THE USE OF INTERACTIVE, COMPUTER-MANAGED INSTRUCTION TO CONTROL THE QUALITY OF SELF-PACED TRAINING WITHOUT REQUIRING THE PRESENCE OF AN INSTRUCTOR

Jesse M. Heines

Educational Services
DIGITAL EQUIPMENT CORPORATION
Bedford, Massachusetts 01730

The cost of computer systems is constantly decreasing, but the cost of training people to run these systems is constantly increasing. To make training cost-effective for its small systems customers, Digital Equipment Corporation has been providing self-paced training packages for several years that can be used by its customers on-site. To control the quality of this training, Digital is now incorporating Computer-Managed Instruction (CMI) into some of these packages. This CMI component uses a sequential probability test algorithm that allows tests to vary in length depending upon the learner's skill level. This algorithm assures that the tests' results are statistically reliable while keeping their lengths as short as possible.

Keywords: Computer-Assisted Testing; Computer-Based Training; Computer-Managed Instruction; Sequential Testing.

PACKAGED CUSTOMER TRAINING

Digital's Educational Services Department has been developing individualized, self-paced training packages since 1975. These packages are designed to teach users to operate their systems without requiring the presence of a Digital instructor. We have found, however, that the use of packaged training at customer sites presents two new problems of its own:

- it is difficult to control the use of these packages on customer sites as well as we can control them in Digital facilites, and
- it is difficult to get accurate feedback on the strengths and weaknesses of these packages from our customers.

Digital's Computer-Based Course Development Group is addressing these problems by writing computer-managed instructional (CMI) materials to run under several of our operating systems. These CMI materials use the customers' computers themselves to control their learning and collect data that we can use to assess the effectiveness of the training packages.

This paper describes a computer-managed instruction program that we have developed which is coupled with a new self-paced training package.

CMI IN A CUSTOMER ENVIRONMENT

Interaction Between CMI and Training Packages

Digital's self-paced training packages are written in a modular format. The modules are arranged in a specific learning hierarchy, based on the prerequisite relationships of their objectives. Each module contains a list of its objectives, text and diagrams to help learners master these objectives, and exercises to be performed both on paper and on a computer system. Each module also has a related module test. The item banks for the module tests are all stored on-line.

Before learners begin work on the training package, they take the pretest for the first module interactively at a computer terminal (see Figure 1). If they can demonstrate mastery on this test, the CMI system branches them to the pretest for the next module in the hierarchy. This loop continues until the learners come to a test on which they cannot demonstrate mastery. At this point, they are directed to study that module off-line, and return to the CMI system when they are ready for the posttest.

An important quality of the CMI approach is that it gets users on-line as soon as possible and therefore has a definite Hawthorne Effect[1]. In the past, customers often just skipped the tests that were included in our training packages, because they felt (erroneously) that testing benefits only the teacher. It is difficult, if not impossible,

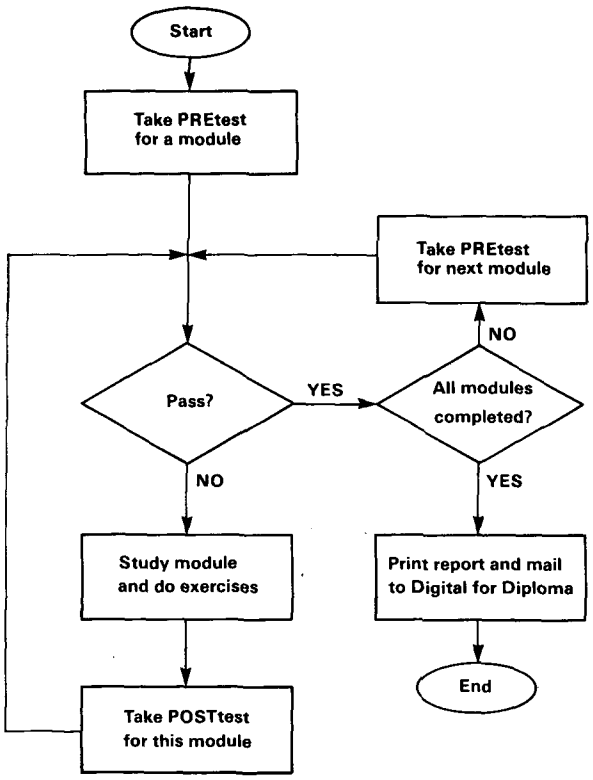

Figure 1. Interaction Between On-Line CMI System and Off-Line Training Package

to change this feeling, but we can capitalize on the Hawthorne Effect to get more of our customers to take the tests. The directions for running the CMI system are provided in the training in cookbook terms, so that even the most inexperienced of our users can get the programs on the air.

General CMI Characteristics

The CMI registration program allows users to register themselves interactively. It records their first and last names (making sure that each is unique) and their addresses.

Users then select a code name by which they will identify themselves in future logins to protect the confidentiality of the data that is stored on their work. This program also allows users to view the status of their work on each of the modules in the course.

The CMI test administration program presents true/false, yes/no, and multiple choice items (with either four or five alternatives). These tests are generated interactively in real time. The items are randomly selected from item banks that are categorized by module and objective. A typical item display is shown in Figure 2.

The system is highly human-engineered to make it as foolproof as possible in a customer environment. For example, it provides the options "SKIP", "QUIT", and "REVIEW" as shown in Figure 2, and it provides "error" messages in plain English: if the user enters "g" in response to a true/false item, the system will print, "Please enter only T, F, SKIP, or QUIT". It will then erase the user's previous response and make room for him or her to enter another one.

THE MASTERY DECISION MODEL

Sequential Testing

Even with a considerable Hawthorne Effect, customers still do not like to be tested. (There is always someone in every customer training course who will say, "I paid my money to be taught, not tested!") It is therefore important to keep the length of the module tests as short as possible. It is useless, however, to make these tests so short that their reliability approaches zero.

Since 1974, all courses developed by Digital's Educational Services Department have been developed using a criterion-referenced philosophy. This philosophy is especially applicable to industrial training, because we are interested in individual

```
MULTIPLE CHOICE.  Enter the letter of the
alternative that BEST answers the question or
completes the sentence in the item below.

Type SKIP if you don't know the answer (counted
   as incorrect).
Type QUIT if you must terminate this test
   before it is completed.
Type REVIEW to see the previous test item
   again.
Press the RETURN key after you type your
   answer.


10.  Which of the following statements will
     cause

                  196

     to be displayed on the terminal?

        A.  PRINT "102 + 94"
        B.  PRINT 102 + 94
        C.  PRINT "102 + 94 = 102+94"
        D.  PRINT "102 + 94" = 102+94


     Your answer?
```

Figure 2. Sample Display of a Multiple Choice Item Presented to a Student

performance rather than a comparison between learners. We therefore required the CMI system to apply this philosophy as well.

Through an examination of educational literature [6], we found that the most highly developed criterion-referenced decision module that takes advantage of the capabilities of interactive computing is one developed by Richard Ferguson [4]. Ferguson's model is based on Wald's sequential probability test ratio [11]. This model allows two criterion scores to be defined, P0 and P1. Both of these scores are expressed in terms of percentages of correct responses.

Learners whose scores are greater than P0 are classified as masters, and learners whose scores are less that P1 are classified as non-masters. Learners whose scores fall between P0 and P1 are presented with another item.

This model also takes into account the probability with which the test developer is willing to allow Type I (false negative) and Type II (false positive) errors to occur[2]. Let us define A as the probability that a Type I error will occur, and B as the probability that a Type II error will occur. The test developer can then assign values to P0, P1, A, and B to determine the learners' mastery state to any desired degree of accuracy.

Computations

Ferguson's scoring algorithm is designed for tests in which the probability of getting an item correct by guessing is the same for all items. Since the CMI system presents true/false, yes/no, and four- and five-alternative multiple choice items, which have varying probabilities of getting them correct by guessing, the algorithm must be modified. Each item is therefore assigned a weight, W, according to the formula:

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

where P is the probability of getting the item correct by guessing. Using this formula, true/false and yes/no items are assigned a weight of .25/.50 or 0.50. Four-alternative multiple choice items are assigned a weight of .25/.25 or 1.0, and five-alternative multiple choice items a weight of .25/.20 or 1.25.

After each test item is administered, the student's score, S is computed using the formula:

$$S = C \times \log(P1/P0) + (T-C) \times \log((1-P1)/(1-P0))$$

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.)

The student is classified as a master and testing is terminated if

$$S \leq \log(B/(1-A))$$

and at least one item has been presented on each objective in the module. If the above inequality is true but all objectives have not been tested, another item is presented. The student is classified as a non-master and testing is terminated if

$$S \geq \log((1-B)/A)$$

regardless of the number of items presented on each objective. If neither of these inequalities is true, that is, if

$$\log(B/(1-A)) < S < \log((1-B)/A)$$

another test item is presented. The system continues in this manner until one of the first two inequalities becomes true or until 30 items have been administered. If no decision can be made after 30 items, the system classifies the student based on the differences between his or her score and the two criteria. The student is classified in the category whose criterion score is closest to his or her computed score after 30 items.

IMPLEMENTATION

Test Parameters

As mentioned previously, the CMI system generates both pretests and posttests. For this reason, it is important to realize that the seriousness of making Type I and Type II errors is different on pretests and posttests. If the system makes a Type II (false positive) error on a pretest, it will tell a student who has not studied the corresponding module to skip instruction that he or she really needs. This same error on a posttest is not as serious, because the student will have already studied the module at least once, and one can assume that at least some minimal learning has taken place. A Type I (false negative) error is never as serious as a Type II error, because this situation simply asks a student to repeat instruction that he or she does not really need. This wastes some time, but one can assume that it does not decrease the learner's proficiency level.

To take the relative importance of these errors into consideration, the CMI system uses the parameters shown in Table 2. These parameters were chosen for the following reasons. First, the pretest and posttest mastery and non-mastery criteria were set to span the percentage score of 70-80% that most criterion-referenced tests use as a mastery level when only one cutting score is employed. Second, the mastery criterion for pretests was increased 5% over that for posttests to reflect a slightly more stringent criterion for mastery if

Table 2. Sequential Testing Parameters for Pretests and Posttests

Parameter	For Pretests	For Posttests
Mastery criterion	0.90	0.85
Non-mastery criterion	0.65	0.60
Prob. of Type I error	0.058	0.104
Prob. of Type II error	0.025	0.050

a module has not yet been studied. The non-mastery criterion for pretests was also increased 5% to keep the differences between these two criteria equal for both types of tests. This was necessary because the difference between the two criteria is itself a factor in determining test length. As the difference increases, the number of test items required to make a decision at any given level of certainty decreases. Conversely, as the difference between the two percentage criterion levels decreases, the number of required test items increases.

Third, the allowable probabilities of Type II (false positive) errors were set to 0.025 and 0.050, respectively, for pretests and posttests. The factor of 2 separating these parameters reflects the relative seriousness of making this type of error on pretests versus its seriousness on posttests. That is, it is estimated that the seriousness of making a Type II error on a pretest is twice as great as that on a posttest, so the allowable probability of this error on pretests was decreased by a factor of 2. Finally, the probabilities of Type I (false negative) errors were derived by computing the highest value that would still require at least three items to be presented before a non-mastery decision is made unless the first two items are both five-alternative multiple choice items (with a weight of 1.25 each). The three item consideration was conceived because it was felt that students would distrust the system if they were judged non-masters after only two items had been presented.

The magnitudes of the error probabilities also warrant some discussion. Ferguson (1970) allowed probabilities of 0.20 and 0.10, respectively, for his Type I and Type II errors. These values reflect the same 2:1 ratio to be used in this study, but their magnitudes are approximately twice those of the ones used in this study. The main reason for selecting lower probabilities is that Ferguson's testing unit was the objective, while the current study's testing unit is the module (a group of up to 20 objectives). It was felt that when working on the module level, the consequences of errors of classification are more serious than at the lower objective lever. Thus, the absolute values of the allowable error probabilities were lowered.

To see how these parameters reflect the mastery decision model in terms of raw scores, refer to Figure 3. Figure 3a shows a graph of the pretest decision rules, while Figure 3b shows the posttest decision rules. Note the difference in the sizes of the two master areas and the specific points labelled. The point labelled "(2.5,0)" in both graphs indicates that the earliest that a non-master decision could be made on either test is after the sum of the weights of all items presented totals at least 2.5. If, at this time, the student has not answered any items correctly, he or she will be classified as a non-master.

In Figure 3a, the point labelled "(11.5,11.5)" indicates that the earliest that a master decision could be made on a pretest is after items having a total weight of 11.5 have been presented and all items have been answered correctly. Contrast this point with the one labelled "(8.5,8.5)" in Figure 3b. The latter indicates that the earliest that a master decision could be made on a posttest is after items having a total weight of at least 8.5 have been presented and answered correctly. Therefore, the posttest mastery criterion is less stringent than the pretest mastery criterion. This relationship is exactly the one desired, because it reflects that an erroneous master decision on a posttest is less serious than that on a pretest.

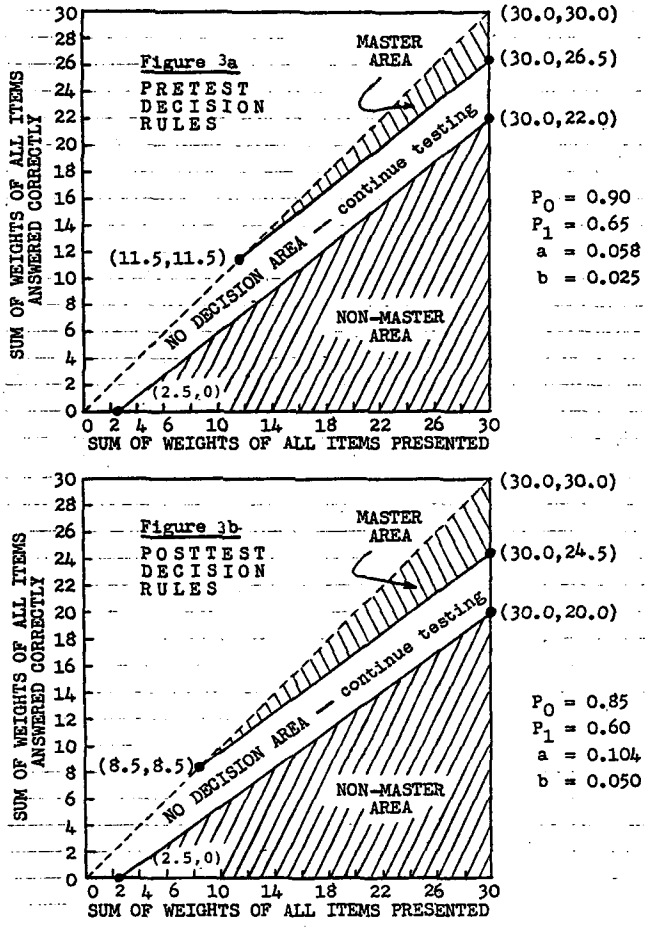

Figure 3. Graphs of the Pretest and Posttest Decision Rules

The Question of Reliability

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 [2, 3, 5, 7, and 8]. 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 4.

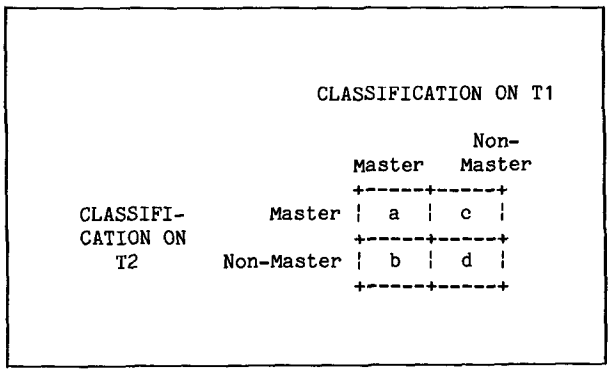

Figure 4. Frequencies of Agreement Between Mastery and Non-Mastery Classifications on Two Sets of Test Data

In this table, 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 [2] 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:

$$PO = \frac{a+d}{a+b+c+d}$$

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

Swaminathan et al. [9] 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:

$$kappa = \frac{PO - Pc}{1 - Pc}$$

where PO is the percentage of agreement, and

$$Pc \text{ is } \frac{(a+c)(a+b)+(b+d)(c+d)}{(a+b+c+d)^2}$$

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:

$$phi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Swezey and Pearlstein suggest that phi $\geq$ 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, PO, 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:

$$G = 2 \times (PO - 0.5)$$

Two examples from Livingston's work suffice to make his point. Consider the data in Figure 5. 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.

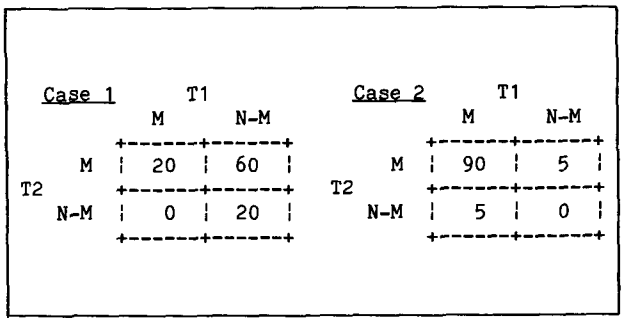

Figure 5. Sample Classification Frequencies [7]

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.

This study assesses criterion-referenced reliability as a reliability of classification using the G index. The two sets of test data used to assess this reliability are the mastery decision made on the normal (variable length) version of a test and that made on the same test when it is extended to 30 items. To do this; every fifth test presented to a particular student is extended to 30 items in length, regardless of the test parameters. When the scoring algorithm makes its initial decision, a tentative mastery classification is recorded. The system continues presenting test items until the maximum of 30 has been presented, at which time the final master classification is recorded. This data is analyzed to determine the percentage of agreement between the two classifications, and the G index will be computed.

Closing the Feedback Loop

The CMI programs are currently implemented on a number of different Digital systems. For this reason, the media on which the CMI programs are distributed and on which student response data is stored varies greatly from system to system. The only common media to these systems is paper, but many have magtapes or floppy diskettes. After users complete the training package and all of the module tests, data on their work is copies to a magtape or floppy diskette or printed on paper and mailed back to Digital. This data allows us to do complete criterion-referenced item analysis on the users' responses and check the status of the users on each module. Users who complete the entire course satisfactorily receive a diploma after their data is analyzed.

FUTURE ENHANCEMENTS

All of the programs that make up this CMI system are written in a subset of the BASIC language. This makes them highly transportable to almost all of Digital's operating systems. In addition, all of the CMI programs and data files for about 800 test items will fit on a single, dual-density diskette (approximately 250K PDP-11 words).

These characteristics make the CMI system applicable to internal and large systems training as well as small systems training, because it is small enough to fit on a diskette yet sophisticated enough to handle more and larger item banks if additional disk space is available. Our future plans include expanding the types of items that the system and mastery algorithms can handle and improving the system's ability to accommodate courses with varying structures.

FOOTNOTES

[1] "The Hawthorne Effect, which was given that label because it was first recognized in a study made at the Hawthorne, Illinois, plant of Western Electric Company, is the tendency of subjects in some experiments to respond the almost any kind of change, apparently due to a feeling of appreciation that someone is paying attention to them." [1]

[2] This study defines a Type I error as a false negative error which occurs when a true master is classified as a non-master by the test. A Type II error is defined as a false positive error which occurs when a true non-master is classified as a master.

REFERENCES CITED

1. BIEHLER, Robert F. Psychology Applied to Teaching. Houghton-Mifflin Company, Boston, 1971.

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

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

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

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

6. HEINES, Jesse M. A review of the literature on criterion-referenced and computer-assisted testing. Boston Univerity, ERIC Document No. ED116633. November, 1975.

7. LIVINGSTON, Steve. Lecture on "statistical concepts in criterion-referenced measurement". Educational Testing Service, Princeton, N.J. February 23, 1976.

8. SUBKOVIAK, Michael J. 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. 1976.

9. SWAMINATHAN, H., R.K. Hambleton, and J.J. Algina. Reliability of criterion-referenced tests: a decision-theoretic formulation. Journal of Educational Measurement 11:263267. 1974.

10. SWEZEY, R.W., and R.B. Pearlstein. Guidebook for Developing Criterion-Referenced Tests. US Army Research Institute for the Behavioral and Social Sciences. 1975.

11. WALD, Abraham. Sequential Analysis. John Wiley and Sons, Inc. 1947.