COMPUTER MANAGED STUDENT ASSESSMENT: A CASE STUDY

Adrian Blunt
Beverley Dent
University of Saskatchewan

Abstract

Computer managed assessment can provide adult and continuing education with a powerful instrument to promote program quality and instructional effectiveness and efficiency. However, few case studies are available to report on the achievements of such systems. This article reports a study of one college's experience with computer managed assessment and makes recommendations for the improvement of first generation computer assessment systems. It was observed that the assessment system was not being used as an instrument for improving program and instructional quality.

Résumé

L'évaluation gérée par ordinateur peut fournir aux services d'andragogie et d'éducation permanente un puissant outil de promotion de la qualité, du rendement et de l'efficacité des programmes et de l'enseignement. Or, il existe peu d'études de cas susceptibles de nous renseigner sur la portée de ces systèmes. Cet article décrit l'incursion d'un collège dans le monde de l'évaluation gérée par ordinateur, et formule quelques recommandations visant l'amélioration de la première génération de systèmes de gestion informatique de l'évaluation. On a notamment observé que ces systèmes sont rarement utilisés pour régénérer la qualité des programmes et de l'enseignement.

The widespread availability of powerful, relatively inexpensive computers provides adult and continuing education systems with an instrument that has great potential for improving programs and supporting learning activities. One important capability of the computer is its capacity for managing student assessment systems. Computers are very useful for constructing and administering tests whereever tests need to be (a) constructed to local specifications, (b) constructed frequently, (c) available in multiple equivalent forms, and (d) available on-demand for use with individual learners. Competency-based education (CBE) and mastery learning systems introduced into adult and continuing education systems over two decades ago are prime situations for the use of computer managed testing because they require two or more of the above conditions (Dunkleberger & Heikkinen, 1983).

Well developed, valid, and reliable tests of achievement and performance are important instructional tools which can provide information that assists instructors to plan and to deliver their instruction effectively and to assess the outcomes of their efforts (Blank, 1982; Gronlund, 1988). Furthermore, research on learning has clearly established that learning is greatly enhanced when learners receive meaningful, immediate, and continuous feedback about their progress (Gagne, 1985). Many continuing education institutions have implemented individualized, competency-based programs; as well, many workplace technical programs use competency based approaches. Thus the topic of assessment is of interest to many adult educators. Major commitments of resources for efficient on-demand assessment systems have been made by some continuing education institutions which have installed computer managed testing systems. However, the experience of these institutions with the implementation of computer based testing has not been adequately reported in the research literature.

Context of the Study

Purpose and Overview

This study was conducted to assess the implementation of one computer-managed adult and continuing education student assessment system at a 10 year old, government proclaimed state-of-the-art[1] college, and to determine whether the system contributed to high quality student assessment and the improvement of instructional programs. The study was conducted within the same paradigm of technical rationality that guided the college's program design, curriculum development work, and daily operations. It is not our intention to directly engage in the debate around the strengths and weaknesses of competency based education (See for example Jackson, 1988) although this study's findings do serve to inform that debate.

First, we review the history of computer managed assessment, describe the study site, and outline the methodology. Next we report our assessment

[1] State of the art when it was opened in 1986.

of the college's test-item bank and computer-generated tests, and present opinions of the system held by students and instructors. We conclude the article with a discussion of our research findings and recommendations for changes to the system.

Background Literature

Over 20 years ago, Lippey (1974) listed eight specific testing functions that could be performed by computers: (a) item banking; (b) item generation; (c) item attribute banking; (d) item selection; (e) test printing; (f) test scoring; (g) maintaining records of students, tests, and items; and (h) diagnosis and remediation. Diagnosis and remediation are provided when a computer is programmed to recommend units of instruction to be studied by a learner who selects incorrect answers on a test.

In the 1970s, computers first began to be used for item analysis and test improvement purposes (Baker, 1974), an application for which they were, and continue to be, ideally suited (Carlson, 1994). By the 1980s four distinct generations of computerized educational measurement had been defined (Bundesen, Inouye, & Olsen, 1988). First generation *computerized testing* is achieved when conventional paper and pencil tests are converted to computer delivery. Second generation *adaptive testing* occurs when items are selected on the basis of item-response theory, with the difficulty of items selected being based on the performance of the examinee to prior questions. Third generation *continuous measurement* embeds assessment procedures within the curriculum, rendering evaluation continuous and unobtrusive. Finally, the fourth generation *intelligent measurement* builds on the instructional and cognitive sciences and uses artificial intelligence to enable an individual learner and teacher to interact through computer software.

By the early 1980s item response theory (Lord, 1980) was well established in the research literature; guidelines for assessing the second generation, computerized adaptive testing had been proposed (Green, Bock, Humphreys, Linn, & Reckase, 1984); and the equivalency of scores from automated and conventional versions of tests was well established (Mazzeo & Harvey, 1988). Also, instructional design and evaluation researchers had made progress towards the evolution of third and fourth generation assessment systems.

However, during the early 1980s, an applications gap occurred between the capacity of the technology and test theory, on the one hand,

and the field of educational practice, on the other hand. Our searches of the literature, including ERIC and education journal indexes, yielded few published reports on the use of computer-based assessment systems and the implementation of each of the four generations of assessment models. No studies of the implementation of the first generation models were located. Research in this area appears to have suddenly ground to a halt, and by the mid 1990s journal articles (See for example Sandals 1992; Carlson, 1994) still tended to outline system guidelines rather than to report the effects of assessment systems on instructional efficiency and effectiveness.

Some important findings were contributed by the few published studies. For example they revealed that the common method of generating parallel tests by random selection of items from a test bank was frequently performed poorly, as most test item banks did not contain a well defined universe of items. Many authors (Baker, 1974; Choppin, 1985; Feuer, 1986) agree that an item bank needs to be assembled using well-developed test specifications and to be more than just a collection of items. "Simply selecting items at random from a collection of items does not insure the creation of randomly parallel tests" (Baker, 1974).

Hsu and Sadock (1985) concluded an assessment of the state of the art of computer-assisted test construction by declaring that, as of the mid-1980s, no studies had been published to demonstrate that the quality of assessment had been improved by computer applications. They concluded that administering tests by computer was only justifiable if it improved the quality of testing, and recommended four means by which qualitative improvements might be achieved: (a) providing immediate feedback to students; (b) adaptive testing, a system whereby the computer selects an item on the basis of the response to the previous item; (c) storing and analyzing test results; and (d) increasing test security.

At present the research literature lacks case studies on the use of computer managed testing in CBE systems. Adult educators still need to know whether, or how, computer managed assessment systems effect the validity of student evaluations and related aspects of program quality—including instruction, instructional materials, educational costs, and the quality of the educational environment overall.

Method

Research Site and Student Assessment System

Northern Community College (NCC), a pseudonym for the research site, commenced operations in 1986 as a government proclaimed state-of-the-art adult and continuing education institution with individualized, competency-based vocational-technical programs and a computer-managed student monitoring and assessment system. Specific features of NCC include individualized modular instruction, year-round operations with continuous student intake and exit, extended daily hours of operation, part-time and full-time enrollments, minimal qualifications for entry, a challenge system for knowledge and performance tests, and a future mission (although not an immediate capacity) to deliver programs by distance education methods to marginalized adults living in remote northern communities and work sites.

NCC delivers 40 programs ranging from trades training (carpentry, electrical, mechanical, etc.) through business training (secretarial, accounting, computer applications, etc.) to semi-professional career programs in fields such as corrections, nursing, childcare, and natural resource management. The curricula for these programs are competency based and designed for delivery by individualized instruction using in-college developed modules, each of which focuses on an area of competency. Within the 40 programs there are approximately 3500 different competencies, all of which can be tested for prerequisite knowledge and performance. Adult basic education programs are offered by the college; however, they are not competency based, and those programs were excluded from this study.

On any given day, about 500 individual students request tests or retests. Instructors therefore need to generate, administer and score a great many different tests on demand across all program areas. NCC uses two types of tests, performance (skills) tests and objective knowledge tests, with the latter being basically paper-and-pencil tests delivered via a computer terminal. The system is therefore an example of a first generation model of computer testing (Bundeson, Inouye, & Olsen, 1988). Most of the objective knowledge tests use a multiple-choice format, and there are approximately 73,000 such items stored in the computer test bank. The computer generates a test by randomly selecting a predetermined number of items from a specified content domain (which is organized by technical area of competency) in the test bank. When a student wishes to challenge a knowledge test for a competency she or he is studying, the student requests the relevant test from the NCC testing center. If the student's fees are paid and all of the required prerequisites have been completed, the student is assigned to a computer terminal through which the test is delivered. The computer scores the test, and within a few minutes the student can retrieve the total score earned from a designated terminal located outside the testing centre. If the student wishes to review the test, an instructor is able to call up the actual test and individual item responses on a terminal in her or his office.

The knowledge test must be taken before the skills test for each particular competency being studied. A student who fails a knowledge test usually returns to the original instructional materials for further study or seeks assistance from an instructor. When ready, the student can take a second knowledge test generated in the same manner as the first. However, a student who fails three consecutive tests is locked out of the system by the computer until an instructor or teaching assistant unlocks it following a review of the student's problems.

Skills tests (frequently check lists or rating scales) are administered and scored by an instructor who enters the results into the computer. Parallel versions of skills tests are kept for subsequent challenges if the student's first challenge is not successful. The computer retains the results of all tests taken but only generates and scores knowledge tests.

Assessment of Test Item Bank

Test improvement processes can be considered from a priori and a posteriori positions. A priori methods assess content validity and technical (or mechanical) construction of test items. Content experts determine content validity by comparing items to course objectives, and test specifications, then applying their own knowledge of the occupation in question. This procedure parallels the use of an advisory panel to develop or validate a job analysis and course objectives. The a posteriori test improvement process is an empirical method that relies on the generation of item statistics, the classical item improvement model (see Popham, 1978). As both approaches rely on the administration of tests to groups of people, preferably large groups, it is necessary to combine the test responses of students in individualized programs over time in order to simulate a group's responses. No attempts have been made to conduct a posteriori test improvement at NCC over the last 10 years.

Twenty years ago Popham (1978) predicted that the a priori method would become the most popular method of test development and improvement, as it is a practical method that yields important information and can be applied before the administration of any tests. It is based on specific test specifications and systematic human judgment. Practising instructors and teachers can be taught to apply this method without any training or background in statistics; however, this does not mean it is a simple mechanistic system, and experts agree (Gronlund, 1988; Popham, 1978) that writing good multiple-choice test items is a demanding exercise.

The NCC computer was used to generate a random sample of 250 items from the 73,000 in the test item bank. Most (242) were multiple-choice items with the few remaining being either binary choice, or fill-in-the-blank items. The multiple-choice items were assessed against 12 a priori criteria for effective test items developed from the literature (Gay, 1980; Gronlund, 1988; Jones & Whittaker, 1975; Nitko, 1996):

1. A single clearly formulated problem is presented in the item stem as a question or as an incomplete statement, but in either case a knowledgeable person can respond to it without looking at the alternatives offered.

2. As much of the wording in the item as is possible is included in the stem. There is no repetitious wording at the beginning of each of the alternatives.

3. Clear, concise, simple, and unambiguous language is used without irrelevant material, difficult words or unnecessary technical terms.

4. The stem is worded in positive terms or, if the negative is used, the negative term is emphasized (for example NO, **not**, <u>never</u>).

5. The item does not use specific determiners (for example never, always, usually, sometimes).

6. Distractors are plausible and attractive to the uninformed as well as homogeneous in type.

7. Alternative answers are grammatically consistent with the stem and parallel in form.

8. There is no similarity of wording between the stem and the correct answer that might provide a clue to the correct answer.

9. The alternatives do not overlap, are not all inclusive, and two or more do not have the same meaning.

10. The correct answer is not stated in greater detail or length than its alternatives.

11. "All of the above" or "None of the above" are not used as alternatives.

12. Items measure higher levels of cognitive knowledge than recall or recognition of facts, or principles. In other words, items determine whether the student understands a concept, is able to apply a principle, or can analyze a problem (higher level learning outcomes than Bloom et. al.'s, 1956 taxonomy level 1).

Unfortunately, two important questions could not be asked: does each item measure an important aspect of the occupation, and does each distractor represent a common student error or misunderstanding? To answer these questions would have required a person, or persons, knowledgeable in each program area represented by the sample and an extensive computer analysis.

One of this study's two investigators rated all of the 242 test items and 78 individual tests analyzed in the entire study. A panel of 11 vocational-technical, undergraduate teacher education and post-secondary certificate students from the University of Saskatchewan, who had recently completed a course in test-construction methods, served as a panel of judges to provide a reliability check on the investigator's ratings. A common sub-sample of 15 items was first rated by each of the 11 judges. This sub-sample contained selected items which contained one or more examples of the types of errors represented by the evaluation criteria. Next, pairs of judges rated five randomly selected sub-samples of 15 items. The panelists' ratings were compared with each other and with the investigator's ratings. Levels of agreement were calculated, and a high level of agreement (80% or higher) indicated items were judged similarly, confirming that the ratings of the investigator were reliable. The procedure also served to demonstrate that, if NCC's instructors were similarly trained, they would have the capacity to assess the quality of the college's test bank.

The sample of 242 multiple-choice test items drawn from the computer bank were then analyzed according to the 12 criteria, listed above, by scoring each item *one* or *zero* for compliance with the item. The results were entered into a spreadsheet for analysis. Only 119 of the 242 items (49.3%) met all of the criteria; 122 items (50.7%) had one or more flaws, and 64 (26.6%) had two or more flaws contributing to the reduced effectiveness of the items (see Table 1). This is a serious problem as only one flaw may permit a student to guess the correct answer. Such a flaw renders the item useless as a

Table 1. *Multiple-Choice Item Flaws*

Number of Flaws	Number of Items	Percent
0	119	49.3
1	58	24.1
2	37	15.4
3	21	8.7
4	4	1.7
5	2	0.8
Total number of items sampled	241	100.0

determinant of whether or not the learner has acquired the knowledge being tested. If a flaw does not entirely give away the answer but permits a student to eliminate one or two of four possible answers, the chance of guessing the correct answer increases. The desired level of difficulty for each norm-referenced test item is 50% (Gronlund, 1988), and it is recommended that test items which have a 75% chance of being guessed successfully not be used.

To check the reliability of the investigator's ratings, the sample of 242 items was broken down into sub-samples for assessment by the panel of 11 judges trained for the task by a university professor teaching a vocational-and-technical, teacher-education student evaluation course. Fifteen items were selected from the first 30 of the computer-drawn sample. These items were chosen by the primary rater and included both well constructed items and items with each type of error represented by the judging criteria. Panelists rated this sub-sample and these results were compared with the primary rater's ratings. The inter-rater agreement on this common sample averaged 83% within a range of 79% to 89%, which provides satisfactory evidence of the inter-judge reliability of ratings.

The remaining 210 items were divided into sub-samples of 15 and assigned to pairs of panelists so each sub-sample was rated by the primary rater and two judges from the panel. All ratings were entered in spreadsheets and compared to arrive at a percentage of agreement. A relatively high level of agreement between raters permits a high degree of confidence in the ratings of the primary rater. The extent of inter-rater agreement averaged 85%, with a range of 73% to 98%, which indicates a very satisfactory level of inter-rater agreement.

The most important—and disturbing—finding from the item analysis was that only 14 of 242 items (6%) sought to test levels of knowledge higher than basic recognition or recall on Bloom et al's (1956) taxonomy of educational outcomes.[2] These 14 items tested only comprehension or application, the second and third levels of the taxonomy, and no items were identified that tested knowledge at levels above the third level of the taxonomy. This finding confirms the frequently expressed criticism of multiple choice test use: that only material and objectives of lesser importance, even trivial, are tested whereas the most important learning outcomes in a program are ignored because of the difficulty in writing multiple choice items at higher levels.

Assessment of Computer Generated Tests

A good item bank and the capability to generate multiple alternative forms of each test are the two major requirements of an assessment system. The NCC system generated a random sample of 39 objective knowledge tests and a second sample of 39 tests intended to be equivalent forms of the first set. These tests were rated against eight criteria developed from the literature (Gay, 1980; Gronlund, 1988; Jones & Whittaker, 1975; Nitko, 1996):

1. Are there sufficient items to establish test reliability and content validity? (10 to 20, never less than 10).
2. Is every item independent of all other items? (One item does not give away the answer to another or is not just a reworded version of another).
3. Are the questions presented in a logical order such as increasing difficulty, or grouped by objective?
4. Does the position of the correct answer vary in the list of distractors?
5. Are the alternatives in a logical order, if one exists?
6. Does the relative length of the correct answer vary?
7. Are there items which test higher levels of learning than knowledge according to Bloom et. al.'s taxonomy?
8. Is each item directly related to the stated learning objective?

The sample of computer generated tests was analyzed in a similar manner to that utilized to analyze the test items. Only 59% of the tests (46 of 78) contained at least 10 items, the minimum number recommended as

[2] Bloom et al's taxonomy of cognitive learning outcomes consists of six major hierachical categories, which comprise what is frequently called the *cognitive domain*. From the lowest to the highest level of knowledge organization the categories are: (a) knowledge, (b) comprehension, (c) application, (d) analysis, (e) synthesis, and (f) evaluation.

necessary to establish test reliability (see Table 2). The number of items varied from 3 to 33 with a mean of 10.

Seventy-five (96%) of the tests had items independent of all others; that is, one item did not provide a clue to the answer for another item. This is not surprising considering the low ratio of items in the test bank to test items drawn; there are approximately three items in the test bank for each one selected. However, the total number of items in the item bank (73,000) divided by the total number of competencies in NCC's programs (3,500) reveals an average of 20.86 items per competency. With an average test of 10 items per competency, the mean ratio of items remaining to items drawn is only two to one. Test experts consider a ratio of ten to one to be the minimum acceptable (Baker, 1974). In this case the findings that 96% of the NCC test bank's items are independent of others in the same test is more likely a measure of the simplicity and inadequacy of the test item bank's size than a product of a well-planned item selection strategy.

Items in the computer test bank are grouped by instructional objective and competency with essentially one objective per competency. There is no logical ordering of items in the bank by level of learning outcome, and with only 6 per cent of all items sampled testing levels of learning higher than recall, one can infer that the item bank consists almost exclusively of low-level knowledge items. As a first generation model the NCC system lacks the capacity to order items by knowledge level or to perform any adaptive testing functions.

Table 2. *Test Analysis Summary*

Criterion	No. of Tests with Criterion Met	Percent
At least ten questions	46	59
Items independent of all others	75	96
Items grouped by objective	78	100
Position of answer varies	76	97
Alternatives in logical order	26	33
Length of correct answer varies	72	92
Test higher levels of learning	5	6
Related to learning objective	78	100
Total number of tests sampled	78	100

The computer managed the task of randomly positioning correct answers, with the position of 97% of correct answers varying from one item to the next, but it was not programmed to place alternatives in a logical order in those cases where a logical order existed or was desirable. The relative length of the correct answer varied 92% of the time. All items were judged to be related to the appropriate learning objective. However, because the raters were not content experts, this finding is best interpreted as an indication that the items contained nothing that a lay observer might detect as a problem. All the test forms reviewed were judged to have an acceptable equivalent form. This observation was not surprising considering that there were so few items for selection at the objective level. If there had been a larger proportion of items testing higher levels of learning and if the tests had been constructed from more adequate test specifications, then these findings would be more meaningful indicators of a quality assessment system. A final major concern was NCC's policy of establishing a test score of 80% for every test without regard to overall test difficulty. This, in our opinion, is an overly simplistic standard having no relationship to employment practice and no empirical justification.

The sample of computer tests was divided into sub-samples and rated by the same judges who rated the test items. The ratings were compared with those of the primary rater in the same manner as that used for the item analysis. All judges rated a common sub-sample of tests, and agreement ranged from 74% to 96% with a mean of 82%. In addition, pairs of judges rated another sub-sample with their ratings being compared to the primary rater and to each other. The average agreement of 83% (range 69 to 98%) observed indicated a very satisfactory level of inter-rater reliability.

Students' Opinions

Data were collected by means of structured interviews with a random sample of 43 students from the student population of approximately 730. Demographic data, including the student's sex, age, and ethnic origin, were noted on the interview schedules and the resulting summaries were compared with the college's demographics. Because the population parameters compared favorably with the sample statistics, the sample was judged to be representative of the college population on the basis of the characteristics selected for comparison including sex, age, and years of education.

The sample of 43 students was interviewed to determine how students used the testing system. A series of Likert scale questions sought opinions on a variety of system-user issues. Most respondents (n = 37; 83%) reported that they *usually* or *always* studied all the material recommended or presented by the study guides before challenging a computer test. The majority (n = 33; 77%) stated it was *usually* or *always* necessary to study all the material before trying a test. Similarly 37 (86%) reported that they *usually* passed the computer tests on their first attempt. Fifty six percent said it was *never* possible to pass a computer test by guessing, but 44% said it was *sometimes* possible to do so. Most students said it was common knowledge that certain modules are easy to pass with little study required. Ninety-five percent agreed it was impossible to cheat on a computer test, although 5 percent thought it might *sometimes* be possible. An on-site examination indicated that the computer testing area was very secure.

Most students, having failed a computer test, either studied the same material again (86%) or went to an instructor for help (65%). Only 21% *usually* and 21% *sometimes* used different materials to study. Thirty-five percent said they can *usually* find alternative materials if they actively seek them. Eighty-four percent found the instructors helpful when they failed a test, and the most common alternative mode of re-test preparation was to get an oral explanation of the area of difficulty from an instructor. According to the students, instructors spend most of their time in interviews with students—so much so, that they may have little time for improving the testing system.

The majority of students (85%) used similar study methods throughout their courses and they recognized that most other students functioned in much the same way as they did themselves. Although the competency based system differed from the traditional education experiences that they had encountered previously, students appeared to have learned a method of coping with it, but not necessarily how to become effective, independent learners. In a traditional setting, the teacher is the primary source of instruction on a group basis and individual reading materials are a secondary source. At NCC, individual reading materials are the primary source of instruction and the instructor becomes the secondary source on an individual basis.

Instructors' Opinions

Ten instructors from a population of 68 were selected at random for interview. The respondents' teaching experience ranged from 0 to 41 years with a mean of 17.5 years, and their years of supervisory experience (supervising other workers) ranged from 0 to 26 with a mean of 5.8 years. Respondents had diverse educational backgrounds; one person lacked high school completion; three had been to technical school, served apprenticeships, and obtained vocational-technical teaching certificates; a fifth had been to technical school but worked in a trade area that was not licensed; and five had earned university degrees. Two of the university graduates had degrees directly applicable to the program in which they were teaching, whereas two others each had two degrees: a B.Ed. and one other degree that did not relate directly to the program in which they were instructing. One person had been to technical school, obtained a journeyman's license, and a B.Ed. degree.

Five instructors had taken one or more university level classes in student evaluation, including four who had attended one or more professional development seminars at NCC. Three persons had no training in evaluation. Nine of the ten had attended at least one professional development seminar, but these seminars were not directly concerned with student evaluation.

Eight of the 10 respondents said they had been involved in developing student tests for use at NCC. When asked about NCC's test development procedures and decision-making about what should be tested and to what standards, the instructors reported that their personal opinions were the primary sources. Two respondents stated that a job analysis carried out elsewhere was the basis for their program's development, whereas four others stated that provincial advisory boards had provided some guidance. Paper-and-pencil tests and hands-on practical testing, either-on-the-job or as simulations, were reported to be the testing alternatives that had been considered. Some perceived that they were locked into the evaluation and testing system implemented at NCC and had no alternatives. Only two persons, both with B.Ed. degrees, had considered or implemented any alternatives testing procedures or amended the existing system in any way.

When asked how standards for testing were determined, the guidelines of provincial advisory boards were mentioned; however, the instructors' opinions were the dominant criteria. Two instructors said that feedback from practitioners in the field was used to help set standards. Checklists for skills tests were commonly used; four instructors indicated that they developed their own, three got them from books, and five from other persons.

The instructors were questioned about basic concepts of evaluation. These were open-ended questions, and the conclusion reached was that only the two instructors with B.Ed. degrees had any understanding of the major concepts of test development including, for example, test specifications, item difficulty, or pilot testing. Most considered the time allowed for a test and the number of correct responses as the only criteria of test difficulty. Instructors indicated that their personal feeling or judgment was the sole basis for deciding whether or not a test discriminated between good and poor performance; they had no empirical basis for their decisions. When asked whether they used test specifications for test development purposes, only three claimed to do so, although one of the three admitted to only doing it mentally. All instructors indicated there was no requirement to pilot test new tests. Nine of ten said that there was no college manual or guidelines for the production of tests; one person identified a draft version of material on testing that had been developed by an NCC program writer. The material in this draft was essentially a summary of the work of Gronlund (1988) and Gagne (1987).

When asked about guidelines for creating items for the test bank, four instructors used the advice of program writers and two used guidelines from Gronlund (1988)—essentially the aforementioned draft guidelines. Most instructors were aware that professional development materials on the writing of test items were available in the NCC library. Four instructors said they regularly modify or revise test bank items, although this may only be one or two items a week. Five instructors reported that they rarely or never modified or revised tests, and those who did revise items acknowledged that it was done usually as a result of student complaints or required curriculum changes. It appears that no a priori test development has been systematically attempted at NCC since the assessment system began operations. The instructors' responses to the interview reveal that they lacked ownership, or a sense of professional responsibility for improving or maintaining NCC's testing or evaluation system; instead they regarded it as an instrument of the college's administration.

Discussion

The NCC testing system is, technically speaking, very simple with a relatively small item bank of approximately 73,000 test items subdivided into item pools for each of the approximately 3,500 curricular competencies. To generate a test the computer uses a simple, stratified random selection procedure. Test items are not currently generated according to carefully developed test specifications, and there is no means of specifying the selection of items on the basis of difficulty or using test results to prescribe remedial instruction. Only a first-generation computer assessment system has been achieved, and that without regard for item quality and for test reliability and validity.

The system is essentially a tool to reduce the clerical effort required to compile and score tests efficiently. By reducing the traditional labor intensive operations of evaluation, individualized CBE is made feasible (that is, affordable in terms of institutional resources). The system also performs very well in maintaining student progress records. Thus, from an administrative point of view, it is a good management tool. From an instructional perspective however, the system is simplistic and plays little or no role in program quality control and instructional improvement; it functions more as a coercive tool for monitoring student compliance, fee payment, and timely progress than it functions as an instrument for motivating and enhancing individual learning performance.

The existing minimum-competency testing and the failure of tests to address higher level learning outcomes is likely to have the effects of promoting educational mediocrity, poor decision-making skills, and poorly motivated future employees. Assessment is always a compromise between the desirable (likely requiring extensive resources) and the practical (requiring limited resources). NCC's system favors the practical. There are no empirical indicators of instructional quality linked to the NCC evaluation system. Although the computer can quickly inform a student whether a test has been passed or failed, no feedback on incorrect answers, guidance for further study, or activities to correct learning deficiencies are provided by the current system. Students go to instructors for this information, which is an inefficient use of instructor time and reduces the time available for other instructional activities.

The self-tests in the instructional modules should serve to help students determine their readiness for the computer tests. However, they do not seem to be effective in doing this for two reasons: (a) the self-tests are no better developed than the computer tests and thus are of limited usefulness, and (b) the students have not, in spite of attending an orientation program, learned how to function well as independent learners.

Overall the data from this study supports Hsu and Sadock's (1985) and Bunderson, Olsen, and Greenberg's (1990) assessments, that computer-managed testing systems have achieved little more than efficient administrative functions for item banking and test scoring and that a major restructuring of current computer-based assessment efforts is required. At NCC the potential for computer-managed testing as a means of improving the quality of training and instruction remains unrealized. Further, the potential of computer-managed testing for contributing to the development of a mastery assessment system wherein mastery is not defined as a minimal competency but includes persónal learning goals and indicators of excellence (Forehand & Bunderson, 1987) remains unrecognized by faculty and the administration. The NCC system, when it was implemented, approached but failed to achieve state-of-the-art capability and has now functioned without any technical upgrade for 10 years. The model has not been adequately sustained or refined to incorporate additional media applications for student assessment with, for example, questions presented not only in text, but supported with images, video, and simulations. Today the system is obsolete and hinders the college's introduction of program improvement strategies.

Implications for Program Improvement

The NCC student assessment system of the 1990s fails to adequately meet the evaluation and testing system goals first outlined in the literature as early as the 1970s and now is defined as a first generation system model. NCC's experience may well be similar to that of other adult and continuing education colleges which implemented computer managed evaluation and testing systems in the 1970s and 1980s. To establish a true state-of-the-art computer managed assessment system at NCC and similar institutions it may be necessary:

1. To examine all items in the computer test bank and rewrite those which have technical flaws;

2. To increase the number of items in the item bank to achieve a ratio of items available for selection to items drawn for any test to at least 10:1;

3. To increase greatly the proportion of items that test higher levels of knowledge;

4. To implement an item selection design based on comprehensive test specifications;

5. To implement adaptive testing procedures (second generation assessment) which allows test items to be selected based on the learner's response to prior items; ideally, feedback would also be provided to students who fail a test by identifying remedial instructional activities, alerting individual

students to areas in which they are weak, and suggesting what they could do for further study; this may require upgrading institutional CBE instructional design systems to provide alternative methods of instruction;

6. To use a combination of computer and multimedia technology to expand the methods of assessment and reduce the reliance on texts and locally written modules in instructional programs;

7. To establish professional development programs to enable instructors to develop knowledge and abilities in test development and to decentralize the use of computer assessment to the department and program level;

8. To transfer responsibility and resources for maintaining and improving the computer-managed evaluation system from administrative to instructional staff;

9. To revise student orientation programs in order to enable students to achieve higher levels of independent learning and to establish incentive programs to eliminate the attitude that the goal is to "beat" the computer; and

10. To reorient the role of the computer-managed evaluation system from monitoring and surveillance of students' progress to the improvement of program and instructional quality.

Conclusions

The existing NCC computer managed student assessment system is, in our opinion, now limiting the college's performance because the system's test item bank is inadequate, the tests generated address low-level learning outcomes, and the location of the system as a function of the college's centralized administration promotes an institutional culture whereby instructional staff have little sense of ownership of the system and neglect to use it for program and instructional quality-improvement purposes.

The claims made by NCC and similar institutions to state-of-the-art learning assessment are unfounded when they are based solely on having a centralized, test-item data bank and the capacity of only a first generation computer assessment model. If Canadian adult and continuing education institutions which currently utilize CBE and computer managed assessment are to achieve the qualitative improvements in graduate capabilities that are being publicly espoused as essential to ensure Canada's place as an efficient player in the "new" economy, they will need to focus on strategies to improve instructional effectiveness. These institutions need to review their use of existing computer managed student assessment procedures and to introduce second generation, at the minimum. Preferably third generation models that have the capacity to select items at appropriate levels of difficulty for individual learners and which can embed assessment in the curriculum should be chosen. This choice can provide continuous, unobtrusive, learning-supportive feedback to the learner as well as data on instructional effectiveness to the instructor.

References

Baker, F. B. (1974). Systems considerations. In G. Lippey (Ed.), *Computer-assisted test construction* (1st ed., pp. 203-230). Englewood Cliffs, NJ: Educational Technology.

Blank, W. E. (1982). *Handbook for developing competency-based training programs*. Englewood Cliffs, NJ: Prentice-Hall.

Bloom, B. S., Engelhart, M. D., Hill, W. H., Furst, E. J., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives. Handbook 1: Cognitive domain* (1st ed.). New York: David McKay.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1988). The four generations of computerized educational measurement. In R. L. Linn (Ed.) *Educational measurement* (3rd. ed., pp. 367-407), New York: Macmillan.

Bunderson, C. V., Olsen, J. B., & Greenberg, A. (1990). *Computers in educational assessment: An opportunity to restructure educational practice.* (ERIC Document Reproduction Service No. ED 340 771)

Carlson, R. D. (1994). Computer adaptive testing: A shift in the evaluation paradigm. *Journal of Educational Technology Systems, 22*(3), 213-224.

Choppin, B. H. L. (1985). Principles of item banking. *Evaluation in Education: An International Review Series, 9*(1), 87-90.

Dunkleberger, G., & Heikkinen, H. (1983). Computer-made tests set mastery learning free. *Curriculum Review, 22*(5), 34-35.

Feuer, D. (1986). Computerized testing: A revolution in the making. *Training, 23*(5), 80-82, 84-86.

Forehand, C. A., & Bunderson, C. V. (1987). *Mastery assessment systems and educational objectives.* Princeton, NJ: Educational Testing Service.

Gagne, R. M. (1985). *The conditions of learning and theory of instruction.* New York: Holt, Rinehart & Winston

Gagne, R. M. (Ed.). (1987). *Instructional technology: Foundations.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gay, L. R. (1980). *Educational evaluation and measurement.* Columbus, OH: Charles E. Merrill.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive testing. *Journal of Educational Measurement, 21*(4), 347-360.

Gronlund, N. E. (1988). *Constructing achievement tests* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Guskey, T. R. (1985). *Implementing mastery learning*. Belmont, CA: Wadsworth.

Hsu, T. C., & Sadock, S. F. (1985). *Computer assisted test construction: The state of the art.* Princeton, NJ: ERIC Clearinghouse on Tests, Measurement and Evaluation. (ERIC Document Reproduction Service No. ED 272 515).

Huck, S. W., Cormier, W. H., & Bounds, W. G. (1996). *Reading statistics and research* (2nd ed.). New York: Harper & Row.

Jackson, N. S. (1988). Competence, curriculum and control. *Journal of Educational Thought, 22*(2a), 247-258.

Jones, A., & Whittaker, P. (1975). *Testing industrial skills.* Toronto: John Wiley & Sons.

Lippey, G. (Ed.). (1974) *Computer-assisted test construction* (1st ed.). Englewood Cliffs, NJ: Educational Technology.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Mazzeo, J., & Harvey, A. L. (1988). *The equivalency of scores from automated and conventional versions of educational and psychological tests: A review of the literature* (Publication No. ETS RR 88-21). Princeton, NJ: Educational Testing Service.

Nitko, A. J. (1996). *Educational tests and measurement: An introduction* (2nd ed.). New York: Harcourt Brace Jovanovich.

Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Sandals, L. H. (1992). An overview of the uses of computer based assessment and diagnosis. *Canadian Journal of Educational Communications, 21*(1), 67-78.