An Overview of the Uses of Computer-Based Assessment and Diagnosis

Lauran H. Sandals

Abstract: This paper presents an overview of the applications of computer based assessment and diagnosis for both educational and psychologlcal placement and Interventions. The paper includes a review and brlef history of computer testing and the antecedents that led to the current acceptance of this medlum as an assessment tool. A rationale for the use of Computer Based Assessment (CBA) and its potentlal advantages in relationship to our current testing practice is also dlscussed. The four generations of (CBA) are presented with a discussion of the strengths and weaknesses of each stage and concludes with some of the issues regarding the construct validity of computer based assessment Instruments vis a vls conventlonal testing practlce.

Résumé: Cet article présente une vue d'ensemble des applications diagnostiques et évaluatives basées sur l'informatique pour le placement et l'intervention pédagogique et psychologique. Après un bref historique du testing par ordinateur, l'article discute des raisons qui ont amené les intervenants à utiliser l'ordinateur en tant qu'outil d'évaluation. tes bases théoriques sur lesquellessont basées l'utilisation du "Computer Based Assessment" (CBA) sont présentées ainsi que les avantages que l'on en retire dans la pratique. De plus, quatre générations de CBA sont discutées au regard de leurs qualités et leurs faiblesses. Enfln, en conclusion, la validité du construit des Instruments d'évaluation informatique est considerée et comparée aux pratiques de testlng conventionnelles.

INTRODUCTION

Testing has been with us since the beginning of recorded history. The Chinese used formal assessment procedures by 1115 years B.C. (Dubois, 1970) in deciding which individuals should be assigned different positions in the Chinese civil service. Throughout time scientists, psychologista, educational diagnosticians, and teachers have looked for better ways than their own feelings to assess an individual's potential in order to provide better educational interventions or treatment programmes. Today a student's ability to enter post-secondary training programmes or different career paths is often determined by national, provincial and state-wide examinations that assess

Canadian Journal of Educational Communication, VOL. 21. NO. 1, PAGES 67 78. ISSN 0710-4340

and rank students on their knowledge of a variety of subjects that are reported to be necessary for post-secondary success. Thus the wide acceptance of tests by society in general has brought us to a time where norm or criterion referenced tests are used: a) to diagnose learning needs; b) to determine eligibility for special programmes; c) to formatively monitor progress; d) to summatively assess student achievement; and e) to assess a student's personality In the past most of these tests were administrated individually by a trained psychologist or educational diagnostician who presented many of the questions verbally or by demonstrating an individual task which the examinee had to replicate or modify or in large group paper and pencil formats with printed booklets.

Society has always tried to improve on the efficiency of such assessment tasks, however, in the 1930's Pressey developed an early testing machine which Skinner revamped in the 1950's into an early commercial success with his original teaching machines that were to test students. These evolved into the earliest instructional-based teaching machines through the use of linear programming techniques. Our continuing acceptance regarding the application of technology and machines in order to lessen an individual's workload has lead through history to the development and use of such things as gears, tractors, and assembly line robots to carry out many tasks that were originally carried out totally by human brawn and brain power. Since becoming an accepted tool in universities and colleges in the mid-1960's computers have become the focus of research in prototype systems that could make use of the computer as an assessment tool that would free the educator or psychologist from certain aspects of the testing environment that could be done as well as or better by a machine. This would leave the psychologist or educator free to work on an individual basis with the client or student in ways which a computer could not. The major limitation regarding this increased use of technology as an assessment tool usually centred on the costs of the machine and the limitations of the programming languages in addition to the problems with either highly graphic material or the need for verbal instruction. However, rapidly emerging technologies are now taking the computer from the research labs and prototype case situations to schools. Many highly optimistic projections for computers in the early 1970's (Knights, Richardson & McNarry, 1973) and the 1980's (Colbourn & Mcleod, 1983) for their widespread use in assessment and diagnosis by the mid to late 1980's will now actually take place in the mid-1990's.

Thus the onset of these technological enhancements and their related psychometric capabilities have now brought us to a point in time where there are some wide uses of certain computer-based assessment and diagnostic packages by the psychology profession. These developments are just being introduced to the education profession at large with specific applications being targeted towards Special Education.

Rational for Computer Based Assessment

Almost all measurements of human performance have to come to grips with the concept of error in assessment. In most instruments there is some variability that is unknown or unpredicted. There are errors such as the different meanings that different individuals make on the interpretation of the same word or phrase. There is human error in the scoring or interpretation of grouped or individual tests. Thus in current assessment practice either through better standardized test procedures, item analysis or statistical tests we are constantly trying to reduce the amount of error one produces in making predictions based on test instruments. In 1985 Poteet and Eaves edited a special issue of ***Diagnostique*** entitled "Perspectives in Special Education Assessment." In Table 1 the author has presented a summary of their 10 major concerns regarding common errors in current assessment practice. Some of these concerns relate to such practical issues as who makes the assessment decision about which instruments are used in the school division. Other issues involve such things as human error in the administration and/or scoring of the test. Many of the issues raised regarding error are more related to common sense. The use of a computer administered version of the same test could

TABLE 1
*Common Errors in Current Assessment Practice**

1. Instruments to be used in the assessment process are often stipulated by administrators of the School system.
2. Educational diagnosticians regularly use instruments for purposes other than those for which they have been validated.
3. Related to Number 2 above is the practice of taking the recommended uses of an instrument at face value.
4. Educational diagnosticians sometimes become caught up in a "drive up window" mentality that leads to the selection of "quick and dirty" instruments.
5. The band wagon effect too often plays a part in instrument selection.
6. During data collection, practitioners can and do commit a number of errors.
7. In Special Education the use of individually administered Instruments is considered the "sine qua non" of assessment practice.
6. Although it seems too elementary to mention, not enough attention is paid to standardized administration rules.
9. Of the mistakes that are made during the use of assessment instruments, perhaps the most common of all is the scoring error.
10. Interpretation of assessment results is considered by many educational diagnosticians to be their most onerous task.

*Note: Adapted from Poteet and Eaves (Eds.). (1964-1965), Perspectives in Special Education Assessment [Special Issue] Diagnostique, 10, 1-4.

possibly compound the error with much more rapidity This may occur due to the fact that a testee may make several mistakes that can not be changed even if they know they are wrong and in the case of an adaptive test the following questions are individualized from prior responses. In addition the diagnostician or psychologist may not review the computer test before continuing on with the computer scoring and possible scale value interpretations and thus report data from a possibly invalid test situation.

Advantages of Computer Based Assessment (CBA)

Many individuals feel that there is a distinct advantage in using the microcomputer as assessment and diagnostic tools for both psychologists, and educational diagnosticians because of the perceived errors in contemporary assessment techniques and the potential overall cost savings. Some of these advantages are adapted and summarized by the author from Poteet and Eaves (1985) in Table 2Aand also by Bunderson, Inouye, and Olsen (1989) in Table 2B (see page 71).

Many of these advantages relate to computers in education in general but many others relate to such issues as item response theory and the practical comparison of test results using paper and pencil administrations vs. the computer vis a vis comparative scores, time on task, cost justification and human time.

These advantages are particularly apparent when one looks at the potential use of these computer based tests from a psychologist's perspective, especially for an individual who may be in private practice. These advantages tend to deal with issues that may not be particularly of interest to educators and diagnosticians in the public school system but at the same time they provide a valid rationale for their continuing use as described by Jackson in (1986) for the American Psychological Association (APA) Scientific affairs office on the use of Computer Based Personality Testing (See Table 3, page 72).

Thus the numerous problems and error in current contemporary assessment practices when compared with the advantages of computer based assessment leads one to believe that the future for computer based assessment is assured. The major impediments to this evolutionary continuum of developments in (CBA) is only limited by (a) the costs of hardware and software, adequate research expertise in the development of these instruments, and (c) the training and professional development of psychologists and educational diagnosticians in the availability and effective use of the (CBA) instruments. The next section will overview the four generations of (CBA) and the relevant issues regarding the construct validity of these automated assessments.

TABLE 2A
*Advantages of Microcomputers and Item Response Theory**

1. They nearly eliminate error in deriving scores.
2. They reduce scoring time by up to 70% or 80%.
3. They provide a simple mechanism for storing and retrieving valuable information.
4. They have intrinsic motivation for the testee.
5. They have the ability to provide immediate feedback to the examinee.
6. They have the speed to handle the evaluation of tests and their items (reliability, item difficulty, biserial cor relations, etc.)
7. They have the ease to store data and to retrieve it when it has to be recalled.
8. They have the capabilities to detect aberrant response patterns,
9. They have the capabilities to provide ongoing group analysis of the test and item bias.
10. They have the capability to evaluate translations of measurement scales to different languages,
11. They have the capability to tailor the test to individual needs.

*(Eaves, 1984-l985, pp. 28-30)

TABLE 2B
Advantages of Computerized Tests and Computerized Adaptive Tests over Paper Based Testing

1. They have enhanced control in presenting item displays. Greater standardization of test administration.
2. They offer improved test security.
3. They can enrich display information.
4. They can provide equivalent scores with reduced testing time.
5. They can improve the obtaining and coding of responses.
6. They can reduce measurement error.
7. They have the ability to measure response latencies for items and components.
8. They provide improved scoring and reporting.
9. They can be automated for individually administered tests.
10. They can obtain records at a central site.
11. They have the ability to construct tests and create items by computer.
12. They have immediate test scoring and feedback.
13. They can provide an increased variety of testing formats. different languages.

**(Bunderson, Inouye & Olsen 1989)

AN OVERVIEW OF THE FOUR GENERATIONS OF COMPUTER BASED ASSESSMENT

In 1990 Bunderson, Inouye and Olsen presented a definitive chapter on the Four Generations of computerized educational measurement. In this part of the paper a brief summary of the major themes of each of these four generations or stages will be presented in order to provide some continuum of the events that have influenced contemporary computer based assessment strategies. The four generations are: 1) Computer Testing (CT); 2) Computer-Adaptive Testing (CAT); 3) Continuous Measurement (CM); and 4) Intelligent Measurement

TABLE 3
*Advantages of Computerized Testing for Personality Testing**

1. It is quite economical particularly in the saving of expensive professional time.
2. Training technical assistants to supervise administration permits considerable savings.
3. The reduction of time between administration and interpretation speeds up feedback to the patient.
4. Virtually all clerical errors are eliminated.
5. There is a considerable gain in reliability of interpretation by using pre-set rules consistently.
6. There is considerable potential for the systematic gathering of normative information as data recording is cheap and accurate.
7. Complex (i.e., non-linear) scoring procedures are much more feasible in the computer environment.
6. Proper human factors concerns will permit a move to special populations some of which are unserviced by the testing field.

* (Jackson, 1986)

Computer Testing (CT). This is where an existing paper, pencil or other conventional tests are transferred to the computer mainly for the technological advantages of the computer but with the original test and
ing almost identical to the non-computer version. Many research studies have been carried out contrasting the equivalence of paper and pencil vs. computerized tests and these are presented in detail by Bunderson, Inouye and Olsen (1989). Suffice it to say that one variable addressed the issue of the type of test (such as Free Response tests, computerized personality tests, aptitude tests, achievement tests, coding skills tests, graphics tests, multiple page tests) vis a vis research results that presented data in three categories (computer tests scores higher than paper administrated, computer teats scores lower than paper administrated and no significant differences between (CT) and paper and pencil tests). The main characteristics of this type of system are computer controlled administration; rapid scoring and reporting, new display and response types; mass storage for displays and item banks; network communications and the utilization of classical test theory.

Computer-Adaptive Testing (CAT). In this situation the major characteristics are all of those in (CT), however, there is a process of adaption throughout the administration of the test. In this computer environment the computer continually checks the testee's responses in order to adapt the presentation of the next item based on the preceding response, or series of responses, or overall response patterns of prior groups of responses. The computer uses floating point arithmetic and high speed processors in order to calibrate all the parameters in making the selection of the next item or group of items. The adaption can take one or more of three possible examples (adapting item

presentation, adapting item presentation times and adapting the content or composition of the item and subsequently adapting the overall test length based on the prior adaptions). It should be noted that the test lengths may be longer but in many cases the (CAT) may present a shorter test if the program assumes the testee either has mastery of a particular set of concepts through a high percentage of correct responses early on in the interaction, or if the testee receives a high percentage of failures early on in the presentation of items. In general the characteristics of (CAT) include all of those in (CT) plus fast floating point calculations for adaptive algorithms that have its theoretical psychometric routes in the field of item response theory and the computer systems that provide item test banks for a multitude ofscience and mathematics tests.

Continuous Measurement (CM). In (CM) the tests use a form of continuous measurement that is embedded in the curriculum in order to measure the changes in the students knowledge and thus to alter instructional interactions accordingly. Measurements include an item, clusters of items, and other exercises and related independent work either on or off the computer. These systems are usually used in what has been typically termed as a "mastery learning" environment where criterion referenced tests are indexed to an individual's educational or behavioural objectives. The curriculum within this type of assessment and measurement usually includes: 1) a course ofobjectives laid out to help the learner attain certain educational goals; and 2) a way of charting an individual's growth through the system either with or without the computer, but more than likely analogous to the previously defined computer managed learning (CML) strategies. The general characteristics of this system includes all of those in (CT) and (CAT) plus the features found in a criterion referenced, computer managed mastery system. The psychometric characteristics includes those of (CAT) and item response theory in addition to clearly stated objectives and the presentation of learning profiles in making computer based assessment decisions. It should be noted that in the area of special education much of the literature on (CM) is reported as Curriculum Based Measurement and the bulk of the research at the elementary and secondary levels has been carried out and reported by Fuchs & Fuchs (1986,1987,1988, 1989).

Intelligent Measurement (IM). Intelligent measurement makes use of most of the general concepts that are presented in CT, CAT, and CM with the significant addition of "knowledge based capability". This type of test is most likely using "artificial intelligence" based concepts in the development of a diagnostic/assessment system that some individuals term expert systems.

Thus (IM) systems are basically the computer based assessments most researchers were hoping that would evolve over the last 25 years of research since we were trying to provide a computer system that could diagnose and assess many educational and psychological concepts as reliably as trained educational diagnosticians and psychologists. One of the biggest differences between (IM) and the preceding three generations is that many different inter-

pretations can be analyzed of a response or series of responses well past that of simple (CAT) measures, Some of these measures have to factor in a summative knowledge base built on the intuitive and subjective experiences of hundreds of educational diagnosticians and psychologists who make everyday use of the manually prepared version of the assessment instrument. Usually the (IM) system will provide the professional with the ability to: a) score complex responses or a series of items; b) to provide interpretations including narrative ones based on a student's or client's profiles on one or more tests; and c) provide advice on either the educational or psychological interventions which the teacher or psychologist may or may not agree with. Thus in general (IM) provides all of the features of the preceding three generations plus knowledge based expert systems. Within (IM) the system uses the knowledge of a number of experts for the scoring, profile interpretations, teaching expertise, possible psychological interventions plus the vast knowledge base of similarly assessed individuals who may be at the same stage in their educational or psychological development.

Thus these four generations of computers have progressed to the point where one supersedes the others. Many important contemporary research projects and relevant commercial projects use one or all four of the previously discussed systems either (CT), (CAT), (CM) or (IM). Because of certain limitations (CT) may be more than adequate in assessing certain achievement skills in a formative setting in education while for another individual (IM) may be necessary for the presentation, scoring and interpretation of a psychologically based personality test.

Whenever an educator or psychologist tries to develop a new form of an old test or to modify an existing one the issue of test reliability and validity comes into question.

Many of the issues regarding the equivalence and comparative nature of the conventional and computer based forms and the generalizability of the results have been addressed by (Greaud and Green, 1986) and (Olsen, Maynes, Slawson and Ho, 1989). In an article "Psychoeducational Testing and the Personal Computer" (Fifield, 1989) presents a strong case for a critical review of either modified or new computer based tests in the area of Technical adequacy under the topics of: 1) reliability; 2) fidelity of administration; 3) alternate forms reliability ; 4) validity; 5) concurrent validity; 6) content validity; 7) external validity; and 8) social validity He makes a strong case regarding the changing role these reliability and validity techniques have in (CBA) and that we have to reconsider how these measures can be applied or even generalized in comparison to our conventional instruments and test procedures. The next part of this paper will discuss the area where the greatest possible changes occur namely in the area of the tests construct validity.

CONSTRUCT VALIDITY OF COMPUTER BASED TESTS

One of the major issues in the field of computer based testing and assessment has to deal with the issue of does a conventional test change when it is reformatted for a computer based presentation even if all of the items and the test itself appear to be identical. One researcher (Green, 1988) addressed these issues primarily in his interpretation whether the construct validity of the test changed from a conventional paper pencil administration to one where it is administered totally on the computer. Some of the main issues addressed had to deal with the following characteristics which may affect the construct validity of the computer based administration. They are: a) Passive omitting b) Back tracking; c) Screen capacity; d) Graphics; e) Responding; f) Time limits: and g) Adaptive tests and related dimensionality. If even one on the topics to be addressed changes when a test is administered with a computer then the tests' prior norms and validity may have to be re- established in its new format.

Passive omitting. On a paper pencil test a respondent can pass on one or two items (for example items two and three) and then he or she can respond to item four and then item five. In fact a respondent can review the whole test before they go back to start answering and filling in responses to questions. In a computer based test (CBT) this cannot be done unless another choice command or control function keys are provided to allow for a "skip"or "next'item" pass etc. Even if this "skip" and "return" function is allowed it places the examinee in a different mental set and it also requires a breakdown of attention to the task on hand (responding to the cognitive nature of the material being evaluated) to mentally rearranging response patterns through different keyboard manipulations.

Back tracking. This occurs when an item has been previously skipped or passed as in passive omitting above or when a student answers a later question (item 10 for example) and now realises that he or she had made a mistake in a prior item (item 3 for example) and that the answer cannot be changed or can only be changed by further mechanical manipulations of the keyboard and the related user software.

Screen capacity. Prior research by human factor specialists (Sandals 1987) state that approximately 64% of the computer screen should be blank when information is presented in a learning or testing situation. Thus there is a chance that some items such as those that include a lot of reading comprehension may not fit on one screen and actually may take up two or three screens before a response can be made. In the paper version all of the information may be included on an 8-1/2 x 11 page.

Graphics. Some of the same issues raised in C above also relates to the size of the screen. Unless the user is using a screen with high resolution colour graphics (such as super VGA) or digitally stored images or laser discs or CD-ROMS then there will be difficulty in presenting many graphics in the same resolution as the original in the printed test booklet. The technology is available to make the reproduction almost 100% accurate, the limitation is the

related high costs for many educational institutions on affording this sophisticated state of the art hardware and software.

Responding. The response in our computer based test usually consists of pressing a key and in most cases this is faster than transferring an answer to an answer sheet and thus this can cause a difference in scores with highly speeded components on tests that may cause vigilance error in the filling in of the answer sheets as reported by (Sandals, 1970). Thus responding may be faster and more accurate through a keyboard, mouse, a light pen, and also the computer may not accept an incorrect answer if it is not in a proper field and thus, as a consequence feedback is given. However, feedback for a misplaced response cannot be provided in a paper pencil test. Thus the whole process of responding may affect the overall test score and the construct validity especially in a speeded test.

Time limits. In most grouped test situations a time limit is given in order to allow a teacher or tester control of the testing situation for the norm of the group. However, in the case of the computer the question is raised whether the customary time limits should be abandoned unless the test has a speeded component which is central to the construct validity of the instrument. Thus the construct validity may change if the computer administration does not have the time limit of the paper pencil version.

Adaptive testing and dimensionality. The major construct validity problem with computer adapted tests (CAT) is that the computer constantly changes the test and the item selections based on the prior response or the prior group of responses. Thus passive omitting is not possible, neither is back tracking or the changing of a prior response. In (CAT) a test item or pattern cannot be changed once an item response has been made or skipped. In addition, the dimensionality and content validity may change since usually no two students get the same test. Thus the usual criterion referenced test decisions can be made but norm referenced comparisons become impossible to report with traditional reporting methods. Two students may go through the exam with one taking only half as many items as another with the items that are actually the same being in the 20% range. The usual interpretation of test results may change since a direct item to item comparison may not take place only the domaincanbecrossvalidated. It may become more difficult to compare student performance when the domain being tested is in Language Arts and Social Studies in comparison to Math and Science where the concepts are more hierarchical and well defined. Thus once a conventional test is placed in a (CAT) mode the construct validity may change significantly as does the content validity and probably most of the reliability of the original test.

Thus these construct validity issues really question whether the computer administration of a conventional test is really measuring the same the original and if not, new norms have to be provided in addition to the new interpretation of the results from the (CBT or CAT).

CONCLUSIONS

This paper has presented an overview of the role computer based assessment and diagnosis has played in both educational and psychological environments. Many who made great predictions in the early 1970s for computer based testing were over optimistic on both the acceptance, funds, research and availability of the hardware and software for the mid 1980s. It is only now that we are seeing the reduction in costs and the research in psychometric theory and expert systems that are needed to make wide ranging applications of computer based assessment and diagnosis a reality. The advent of interactive cd's and CD-ROM's are now going to allow us to provide verbal instructions and graphics and pictures that provide a realistic alternative to conventional individualized assessment instruments. Again the whole issue of the use and misuse of computers in education will come into play if those in power make some of the same mistakes that computers educators did from 19751983. In addition, if the concerns outlined in Table 1 are readdressed then there are many potential benefits for society in general in the use of computer based assessment and diagnosis.

REFERENCES

Bunderson, C., Inouye, D., & Olsen, J. (1989). The four generations of computerized education measurement. In Robert Linn (Ed.), Educational measurement, (3rd edition), 367-407. NY: American Council of Education/ MacMillan.

Colbourn,M. J., &McLeod, J. (1983). The potential and feasibility of computer-guided educational diagnosis. In R.E.A. Mason (Ed.), ***Information processing,*** 83. North-Holland: Elsevier Science Publishers.

Dubois, PH. (1970). A history of psychological testing. Boston: Allyn & Bacon.

Eaves, C. R. (1984-1985). Educational assessment in the United States, ***Diagnostique 10, 5-39.***

Fifield, M. B. (1989) Psychoeducational testing and the personal computer. ***Journal of Special Education Technology, 9(3),*** 136-143.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation on student achievement:Ameta-analysis. ***Exceptional Children, 53,199-208.***

Fuchs, L. S. (1987). Curriculum-basedmeasurementfor instructional program development. ***Teaching Exceptional Children,*** 20(l), 42-44.

Fuchs, L. S. (1988). Developing computer-managed instruction on teacher's implementation of systematic monitoring programs and student achievement. ***Journal of Educational Research, 81,294-304.***

Fuchs, L. S., & Fuchs, D. (1989). Enhancing curriculum-based measurement through computer applications: Review of research and practice. ***School Psychology Review, 18, 318-327.***

Greaud, V., & Green, B.F. (1986) Equivalence of conventional and computer presentation of speed teats. ***Applied Psychological Measurement, 10, 23-24.***

Green, B. F. (1988). Construct validity of computer-based tests. In H. Wainer, & H.I. Braun (Eds.), ***Test validity.*** Hillsdale, N.J.: Erlbaum.

Jackson, D. N. (1986). ***Computer based personality testing.*** Washington D.C.: APA.

Knights, R. M., Richardson, D. H., & McNarry, L. D. (1973) Automated vs clinical administration of the Peabody Picture Vocabulary Test and the Coloured Progressive Matrices. ***American Board of Mental Deficiency, 78,*** 223-225.

Olsen, J. B., Maynes, D. D., Slawson, D., & Ho, K. (1989). Comparisons of paper-administered, computer-administered and computerized adaptive achievement tests. ***Journal of Educational Computing Research, 5(3),*** 311-326.

Poteet, J. A., & Eaves, R. C. (Eds.). (1984-1985). Monograph: Perspectives in special education assessment [Special Issue]. ***Diugnostique,*** 10(1-4).

Sandals, L. H. (1970). ***Vigilance errors on a search examination.*** Unpublished masters thesis, Xavier University, Cincinnati.

Sandals, L. H. (1987). The role of screen design graphics, colour and sound in computer based learning. How much is too much or too little. ***Proceedings of the International Conference on Computer Assisted Learning in Post Secondary Education.*** (pgs. 189-196). Calgary, Alberta, Canada: University of Calgary.

AUTHOR

Lauran H. Sandals is a Professor in the Department of Educational Psychology at the University of Calgary, Education Tower, Calgary, Alberta T2N lN4.