DOCUMENT RESUME

ED 116 633 IR 002 872

AUTHOR Heines, Jesse M.
TITLE An Examination of the Literature on Criterion-Referenced and Computer-Assisted Testing.
PUB DATE Nov 75
NOTE 45p.

EDRS PRICE MF-$0.76 HC-$1.95 Plus Postage
DESCRIPTORS *Computer Programs; *Criterion Referenced Tests; Educational Testing; Information Processing; Instrumentation; Item Banks; Literature Reviews; Measurement Techniques; Norm Referenced Tests; State of the Art Reviews; Test Construction; *Testing; Testing Programs
IDENTIFIERS Automatic Examination Generator; Classroom Teacher Support System; *Computer Assisted Testing; Domain Referenced Testing; Educational Testing Service; Mentrex Enterprises

ABSTRACT

Criterion-referenced testing (CRT) is defined as a method of ascertaining an individual's status with respect to some performance standard. Computer-assisted testing (CAT) is a method of constructing tests using a variety of computer techniques such as a single test computer printouts, stored item banks, teacher specified criteria, machine readable answer sheets, etc. After an examination of the literature on both subjects, the conclusion reached is that CRT and CAT may help each other in the following ways: (1) item generation techniques may be refined to allow more comprehensive evaluation of domains by making more items available; (2) item sampling algorithms may be used to achieve more representative tests from existing domains; (3) branching tests may be utilized to arrive at the most cost-effective method for evaluating performance; (4) test models may be simulated to ascertain their feasibility; (5) mathematical models may be developed to help define and standardize the criteria by which performance is judged; and (6) CRT can be used more widely as a valid theory to aid the design of CAT systems. There is a 16 page annotated bibliography divided into two separate subject divisions. (Author/NR)

AN EXAMINATION OF THE LITERATURE ON CRITERION-REFERENCED AND COMPUTER-ASSISTED TESTING

by

Jesse M. Heines

a report submitted to

Dr. Louis P. Aikman

on an Individual Study Project in Learning Resources

SED EM 902 M

Boston University School of Education
Department of Educational Media and Technology

November, 1975

TABLE OF CONTENTS

	Page
LIST OF TABLES	3
LIST OF FIGURES	3
PURPOSE AND ORGAINIZATION OF THIS REPORT	4
CRITERION-REFERENCED TESTING	5
THEORY	5
MODELS	6
The Dichotomous Outcomes Model	6
Domain-Referenced Testing	7
Mathematical Interpretations	8
COMPUTER-ASSISTED TESTING	15
INTRODUCTION	15
TEST PRINTING SYSTEMS	16
TEST CONSTRUCTION SYSTEMS USING ITEM BANKS	19
ALGORITHMIC APPROACHES TO ITEM GENERATION	20
INTERACTIVE, BRANCHING TESTS	22
CONCLUSIONS	26
ANNOTATED BIBLIOGRAPHY	28
INTRODUCTION	28
SECTION A Criterion-Referenced Testing	28
SECTION B Computer-Assisted Testing	34

LIST OF TABLES

Table		Page
1	Pools for Constructing Items on Electron Configurations	21

LIST OF FIGURES

Figure		Page
1	Hively's Domain-Referenced Testing Model . . .	9
2	Ferguson's Method for Determining Proficiency on a Criterion-Referenced Test	13
3	Hansen's Sequential Item Tree Network	23

PURPOSE AND ORGANIZATION OF THIS REPORT

Literature on both criterion-referenced testing (CRT) and computer-assisted testing (CAT) is abundant. Relatively few researchers, however, have attempted to synthesize these two fields. The author examined literature from both fields in an effort to:

(1) identify studies that have used CRT models in designing CAT systems,

(2) gain a thorough understanding of the test administration and analysis procedures used in those studies, and

(3) discover other facets of CRT models that might be realized through CAT techniques.

This paper reports on the literature search conducted by the author by discussing representative studies in both fields and noting additional research efforts in an annotated bibliography. The report begins by discussing articles on CRT theories and models. These articles provide a background for examining the second set of papers: reports on existing CAT systems. Conclusions are drawn about the states of the art for both CRT and CAT, and comments made on areas in which the two fields might complement each other.

CRITERION-REFERENCED TESTING

Theory

Criterion-referenced testing (CRT) is perhaps the most significant development in the evaluation of instruction since norm-referenced testing (NRT) was implemented on a large scale in the early 1900's. CRT differs from NRT in the following way:

> Norm-referenced measures are those which are used to ascertain an individual's performance in relation to the performance of other individuals on the same measuring device. . . Criterion-referenced measures [are used] to ascertain an individual's status with respect to some criterion, i.e., performance standard." (Popham and Husek, 1969)

The former are used to make decisions about <u>individuals</u>; the latter, about individuals and <u>treatments</u>. Glaser (1963) adds that NRT provides "information about the capability of a student compared with the capabilities of other students", while CRT provides "explicit information on what the individual can and cannot do".

Cox (1971) feels that "it is possible for a single test to yield both norm-referenced and criterion-referenced information". This posture appears to oppose that held by Glaser, who feels that the choice of items differentiates test design. Many researchers (Adams, 1974; Cox, 1971;

Glaser, 1963; Popham and Husek, 1969) do agree, however, that traditional item analysis information (difficulty and discrimination indices) and test characteristics (reliability and validity) have different meanings in CRT than they do in NRT. That is, decisions on the value of a given item or the worth of a given test would be different in the two applications. For example, Cox and Glaser both note that NRT items must discriminate between individuals on a single test. Therefore, items with difficulty levels of 1,00 or discrimination indices of 0,00 are useless in a norm-referenced test. A criterion-referenced test, however, is designed to make in "generally difficult for those taking it before training and generally easy after training" (Glaser, 1963). Therefore, items that are useless in NRT would be retained in CRT if they are answered correctly after training but answered incorrectly before, i.e., if they provide pretest/posttest discrimination.

Models

The Dichotomous Outcomes Model. The ideal CRT is one which yields a single, unambiguous answer to the question: does the learner possess the skill being tested? This ideal is well described by Adams (1974) as the "Dichotomous Outcomes Model" (DOM). In this model, a learner may be either in the mastery state or the non-mastery state, exclusively. On an ideal, valid test

item, the learner will give a correct response if he/she is in the mastery state and an incorrect response if he/she is in the non-mastery state. Adams states that an "error of testing occurs whenever learner performance on an item does not reflect his true competence in the trait in question".

Thus, two types of errors can occur. A Type I error (in Adams' scheme [1]) occurs when the learner is in the non-mastery state but gives a correct response on a valid item. A Type II error occurs when the learner is in the mastery state but gives an incorrect response on a valid item. The goal of the test designer, therefore, is to minimize the probability of these errors by requiring the learner to respond to a sufficiently large number of items to assure reliability, yet to maximize the cost effectiveness of the testing procedure by keeping the number of items as small as possible. A CAT system that realizes these goals has been designed by Ferguson (1971) and will be discussed later in this paper.

Domain-Referenced Testing. An important field that is a sibling to CRT is Domain-Referenced Testing (DRT). Hively (1974a) differentiates the two as follows:

> The world of psychometrics may be seen as a contrast between Domain-Referenced Testing and Norm-Referenced Testing. The distinction is essentially the same as the one Robert Glaser made between

---

1. These two types of errors are also described by Ferguson (1971), but the numbers of the types are switched. That is, Adams' Type I error corresponds to Ferguson's Type II, and Adams' Type II corresponds to Ferguson's Type I.

> Norm-Referenced Testing and Criterion-Referenced Testing. But the term "criterion" lends itself to misinterpretation. It carries surplus associations to mastery learning that are best avoided by using the more general term "domain" instead. Most people who talk about Criterion-Referenced Testing assume that the technology of Domain-Referenced Testing exists, but they often do not fully recognize what that would imply. (page 5)

Hively further clarifies DRT theory with the diagram in Figure 1.

It is this author's opinion that the distinction between CRT and DRT is most important when working with the cognitive and affective domains, where the universe of target behaviors can indeed be abstract and infinite. In the psychomotor domain, and even in some applications in the cognitve domain, the universe of target behaviors can usually be much more clearly defined and approach a concrete domain, thereby minimizing the distinction between CRT and DRT for these behaviors. The problem seems to one of the preciseness with which the behavioral objective can be stated.

Hively (1974a) and Baker (1974) both emphasize the importance of transfer in constructing items for inclusion in a test domain. The goal of the DRT constructor, according to Hively, is "to create an extensive pool of items that represents, in miniature, the basic characteristics of some important part of the orginal universe of knowledge. . . The basic notions that guide this activity are those of <u>generalization</u>, <u>transfer</u>, and <u>subject matter structure</u>".

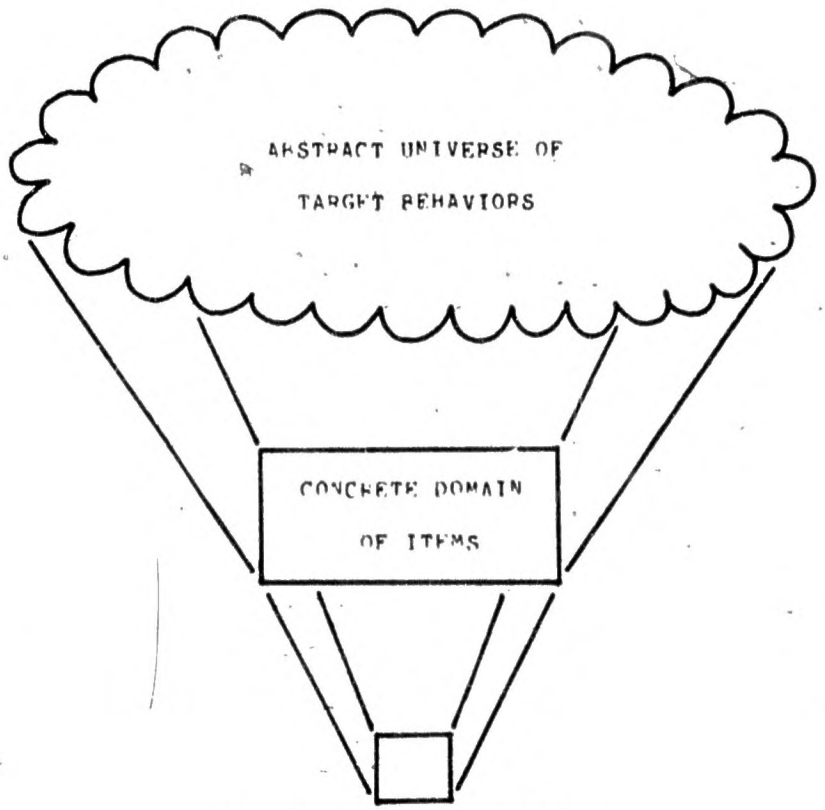

Figure 1

Hively's Domain-Reference Testing Model

(after Hively, 1974b)

Mathematical interpretations. Millman (1974) and Ferguson (1971) have both worked to interpret CRT models into mathematical terms. Their work provides means for implementing the DRT and DOM models in real testing situations.

Millman (1974) models potential testing situations as a three-dimensional matrix of items, examinees, and occassions. Items are performances that are "unambiguously scoreable as either correct, incorrect, or not attempted", i.e., their outcomes are dichotomous. Occassions are "observations designed to detect the growth or change in which we are interested". When examinations are scored, the percent of items correct is judged against a passing standard, but allowance is mad for the error of testing by computing the "Uncertainty Band" (UB) as follows:

$$[1] \qquad UB = 2\sqrt{\left[\frac{N-n}{N-1}\right]\frac{P_0(1-P_0)}{n}}$$

where N is the number of items in the domain,
n is the number of items in the test, and
P_0 is the passing standard in percent.

It is interesting to note that as the number of items in the domain (N) approaches infinity, the term [(N-n)/(N-1)] approaches 1, and Equation [1] then simplifies to:

$$[2] \qquad UB = 2\sqrt{\frac{P_0(1-P_0)}{n}}$$

Millman claims that "when scores fall outside of the Uncertainty Band, correct decisions [on the learner's

mastery state) are made over 95% of the time".

Ferguson (1971) developed a much more generalized mathematical interpretation of the DOM. His interpretation uses two test scores that are each percentages of correct responses expressed as decimals, p_0 and p_1. A learner is said to have "sufficient proficiency" (mastery) on the skill being tested if his/her score is greater than p, and "insufficient proficiency" (non-mastery) if the score is less than p.

Ferguson then identified the two types of errors discussed by Adams [2]. He defined α as the probability that a Type I error will occur, that is, the probability that a learner with sufficient proficiency will be incorrectly classified as having insufficient proficiency by the test results. The probability that a Type II error will occur was defined as β.

The test administrator or developer could then assign values to p_0, p_1, α, and β and determine the learner's proficiency to any desired degree of accuracy as follows. After each item is administered, a score, S, is computed using the formula:

[3] $$S = c \cdot \log\frac{p_1}{p_0} + w \cdot \log\frac{1-p_1}{1-p_0}$$

where c is the number of items answered correctly, and

2. Note once again that Ferguson's Type I error corresponds to Adams' Type II, and Ferguson's Type II corresponds to Adams' Type I.

w is the number of items answered incorrectly.

If the learner has sufficient proficiency,

$$[4] \qquad S \leq \log\frac{\beta}{1-\alpha}$$

If the learner has insufficient proficiency,

$$[5] \qquad S > \log\frac{1-\beta}{\alpha}$$

If neither inequality [4] nor [5] is true, i.e., if

$$[6] \qquad \log\frac{\beta}{1-\alpha} < S < \log\frac{1-\beta}{\alpha}$$

another test item is administered.

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

$p_0 = .85$

$p_1 = .60$

$\alpha = .20$

$\beta = .10$

With 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 p_0, p_1, α, and β chosen. At least two items must be answered incorrectly for alearner to be classified as possessing insufficient proficiency, and at least six must be answered correctly for the opposite

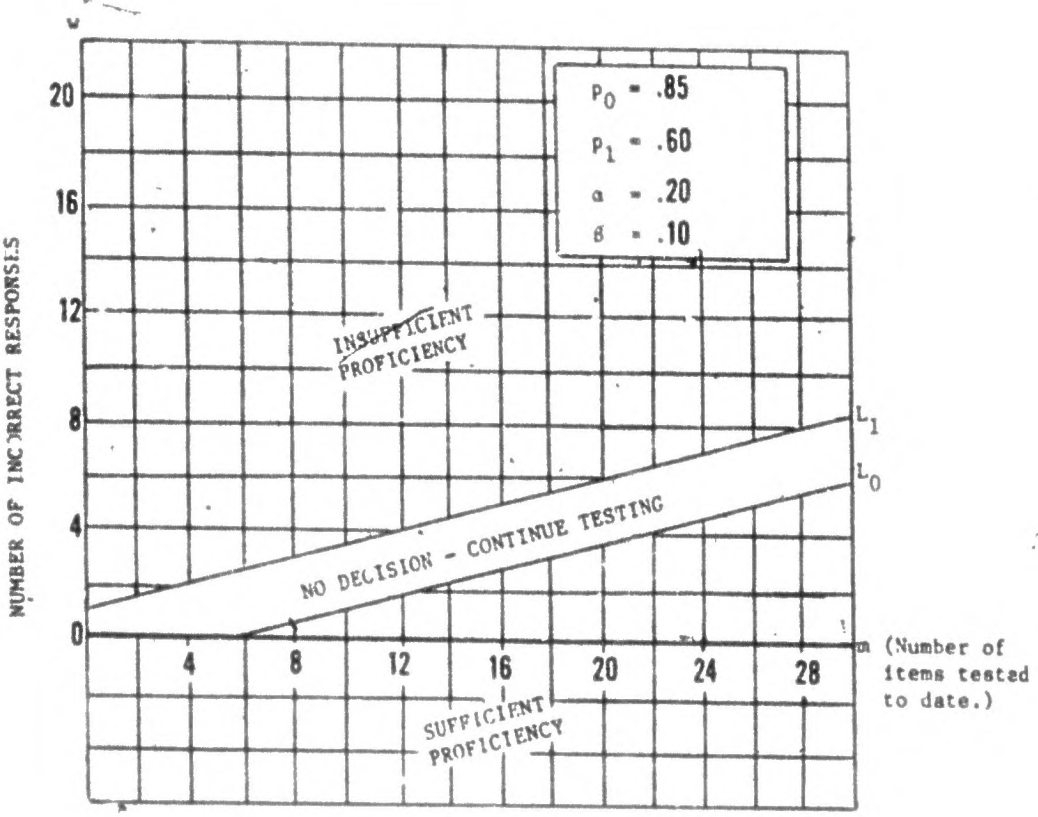

Figure 2

Ferguson's Method for Determining Proficiency on a Criterion-Referenced Test

(Ferguson, 1971, p. 30)

classification to be made. By changing the values of the variables, the position of the Uncertainty Band may be altered. The implementation by Ferguson of these mathematical scoring algorithms into a sophisticated CAT system is discussed later in this report.

COMPUTER-ASSISTED TESTING

Introduction

Computer-assisted testing (CAT) is one of the fastest growing applications of instructional computing. Constructing tests by computer is a relatively straightforward process and can be shown to be cost-effective (Ansfield, 1973; Menne and Lustgraaf, 1974; Prosser, 1975). Lippey (1973) enumerates the major benefits of CAT as follows:

(1) reduces clerical chores required of an instructor,

(2) provides error-free text,

(3) allows the educator to concentrate on content rather that the mechanical aspects of test construction,

(4) eliminates the problem of securing test items from premature release if the item bank is sufficiently large, and

(5) centralized collection of items allows input from many users, thus improving the quality of the items through experience.

A large variety of CAT systems are currently in use, from those that store only item characteristics (ETS, 1974)

to those that construct and administer tests through an interactive terminal (Ferguson, 1971). The systems discussed in this report are grouped into four major categories by their apparent level of sophistication. Systems at the first three levels generally employ batch processing and include, respectively, systems that store and print teacher-constructed exams, those that automatically construct exams from a given item bank, and those that employ an algorithmic approach to item construction. The fourth level is characterized by interactive systems that make use of branching tests to control the sequence in which items are presented to the student.

Test Printing Systems

The simplest type of CAT system is one which does the job of a secretary by printing test questions selected by an instructor (Remondini, 1973). The items to be printed may be stored in any machine readable format, e.g., magnetic tape, disk, or punched cards. In Remondini's system, the computer produces a single copy of the test. This is photocopied and transferred onto ditto masters for duplication. The answer sheets are corrected by a mark sense device and the computer is then used to produce an item analysis and update the statistical data for each item on punched cards.

Salisnjack (1973) uses a system almost identical to

Remondini's. He claims that it only takes 25 minutes to prepare two forms of a 75-item multiple choice test with the aid of the computer. Salisnjack finds that his CAT system controls the cost of test construction, solves the problem of cheating, and reduces the "edge" provided by fraternity test files. He comments, however, that "attempts at making the complete data bank available to all students as a study guide so far have been unsuccessful--the cost of providing individual copies is too high, and copies placed in the library tend to disappear".

MENTREX Enterprises in Los Angeles is a commercial company that provides test construction services similar to those offered by the systems of Remondini and Salisnjack (Libaw, 1973). Users request tests through the mail by selecting questions from a "catalog" supplied by the company. The system can produce several forms of the same test by "scrambling" the items or select items for the test based on "keys" specified by the user. Test masters are returned ready for duplication, along with an answer key and machine readable answer sheets. Answer sheets are later returned to MENTREX for item analysis.

Educational Testing Service (1974) is a unique user of CAT due to the sheer size of their operation. They have stated that there are two tasks that are necessary before they can implement large scale CAT use, and they do not yet see these tasks as part of the current state of the CAT art. These tasks are:

(1) "the development of detailed item classification systems", and

(2) "delineation of the professional judgements made in building a test from a group of itmes in detailed content, ability, and statistical specifications in terms precise enough to be translated into computer programs",

ETS currently uses a CAT system to help select items from their huge data banks. The system does not print tests, but simply returns item numbers that fit specified characteristics. Their computer records on each item include:

(1) the item ID number,

(2) its classification,

(3) a history of its use,

(4) up to five sets of statistics,

(5) codes for security level and current activity, and

(6) twelve 15-character keywords.

It is interesting to note that ETS sees the demand for large national selection tests as diminishing. They feel that interactive testing is required for the future, with tests for guidance, placement, and evaluation. Their paper states that the technology for such systems exists, but that development funds are needed to make them cost-effective.

Test Construction Systems Using Item Banks

The second level of CAT is characterized by systems that construct tests from stored item banks. In addition to the benefits noted earlier, these systems provide a means for generating multiple forms of the same test. Jensen (1973) has used such a system to generate 4000 different forms for a class of 1500 students. He achieves criterion-referencing by allowing students to take a test on a specific topic as often as they like and counts only the highest grade. His philosophy in this approach is that "...one should ask only what one wishes the student to know, but ask it in so many different ways that the student cannot learn the items without learning the concept".

Prosser (1973) describes a similar test construction system but includes some figures on its cost. This system selects items from predefined "groups" that are specified by the user. To produce 1000 3-page tests, the system requires 20 seconds of CPU this and three hours of printer time, making the cost of each form about five cents.

The Classroom Teacher Support System (CTSS) was designed by IBM for the Los Angeles Unified School District (Toggenburger, 1973). This system constructs multiple choice exams according to teacher specified criteria such as course, category, difficulty level, behavioral level, and keywords. The system can also work with "macro" items, i.e., stories or documents followed by two to nine related

questions. Toggenburger reports that CTSS currently uses an American History item bank of 8000 items that were written by 20 teachers over the period of one summer.

Ansfield (1973) has developed a system similar to CTSS called the Automatic Examination Generator (AEG). Ansfield's report on AEG includes data on cost: the total computer expense for producing four versions of a 70-item objective test with answer keys is $1.75.

One last item banking system with a somewhat unique character is one developed by Cohen and Cohen (1973). The main purpose of this CAT system is to assure no overlap in the items presented on successive administrations of a test for any one student. Cohen and Cohen have developed two versions of this system, one for batch processing in COBOL, and one for interactive processing in FORTRAN.

Algorithmic Approaches to Item Construction

Olympia (1975) contends that standard item banking has three disadvantages:

(1) it lacks repeatability (unless the item bank is extremely large), especially when a given item always appears in a test exactly as it is stored,

(2) it requires a large amount of construction time and storage to create a usable bank, and

(3) it discourages the sharing of one program by various disciplines [3].

To overcome these drawbacks, Olympia devised a system for storing examination items in three "pools": a keyphrase pool, a statementphrase pool, and a distractor pool. The system constructs an item by joining one member of the keyphrase pool with one member of the statementphrase pool and then selecting a list of answers (including the correct answer) from the distractor pool. As an example, three pools for constructiong items dealing with electron configurations are shown in Table 1.

Table 1

Pools for Constructing Items on Electron Configurations

(after Olympia, 1975)

Keyphrase Pool	Statementphrase Pool	Distractor Pool
Chlorine	has how many valence electrons?	0
Oxygen	has how many L-shell electrons?	1
Hydrogen	needs how many more electrons in order to have an inert gas structure?	2
Magnesium		3
Helium		4
		5
		6
		7

Denney (1973) describes a system similar to Olympia's. This system stores a multiple choice question as a stem with up to seven distractors. With this data, the

3. This author feels that the example systems discussed in the previous two categories demonstrate capabilities which clearly contradict Olympia's third objection.

computer can construct 245 different questions consisting of a correct choice and four distractors. If the order of the five alternatives is randomized, up to 29,400 different variations of the same question can be generated.

Heines (1974) created an interactive CAT system that randomly generates data to complete item forms or selects one of four previously defined item variations. Regardless of the item generation scheme, the system assured that no student would be presented with the same item on successive administrations of the test. This system is also interesting in that it was introduced by instructions on audio cassette, tied to diagrams presented via a slide projector under student control, and designed to provide and interactive environment for the instructor as well as the student.

Interactive, Branching Tests

Ferguson (1971) defines a branching test as "any instrument designed to measure a set of skills or objectives by routing the examinee to items neither too easy nor too difficult for him to solve". A simple example of this technique was developed by Hansen (1969) and is shown in Figure 3. In this scheme, Item 1 is presented to the student and he/she is then branched to Item 2 if Item 1 is answered correctly and Item 6 if it is answered incorrectly. Item 1 is designed to have a difficulty index of .50, and

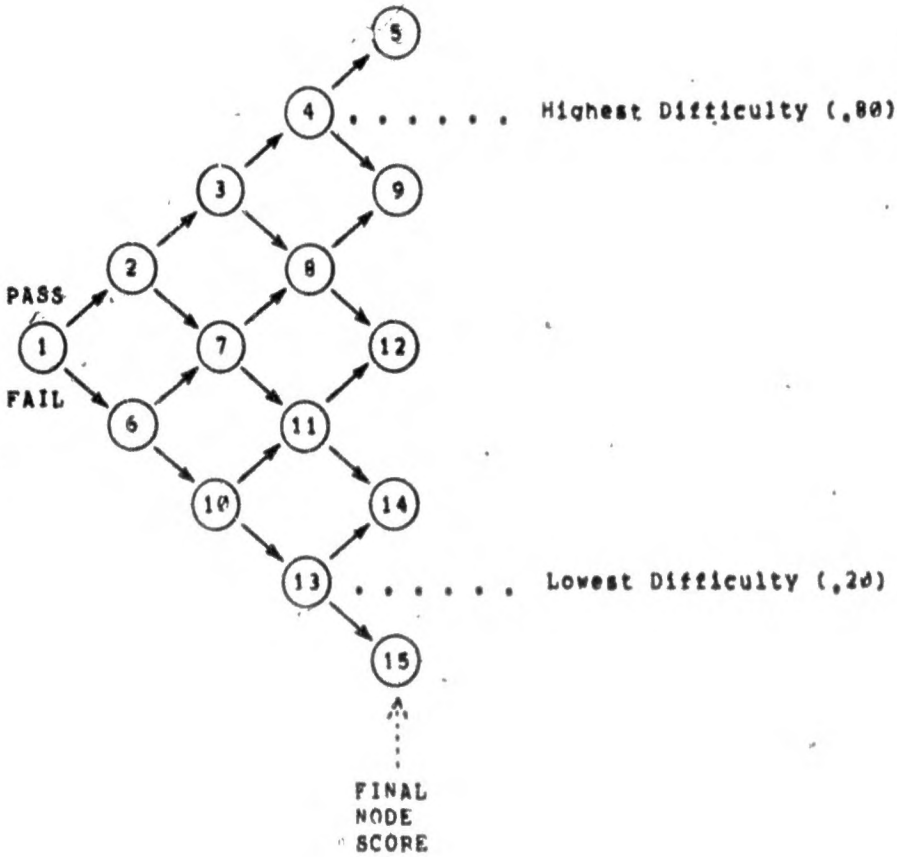

Figure 3

Hansen's Sequential Item Tree Network

(Hansen, 1969, p. 212)

each successive item is designed to have a difficulty differential of +.10 from the preceding item. Thus, the most difficult item in the tree (Item 4) will have a difficulty index of .80, and the easiest item (Item 13) will have an index of .20. Hansen found that this scheme is significantly more reliable than the traditional classroom test and is effective at reducing test anxiety.

The criterion-referencing aspect of Ferguson's work (1971) has already been discussed at length. By comparing Ferguson's work to that of the other CAT researchers discussed so far, it can be seen that Ferguson is one of the only researchers to have created a CAT system as a means for implementing a well-developed theory of evaluation. This system tested objectives in the IPI (Individually Prescribed Instruction) Mathematics curriculum, a program that already made use of comprehensive paper-and-pencil testing and therefore provided a useful measure of the system's success. Ferguson administered tests that utilized his item sampling and evaluation techniques (discussed previously) and then branched students to test items on either more advanced or preliminary objectives based on the results. By this process, Ferguson was able to pinpoint a student's competency level with any desired accuracy and then prescribe instruction to fit the student's needs. Ferguson found that his branching CAT system yielded classification decisions that were "consistent with subsequent paper-and-pencil test outcomes approximately 99% of the time". He conjectured that "by employing an item sampling

technique that permits control over classification errors, the CAT model may increase reliability".

Ferguson discussed three "suggested refinements" to his model. First, he felt that testing must be representative and that this was not always guaranteed by random sampling. He therefore recommended a combination of randomly constructed items with domain-referenced item forms. Second, Ferguson felt that research is needed to achieve a compromise between minimizing Type II errors, which he considers the more serious (these occur when the examinee is a non-master but the test results indicate mastery), and reducing the number of items presented (for expediency) by allowing the error parameters to increase. Third, he noted that all examinees started at the same point, and therefore highly competent examinees did problems that were too easy while incompetent examinees did ones that were too hard. He suggests that examinees might be allowed to choose their own starting points. Ferguson concludes, "By tailoring the test to individuals, fewer objectives need to be tested and the objectives that are tested are less subject to errors of proficiency classification".

CONCLUSIONS

The fields of criterion-referenced and computer-assisted testing are still in their infancies. The literature examined for this report shows that considerable differences of opinion exist on the meaning and uses of CRT, and that very few CAT systems are based in sound theories of evaluation. O'Reilly, Gorth, and Pinsky (1973) comment on the current state of the CAT art as follows:

> [Current CAT efforts] tend to be largely superficial, poorly grounded in relevant evaluation models and test theory and tend to continue a questionable school and classroom practice. . . [They] focus on the mechanics of test production via machine, a tendency which works against the need to maintain precise relationship between the intent of instruction and the measurement process. (page 34)

This author feels that CRT and CAT may help each other mature in several ways:

1) Item generation techniques may be refined to allow more comprehensive evaluation of domains by making more items available.

2) Item sampling algorithms may be used to achieve more representative tests from existing domains.

3) Branching test may be utilized to arrive at the most cost-effective method for evaluating performance.

4) Test models may be simulated to ascertain their

feasibility.

(5) Mathematical models may be developed to help define and standardize the criteria by which performance is judged.

(6) CRT can be used more widely as a valid theory to aid the design of CAT systems.

At present, CRT is lacking in demonstratable, practical applications, while CAT is lacking in sound instructional theory. Researchers who synthesize the best characteristics of these two fields may find that they complement each other smoothly and can contribute heavily to each other's development.

ANNOTATED BIBLIOGRAPHY

Introduction

This bibliography is broken down into two sections, CRT and CAT. The only papers that really fall into both categories are those by Ferguson, and these are categorized under CRT. Within each category, all papers are listed in alphabetical order.

SECTION A: Criterion-Referenced Instruction

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

   Defines the Dichotomous Outcomes Model and its implications for CRT. Differentiates the two types of errors of testing that can occur and explains their relationships. (See also Ferguson on errors of testing.) Philosophy of test design: value to be maximized is cost effectiveness; value to be minimized is "regret" (cost of classifying a master as a non-master and vice versa).

2. Baker, Eva L. Beyond objectives: domain-referenced tests for evaluation and instructional improvement. Educational Technology 14(6):10-16, June, 1974.

   Argues that objectives consist of substance and form, the former defining "the content to which the learner is to respond" and the latter how the learner is to display what he/she learned. Claims that

"overemphasizing either . . . may inhibit the improvement of instructional practice". Feels that "most objectives do not present sufficient cues regarding what a teacher should alter in instruction to facilitate improved learning", but "DRT can supply both the data needed for assessment of instructional programs and information suitable for feedback to teachers to facilitate planning". Key is transfer.

States that domains should be prepared with the following considerations:

(1) Domain description: "a general, but operational, statement of the behavior and content upon which the test focuses".

(2) Content limits: "a set of rules of content eligible for inclusion in the test items or in instruction".

(3) Criteria for constructed responses: "rules by which the adequacy of responses to the item can be judged".

(4) Distractor domain: "specifies the rules for inclusion of wrong-answer alternatives".

(5) Format: "a description of the form in which the items will be presented to students".

(6) Directions: "a facsimile of direction provided the learner in the test situation".

(7) Sample item: "intended as a representative of the class of responses desired".

3. Cox, Richard C. Evaluative aspects of criterion-referenced measures. In Popham, W.J. (ed.), Criterion-Referenced Measurement, pp. 67-75. Educational Technology Publications, Englewood Cliffs, N.J., 1971.

Discusses uses of reliability, validity, and item analysis data in CRT. Cites two methods of item analysis: (1) upper and lower thirds (traditional method), and (2) percent passing on posttest minus percent passing of pretest. Also discusses the sequentially scaled achievement test, where a pupil answers all questions up to a certain point correctly (his/her level of achievement), and misses all items beyond that point. (This is the "ideal" test described by Popham and Husek.)

4. Ferguson, Richard L. Computer-assisted criterion-referenced testing. Learning Research and Development

Center, University of Pittsburgh, Working Paper 49, March, 1970.

Background paper for Ferguson's later work (1971). Describes his item generation, branching, and test scoring algorithms, and provides data on a comparison of the computer-assisted test with the corresponding paper-and-pencil tests. Branching technique involves comparison of the percentage of items answered correctly on a given objective (p) with the passing criterion for that objective (p_0). If the objective is mastered and $p \geq (1 - .5*p_0)$, the student is branched to the most difficult untested objective in the sequence. If the objective is mastered but $p < (1 - .5*p_0)$, the student is branched to a more difficult objective midway between those not already tested. If the objective is not mastered, a similar procedure is used to branch the student to an easier objective. Ferguson reports on a simulation comparing this branching technique to two others: (1) branching up one objective in the sequence if the objective currently being tested was mastered and down two objectives if it was not, and (2) branching up two and down one (these techniques are similar to Hansen's). Results showed that Ferguson's branching technique required fewer test items than the other two in almost all cases.

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

A more comprehensive discussion of the 1970 work, describing all aspects of Ferguson's test model and its use in the IPI Mathematics Curriculum. Presents prior research on branching tests and full capabilites of the current CAT system. (Specific aspects of this work are described in detail in the body of this report.)

6. Garvin, Alfred D. The applicability of criterion-referenced measurement by content area and level. In Popham, W. James (ed.), Criterion-Referenced Measurement, pp. 55-63. Educational Technology Publications,

Englewood Cliffs, N.J., 1971.

A slightly humorous view of CRT, proposing that some subjects, e.g., English, need not have criterion levels that everyone must master. Proposes the following "general principles" on the applicability of criterion-referenced measurement (CRM) to various content areas:

(1) "Unless at least one of the instructional objectives of a unit envisions a task that must subsequently be performed at a specified level of competence in at lease some situation, CRM is irrelevant because there is no criterion."

(2) "If public safety, economic responsibility, or other ethical considerations demand that certain tasks be performed only by those 'qualified' for them by formal instruction, them CRM of the outcomes of such instruction is clearly indicated."

(3) "In any instructional sequence where the content is inherently cumulative and the rigor is progressively greater, CRM should be used to control entry to successive units."

(4) "There are certain content areas to which criteria do apply but not everyone need meet them."

7. Glaser, Robert. Instructional technology and the measurement of learning outcomes: some questions. American Psychologist 18:519-521, 1963.

Defines norm-referenced and criterion-referenced measures and the uses of achievement measures in general. Contends that the diference between the two types of measures is determined by the selection of items and discusses the implications of this contention on the interpretation of observed discrimination indices.

8. Glaser, Robert. A criterion-referenced test. In Popham, W. James (ed.), Criterion-Referenced Measurement, pp. 41-51. Educational Technology Publications, Englewood Cliffs, N.J. 1971.

An extremely detailed discussion of the characteristics of CRT and its differences from NRT. Contends that "the distinction is found by examining (a) the purpose for which the test was constructed, (b) the manner in which it was constructed, (c) the

specificity of the information yielded about the comain of instructionally relevant tasks, (d) the generalizability of test performance information to the domain, and (e) the use to be made of the obtained test information". Also includes a fascinating reprint from Edward L. Thorndike's classic work Educational Psychology (1913) which shows that the problem of establishing criteria against which to measure student achievement is indeed a basic one in instructional theory.

9. Hambleton, Ronald K., and William P. Gorth. Criterion-referenced testing: issues and applications, University of Massachusetts School of Education, Amherst, September, 1971.

Defines reliability, validity, and item analysis and discusses the use of each in criterion-referenced measures. Describes the uses of test results for (1) individual assessment, (2) teaching material assessment, and (3) evaluative material assessment (implicitly). Presents descriptions of two criterion-referenced measurement systems, one of which is a CAT system. Contains comprehensive bibliography.

10. Hively, Wells. Introduction to domain-referenced testing. Educational Technology 14(6):5-10, June 1974a.

Defines DRT and contrasts it with CRT. Defines reliability and validity in DRT as the "accuracy with which one can estimate the probabilities of correct performance within a concrete domain" and the "sucess of generalization from performance on a concrete domain to performance in the larger universe of knowledge from which the domain was generated", respectively. Differentiates NRT and DRT and points out that both are useful in different applications. States that item analysis as used in NRT does not consider the validity of items, and suggests that items on norm-referenced tests "may be selected for their ease of administration rather than for their formal correspondance to the original universe".

11. Hively, Wells. Some comments on this issue.

Educational Technology 14(6):60-64, June 1974b.

Comments on all articles in this special issue of Educational Technology magazine on DRT, clarifying some points and contesting others. Provides a clear understanding of DRT theory and an interesting discussion of many facets of this work.

12. Lindvall, C.M., and Anthony J. Nitko, Criterion-referenced testing and the individualization of instruction, Learning Research and Development Center, University of pittsburgh, paper presented at the annual meeting of the National Council on Measurement in Education, February, 1969.

Excellent discussion of the differences between CRT and NRT with the characteristics of CRT clearly presented. Concise and easy to understand.

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

Presents a mathematical model similar to Fergusons in which an "Uncertainty Band" is computed and used to judge the reliability of a DRT. This UB is a function of the number of items in the domain, the number of items in the test, and the passing standard in percent. Generalizes the computation for situations in which subtests are used. Discusses the purposes of comparing scores on two or more DRT's and enumerates sampling considerations for constructiong DRT's.

14. Popham, W. James, Indices of adequacy for crierion-referenced test items. In Popham, W. James (ed.), Criterion-Referenced Measurement, Educational Technology Publications, Englewood Cliffs, N.J., pp. 79-98, 1971.

A complex discussion of various techniques for assessing statistical characteristics for CRT, using the SWRL and PROBE projects (Southwest Regional

Laboratory and UCLA, respectively) as examples. Highly technical, showing the results of using different statistical techniques to analyze the same sets of data.

15. Popham, W. James, and T.R. Husek. Implications of criterion-referenced measurement. Journal of Educational Measurement 6(1):1-9, 1969.

Describes the differences between CRT and NRT and contends that the traditional methods for computing the reliability and validity of a test are not appropriate for CRT because these measures are based on variability of test scores. Presents a similar argument against traditional item analysis techniques, but admits that "as data-processing becomes increasingly automated and less expensive, such analyses would seem warranted inn situations where the effort is not immense". Defines an ideal CRT as one which has a one-to-one correlation between score and response pattern, i.e., each score may only be achieved in one way. (A means for realizing this type of test described by Cox.) Recognizes the more typical type of CRT as a DRT.

SECTION B: Computer-Assisted Testing

16. Ansfield, Paul J. A user oriented computing procedure for compiling and generating examinations. Educational Technology 13(3):12-13, March 1973.

Description of a system in use at the University of Wisconsin. 360/40-based, using files on magtape. Items may be multiple choice, true/false, or "macro". Instructor input: exam title and date, specific or random item selection, specification of instruction sets to be used, and number of arrangements for multiple forms. Banks currently available in psychology and sociology, with business, biology, and physics planned.

17. Baker, Frank B. An interactive approach to test construction. Educational Technology 13(3):13-15,

March 1973.

This system allows interactive exploration of items at a computer terminal by searching its data bank for keywords that match the user's input. For each item, the system stores: the item itself, its ID, a set of keywords, a code indicating its most recent usage, item analysis results, the total number of times that the item has been used, and a link to the previous version of the item. Items are screened by the parameters supplied by the interactive user and a table is generated containing the number of items requested per area, the number found per area, and the predicted test mean, reliability, and variance. Maintenance and analysis functions are performed in batch mode from card input.

18. Brown, Williard A. Improvement of testing and course evaluation. Journal of Research in Science Teaching 5:240-243, 1967-1968.

Description of a system designed to help detect (1) trivial distractors, (2) erroneous answers supplied by the instructor, (3) inconsistent answers between two forms, (4) answers with no logical distinction, (5) bad question stems, (6) crucial misspellings, and (7) trivial questions. System scores and sorts mark-sense answer cards, computes norm statistics, and performs an item analysis on up to four multiple forms of a single test.

19. Brown, Williard A. A computer examination compositor for the IBM 360/40. Western Washington State College, 1972.

Description of a system that performs the following services: (1) stores questions on disk in compact format, (2) outputs card images for the compressed files, (3) allows updates to question files in batch mode, (4) produces a catalog of questions from the disk file, (5) produces page files of composed exams for output in upper and lower case on the IBM 2741 communications terminal, (6) produces similar page files of exam answers, (7) formats output or allows this feature to be overridden, and (8) provides for multiple testing techniques.

20. Buckley-Sharp, M.D. A multiple choice question banking system. Educational Technology 13(3):16-18, March 1973.

An IBM 360/65-based batch system used in medical colleges in the United Kingdom for filing and printing multiple choice questions. Questions are banked after they are used and validated, allowing other instructors to access them. Advantages cited are saving of instructor time and encouragement of open release of evaluative materials to students. Contends that the latter yields better, more directed learning.

21. Cohen, Perrin S., and Leila r. Cohen. Computer generated tests for a student paced course. Educational Technology 13(3):18-19, March 1973.

This system is designed to prevent overlap in the test items chosen for successive administrations of a randomly generated test for any one student. The exams may be generated in batch mode by COBOL programs or interactive mode by FORTRAN programs. System allows multiple choice, true/false, "identify and define", and "graph" questions. The authors see the advantages of their system as rapid exam generation and elimination of biases due to ordering of questions since each question is randomly generated.

22. Denney, Cecil. There is more to a test pool than data collection. Educational Technology 13(3):19-20, March 1973.

An APL-based system that combines data banking and algorithmic approaches to test construction. A completely interactive systems that includes means for retrieving and editing objectives, activities, and resources. A tutorial CAI program is available that guides teachers through learning the system's use. Output includes test copy, student response sheet; answer key, and diagnostic information for both the student and teacher.

Concludes with the following comments on implementation of CAT:

(1) "...any innovation in education must allow a teacher to begin at his own level of professional skill and grow into its application as his skills improve."

(2) Quality is of the utmost importance.

(3) "...no matter how great the system may seem to the originators and no matter how enthusiastically they are able to describe it, its ultimate success will be determined on the basis of whether the apparent value received is greater than the perceived effort of using it. Technology in education must be a serving tool, not an end in itself."

23. Dudley, Thomas J. How the computer assists in pacing and testing students' progress. Educational Technology 13(3):21-22, March 1973.

A test banking and statistics storing system that offers only upper case output and no graphics, but allows questions to be categorized by objectives up to three levels. Test is printed by the computer and the students' responses are entered through keypunching. Self-pacing aspect is achieved through individualization in a linear, modular program that requires that the student must pass one test before proceeding to the next unit.

24. Educational Testing Service. Computer-assisted assembly of tests at ETS. A paper presented at a conference on computer-assisted test construction, San Diego, California, October, 1974.

A description of a system in use at ETS to store item characteristics. Primary output is a list of item numbers, and the items are retrieved manually, ordered, typed, and printed. A prototype system after that of Willard Brown is being experimented with for storage and retrieval of whole items. Cuurent problems with this system is that it has limited graphics and item formats, and the high-speed print-outs are not of acceptable quality for reproduction. This paper also includes criteria which ETS sees as necessary for a CAT system that they can use to implement large scale computer-assisted test construction, and a statement that interactive testing will be the way of the future.

25. Hansen, Duncan N. An investigation of computer-based science teaching. In Richard C. Atkinson and H.A.

Wilson (ed.), Computer-Assisted Instruction: A Book of Readings, Academic Press, New York, N.Y., pp. 209-226, 1969.

Describes a simple branching test technique called the item tree network and discusses four ways of scoring tests based on this structure. Found no significant difference in administration time between a 17-item CAT and a 20-item conventional test. All four scoring schemes yielded similar results, and each yielded a reliability coefficient that was significantly higher than a comparable classroom test. Hypothesized that this increased reliability might be due to increased dispersion at the upper and lower ends of the scale. Found that student attitude towards the CAT system was positive, and therefore suggested that the system was feasible for reducing test anxiety.

26. Hazlett, C.B. MEDSIRCH: multiple choice test items. Educational Technology, 13(3):24-26, March 1973.

Informative, detailed, step-by-step description of how this question retrieval system works. Uses 57 descriptors including subject areas and statistics. A FORTRAN-based batch system with punched card entry but many utility programs available. Paper includes a flowchart of the program's operation. User can supply question ID's or a "profile" (list of descriptors desired or not desired). Used in medical colleges in Alberta and other Canadian sites.

27. Heines, Jesse M. An interactive, computer-managed model for the evaluation of audio-tutorial instruction. Unpublished Master's Thesis, College of Education, University of Maine, Orono, Maine, May, 1974.

Analysis of the use of a BASIC language CAT system to evaluate students in an audio-tutorial course in physical science for non-science majors. Provides completely interactive environments for the instructor as well as the students. Analysis includes data on the system's cost, the time required by students to master its use, and the effectiveness of its test items in assessing student learning. Appendices include transcripts of actual sessions at the terminal and listings of the programs used.

28. Hsu, Tse-Chi, and Marthena Carlson. Test construction aspects of the computer assisted testing model. Educational Technology 13(3):26-27, March 1973.

An interactive system for the DECsystem-10 written in FORTRAN. Item forms generated for the IPI math curriculum similar to those generated by Ferguson. Statistics are generated for the item forms, not for the individual items.

29. Jensen, Donald D. Toward efficient, effective, and humane instruction in large classes: student scheduled involvement in films, discussions and computer generated repeatable exams. Educational Technology 13(3):28-29, March 1973.

Report on the use of a CAT system to generate a large number of forms for the same test in a large enrollment (1500 students) course. The course is self-paced within a week's time, i.e., one unit must be completed each week. Tests are given every one or two weeks. Since CAT and the opportunity to retake an exam have been implemented, student enrollment has doubled in three years, the modal grade has increased to an "A", and the failure rate has dropped to less than 10%. Recommends as 1# to 1 ratio in the item bank size to the number of items to be included on any one test.

30. Libaw, Frieda B. Constructing tests with the MENTREX tutorial testing system. Educational Technology 13(3):30-31, March 1973.

Description of a commercially available test construction service from MENTREX Enterprises in Los Angeles. The system can produce scrambled tests, sort and select items on keys, and sort and select items on a two-dimensional matrix of keys. The system currently handles only multiple choice items. Users request tests through the mail from a "catalog" of test questions. MENTREX returns an assembled test on ditto masters or ready for offset printing, an answer key, and machine readable answer sheets. The answer sheets are returned to METREX for item analysis. Items can be augmented by text or graphics keyed to the question.

31. Lippey, Gerald. The computer can support test construction in a variety of ways. Educational Technology 13(3):10-12, March 1973.

Introductory article to this special issue of Educational Technology magazine on CAT. Describes the variety of ways in which CAT has been implemented and summarizes the benefits that CAT offers. Claims that development of CAT systems is stimulated by classroom teachers rather than professional innovators and that the activity is seldom supported by special funding, and contends, therefore, that CAT satisfies educators' needs, is financially feasible (has a high value-to-cost ratio), and that CAT applications will grow.

32. Wenne, John W., and Paul Lutgraef. Computer-assisted test assembly at Iowa State University. Paper presented at a conference on computer-assisted test construction, San Diego, California, October 1974.

PL/I-based system that currently stores 13,000 items on a dedicated disk pack using 6.4 million characters of storage. Each item requires about 500 characters of storage including about 174 characters for classifiers and usage statistics. System design considerations were:

(1) that each instructor must be able to use his/her own item indexing scheme,

(2) that cost must be minimized, dictating that the system must be operable by clerical personnel, and

(3) that the system must allow for the inclusion of item statistics.

Most functions are done overnight at a cost of one cent per item generated with 15 minutes of clerical time required to set up the program run. Currently a batch system, and thus the cost of clerical time is greater than the computer cost and delays are long. Conjectures that an interactive system would result in higher computer costs and shorter delays with the same amount of clerical time.

33. Olympia, P.L., Jr. Computer generation of truly repeatable examinations. Educational Technology 15(6):

53-55, June 1975.

Discusses the disadvantages of standard item banking systems and presents an algorithmic approach to item generation. Questions stored as a keyphrase pool, statementphrase pool, and distractor pool. Item constructed by joining one member of the keyphrase and statementphrase pools and then constructing a list of alternatives from the distractor pool. Can be used for multiple choice, true/false, completion, and matching items.

34. O'Reilly, Robert P., William P. Gorth, and Paul Pinsky. Computer-assisted test construction: an effort based on an evaluation methodology. Educational Technology 13(3):32-34, March 1973.

Argues that the current state of the CAT art is preoccupied with the mechanics of test construction rather than relevant evaluation models. Describes the Comprehensive Achievement Monitoring (CAM) System. This system is designed to help instructors perform the following functions:

(1) Classification: assign or eliminate students to or from a treatment due to scarce resources.

(2) Summative evaluation on programs: select treatment A over treatment B on effect or efficiency.

(3) Formative evaluation on programs: redesign component A of treatment B to meet specifications.

(4) Instructional management: place student in component A of program B; repeat student in A; etc.

(5) Curriculum validation: remove objective C from programs A and B at level 1; place at level 2.

System currently uses paper-based objectives and test item banks. Computer schedules tests, indexes current objectives and item banks, finds item numbers, and constructs forms consisting of item numbers. (Similar to ETS usage.) No immediate plans for expanded computer use.

35. Prosser, Franklin. Repeatable tests. Educational Technology 13(3):34-35, March 1973.

A FORTRAN-based system in use at Indiana University. Item pools available in English, geography, home economics, chemistry, economics, statistics, psychology, speech therapy, accounting, and education. System provides many forms of the same test

printed individually on a line printer. 20 seconds of CPU time required to generate 1000 3-page tests, with 3 hours required to print them. Cost per test is approximately 5 cents.

36. Remondini, David J. Test Item System: a method of computer assisted test assembly. Educational Technology 13(3):35-37, March 1973.

A typical batch CAT system. Steps followed in preparing, administering, and analyzing a test are:
(1) question selection (currently manual)
(2) test form and answer cards prepared by computer and printed on line printer
(3) editing, e.g., addition of graphic material
(4) duplication by Xerox and Thermofax processing
(5) test administration
(6) test scoring by mark sense computer cards
(7) record updating by computer onto cards and item analysis listing printed
Questions classified by CUEBS categories (Commision of Undergraduate Education in Biological Sciences)

37. Remondini, David J., and John E. Miller. A computerized system for preparation of tests in academic disciplines. Proceedings of a conference on computers in the undergraduate curricula, pp. 7.24-7.30. The University of Iowa, Iowa City, Iowa. June 1970.

Description of a FORTRAN-based system built for the IBM 1130. This system is essentially the same as the one described in Remondini's 1973 article. Control card for each question includes biology subject area, organization level, behavioral objective level (re Bloom), and difficulty index.

38. Salisnjack, Julian. Computer aided test preparation: six years of experience. Educational Technology 13(3):37-38, March 1973.

Another typical system almost identical to Remondini's. Questions stored on disk and manually selected for inclusion on a test. Multiple choice, true/false, short answer, fill in, matching, and "enhanced" (diagram-oriented) items allowed. Discusses costs, solving of cheating problem, and reduction of "edge" of fraternity test files.

39. Schonberger, Richard J. Modular instruction with computer assembled repeatable exams: second generation. Educational Technology 15(2):36-38, February 1975.

Recommends the following principles for achieving success with CAT:

(1) Make a good test item pool.

(a) Use a large number of items, i.e., a 10 to 1 ratio of items to presentations. (This is also recommended by Prosser.) Avoids students relying on co-ops in libraries and dorms.

(b) Force yourself to test for concepts by giving open book exams.

(c) Construct the item bank yourself, i.e., do not deligate the responsibility to graduate students.

(2) Provide fast reinforcement.

(3) Provide flexibility by letting students retake exams and use a contract approach to grading.

40. Sivertson, Sigurd E., Richard H. Hansen, and Adeline O. Schoenenberger. Computerized test bank for clinical medicine. Educational Technology 13(3):38-39, March 1975.

System designed "to identify the continuing education needs of physicians". Tests constructed to match physicians' specialties and backgrounds. Yields relative scores on different parts of the test to show areas in which physician's knowledge is deficient.

41. Stodola, Quentin. Use of computer assembled tests in the California State University and College System.

Educational Technology 13(3):40-41, March 1973.

Comments: "Computer assisted test assembly has sometimes been called 'a poor man's CAI'. Computer assisted test assembly provides some of the advantages of CAI, such as drill, student self-pacing and reference to study aids, but without the high cost of terminals for individual students and without the need for writing highly sophisticated and complex learning programs, which, incidentally, thus far have generally not been written. . . The computer operation of assembling and scoring tests works satisfactorily. The problem is now to create a sufficient number of useful question banks and to orient instructors to their use."

Mentions exact areas in which questions have been developed for use in the CSU system. Project received $38,000 to fund the categorization, editing, and keypunching of 9000 items originally collect by ETS. Additional $26,000 was used for development of a pre-calculus bank.

42. Toggenburger, Frank. Classroom teacher support system. Educational Technology 13(3):42-43, March 1973.

Desciption of a system in use by the Los Angeles Unified School District. Currently in use with a U.S. History item bank. 20 teachers developed the 8000 items currently available over the period of one summer. "Exercises" are requested by filling out an optical scan form. The generated exercises are stored for later modification and student response checking.

43. Vickers, F.D. Creative test generators. Educational Technology 13(3):43-44, March 1973.

Only article in this special issue of Educational Technology magazine that describes test generation without an item bank. Used to generate tests for a FORTRAN course. Two examples, first very simple and provides excellent demonstration of creative test generation. Second example written in SNOBOL and includes a text formatting capability. Very informative examples.