This paper describes computer-administered testing (CAT) whereby the entire testing process is automated; it does not discuss ‘paper and pencil’ tests which have been constructed, or which are scored, by a computer system. Further, it focuses on the testing environment and not on test content and only discusses CAT which is predicated on the same basic principles as traditional non-computer testing procedures. Three basic principles of well-designed CAT systems are discussed:

- students should not be disadvantaged in comparison with non-computer based testing;
- full use should be made of the computer-based medium, particularly with reference to the types of student interactions; and,
- instructor control over the testing environment should be maximised.

Also discussed is the fact that the very features that contribute to good CAT also make it difficult to conduct valid statistical analyses of either the test items or the test results.

INTRODUCTION

Testing is used for a variety of purposes: determining what a student knows, assigning grades, determining University entry, deciding who should be employed etc. Computers have often been used in the testing process as an aid to the construction and scoring of paper-and-pencil tests. Interest in the use of computers to administer tests has received increasing attention in recent times. The term ‘computer-administered testing’ is used to refer to a completely automated assessment process whereby a test is constructed using the computer, the student completes the test at the computer, and the student responses are concurrently scored by the computer. This paper discusses CAT which is predicated on the same basic principles as traditional non-computer testing procedures. It does not describe more complex testing approaches such as ‘adaptive testing’ and ‘admissible probability measures testing’, both of which can only be implemented in computer-based environments.

Research has shown a student preference for CAT over conventional paper and pencil testing provided the testing environment and the test itself are both well-designed (Anderson & Trollip 1982; Bugbee & Bernt 1990). In addition to the obvious savings in instructor time and effort offered by CAT, it also offers instructors a greater ability to tailor tests to suit individual students or groups of students. For example, instructors may easily select appropriate questions from item banks and determine their order of presentation. By controlling the choice of equivalent test items for particular objectives, instructors can define different tests which are equivalent in terms of their level of difficulty and the objectives they measure. Another major advantage of CAT derives from the concurrent record keeping that can occur—this enables feedback to be made available to both the student and the instructor at any time.

Some valuable forms of interaction available in computer-based environments are not possible in traditional testing situations. Examples of these are the ability to drag screen objects using a mouse and the ability to alter variables in a simulated process. Using these types of interaction, test items can be created which require students to demonstrate complex cognitive and psychomotor problem solving skills.

Provided CAT systems are well designed their disadvantages when compared with paper and pencil tests are few. Apart from logistical disadvantages (most notably the need for extensive computer facilities), the
other major disadvantages are a greater level of anxiety among students not comfortable working with computers (a rapidly diminishing concern), some difficulties in conducting valid psychometric analyses, the inability to adequately judge extended or open-ended responses and the dependence on machinery and power availability. Some suggestions for dealing with these disadvantages, together with recommendations for research related to them, are given later in the paper.

DESIRABLE FEATURES OF CAT SYSTEMS

Since most desirable characteristics of any testing environment depend upon the purpose of the test and the goals of the instructor, a fundamental attribute of a good CAT system is sufficient flexibility to allow instructors to control a variety of parameters in order to ensure that the test is compatible with the purpose and goals. A second important attribute of a good CAT system is that it does not disadvantage students when compared with noncomputer-based testing, particularly when test anxiety is a significant issue among students. Unfortunately, most existing examples of CAT systems fail this criterion, some rather badly. Many fail to provide the student with such basic capabilities as previewing, reanswering and easy movement among test items. It is strongly argued that CAT systems should provide these options for students and that they must, in a more general sense, provide at least the level of flexibility that students are accustomed to in traditional paper and pencil tests. The American Psychological Association (APA 1986) guidelines support this view and state that

computerized administration [of tests] normally should provide test takers with at least the same degree of feedback and editorial control regarding their responses that they would experience in traditional testing formats. (p. 12)

The following three principles are suggested by Anderson and Trollip (1982) as the basis for designing CAT systems:

• ensure easy access to needed information
• maximise user control
• install safety barriers and nets.

These principles apply equally to both instructors and students. The first principle can be exemplified by the need for instructors to access test results and the need for students to get directions on how to use the testing system. Examples of the second principle are the instructor's ability to alter parameters such as type of feedback and item randomisation and the students ability to change answers at any time. Examples of the third principle are the need to confirm significant decisions such as an instructor's indication that records are to be deleted or a student's indication that they wish to finish the test.

Instructor options

CAT systems should allow instructors to set a range of parameters such as student access, time limits, the pass mark, resources allowed and the availability of student options such as previewing, marking for review and sample question practice. Instructors should also have an option which enables them to 'test the test' at any time (without having to enter the system as a student) in order to ensure that the items are correct and that the test is working properly. No records should be kept by the system in this case. Instructors should have easy access to detailed results for each student, summary statistics for the test and student comments. The results of statistical item analysis procedures should also be available to enable item performance as well as student performance to be studied. The capability of printing this information for instructors is desirable.

Network environments

In network environments instructors should be able to view status information (e.g. questions answered, time remaining) on all students and be able to communicate to students via electronic mail. This latter facility is particularly useful if an error in the test is detected and the instructor wishes to notify all students. Also, circumstances arise where instructors need to be able to delete or manipulate student records. Appropriate warnings ('safety nets') are necessary if such manipulations will permanently alter stored information.

Prior to testing

Before the testing process commences, students should be presented with clear directions concerning the use of the computer and the use of the testing system. This type of help should also be available during the testing process. The basic requirements for this stage are that the system be easy to use and that it be obvious how to start the test. Students should be presented with information such as the required passing score, the resources permitted (e.g. books, clocks etc), whether sample (practice) questions are available, whether the test questions can be previewed and how much time is allowed for both previewing and testing. The opportunity to practise on sample questions prior to taking the actual test is a particularly important feature for students not familiar with the type of computer-based testing system and/or the type of interactions that will be required of them during the test. This feature helps to ensure a positive reaction to the test by students as it does the implementation of appropriate 'safety barriers and nets' such as the prevention of an accidental start (or finish) to a test by requiring confirmation of the relevant actions.

Screen design

Once the test begins, the student should be presented with a carefully designed screen showing the first question, relevant status information (such as current question number, number of questions answered and time remaining) and navigation controls. Well-designed CAT systems usually partition the screen area to accommodate this type of information, taking account of the fact that while content related to a particular test item will change for each question, the basic layout of the status and navigation information will remain constant. Regardless of the exact nature of the screen layout used by a particular CAT system, it should remain consistent throughout the test.

Changing responses

The procedures for responding to test questions should be flexible enough to allow students to change their response or to not respond at all. Two situations arise with respect to changing responses: the first occurs when the student simply alters a response while the question is still on the screen; the second occurs when a student wishes to return to a question answered earlier and change their response. The first situation should always be possible whereas the second should be possible provided the CAT system has not been configured by the instructor to allow feedback after each question. An inability to change answers
at all can be unfair to students and can affect the reliability and validity of the test. It is worth noting that allowing multiple attempts at questions makes it difficult to use adaptive testing where item presentation is dependent on previous responses.

Judging responses
When judging responses, format errors such as entering a numerical choice that does not exist in the question, clicking on an inappropriate area of the screen or dragging an object to an ‘out-of-bounds’ area need not be accepted, thus allowing the student another response. This is a further example of an advantage CAT has over traditional testing; in the latter, inappropriate responses are only detected at the marking stage and must then be recorded as wrong answers.

Feedback
CAT systems should allow the instructor to decide whether feedback is to be provided or omitted, and if provided, the type and timing of the feedback to be used. It should be possible to set the type of feedback (e.g. indicating correctness or providing an explanation) as well as the timing of the feedback (e.g. after each question, on completion of the test or not at all). Status information such as the current question number, the number of questions answered and the time remaining should be constantly updated and always visible. However, the displaying of status information such as the progressive test score should be determined by the instructor.

Navigation
Navigation through the test questions should, in general, not be constrained, with students having the ability to move easily through the questions in order (forward or reverse) or by directly ‘jumping’ to a nominated question. It is also desirable to allow students to ‘mark’ questions for return at a later date without having to search the whole test for them. Such flexible navigation must, however, be compatible with the level of feedback being given. Unrestricted movement between questions, with the corresponding ability to change previous responses to questions, is clearly not compatible with the giving of feedback after each question. While flexible navigation options are generally desirable for standard CAT, they are not suited to special types of computerised testing such as adaptive testing where the set of items which constitute the test is not pre-defined and is determined as the student progresses, the items being selected or determined by the student’s performance level.

Student comments
An option for enabling students to make comments at any time, which can be later read by the instructor, enables the gathering of useful information and allows students to express frustrations, raise objections and make complaints. Such comment files should be routinely checked by instructors and may in some cases affect final marks (for example, a comment may bring to light the fact that a particular question was ambiguous).

Termination
In addition to the automatic termination of a test based on the expiration of the time available, tests should also be able to be terminated at any time by either the student or the teacher. Teacher initiated termination may be necessary for several reasons (e.g. when a student is found to be cheating) and student initiated termination is necessary to enable students to finish when they wish to do so (this is essential in situations which allow students to ‘browse’ through questions in a flexible navigation environment). If termination is student requested the testing system should provide information on questions not answered and questions marked for review before asking the student to confirm termination. The likelihood of accidental termination by either the student or the teacher should be reduced by always requiring a confirmation procedure.

Data gathering
No termination procedures should result in the loss of data. To ensure this and to help overcome the disadvantages associated with the dependence of CAT on machinery and power availability, student and system files should be updated after each response so that no data is lost if the test is terminated because of machine or power failure. Further, the system should be capable of resuming at the point of termination with all aspects of the original environment restored (e.g. the clock showing the correct time remaining). As a minimum the system should collect the following data for each test item: the question number, an item objective code, the student’s response, the correctness of the response and the correct answer.

Time limits
While time limits are often appropriate for norm-referenced tests, time limits which may affect performance are not generally recommended for criterion-referenced tests. In tests with time limits a warning should be given as the time limit approaches.

After the test
After the testing process has concluded, students should normally be presented with a screen detailing their performance on all questions, together with an overall performance indicator. The capability of printing out this information for students is desirable. However, as with other options, the performance data to be shown and printed should be determined by the instructor. The option for enabling students to make comments should remain active after the test to allow student comments of a more general nature. Information concerning relevant on-line and off-line learning resources should be accessible to students who have just completed the test and seen their results.

Item banks
A major advantage of CAT is the capability of having item banks of questions from which the actual test questions can be easily drawn. This allows the creation of different but statistically equivalent tests, a useful feature when students are taking a test in close proximity or when they are taking a test at different times. Since CAT systems offer the possibility of constructing tests by randomly selecting items from item banks, students who fail a test may see some of the same items again in another attempt. This problem is more serious when feedback is given after each question. To overcome this, CAT systems should allow instructors to select a specific item bank, or set of item banks, from which the items for a test will be drawn. Further, after the test items have been determined, the order of presentation of questions should be under instructor control with a randomising option.

Security
Access to the system should only be possible by authorised instructors and students. Students should only be able to view data on their own performance
and perhaps an indication of relative class standing. These requirements are relatively easy to meet with a network of microcomputers but for CAT using floppy disks, data encryption methods may be required to ensure security. Any system should attempt to ensure that a given student takes the right test and that the right student takes the test. The former task is not difficult and is usually handled by asking students for their name and/or an identification number. The latter task, which equates to cheating, is more difficult to handle without some form of human intervention. For students doing the test on site and under supervision, the procedures are the same as for a traditional test. If students are tested off site (typically at locations some form of human intervention is normally required.

**PSYCHOMETRIC ISSUES**

**Item types**

Different item types are normally used to test different types of learning. The 'selected-response' items commonly found in CAT (e.g. multiple choice, true/false, matching) require recognition but not recall; only 'constructed-response' items should be used to test objectives requiring recall of facts. An exciting characteristic of state-of-the-art CAT systems is the opportunity they provide for testing through innovative ways of presenting and answering questions. For example, answers to questions may be given by one or more of the following means: 'task entry' (typing on the computer keyboard), 'click/touch' (using a mouse or touch screen to identify areas or objects), 'move object' (directly manipulating objects on the screen using a mouse or finger), 'pull-down menus' (selecting from items in a temporary list that overlays the screen on demand), 'key press' (pressing a key on the computer keyboard), and 'pre-set simulations' (computer simulations of an action observed by students as an integral part of a test question).

The computer's interactive and multimedia capabilities enable the use of new creative testing strategies which have the potential to improve the validity of test items. For example, the use of constructed response items based on simulation techniques offer students tasks which are more realistic and closer to those they encounter in education and work settings. The use of non-multiple choice items which can be reliably and accurately scored by a computer makes it possible to broaden the scope of standardised testing as well as diagnostic testing. However, improvements are needed in the natural language processing and artificial intelligence capabilities of standard microcomputer systems before the risk of adversely affecting the reliability measures and discrimination indices for all but the simplest of constructed-response items can be kept to an acceptable level.

**Item contamination**

Information about reliability, validity, item analysis and norms should be collected as part of any test development process. However, the very features that are considered advantages of CAT tend to be those which make it difficult to conduct valid statistical analyses of test results and test items. Some test strategies may 'contaminate' items and render comparison groups 'non-equivalent' for the purpose of item analyses. This is often the reason why the ability to give feedback after each question, a unique advantage in learning environments, is not always recommended for CAT. Two examples where such problems arise are: when feedback for one item provides a clue to the answer for another item; and when consistent negative feedback (concerning item correctness and/or progressive test score, for example) has a detrimental effect on student motivation. Further to this latter example, since the converse is also true, feedback after each question may cause differential effects between low and high achieving students. Such problems can adversely affect the validity of psychometric analyses related to the test performance data.

**Test equivalence**

For a given testing session, students may encounter different items (when the items are randomly selected from item banks), a different order of items (when randomly presented) and/or a different number of items (when the test is discontinued after a pass/fail status is determined). Students have not attempted equivalent tests unless the number of items is the same and equivalent items of equal difficulty and discrimination are used. If a test is discontinued (a common feature of mastery tests) there is often no sensible 'total test score' upon which to base statistical analyses. Item Response theory can handle these difficulties but only if estimates of item parameters can be obtained from large samples.

**CONCLUSION**

Tests are likely to continue to remain a commonly used means of assessing learning. Using well-designed CAT systems can provide relief to instructors and may well provide higher quality tests to students. The major implications for educational researchers is that a range of new questions has arisen from the unique features of CAT. When should student performance feedback be given? Does the use of simulation techniques result in 'better' test items? Are new statistical analysis procedures required? And so on.

Given the lack of guidance available from the research literature, the varying purposes of tests and the differing goals of instructors, the major implications for CAT system developers are threefold:

- students should not be disadvantaged in comparison with non-computer based testing;
- full use should be made of the computer-based medium, particularly with reference to the types of student interactions; and,
- instructor control over the testing environment should be maximised.

**REFERENCES**


