Computer generated repeatable tests

by FRANKLIN PROSSER

Indiana University
Bloomington, Indiana

and

DONALD D. JENSEN

University of Nebraska
Lincoln, Nebraska

INTRODUCTION

While we wait for Computer-Assisted Instruction to revolutionize teaching practices, a number of more tractable computer techniques are proving useful in dealing with large university classes. One of these, the use of Computer Generated Repeatable Tests (CGRT), is the topic of this paper. The concept of repeatable testing has intrigued teachers for years. The idea is to have the flexibility in class examination procedures to test a student repeatedly over a section of material, and at the student's own pace. The implementation of such a plan involves producing, administering, and grading large numbers of different tests over the same material. In classes of moderate or large size, the mechanics of such procedures has defeated even the most dedicated instructors. The present computer augmented procedure—a result of collaboration between a psychology professor (DDJ) and a computer professional—overcomes the difficulties by fully automating the preparation and grading of individualized tests. Before discussing the technical details of the CGRT process, we will briefly give the rationale behind repeatable testing.

THE PROBLEM

Large classes are becoming an increasingly important part of American higher education. This trend is disturbing because available evidence indicates that the conventional large lecture class is an unsatisfactory educational system, one that is disliked by students and considered educationally ineffective by professors. The difficulties with large classes appear to stem from the inflexibility of many of the educational activities in the classes, and from the fact that the student's performance is monitored infrequently and often inadequately. The student has only a passive role in a large lecture section; the lecturer lectures and the student (at best) listens. Opportunities for the student to creatively display knowledge of the subject matter are limited, since the time required to grade essay exams or individual projects dictates that these be given very infrequently, if at all. Often the only measures of the student's knowledge of course material are the examinations he takes infrequently during the course.

Of the several inadequacies of large class instruction, the examination procedures are perhaps most seriously deficient. The typical examination administered to a large class consists of objective questions of the true-false and multiple choice types. It is administered at one fixed time only. Several days elapse before information on the results is available to the students, and often the only information provided to the student is the total score, which is of little use to him in guiding his study. Such exams commonly are given infrequently and therefore cover an extensive amount of material.

There are several objectionable consequences of these procedures. First, the conventional large class examination is not adequate to motivate routine study. The infrequency of testing gives rise to the well known "loaf and cram" pattern of study. Second, conventional infrequent exams provoke excessive anxiety in the student. The exam covers a large amount of course material and accounts for a substantial part of a student's grade. A student who is not well up on his reading or is just not well will be very anxious about his performance on the exam; further, a poor performance may cause him to despair and abandon meaningful

study in the course. Third, students exposed to multiple choice test questions become testwise. Unless test items are extremely well formulated, students develop the ability to *recognize* answers from among the choices, and they are thus not required to *recall* course material in order to construct a response to the question. Unfortunately, this recognition ability is not likely to result in a significant increase in active vocabulary or intellectual competence. Fourth, the necessity of writing new test questions semester after semester means that only in rare instances does the instructor have *a priori* objective evidence that each exam question validly discriminates between good and bad student performance.

A SOLUTION

The goal is to find a practical and economic way to overcome the principal deficiencies of conventional examinations for large classes. This necessitates using only those facilities readily available at most academic institutions. There are three fundamental ways in which the Computer Generated Repeatable Testing method departs from a conventional approach. First, students may be examined *more frequently*. This encourages students to keep up with their course work and allows them to evaluate their performance frequently. Second, the tests provide *immediate feedback* to the student. When a student turns in his answer sheet, he keeps his test questions and receives a matching list of correct answers to the items on the test along with study aids such as textbook page references.

Third, and most important, examinations have been made *repeatable*. The digital computer is used to generate individualized repeatable tests. Large numbers of unique but equivalent tests are generated by a computer program which takes stratified random samples from an item pool and prints out questions and answers in a format appropriate to test taking and machine or hand grading. Since the instructor may prepare large quantities of individualized tests, he may allow the student to take tests over an examination unit as often and whenever the student desires. Students are free to take a test and find out thereby what they have not mastered, review that material, and try again. In this way they can work up to a high level of proficiency in the content of the course.

Conventional examinations provide only a single opportunity to show knowledge, and diagnostic information cannot be used to improve one's grade. Our method of computer generated repeatable examinations encourages the student to use diagnostic information and to restudy material he has not initially mastered, and it thereby decreases the student's aversion of the examination process while maintaining appropriate demands for the mastery of the course content. The technique of repeatable testing, of allowing more than one chance to demonstrate competence with material, is an exceedingly attractive instructional procedure. In a single stroke it releases the professor from his conflict between being excessively demanding and transmitting important information; he can expect and demand that the student master the material because the student can work up to mastery through a series of study sessions and examinations. Similarly it releases the student from his conflict between fear and fatigue; he has the opportunity to set a humane pace of study for himself, because if that pace is insufficient to obtain a satisfying grade on the first test over a unit of material, he can increase his pace in order to succeed on subsequent tests covering that material. Work rather than worry is elicited from the student.

One should not consider the CGRT system to be a kind of Computer-Assisted Instruction,[1] since students are not in interaction with a computer. It resembles more what Cooley and Glaser[2] have termed "computer-managed instruction" since the computer is used to facilitate instructional examination processes. We strongly point out that CGRT does not suffer from the present serious technological and economic disadvantages of Computer-Assisted Instruction, but nevertheless shares many of its educational advantages.

THE CGRT PROCESS

The Computer Generated Repeatable Testing process typically consists of four steps: (1) developing pools of test items, (2) producing tests, (3) administering the tests, and (4) scoring the tests. The second and fourth steps are managed by computer, while the execution of the first and third steps is strongly influenced by the computerized nature of the process.

Developing test item pools

For each exam, the course instructor develops a pool of items (test questions) which forms the data base from which tests are prepared. This is a rather formidable step. Our experience indicates that one should have about six to ten items in the pool for each question on an exam to assure adequate variation on the individual tests. An instructor planning to give eight exams of twenty questions each should construct about fifteen hundred individual items for his course. This work, which is every bit as tedious and time consuming as it sounds, should be done prior to the first semester in

which repeatable testing is to be used in the course. Fortunately, the item pools, once developed, are rather permanent, especially for the basic college undergraduate courses that are the most likely candidates for this computerized testing scheme. Only relatively minor alterations to the item pools are needed to accommodate other instructors, changes of texts, etc., that may occur in subsequent semesters. Further, textbook publishers often have compendiums of test questions for their popular texts. Reusing test questions semester after semester, or even making the entire item pool available to students, is not a disadvantage under our procedure, and in fact is likely to be distinctly advantageous!

Since items will ultimately appear on computer generated tests, the form of the items must conform to the requirements of present computer printing technology. Normally, items may consist of upper-case letters, numbers, and the usual special characters available on modern high-speed line printers. Diagrams, pictures, and other graphic aids usually cannot be printed directly, although the instructor may easily include these by providing the student with a supplementary sheet of diagrams to accompany the tests.

If the tests are to be graded manually, technology imposes no limitation on the structure of the answer to an item. The test questions may elicit objective or subjective responses from the student. On the other hand, if the instructor wishes to use mechanical grading techniques, he must provide for a *single-character* response for each item, because of restrictions imposed by the optical mark sense form readers usually available in universities. While this requirement may appear to be a severe limitation, it in fact allows considerable freedom in the form of objective test items. True-false and multiple choice items call for single character responses. Key-word, fill-in, and other forms resulting in a definite numeric or symbolic answer may easily be reduced to a single character response using the following convention: In such a question the form of the answer is indicated by a series of dots which includes one asterisk. The student will construct his symbolic or numeric answer to the question, and will record as his response on his mark sense form the single character selected by the position of the asterisk in the string of dots. For example, $..*.$ means code the third letter or digit of the answer, $*...$ means the first letter or digit, and so forth. The notation $.*..$ appearing in a fill-in-the-blank question calling for the answer "INTEGRAL" would require the student to mark the "N" space on his answer sheet. Students describe such alphabetically or numerically coded items as being hard but fair. The student cannot answer such an item unless he has mastered the basic concepts and vocabulary. Recall is emphasized; simple recognition is subordinated.

In addition to the question part of an item, which the student sees when he takes a test, each item also has an answer part to allow machine grading and to provide information to the student after testing. The answer part of an item may contain, in addition to an answer character, any relevant information, such as the full symbolic or numeric answer, textbook page references, and other diagnostic aids for the student.

After the instructor has developed a section of his test item pool, he will have it punched onto punch cards or entered into an appropriate editable data file system. To facilitate the selection of items for an individual test and to maintain order among the large item pools the instructor classifies his items into sets, the items within a set are given distinct unit numbers, and cards or lines for the question part and the answer part of each item are numbered serially. Usually a set will consist of those items that test similar material. The use of set numbers is explained in the next section.

Producing tests

The individualized tests are generated on a digital computer using a computer program GENERATOR. This program, which is described in more detail in a later section, reads the item pool for a particular exam, checks the input data for proper sequencing and correct format, reads information describing the tests to be generated (number of tests, number of questions per test, etc.), generates and prints the individual tests, and punches a small answer summary deck for use in mechanized grading. The appearance of the tests is similar to the photo-reduced sample in Figure 1. Each test is individually numbered and has questions on the left part of the line printer page and answers on the right. The item identification numbers for each question appear in the answer part for reference. The instructor will of course separate the answer part from the question part prior to giving a test to the student.

The computer program selects items for a test by randomly choosing an item from each set. The order of choosing sets is also randomized. No item is used more than once per test. The digital computer is vital to test production, since the random item selection, formatting, and printing of large numbers of individualized tests is beyond the capacity of nonautomated operations. In the sample test in Figure 1 each item is assigned equal weight. The instructor may also assign weights (point values) to sets of items, thus allowing him to emphasize particular topics or award points based on the difficulty of items.

The computer time required to generate the tests is very small; the time required to print tests is, however,

```
EXAM NUMBER  03.   FORM NUMBER 0704
CGRT SAMPLE TEST... ECONOMICS E201 DATA

QUESTION   1
WHAT ANTITRUST LAW FIRST EXEMPTED LABOR UNIONS FROM PROSECUTION AS
CONSPIRACIES IN RESTRAINT OF TRADE ..*.. .....
--------------------------------------------------------------------------------
QUESTION   2
TRUE-FALSE.  COMMERCIAL BANKS PREFER TERM LOANS OF SEVERAL YEARS
DURATION RATHER THAN SEASONAL LOANS WHICH ARE PAID OFF IN A SHORT
PERIOD OF TIME .
--------------------------------------------------------------------------------
QUESTION   3
CORPORATE BONDS WHOSE INTEREST IS PAYABLE ONLY IF EARNINGS ARE
LARGE ENOUGH ARE CALLED  ..*..  BONDS.
--------------------------------------------------------------------------------
QUESTION   4
THE TRUE ADDITIONAL BURDEN OF MONOPOLY IS THE CONTRIVED DIVERGENCE
BETWEEN .*... AND MARGINAL COST.
--------------------------------------------------------------------------------
QUESTION   5
THE FEDERAL CORPORATE INCOME TAX IS BASED ON A FIRM'S  (A) GROSS
RECEIPTS  (B) DIVIDENDS  (C) CASH RECEIPTS  (D) RETAINED EARNINGS
(E) PROFITS
--------------------------------------------------------------------------------
QUESTION   6
THE SHUTDOWN POINT OF LONG-RUN NO-PROFIT COMPETITIVE EQUILIBRIUM OCCURS
  (A) AT MINIMUM LONG-RUN AVERAGE COST  (B) AT MINIMUM LONG-RUN AVERAGE
VARIABLE COST  (C) WHERE PRICE IS EQUAL TO MARGINAL COST  (D) AT
MINIMUM LONG-RUN MARGINAL COST  (E) NONE OF THE ABOVE
--------------------------------------------------------------------------------
QUESTION   7
WHICH OF THE COST CURVES SLOPES STEADILY DOWNWARD ON A GRAPH RELATING
COST AND OUTPUT.  (A) TOTAL COST  (B) VARIABLE COST  (C) FIXED COST (D)
 AVERAGE VARIABLE COST  (E) AVERAGE FIXED COST
--------------------------------------------------------------------------------
QUESTION   8
AN OLIGOPOLISTIC MARKET SITUATION CONSISTING OF TWO SELLERS IS KNOWN AS
A (AN) .*...
--------------------------------------------------------------------------------
QUESTION   9
A SCHEDULE RELATING A FIRM'S TOTAL COST TO OUTPUT IS THE RESULT OF  (A)
 PRICES OF FACTOR INPUTS  (B) ENGINEERING TECHNOLOGY  (C) ECONOMIC
DECISIONS MINIMIZING EXPENSE FOR EACH LEVEL OF OUTPUT  (D) ALL OF THE
ABOVE  (E) NONE OF THE ABOVE
--------------------------------------------------------------------------------
QUESTION  10
THE LOWEST AGGREGATE DOLLAR EXPENSE NEEDED TO PRODUCE EACH LEVEL OF
OUTPUT IS CALLED ..*.. .....
--------------------------------------------------------------------------------
END OF TEST              EXAM NUMBER  03.   FORM NUMBER 0704
```

```
EXAM NUMBER  03.   FORM NUMBER 0704
CGRT SAMPLE TEST... ECONOMICS E201 DATA

QUESTION   1    SET0430. ITEM02

A. ..*.. ..... CLAYTON ACT (1914)  P.
500
--------------------------------------------------------------------------------
QUESTION   2    SET0401. ITEM06

F  P. 78
--------------------------------------------------------------------------------
QUESTION   3    SET0422. ITEM04

C  ..*..  INCOME  P. 85
--------------------------------------------------------------------------------
QUESTION   4    SET0429. ITEM01

R .*... PRICE  P. 492
--------------------------------------------------------------------------------
QUESTION   5    SET0407. ITEM09

E  P. 84
--------------------------------------------------------------------------------
QUESTION   6    SET0411. ITEM04

B  P. 458
--------------------------------------------------------------------------------
QUESTION   7    SET0410. ITEM06

E  P. 455
--------------------------------------------------------------------------------
QUESTION   8    SET0428. ITEM03

U .*... DUOPOLY  P. 486
--------------------------------------------------------------------------------
QUESTION   9    SET0409. ITEM04

D  P. 453
--------------------------------------------------------------------------------
QUESTION  10    SET0424. ITEM02

T ..*.. ..... TOTAL COST  P. 455
--------------------------------------------------------------------------------
END OF EXAM  03.   FORM NUMBER 0704
```

Figure 1—Sample individualized computer generated test with answers attached

substantial. Typical times on the Indiana University CDC 3600 computer system are about four minutes of computer time (of which about 20 seconds are for item selection) to generate 1,000 three-page tests, and about three hours of printer time to print them. As we show later, the total cost per test is about 5¢. This compares well with the 5¢ cost per test for conventional exams using standard office facilities!

To avoid grief caused by possible computer delays and human errors, an instructor should submit his test production runs to the computing facility well in advance of his need. With many hundreds of students eager and ready to be examined on the course material, the instructor should risk no delays in preparing the tests. He need take no special precautions against pilfering of tests or even of listings of entire test item pools. The tests are individualized, and the item pools are large enough that the memorizing of the whole question pool is not a fruitful approach. (Indeed, as we implied earlier, a potentially useful study aid is to make the entire pool of test questions and answers available to the students prior to examination times).

Administering the tests

The instructor decides for himself how and when to test his students. He may give tests in class or at other scheduled times; or, more flexibly, he may allow his students to choose their own times for testing. A combination of in-class testing followed by opportunities for student-scheduled retesting appears to be useful. Such options depend on the instructor's preference and the availability of testing room space and personnel.

A student taking a test usually obtains an individualized test (with answer part removed) and a mark sense form and special pencil. He takes a seat in the testing area and immediately enters on his mark sense form his student identification number (social security number or other agreed-upon identifier), the exam number, and his individual test number. The student then marks his answers on his test, and for each question enters the appropriate single-character response on his mark sense form. After completing a test, the student exchanges his mark sense form for the answer part of his individual test. The mark sense form is kept by the proctor for later grading. The student, having the correct answers in hand, can immediately determine his errors, and is stimulated to improve his knowledge of weak areas. Since the tests are individualized, the student may repeat the examination at later times, within the constraints imposed by the instructor.

Scoring tests

The instructor and his assistants may of course grade tests manually if they desire. However, computerized grading of the individualized tests is usually desirable, and may be performed using the information on the student's mark sense forms. Since the student has received the answer part of his test in exchange for his filled in mark sense form, there is no necessity for undue haste in grading the tests. The instructor or his assistant will, whenever convenient, have the information on the mark sense forms transformed to punch cards on an optical mark sense form reader. This step is required to obtain a form of input acceptable to the typical academic computing facility; one can bypass this step if optical mark sense form reading equipment is attached directly to his institution's computing equipment.

Scoring of the student responses for an exam is done by computer using a program GRADER. Input to this program is the answer summary deck punched by program GENERATOR when the tests were prepared, and the student response cards derived from the mark sense forms. Output of this program is a roster of student ID's and test scores and a punch card deck of the high score for each student for this exam.

Most academic institutions have available cumulative grading computer programs. These permit exam grades to be accumulated, and aid in the eventual preparation of final grades by generating score distributions and other statistics. The card deck prepared by GRADER is for use with such cumulative grading systems.

As a followup of test scoring, we are developing an item analysis procedure for CGRT. Since the item pools tend to be reused many times, such an item analysis will aid in the detection of defective test items and will assist the instructor in polishing his item pool.

THE COMPUTING PROCESSES

The test producing program GENERATOR and the scoring program GRADER are written in Fortran. Virtually all academic computing centers have well-maintained Fortran compilers that produce a fairly good quality of object code. We have several versions of the CGRT programs: well-documented ANSI Fortran versions designed to run on all commonly-available computers, and specialized versions of GENERATOR for the CDC 3600 and for the CDC 6600. The specialized versions utilize CDC extensions of ANSI Fortran to decrease the execution time dramatically by bypassing the repetitive processing of format statements during test printing. Since GENERATOR is completely output-bound, we anticipate that many potential users of the ANSI Fortran version would wish to discuss modification of the program with their systems people to take advantage of local extensions to Fortran output facilities.

GENERATOR reads the test item pool, checks each record for consistency of identification information, and creates a condensed file of the test item questions and answers, partially formatted for output. This item file is kept in primary storage. A directory of the origins of sets and individual items is formed to provide rapid retrieval of item information for test printing. GENERATOR then reads directives for test production: an arbitrary header line for all tests, an exam number, the number of individualized tests wanted, the number of items (questions) on each test, the test number of the first test (others are numbered sequentially from the starting number), and possibly other data to select additional options. A file is written for punching which records the set number, item number, and answer character of all test items. This deck, which is typically about 25 cards, is used by GRADER to regenerate the sequence of items and answers in a given test.

Then for each test, the program selects question items and writes the test onto an output file. For each test, a pseudo-random number generator is initialized with a unique but reproducible number. Using the "random" but reproducible sequence of numbers from the random number routine, GENERATOR determines the order of questions on a test by random selection without replacement of sets followed by random selection without replacement of an item from each set until the requisite number of questions is chosen. If an item from each set is used and questions remain to be chosen, the process repeats. When selection is complete, the question

TABLE I—Cost Analysis of CGRT[a]

ITEM	COST
Punch cards for item pools (One-time expense)[b]	$ .40
Punch cards for student responses: 1000 cards	1.00
Printer paper: 3000 sheets	8.30
Mark sense forms: 1000 forms	8.80
Keypunching services for item pool punching (one-time expense)[b]	6.50
Computer charges for test production and grading: about 5 minutes @ $200 per hour[c]	16.70
High speed line printer and controller rental and maintenance: @ $1800 per month[d]	6.00 to $12.00
Total expenses	$47.70 to $53.70
AVERAGE COST PER TEST: 4.8¢ to 5.4¢	

[a] for 1000 three-page tests.
[b] prorated over four semesters.
[c] Indiana University CDC 3600 system.
[d] CDC 512 printer system.

and answer text for each selected item is formatted and written. Program control then returns to prepare the next test.

GENERATOR also has several optional facilities, such as multiple copies of tests, an answer summary for the instructor, and a method of assigning point values to items to allow weighting of the items during scoring.

Since the amount of output is substantial and on-line secondary storage is limited, most people will find it convenient to write the tests as blocked records on a magnetic tape. The computer center staff may then print the tape at a convenient time.

GRADER accepts as input the item pool answer summary deck punched by GENERATOR and the student responses punched from mark sense forms. To grade a student's response to a particular test, GRADER uses the same item selection algorithm as GENERATOR to recreate the same sequence of items and answer characters. The student's score is formed as the sum of the values of each correctly answered item. The score, the test number, and the student's ID number are saved. When all student responses have been graded, GRADER sorts the ID's and test scores, and a roster is produced showing for each student his scores, highest first, and the test numbers. As a final step, a punch card summary of the roster is prepared for use in possible later cumulative grading operations.

THE ECONOMICS OF CGRT

At first glance a procedure that uses a computer for test preparation and for printing of individualized tests appears economically unsound. This is very definitely not the case. In Tables I and II, which are cost analyses for the preparation, printing, and scoring of 1000 three-page tests, we have attempted to itemize expenses in a similar manner for both CGRT and the conventional method. Therefore, the cost of a computer line printer and associated equipment has been treated as a separate entry, rather than included in general computer charges. We have assumed that such expenses as the initial keypunching of item pools are distributed over four semesters.

The analyses show that both CGRT and conventionally prepared tests cost about 5¢ per test. While the estimate for conventional exams is fairly accurate, changing some of the assumptions in the CGRT analysis may alter the estimate by perhaps up to two cents per test. Also, under repeatable testing, students tend to take more than one repeatable test over each examination unit. In any event, the cost of repeatable tests is in the same range as conventional tests. More important, the expenses of the CGRT process are a very minor item in the cost of educating the student, amounting to $.50 to $1.50 per student per course. This is inexpensive education!

SUMMARY

Computer Generated Repeatable Testing works. It has been used in numerous courses for nearly three years at Indiana University, and it is also in use at the University of Nebraska, Illinois Institute of Technology, Indiana University-Purdue University at Indianapolis, and

TABLE II—Cost Analysis of Conventionally Prepared Tests.[a]

ITEM	COST
Paper: 3000 sheets	$ 6.00
Mark sense forms: 1000 forms	8.80
Punch cards for student responses: 1000 cards	1.00
Clerical services @ $4.00 per hour:[b]	
Typing: 1½ hours	6.00
Multilithing: 2 hours	8.00
Collating and stapling: 4½ hours	18.00
Computer charges for grading: about 1 minute @ $200 per hour[c]	3.30
Total expenses	$51.10
AVERAGE COST PER TEST: 5.1¢	

[a] for 1000 three-page tests.
[b] estimates supplied by Indiana University Chemistry Department.
[c] Indiana University CDC 3600 system.

other places. The method has been enthusiastically used by instructors of undergraduate courses in such varied disciplines as psychology, chemistry, computer science, economics, English, speech therapy, home economics, accounting, and education.

In general, students have been highly satisfied with the repeatable testing method. Their mood is one of alertness rather than anxiety. They are relaxed during examinations, and their morale is good. The undergraduate counselling units of Indiana University have received numerous student comments favorable to CGRT.

An unexpected result in some of the CGRT courses has been the students' excellent performance on technical material not discussed in class. Repeatable examinations appear to provide a stimulus and a way to master material typically neglected by students in conventional courses. Although we have only a little data taken under properly controlled conditions, indications from several common achievement tests given at Indiana University are that overall student achievement in repeatably tested sections is higher than in conventionally tested sections of the same course.[3] All available evidence suggests that a system of frequent and repeatable examinations provide an excellent atmosphere for scholarly activities of beginning students.

We hope that many readers will wish to try CGRT or suggest its use to their non-computer-oriented colleagues. The computer programs and ample documentation are available from Franklin Prosser.

REFERENCES

1 P SUPPES M MORNINGSTAR
Computer-assisted instruction
Science Volume 166 pp 343–350 1969

2 W C COOLEY R GLASER
The computer and individualized instruction
Science Volume 166 pp 574–582 1969

3 D JENSEN F PROSSER
Computer-generated, repeatable examinations and large class instruction
Presented at Midwestern Psychological Association Meeting Chicago Illinois 1969