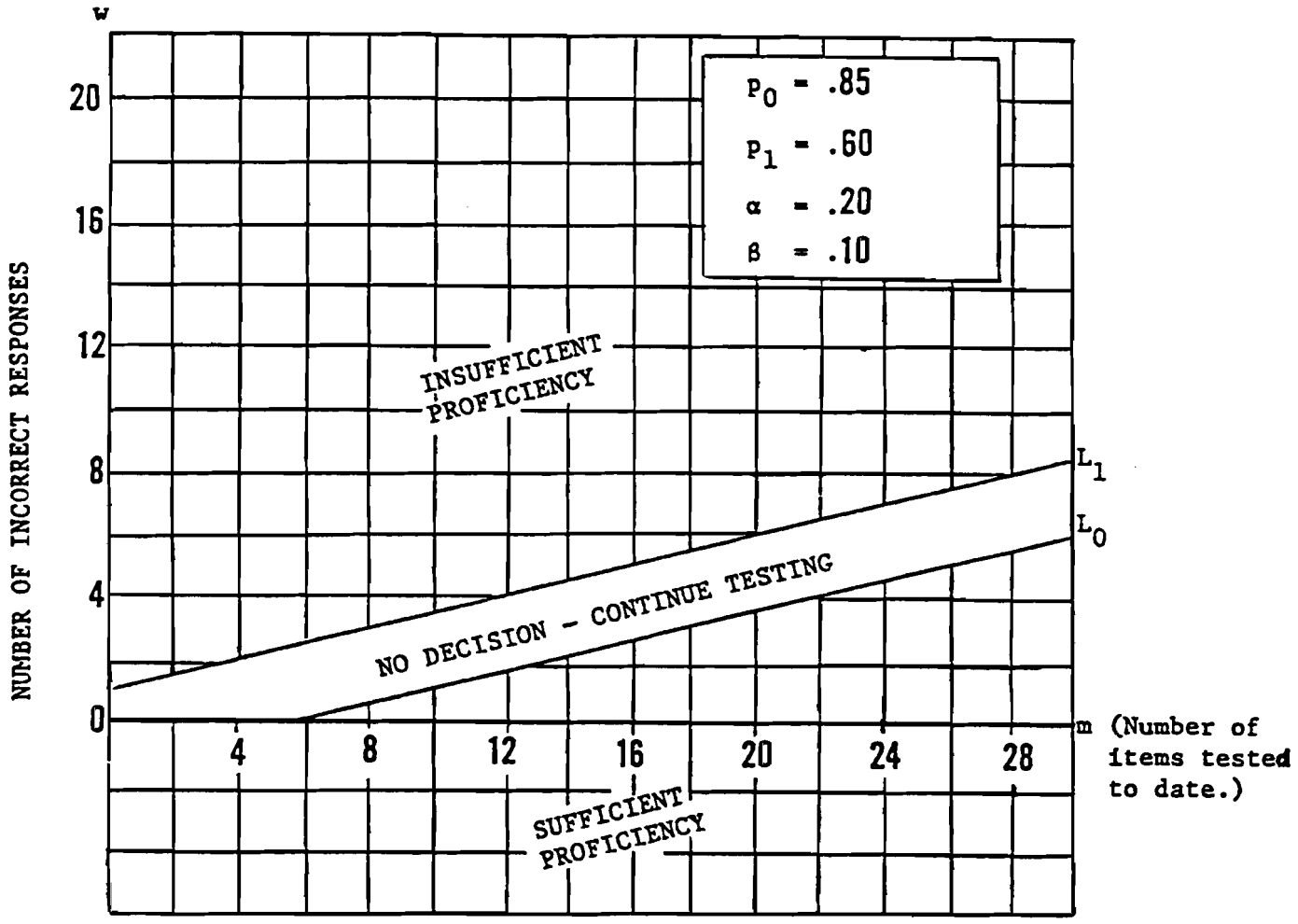
As an example of Ferguson's scheme, consider an exam with the following parameters:

```
PØ = Ø.85
P1 = Ø.6Ø
a  = Ø.2Ø
b  = Ø.1Ø
```

Using these values, the graph in Figure 2 can be constructed to illustrate how a learner's test results would be used in determining proficiency. Note that the learner's proficiency state cannot be classified after just one response is made due to the position of the "Uncertainty Band" for the values of PØ, P1, a, and b chosen. At least two items must be presented and answered incorrectly for a learner to be classified as possessing insufficient proficiency, and at least six must be presented and answered correctly for the opposite classification to be made. By changing the values of these four parameters, the position of the "Uncertainty Band" can be altered. This model forms the basis of the mastery algorithm implemented for this study.

Reliability Models

The concept of criterion-referenced reliability as a measure of the consistency of mastery and non-mastery classifications is one which has received considerable support (Carver, 1970; Hambleton and Novick, 1972; Hansen *et al.*, 1977; Livingston, 1976; Subkovniak, 1976, 1978; Curlette, 1977). Such measures require two sets of test data. The frequencies of agreement between the classification decisions made by both sets of test data may then be represented in a 2x2 table as shown in Figure 3.



H_0: p=.85 (Student has sufficient proficiency, omit instruction)

H_1: p=.60 (Student does not have sufficient proficiency, give instruction)

Figure 2. **Ferguson's method for determining proficiency on a criterion-referenced test. (Ferguson, 1971)**

		CLASSIFICATION ON T1	
		Master	Non-Master
CLASSIFICATION ON T2	Master	A	C
	Non-Master	B	D

Figure 3. **Frequencies of agreement between mastery and non-mastery classifications on two sets of test data.**

In the table in Figure 3, A is the number of students who were classified as masters on both T1 and T2, and D is the number who were classified as non-masters on both tests. As these frequencies increase, the more the two sets of data agree and the higher the reliability of classification. Conversely, B and C are the disagreement frequencies, and as they increase the reliability of classification decreases.

Carver (1970) points out that reliability of classification does not depend on score variability, and is therefore useful in assessing the reliability of criterion-referenced tests. The simplest expression of a reliability coefficient based on this concept is the percentage of cases in which both sets of data agree, namely:

$$P\emptyset = \frac{A+D}{A+B+C+D}$$

This measure varies between 0 and 1 and is referred to as the "percentage of agreement."

Swaminathan *et al.* (1974) prefer using a refinement of the percentage of agreement known as the kappa coefficient. This expression attempts to correct the percentage of agreement for chance. The computation is:

$$\text{kappa} = \frac{P\emptyset - Pc}{1 - Pc}$$

where $P\emptyset$ is the percentage of agreement

$$Pc \text{ is } \frac{(A+C)(A+B)+(B+D)(C+D)}{(A+B+C+D)^2}$$

Swezey and Pearlstein (1975) prefer a slightly more sophisticated expression called the phi coefficient. This coefficient is really the correlation of two sets of test data using 0 as the non-mastery score and 1 as the mastery score. The computation is:

$$\text{phi} = \frac{AD - BC}{\sqrt{(A+B)(A+C)(B+D)(C+D)}}$$

Swezey and Pearlstein suggest that phi > 0.5 represents "sufficient reliability," while phi < 0.5 represents "insufficient reliability." Note that if B = C, kappa = phi.

Livingston (1976) analyzed these computations and suggested yet a fourth coefficient. His purpose was to modify the simple percentage of agreement, PØ, so that it varies between −1 and +1 (like the kappa and phi coefficients) and to show that this new coefficient, the G index, more logically reflects the reliability of classification. The computation is:

$$G = 2 \times (P\text{Ø} - \text{Ø}.5)$$

Two examples from Livingston's work suffice to make his point. Consider the data in Figure 4.

Case 1

T2 \ T1	M	N-M
M	2Ø	6Ø
N-M	Ø	2Ø

Case 2

T2 \ T2	M	N-M
M	9Ø	5
N-M	5	Ø

Figure 4. **Sample classification frequencies.** (after Livingston, 1976)

Livingston argues that the data in Case 1 clearly show that, in most cases, T1 and T2 do not agree. Yet the kappa and phi coefficients for these data are +0.12 and +0.25, respectively, which are small but definitely positive. The corresponding G index for the data in Case 1 is -0.20, which, Livingston argues, more accurately indicates the disagreement because it is negative.

The data in Case 2 are even more striking: T1 and T2 agree in 90% of the testing cases, yet the kappa and phi coefficients are both −0.05. The corresponding G index is 0.80. Here again, Livingston argues, the G index more accurately reflects the correlation of classification because it is positive.

Subkoviak (1978) found that for all four reliability computations, reliability estimates stabilize as test length increases. All four of these indices are reported for the data gathered during this study.

IMPLEMENTATION OF THE STUDY

An interactive, computer-assisted testing (CAT) program was written to evaluate learning in a self-paced course on BASIC language programming. Students worked through the course as shown in Figure 5. Before studying each module, students were given the opportunity to take a pretest. If they could demonstrate mastery on this test, they were branched to the pretest for the next module in the hierarchy. This loop continued until students came to a test on which they could not demonstrate mastery. At this point, they were instructed to study that module off-line, and to return to the CAT program when they were ready for the posttest.

The Mastery Algorithm

The CAT program employed Ferguson's sequential probability scoring algorithm with minor modifications. Ferguson's algorithm was designed for tests in which the probability of getting an item correct by guessing is the same for all items. The CAT program developed for this study presents true/false, yes/no, and four- and five-alternative multiple choice items, which have varying probabilities of getting them correct by guessing. Therefore, the algorithm had to be modified. Each item was assigned a weight, `W`, according to the formula:

$$W = \frac{.25}{Pg}$$

where `Pg` is the probability of getting the item correct by guessing

Using this formula, true/false and yes/no items were assigned a weight of .25/.50 or .50. Four-alternative multiple choice items were assigned a weight of .25/.25 or 1.00, and five-alternative multiple choice items a weight of .25/.20 or 1.25.

After each test item was administered, the student's score, `S`, was computed using the following version of Ferguson's formula:

```
S = C x log(P1/PØ) + (T-C) x log((1-P1)/(1-PØ))
```

where `C` is the sum of the weights of the items answered correctly, and
`T` is the sum of the weights of all items that have been presented (thus, `T-C` is the sum of the weights of the items answered incorrectly)

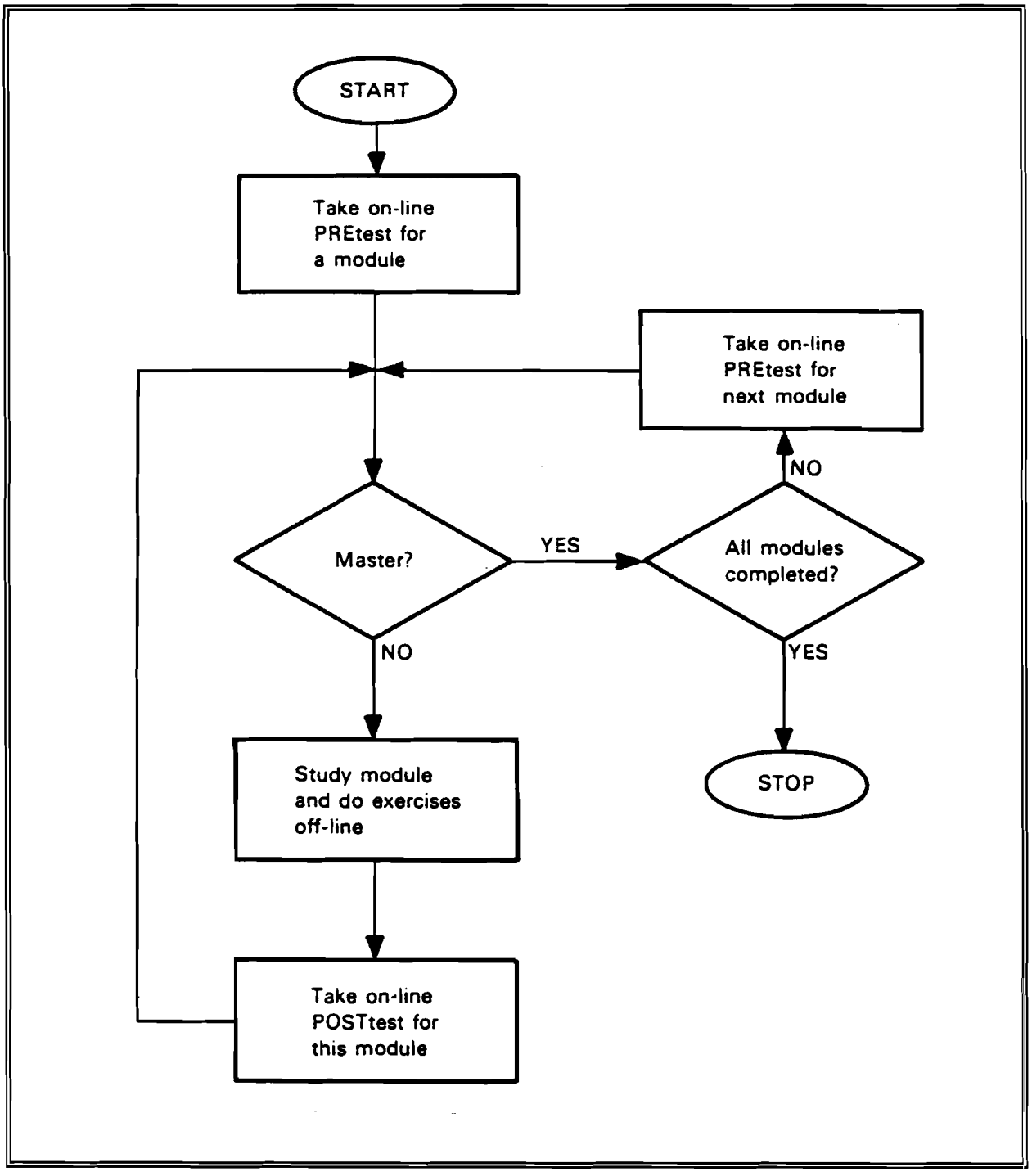


Figure 5. **Relation between on-line CAT program and off-line learning modules.**

This score was compared to the logarithmic functions described earlier, and a decision was made to classify the student as a master or non-master or to continue testing. If a mastery decision still could not be made after 30 items had been administered, the system classified the student based on the differences between his or her score and the two criteria. The student was classified in the category whose criterion score was closest to his or her computed score after 30 items.

The CAT programs recorded the following data so that the effect of the sequential testing model on testing time could be evaluated:

(1) the total amount of time that each student was logged in and running the CAT programs

(2) the number of times that each student logged into the CAT system

(3) the total amount of time that each student spent taking tests

(4) the number of test items answered by each student

(5) tallies of the number of tests that contained each possible number of items (1-30) for:

- pretests that resulted in master classifications
- pretests that resulted in non-master classifications
- posttests that resulted in master classifications
- posttests that resulted in non-master classifications

The Reliability Algorithm

The study assessed criterion-referenced test reliability as a reliability of classification. The two sets of test data used to assess this reliability were the mastery decisions made on the normal (variable length) versions of specific tests and those made on the same tests when they were extended to 30 items. This is a within-subject design because it yields two measures for a single student on a single test.

To do this, every fifth test presented to any particular student was extended to 30 items in length, regardless of the test parameters. When the scoring algorithm made its initial decision, a tentative mastery classification was recorded. The system then continued presenting test items until the maximum of 30 had been presented, at which time the final master classification was recorded. (The transition from short test to long test was imperceptible to the student being tested.) This data was analyzed to determine the percentage of agreement between the two classifications, and the kappa, phi, and G indices. In addition, the CAT programs recorded:

- the amount of time spent on the shorter portion of each extended test and the amount of time spent on the entire 30-item test

- the number of items at which testing could have been terminated for each extended test

These data made it possible to assess the time advantage gained by using the sequential testing mastery algorithm over the fixed length 30-item tests.

Results of the Study

Data on the use of the CAT system were collected from four sites:

- Falmouth High School
  Falmouth, Massachusetts
- Rhode Island Junior College
  Lincoln, Rhode Island
- Wachusett Regional School District
  Holden, Massachusetts
- Watertown Senior High School
  Watertown, South Dakota

The system was used mainly by adults for teacher training.

Test Length Data

The CAT system stored two types of data relating to this subproblem. The first is test time data that are summarized in Table 1. The average time per test was computed by dividing the total student testing time by the number of tests taken. The login overhead was computed by dividing the difference between the total student time on-line and the total student testing time by the number of test items presented.

The data in Table 1 show that:

- students typically required 9.6 minutes to take one test
- students typically took only one or two tests each time they logged in to the CMI system (1400 tests taken with 1405 logins)
- students typically spent about half as much time logging into the system and displaying their status as they did taking tests (48.9% login overhead)

One may therefore conservatively estimate that students typically required approximately 15 minutes to log into the system and take a test. This figure is highly supportive of the contention that use of the CAT system required a minimum amount of time, especially

Table 1. **Summary test time and related data for each site.**

	All Sites Combined	Wachusett Regional School District	Rhode Island Junior College	Falmouth High School	Watertown Junior High School
Total Student Time On-Line (hours:minutes)	332:53	21:05	139:02	41:37	131:09
Number of Student Logins	1405	94	388	264	659
Total Student Testing Time (hours:minutes)	223:34	12:31	97:37	33:05	80:22
Number of Tests Taken	708	108	465	213	614
Number of Test Items Presented	21014	1433	7278	3600	8703
Average Time Per Test (minutes)	9.6m	7.0m	12.6m	9.3m	7.9m
Average Login Overhead (percent)	48.9%	68.4%	42.4%	25.8%	63.2%
Average Time Per Item (seconds)	38.3s	31.4s	48.3s	33.0s	33.2s

when one considers that the test time data include tests extended arbitrarily to 30 items (about 20% of all test administered) to allow assessment of reliability.

The second type of data is observed test length data. To assess the effect of the sequential testing mastery algorithm, the system stored tallies of the number of tests that resulted in each possible test length (1-30 items). These data are shown in Table 2, broken down by pretests and posttests and masters and non-masters.

The data in Table 2 show that, in 71.5% of the tests that resulted in mastery classifications and in 99.2% of the tests that resulted in non-mastery classifications, the sequential testing algorithm was able to terminate the test before it reached 30 items in length. This indicates that sequential testing contributed significantly to reducing test lengths and shortening test time.

Table 2. **Summary test length data for all sites combined.**

Test Length in No. Items	No. of Pretest Masters	No. of Pretest Non-Masters	No. of Posttest Masters	No. of Posttest Non-Masters	Total No. of Masters	Total No. of Non-Masters
1	Ø	Ø	Ø	Ø	Ø	Ø
2	Ø	8	Ø	14	Ø	22
3	Ø	22	Ø	2Ø	Ø	42
4	Ø	25	Ø	34	Ø	59
5	Ø	21	Ø	29	Ø	5Ø
6	Ø	18	Ø	2Ø	Ø	38
7	Ø	2Ø	Ø	21	Ø	41
8	Ø	1Ø	5	22	5	32
9	Ø	15	14	24	14	39
1Ø	Ø	1Ø	17	15	17	25
11	2	8	13	17	15	25
12	1	6	11	12	12	18
13	1	6	18	1Ø	19	16
14	1	1Ø	13	11	14	21
15	1	3	12	3	13	6
16	Ø	5	11	8	11	13
17	2	3	1Ø	11	12	14
18	Ø	8	15	9	15	17
19	Ø	7	12	9	12	16
2Ø	1	5	7	6	8	11
21	1	4	9	8	1Ø	12
22	4	5	3	6	7	11
23	1	3	8	5	9	8
24	1	1	12	2	13	3
25	2	2	8	Ø	1Ø	2
26	2	5	9	4	11	9
27	1	2	4	1Ø	5	12
28	2	7	7	9	9	16
29	5	3	7	5	12	8
3Ø	19	2	82	3	1Ø1	5
Totals	47	244	3Ø7	347	354	591

The median test lengths for each of the four types of tests are shown in Table 3. They were all less than 30 items, and vary somewhat linearly with the various certainty criteria (error probabilities) set by the author. Further data on the 155 tests extended to 30 items indicate that, on the average, these extended test could have been terminated after 19.1 items had been presented if the sequential probability test ratio had been applied. These early

Table 3. **A priori error probabilities and a posteriori median test lengths.**

Test Type and Classification	Critical Score	Allowable Error Probability	Median Test Length
Pretest Master	9Ø%	.Ø25	29
Posttest Master	85%	.Ø5Ø	2Ø
Pretest Non-Master	65%	.Ø58	8
Posttest Non-Master	6Ø%	.1Ø4	9

terminations would have resulted in an average time saving of 5.7 minutes on each extended test.

Test Reliability Data

Figure 6 presents the test reliability data for all sites combined. These data show that there were a total of 155 extended tests and that in 150 (96.8%) of these the decisions on the short tests and extended tests agreed. In 4 cases, the system would have made a Type I or false negative error (by classifying a true master as a non-master if the early decision had been allowed to stand. In addition, the system would have made 1 Type II (false positive) error if it had accepted its early decisions. That is, it would never have classified a true non-master as a master.

		EARLY DECISION	
		Master	Non-Master
EXTENDED DECISION	Master	56	4
	Non-Master	1	94

Figure 6. **Test reliability data for all sites combined on modules 2-16.**

The corresponding reliability indices for the data in Figure 6 are as follows:

```
Percentage of Agreement = Ø.968
                 Kappa = Ø.931
                   Phi = Ø.932
                     G = Ø.935
```

These indices indicate that the sequential probability test ratio yielded highly reliable classifications, even when the tests were shortened.

CONCLUSIONS AND DIRECTIONS FOR FUTURE STUDY

The study showed that sequential testing can be used to reduce test lengths significantly without sacrificing test reliability. In addition, it showed that the amount of test length reduction and reliability indices desired can be controlled by using Ferguson's model.

The observed test length data indicates that the four sequential testing parameters (PØ, P1, a, and b) may have been a bit too stringent, especially for pretests on which mastery decisions were made, which had a median length of 29 items. Further research should be conducted into varying the parameters so that the balance between test length and reliability can be optimized.

As shown in Table 3, the critical scores (PØ and P1) were set at different values for pretests and posttests. This introduced an unnecessary complication in analyzing the test length data because it made it difficult to ascertain how much of the differences in median test lengths was attributable to differences in error parameters alone. A further study might be done in which PØ and P1 are held constant on the two tests and only a and b are changed.

Since the test item banks were not precalibrated before they were used at the test sites, it was impossible to weight the items based on observed difficulty indices. A weighting algorithm based on item type was therefore implemented. This algorithm has been criticized on the grounds that 4-alternative multiple choice items may not prove twice as difficult as true/false or yes/no items. Further analysis of this criticism is warranted, including the more global question of using several types of items in a single sequential probability test.

Another issue related to weighting is the varying importance of objectives. In the current study, the importance of each objective (and therefore the probability of selecting an item for that objective) was equal. A further study might modify the existing software to all objectives, as well as items to be weighted.

REFERENCES CITED AND RELATED READINGS

Adams, E.N., 1974. On scoring a mastery learning control test. *Journal of Computer-Based Instruction* 1(2):50-58.

Carver, R.P., 1970. Special problems in measuring change with psychometric devices. In *Evaluative Research: Statistics and Methods*. American Institute for Research, Washington, D.C.

Curlette, William L., 1977. Assessing the reliability of criterion-referenced tests. Paper presented at a conference on Innovative Education: Preservice through Inservice, Atlanta, GA.

Emrick, J.A., 1971. An evaluation model for mastery testing. *Journal of Educational Measurement* 8:321-326.

Ferguson, Richard L., 1971. Computer assistance for individualized measurement. Learning Research and Development Center, University of Pittsburgh.

Ferguson, Richard L., and Melvin R. Novick, 1973. The implementation of a Bayesian system for decision analysis in a program of Individually Prescribed Instruction. American College Testing Program, Iowa City, Iowa, Research Report No. 60.

Hambleton, Ronald K., and Melvin R. Novick, 1972. Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement* 9:27-29.

Hansen, Duncan N., Steven Ross, and Dickie A. Harris, 1977. Flexilevel adaptive testing paradigm: hierarchical concept structures. Technical Training Division, Lowry Air Force Base, Colorado. AF-HRL-TR-77-35 (II).

Livingston, S., 1976. Lecture on "statistical concepts in criterion-referenced measurement." Educational Testing Service, Princeton, N.J.

Meskauskas, John A., 1976. Evaluation models for criterion-referenced testing: views regarding mastery and standard setting. *Review of* Educational Research *36(1):133-158.*

Millman, Jason, 1974. Sampling plans for domain-referenced tests. *Educational Technology* 14(6):17-21.

Subkoviak, Michael J., 1976. Estimating the reliability from a single administration of a criterion-referenced test. Paper presented at a conference of the American Education Research Association, San Francisco, CA.

Subkoviak, Michael J., 1978. Empirical investigation of procedures for estimating reliability of master tests. *Journal of Educational Measurement* 15(2):111-116.

Swaminathan, H., R.K. Hambleton, and J.J. Algina, 1974. Reliability of criterion-referenced tests: a decision-theoretic formulation. *Journal of* Educational Measurement *11:263-267.*

Swezey, R.W., and R.B. Pearlstein, 1975. *Guidebook for Developing Criterion-Referenced Tests.* U.S. Army Research Institute for the Behavioral and Social Sciences.

Wald, Abraham, 1947. *Sequential Analysis.* John Wiley and Sons, Inc.