International Journal of Education and Development using Information and Communication Technology (IJEDICT), 2006, Vol. 2, Issue 4, pp. 44-66.

Applications of computer-aided assessment in the diagnosis of science learning and teaching

J. W. F. Muwanga-Zake
University of KwaZulu Natal, Durban, South Africa

ABSTRACT

This paper reports on a qualitative evaluation using questionnaires and interviews in South African Grade 10 classes on the diagnostic value of Computer-Aided Assessment (CAA). A two-stage evaluation was necessary: the first stage involved validation of diagnostic test items; and the second stage evaluated the diagnostic value of data that CAA produces. While results confirmed earlier findings about the advantages of CAA, the diagnostic and remediation potential of CAA data depended upon teachers' capacity to set diagnostic test items particularly in a multiple-choice format, teachers' ability to interpret data produced by CAA, teachers' skills in remedying their classroom as well as learners' problems, the quality of the test items, and the learning as well as the teaching strategies.

Key words: *computer-aided assessment; diagnostic assessment*

INTRODUCTION

Context

This evaluation was conducted in East London, South Africa (SA) during 1998 and 2001, with a science teacher in each of 2 schools, and Grade 10 learners. The number of computers at each school (13 computers in one school and 14 computers in the other school) limited the number of learners who participated to 27. School 1 was private and focussed on science, mathematics and technology, and School 2 was a government formerly whites-only school with a technical-based curriculum. Learners were of mixed races but predominantly used English as a Second Language. These two schools were selected because their curriculum was predominantly science-based, and had over 40 Grade 10 science learners per class. The schools were in the City of East London, and so easily accessible physically and by telephone.

The problems

The poor performance in science in South African schools is a matter of concern (E.g., Ogunniyi, 1997, 2000; Manzini, 2000; McComas, Clough & Almazroa, 1998; Muwanga-Zake, 2000, 2004), but computer technology could help to solve some of the problems in science classrooms. For example, in a survey of 26 senior schools in SA, each of the 42% of teachers surveyed, engaged over 51 learners in a science class (Muwanga-Zake, 2004: 15). Teachers face difficulties in diagnosing learning problems of large numbers of learners (Stiggins, 2001) and CAA could provide a remedy.

Rationale

One of the ways of improving performance in science is to use diagnostic assessment to scrutinise learning difficulties, so that appropriate remedial help and guidance can be provided. CAA has been found to provide data rich enough for diagnosing learners' problems (Bright, 1987).

The rationale of this study was to test the potential of CAA towards easing the labour of diagnostic assessment in classrooms where large numbers of learners per teacher are common, and where reviews of CAA are scarce. Additionally, a new curriculum in SA (the National Curriculum Statement) introduced rubrics as a method of scoring learners' achievements. The rubrics comprise possible or common alternative constructs that learners make, which are too many for paper-pen marking. These alternative constructs are however possible distracters in a Multiple-Choice Question, which are easily set upon CAA.

LITERATURE REVIEW

Assessment

Assessment in the National Curriculum Statement is an integral part of teaching and learning (DoE, 2005: 7), I.e., assessment and therefore, diagnosis appear to be in concert with learning theories (Madaus & Kellaghan, 1992; Gipps, 1996). For example, … *any method used to better understand the current knowledge that a learner possesses* (Dietel, Herman, & Knuth, 1991), is relatively behavioural. Madaus & Kellaghan (1992: 120) gives a similar definition. On the other hand a process-oriented, and therefore possibly constructivist definition is *… a systematic collection of information about learning and other variables associated with particular learning experiences* (Tamir, 1996: 94). In this paper I will take assessment to be a measurement of learning achievements against the objectives of a course.
Behavioural assessment seeks measurable objective knowledge and skills (Hannafin, Kim, & Kim, 2004: 13) under stated rules such as time (Birenbaum, 1996: 6). Each question and answer deals with one simple low cognitive concept (Scott, in association with Dyson & Gater*,* 1987: 19; Fuchs, 1995: 1-2; Gipps, 1996) and the item mark has to correlate positively with the total score, other wise it would be discarded (Gipps, 1996: 252-253). Similar statistics mean the same thing and so, behavioural results can be compared (Hannafin, Kim, & Kim, 2004: 13).

Higher order items fall into the cognitive assessment category, and differ from behavioural items in finding out the quality of an individual's cognitive constructs (i.e. criterion-referenced), under relatively *unregulated conditions* that do *not produce 'well-behaved' statistics;* it aims at improving instead of *sentencing* a learner (Wood as cited in Gipps, 1996: 255). Similar views appear in Jonassen, Howland, Moore, & Marra (2003: 228). Such assessment arises naturally in situations that an individual has experienced (Salviati as cited in Cunningham, 1991: 15), and so suffers invalidity on account of cognitive development in different environments and cultures (Mwamwenda, 1993).

Similarly but more subjectively, constructivists desire a lesson-integrated and contextual assessment that measures a learner's processes, interpretations and constructs (Hannafin *et al.,* 2004: 13; The Maricopa Center for Learning & Instruction, 2000; Pachler & Byrom, 1999: 126; Birenbaum, 1996: 6 – 7), for example, from practical tasks (Ryan & DeMark, 2002: 67). Think-aloud protocols or dialogue with the learner, asking a learner what s/he is doing and why, could be used. Jonassen *et al.* (2003: 230-232) recommends a rubric, which *takes the form of* learner-teacher well-defined negotiated *sets of scales* of expected performances for constructivist assessment.

The question of motivation

Assessment should not be intimidating and should rather motivate learners (Stiggins, 2002). However, there are fears of learner de-motivation when it is informal (Hickey, Kindfield, Horwitz, & Christie, 2003: 529).

Diagnostic assessment

There are a wide variety of definitions of diagnostic assessment, possibly as many as the number of definitions of learning, although most of the definitions concern identifying learning potential and difficulties (E.g., Lawton & Gordon, 1996: 88; Tuckman as cited in Fraser, 1991: 5; Bright, 1987: 71; Fuchs, 1995: 1). Additionally, Hein & Lee (2000: 3) and Bright (1987: 83) suggest conceptual pre and post diagnosis when planning a lesson. Such diagnosis identifies and analyses abilities, difficulties, and incorrect conceptions, and then provide appropriate remediation to learners (e.g., Hein & Lee, 2000: 3; Linn, 2002: 40; Harlen, 2000; Little & Wolf, 1996: xi). I however, derive a definition for this paper of diagnostic assessment from Black & Dockrell (as cited in Fraser, 1991: 5) and the Maricopa Center for Learning & Instruction (2000): Diagnostic assessment is a means by which a teacher and a learner iteratively and mutually agree to collaborate in monitoring a learner's conceptions and in deciding upon subsequent remediation. The definition takes recognition of Pollitt's (1990: 879) advice of integrating diagnostic assessment with teaching and attempts to ensure that gaps in teaching and learning are filled.

Amir & Tamir (1994: 94) suggest that *descriptions of* popular *misconceptions are very important starting points* during diagnosis, and such descriptions could derive from *interviewing a small number of students.* However, ensuring that standards and conceptions are appropriate and that the conceptions are tested at an acceptable standard and in prescribed ways entail validating the test items. Items must focus on just one aspect of learning at a time and should have a high degree of sensitivity with regard to the variety of possible misconceptions or alternative conceptions. Furthermore, a number of questions on a single aspect of learning could be used to pinpoint misconceptions and isolate the most popular ones, while at the same time ascertaining that a ‘wrong’ or 'right' answer is not by chance through guessing, errors (consistent pattern of mistakes [Bright, 1987: 72]) or mistakes (incorrect answers [Bright, 1987: 72]), but are respectively due to lack of, or adequate understanding.

Validation of diagnostic items

Since diagnostic assessment assumes that test items are valid, Linn (2002: 27) as well as Ryan & DeMark (2002: 67) believe that validity is the most important consideration in evaluating the quality of the uses, and interpretations of test results. Among many others, I have considered Haladyna's (2002: 94) and Linn's (2002: 28-33) description of validation as a process that evaluates the degree to which theory and evidence support the interpretation or purpose of a test-score. In concert, this paper considers validity to be the extent to which a test accurately measures what the users understand and want it to measure. That is, test items should be of appropriate standards and be presented in a way acceptable to most of the stakeholders concerned about a learning exercise to the extent that the inferences from test results should support the purposes of a test. This requires users to set priorities or pose the most critical validity questions (Linn, 2002: 46), which in this study were limited to:

- **General validity:** accuracy of items and scores; roles of participants; aims for the test (Ryan & DeMark, 2002: 67; Linn, 2002: 27-46; Haladyna, 2002: 94).
- **Construct validity**: the extent to which a test measures an intended characteristic or construct (Gay & Airasian, 2000: 169).
- **Content validity‘ or curricular validity’**: teaching methods; the Outcomes justified & accurately assessed (e.g., content, Grade) (Gay & Airasian, 2000: 163).
- **Face validity (Fairness**): items *appears to measure what it claims* (Gay & Airasian, 2000: 164); not biased; e.g., gender, language and values concerns (DoE, 1998)*; relevant & adequate coverage* (Taiwo, 1995: 7)

- **Technical validity:** refer to the appropriateness of the technology used in presenting the test. For instance, the test is technically valid if it affords the learners freedom of starting with any question.

Computer-Aided Assessment

There are terminologies some times synonymously used with Computer-Aided Assessment. These include: (1) Computer-Assisted Assessment; (2) Computer-Mediated Assessment (CMA); (3) Computer-Based Assessment (CBA); (4) e-assessment; and (4) online assessment, which is simply assessment which requires the use of the internet (Wise & Plake, 1990; Sandals, 1992; Thelwall, 2000; Gretes & Green, 2000; Bojic, 1995). Computer Assisted Assessment is used in similar ways with CMA and refers to any application of computers within the assessment process; these are therefore synonyms with a newer buzz word, e-assessment, where information technology is used for assessment-related activities. The computer plays no part in the actual assessment but merely facilitates the capture and transfer of responses. Computer-Based Assessment refers to assessment via the computer (Bojic, 1995). The computer does all processes such as marking, recording, feedback, and analyses (Gretes & Green, 2000: 47). Thus, Thelwall (2000: 38-39) adds that CBA is used for examinations and diagnostic assessment. Bojic (1995) defines Computer Aided Testing as the use of technology *to manage or support the assessment process.*

This study uses the term Computer-Aided Assessment (CAA) to include the administration, marking, storing, and processing of learners' assessment records, arguing that the term 'aided' signifies the computer's role as an aiding tool besides other tools of assessment such as dialogue between a teacher and a learner.

CAA software

Authors such as Oliver (2000: 2) and the Internet (e.g., Software Reviews) provide links to commercial and free CAA software along with reviews. This evaluation used Question mark (QM) *Designer* 1993 that could run on Windows 3.11 platform and later Window versions. Question Mark has been evaluated (for example, Knight & Brown, 2000: 2, 5;; Gretes & Green, 2000: 47; Thelwall, 2000, 46-47; Croft, Danson, Dawson, & Ward (2001: 53) before but, as with other CAA software results were inconclusive.

CAA software might be the answer to saving time and energy for teachers to make assessment effective and efficient, as well as about enabling learner self-assessment. Most of the CAA software provides immediate self-assessment and feedback opportunities that enable learning or teaching diagnosis (Oliver, 2000: 1; Thelwall, 2000: 40, 45, 46; Gretes & Green, 2000: 46; Croft *et al.*, 2001: 62; Hickey, *et al.*, 2003: 531; The New Zealand Council for Educational Research, 2001). QM, in particular, has two diagnostic facilities: QM Designer provides shells for a range of Multiple Choice Questions (MCQs), including hot spot, word-match, and multiple choices, with options on repeating, showing marks, and the duration and time the test is done; and, QM Reporter, which instantly analyse performance for every learner, class, item, and test, and shows choices each learner makes in a MCQ (Thelwall, 2000: 41), whatever the number of learners.

However, in applying CAA, the evaluator has to watch for computer anxiety or attitudes or lack of computer skills among learners and teachers, and the difficulty to assess higher order thinking skills besides, high cost of computers and of CAA software, as well as computer system quality and maintanance (Thelwall, 2000: 40; Oliver (2000: 2).

Using CAA for diagnosing

Bright (1987: 75) explains that using computers to diagnose is *most appropriate* when a learner *has repeatedly been unable to learn,* such that there is suspicion of *fundamental misunderstanding* that needs to be corrected, presumably by the use of CAA. Bright (1987: 77-79) advises that diagnosis is made easier because CAA can match particular responses with errors that were pre-identified and stored in the program, thus CAA allows remediation using records of the learner's performance.

Multiple-Choice Questions (MCQs)

The definition of MCQ is adopted from the Department of Computer-Based Education, University of Cape Town (2000, section 2.1); a question in which [learners] are asked to select one alternative from a given list of alternatives in response to a 'question stem'.

Why focus on MCQs?

SA uses MCQs extensively in assessing science for their advantages (E.g., the Department of Computer-Based Education, University of Cape Town, 2000: section 2.2; Tamir, 1996: 96) such as:

a. Assessing very large numbers of candidates;
b. Reducing problems due to language, since answers are normally either provided or short;
c. Easier to incorporate into CAA, and to analyse statistically
d. Possibility of dealing with a wider range of topics and cognitive levels in a short time; and
e. Easier and accurate marking, as well as administration.

For example, a science Matric examination can comprise 25% MCQ worth of marks. MCQ comprised a stem with more than one possible answer.

THE STUDY QUESTIONS

The main question of concern was: **What kind of diagnostic information does computer-aided assessment (CAA) provide on learners' current understanding of the science concepts at grade 10 level**? I unpacked this question into the following subsidiary questions:

i. How technically sound is the test?
The test set and presented by use of CAA is in this paper called the Computer-Aided Test (CAT). This excludes CAA uses besides the technical aspect of CAA presenting a test. In this study, technical aspects were limited to screen design, computer capacity, graphics, method of responding, and time limits (Sandals, 1992: 75-76). The CAT was presented to an Instructional Designer expert for technical validation.

ii. What is the quality of data that CAA provides?
The quality of data that the CAA provides was examined using three questions (Table 1).

Table 1: *Questions, methods, and respondents*

Questions	**Primary Source**	**Data collection methods and Respondents**	**Analysis of interview data**
i. What are the learners' results on the CAT?	The CAT	Learners did the CAT; interacted with learners while they did the CAT; and interviewed a focus group of learners afterwards	Constant comparative / some discourse analysis
ii. What information does CAA provide that is useful for diagnosing learners' knowledge?	CAA Reporter	Interviewed the 2 Teachers about CAA reports on learners' performance	
iii. How well do the results indicate the problems that learners have with the topic tested by the CAT?	Analysis of CAT results by CAA Reporter	Interviewed the 2 Teachers about CAA Reporter analyses	

It should be noted that the study had to use validated test items. Therefore, the study comprised two levels of evaluation; validation of test items, and the evaluation of data provided by CAA. These are outlined below in the respective order.

THE FIRST LEVEL OF EVALUATION: VALIDATING DIAGNOSTIC MULTIPLE-CHOICE QUESTIONS (MCQs)

Setting the first test items

Teacher 1 set Test 1. The analysis of the teacher's test, on the basis of Bloom's taxonomy of learning objectives and its content is shown in Table 1 below. Table 1 shows that the test covered a wide variety of concepts; it looked like a revision for the years' work.

Table 2 below indicates that the test comprised mainly recall such as Question 1 and 2 below.

1. *The following statement on electrification is not true:*
   a. *It occurs according to the Law of Conservation of charge;*
   b. *It occurs through friction*
   c. *It occurs when electrons are transferred from one object to another*
   d. *It is the result of creation of some charges and the destruction of others*

2. *The factor that does not influence the resistance of a conductor is _____*
   a. Mass b. Temperature c. Diameter d. Type of material
   Answers: 1) d 2) a

# Note the use of negative terms in both questions.

Table 2: *Test 1 set by the science teachers at East London: Table of specifications (e.g., Taiwo, 1995: 41)*

Cognitive Domain	No. of items		Science topics						
		Marks	Roots	Cell structure	Acids	Electricity	Electrostatics	Decomposition	Force / Pressure
Knowledge	7	14	(9), (10)	(8)	(6)	(2)		(5), (7)	
Comprehension	1	2					(1)		
Application	-	0							
Analysis	2	4							(3), (4)
Synthesis	-	0							
Evaluation	-	0							
Total	**10**	**20**	**2**	**1**	**1**	**1**	**1**	**2**	**2**

# Note: numerals in brackets indicate the numbering of the question

Validation processes

Revising Test 1 items to produce Test 2

The teachers and I improved Test 1 to Test 2. We followed principles (e.g., in Croft *et al.,* 2001: 58) as follows: English was at a level for Black learners; Each item dealt with one clearly stated problem; Short stems; Negative stems were avoided, and highlighted if used; The alternatives were clearly correct with attractive distracters; and clues were avoided. For example, Questions 6 below replaced Questions 1 and 2 in the above. Test 2 also had 15 items.

> ***Select THREE factors, which influence electrical resistance of a conductor from the following****: Diameter, type of material, strength of electrical current, strength of potential difference, temperature, mass*

The class teachers and I validated the test items by doing Test 2, and then I gave Test 2 to the Subject Adviser from the South African Department of Education for further construct and content validation. The test items were accompanied by questionnaire about the validity of Test 2 items. I held an interview with the subject adviser on collecting the questionnaires, specifically focussing on answers the subject adviser provided to the questionnaire. The interview also sought ideas for improving the test items.

Findings by the subject advisor

The subject advisor found Test 2 to cover content well and to be very accurate, with suitable English vocabulary, for 15 – 16 years age range, and for Grades: 8 – 10. He complained of the high proportion of recall items.

Test 2 results

Learners did Test 2. Results from Test 2 indicated more conceptual problems in electricity than in other areas. Therefore, the teachers and I set a diagnostic Test 3 on electricity.

Setting diagnostic items on electricity
We identified learner problems about electricity by checking for errors and mistakes by allowing learners to do many questions about electricity that tested for understanding at different difficulty levels (Tamir, 1996: 96-97, 107) as follows:

- Using misconceptions from Test 2 and processes that learners might have used for constructing distracters, on the basis of *'correct – best answer'* with some *factually correct information*
- Ranking items and giving reasons for the ranking - candidates were provided with possible reasons in the defence of their reasoning and understanding
- Using paired-problem-solving activities, in which for example, a concept is needed before calculations.
- Penalising by awarding a negative mark for ludicrous choices that show complete misunderstanding
- In 'confidence in chosen response'; learners were asked to choose the best answer, state how sure they were, and then to choose the second best answer, etc. Each confidence level was given a different mark. E.g., Correct sure = 2 points; Correct not sure = 1 point; etc.
- Ask learners to give justification for their choice
- Providing data for learners to describe (analyse, and evaluate).

A sample of diagnostic questions, which the teachers and I set in Test 3

Battery A is marked 8 Volts and sends out 2 Coulombs in one second, Battery B is marked 4 Volts and sends out 4 Coulombs in one second, and Battery C is marked 16 Volts and sends out 2 Coulombs in two seconds. From the list below, select FOUR statements that are true.

a. Coulomb in Battery C has more energy than a coulomb from Battery A and Battery B. (3)
b. Battery C is pushing out the lowest current (1)
c. Battery B could be experiencing the lowest resistance (2)
d. Battery B is the smallest in size because it is producing the lowest Volts (0)
e. Battery B will be most powerful after an hour (-1)
f. All batteries produce the same amount energy in one second. (0)
g. The three batteries are producing the same power (4)
h. The current is highest through Battery C because it has the highest Volts (0)

In a higher resistance fewer charge pass through, and so the potential difference across that resistor will have to be low.

a. Strongly agree (-2); b. Agree (-1); c. Disagree (1); d. Strongly disagree (2)
b. Neither agree or disagree (0)

Conclusion on the first level of evaluation

The first level evaluation which involved validation of test items, revealed incapacity of teachers to set diagnostic science test items at Grade 10 level, especially in the MCQ format. For example, Test 1 was short of diagnostic test question, and included negative statements, against recommendations of setting MCQs. The Subject Advisor alluded to the teachers' incapacity in the complaint about a high proportion of memory test items in Test 2. I had to provide guidance to these teachers in setting diagnostic test items in Test 3. Thus, it might be necessary to train teachers further about diagnostic assessment.

THE SECOND LEVEL OF THIS EVALUATION – THE DIAGNOSTIC VALUE OF CAA

Test 2 and Test 3 were used in the CAT – i.e., were typed into CAA and given to learners to do. The second level of evaluation started with learners doing the CAT, by which the values teachers attached to CAA were accessed. Therefore, the rest of this paper is about this second level.

Methodology used in the application of a CAT in two schools

QM documents the procedures to install and use at http://www.qmark.com/. Some of the aspects of technical validity are controlled by QM Designer, which we used to set the CAT with the following options:

- Learners were free to start with any question, or skip questions and could browse through the test back and forth, including the freedom to change answers.
- Captions: The screen showed the user’s name, the name of the test, number of questions attempted so far, and time left.
- Other settings: A time limit of one hour, a user could escape from the test at any time, and answers were automatically saved upon the hard drive.

The teacher responsible for the computer laboratory, the science teacher and I loaded a validated Test 2 on the LAN and learners did the test. We made Test 2 available to learners on a Local Area Network (LAN) as advised in Oliver (2000: 1), Thelwall (2000: 39), and Gretes & Green (2000). Learners 'logged-on’ school computers in a laboratory to do the Computer-Aided Test (CAT). Test 3 on electricity was set on the basis of QM reporter data, which revealed learners weaknesses in electricity in Test 2.

Test 3 appeared high levelled to learners (and to the teachers) that we had to discuss and help learners as they did the test. We also set QM Designer to allow learners to attempt the test any number of times, and to show correct answers and marks after the first attempt. With reports on Test 3 from QM Reporter, I interviewed the teachers about the data, their experiences with the CAT and on how they could use the QM reports. I also interviewed learners after they had finished doing the test for face validity.

Findings and analysis of findings

Technical validity of the CAT
An ID expert found that the CAT was technically valid except for the lack of dialogue boxes to help in case of need, and the fact that QM filled the whole screen without room for accessing other computer activities.

Face validity (Fairness) of the CAT
All participants were asked to check for gender sensitivity in terms of language, design aspects, and contexts used. Teachers and a focus group of learners after they did the CAT approved the CAT.

Data produced by QM reporter
Among the numerous reports that QM reporter produces a test report (summary and list report), item analysis report, and learner report are relevant for diagnosis. I give examples below.

i) An example of a Summary Report

***Summary Report** 2000/10/25; Test name: College; _______; Number of users: ____;*
Score: Maximum - 75 %; Minimum - 33 %; Average: 49 %; Standard deviation: 14.00 %;
Time to complete test: Maximum: - 19:55; Minimum: 0:03; Average: 14:50

ii. An example of a List report

On addition to the information below, QM reporter included the time and duration of doing a test by each learner. QM also kept a record of every attempt each learner made.

***List Report** 2000/10/25*

Score	*Test name*	*User name*	*Date*	*Duration (minutes)*
33 %	*College*	*L1*	*2001/10/25*	*18.33*
33 %	*College*	*L2*	*2001/10/25*	*14.16*
73 %	*College*	*L13*	*2001/10/24*	*10.03*
75 %	*College*	*L14*	*2001/10/25*	*13.43*

QM arranged the report according to a teacher's needs; for example, according to the marks obtained (as in the above), name of test, alphabetical order of the learners, date, and time.

An example of a response item analysis by QM (Test 2 Question 6)

The QM item analysis includes 'facility' and 'discrimination' that are important in setting diagnostic questions:

Facility is the difficulty level of the question, ranging from 0.0 to 1.0, and is calculated as the average score for the question divided by the maximum achievable score. A facility of 0.0 indicates a very hard question (no-one got it right), while a facility of 1.0 shows a very easy question (no-one got it wrong). Questions that have facilities of more than 0.75 or less than 0.25 are less effective in differentiating users. An ideal facility is 0.5.

Discrimination ranges from -1.0 to +1.0, and is a Pearson product-moment correlation between the item and the test score. A correlation close to +1.0 means that the question is measuring the same thing as the test. A low correlation shows that getting the question right is not related to a good test score. A negative correlation means that getting the question right associates with a low-test score. Users who answer correctly questions with a discrimination of less than 0.25 up to -1 do not necessarily do well in the test as a whole.

***Table 3: Test 2 Question 6**: Select THREE factors, which influence electrical resistance from the list below:*
Number of times question answered: 14; Average score: 3.43; Maximum: 6; Minimum: 2; Standard deviation: 1.22

% of learners	*Score /choice*	*Choice*
21%	*4*	*"3 items selected: Diameter, Type of material, The strength of electrical current"*
7%	*6*	*3 items selected: Temperature, Diameter, Type of material"*
29%	*2*	*"3 items selected: Temperature, Mass, The strength of electrical current"*
29%	*4*	*"3 items selected: Temperature, Type of material, The strength of electrical current"*
7%	*2*	*"3 items selected: Type of material, Mass, The strength of electrical current"*
7%	*4*	*"3 items selected: Temperature, Diameter, The strength of electrical current"*

Note that only 7% got this question correct, and that overall the rest (i.e., 93%) believed that 'the strength of current' influenced resistance". This data supported us in focussing upon electricity in Test 3.

Table 4: Test 3 Question 4 - First attempt:
*Multiple Choice - " **The Ohm is the unit of** -----"*
Number of times question answered: 14
Average score: 1.60 Maximum: 2 Minimum : 0
Standard deviation: 0.84 Facility: 0.80 Discrimination: 0.65

% of learners	*Score /choice*	*Choice*
80%	*2*	*Resistance*
10%	*0*	*Charge*
10%	*0*	*Potential difference*

The item analysis also shows the change in choices and improvement as learners repeated the test, e.g:

***Test 3 Question 10 "Siphokazi connects resistors of 4 Ohm and 2 Ohm in parallel, while Olwethu connects resistor of 4 Ohm and 2 Ohm in series. What can you say about Siphokazi's and Olwethu's total resistance?"** Maximum: 3; Minimum: -2*

***Table 5: First attempt**: Average score: 1.20; Standard deviation: 2.04; Facility: 0.40; Discrimination: 0.49*

% of learners	*Score /choice*	*Choice*
20%	*-2*	*"The strength of the resistors will depend upon the current"*
40%	*3*	*"Siphokazi has the lowest resistance"*

***Table 6: Second attempt**: Average score : 1.67; Standard deviation: 2.16; Facility : 0.56; Discrimination: 0.80*

% of learners	***Score /choice***	***Choice***
17%	*-2*	*"The strength of the resistors will depend upon the current "*
67%	*3*	*"Siphokazi has the lowest resistance"*

QM Reporter shows that a lower percentage (17%) of learners chose the wrong answer, while an increased percentage (67%) of learners chose the correct option in the second attempt. The average score improved from 1.20/3.0 = (40%) to 1.67/3.0 = (56%).

In the above, note that learners scored better in a memory Question 4, with an average of 80% (facility = 0.8) (or average mark 1.6 out of a maximum of 2.0) than in a more challenging Question 3 with a score of 40% (Facility = 0.4) (or average mark of 1.2 out of a maximum of 3) in the first attempt.

The learner report

QM Reporter does not show the whole statement of the answers but shows the choices made by each learner, the marks each learner obtained for each question, the total percentage, the maximum mark obtainable for each question, and the overall percentage. For example:

Examples of a Learner Report (from Test 2)	
User name: L1	***Score***
6 Multiple Response "Select THREE factors, which inf"	
Answer given: 3 items selected:	
Temperature, Mass, The strength of electrical current	*2/6*

Teachers' responses to the interview

I have derived three categories of themes below to indicate the sense teachers made of CAA.

i. Diagnostic values teachers attach to CAA

It is remarkable that teachers avoided the word 'diagnosis', even in questions about the diagnostic value of CAA. The diagnostic value teachers attached to CAA was deduced from their statements as shown below.

Table 7: *Diagnostic values teachers attach to CAA*

Theme	Statements
1. Identifying problems	• Tells me the errors and mistakes that learners make; immediate feedback; Can show where the problem is and what kind of mistakes
	• Make sure about the problems; help more needy students
2. Reveals learner's thinking	• Can show how the learner thinks
3. Revision	• For revision

ii. Other values associated with CAA mentioned by the teachers

T2 was more impressed by the self-assessment opportunity that the CAT offered to learners, such that the test can '*be done unsupervised*'. He also noted the possibilities it offered such as testing many learners, while enabling attention to *'only those students who don't understand'*. Responses indicate that the CAT was fun: i.e., the CAT was interesting, exciting to learners, convenient, and made learners relaxed.

Table 8: *Other values associated with CAA mentioned by the teachers*

Theme	Statements
4. Contributes to motivation	• Interesting, students are excited; more relaxed; encourages learners to do the test
5. Eases the assessment of large numbers of learners	• Is helpful – saves paper, marking time, immediate feedback
	• With few teachers, can cover more ground in big classes; takes burden off the teacher; saves time
6. Self-assessment	• Be done unsupervised; Encourages self-learning
7. Helpful	• Convenient

iii. Problems teachers associated with CAA

It can be noted that all the problems were raised by T2, who was teaching in a more advantaged school. For example, he stated that computers were not enough to have effective CAA, and that CAA and computers were expensive. Another inhibiting factor was the need for training. Teacher 2 identified the need for skills to structure questions for diagnosis with particular reference to what a teacher wanted to achieve from the diagnostic exercise, and the need for many questions.

Table 9: *Problems teachers associated with CAA*

Theme	Statements
8. Poor training	• Structure questions; Require more questions; how to set questions; need how to interpret and to use data
9. Economic disadvantage	• Computers not enough; Requires a lot money; Requires finance; Need for more computers and CAA software

Learners' opinions after using the cat – was test 2 fair to learners?
A sample of 14 learners' responses to the fairness of Test 2, which they did using computers appear in Appendix IV. This section provides insight regarding the 'face validity of Test 2. As far as the fairness of the CAT was concerned, the following are themes that learners made (individual learners appear in brackets)

1. Good (L1 - good, L8 – nice, L12 – okay): Fairness = Face validity
2. Interesting (L2, L3, & L4) + Enjoyed (L5) + Encourages study (L10): Motivating
3. Required thinking (L2): Diagnostic
4. Difficult (L1, L3, L6, L8, L11, & L14): Invalid (Content & Curricula)
5. It helps one to revise (L5, L9): Diagnostic
6. Very easy (L9) = 1 This particular learner obtained low marks!

There was no apparent relationship between 'fairness', 'time a learner took to complete the test', and the 'total marks' a learner earned. Furthermore, learners tried to respond in English, which was a 'second language' to them such that their vocabulary was limited. I had to probe for clarity, sometimes using vernacular and a 'discourse' besides a constant comparative analysis for some learners' responses.

Learners agreed with their teachers that CAA helped them to revise, and were excited about using the computer, getting marks immediately, being able to re-do the test, and knowing the time left to do the test. However, the most popular theme indicated that learners found the CAT difficult (six learners). Consider this comment in light of an average time of 14.5 minutes (with a maximum mark of 75%, a standard deviation of 14%, and an average mark of 49%) for a test set to be done in an hour.

DISCUSSION

Diagnostic information provided by CAA (QM Reporter)

Note that other methods of assessment such as a pen-and–paper test can yield data for diagnostic assessment. But, similar to Bright (1987:72), I have experienced strenuously the analysis of results from a diagnostic paper-and-pen test in a Carnegie Project at CASME. CASME had to employ somebody to process just over 200 scripts each with five diagnostic questions for a whole year. Secondly, it is important to differentiate between validity of the questions and the contribution to diagnosis that CAA makes. I am commenting on the later.

QM Reporter provided valid and relevant evidence for interpretations of performance (Linn, 2002: 40), especially with regard to identifying learners' problems in electricity. QM Reporter promptly analysed all learners' performances, and provided data that helped in identifying recurring mistakes and errors for each individual learner (The New Zealand Council for Educational Research, 2001). QM Reporter diagnostic data not normally available from a paper-and-pen test included standard deviation, facility, and discrimination for each question. On addition, QM Reporter indicated the choices that individual learners made, and the number of learners who made a particular choice. However, as one teacher pointed out, it was clear that CAA does not diagnose but only provides data immediately to enhance diagnosis.

The learners' use of CAA to diagnose their problems

Learners realised the usefulness of CAA, mainly because QM provided opportunities for *ipsative* (successive learner performance), and criterion besides norm referencing. Results indicated that learners improved with practice (see item analysis) such that, possibly, the uncontrolled number of testing served recommendations for *regular self-assessment*, and *descriptive* feedback (Stiggins, 2002), which enabled learners to diagnose, and remedy mistakes themselves, through dialogue with or without assistance from the teachers; thus, self-evaluation was a possibility at a pace learners desired (Little & Wolf, 1996: xi; Tamir, 1996: 98-99). Learners also wanted the CAT to reveal answers, which was activated in Test 2. (L2 – '… *answers were not revealed at the end. I wanted to see where I went wrong').* Hence, learners could *watch themselves grow over time and thus feel in charge of their own success,* which they were eager to communicate to their teachers (Stiggins, 2002).

The increase in marks due to repeated item attempts (e.g., of Question 10, Test 3) could have indicated either learning the test or better understanding, not withstanding other factors such as improved familiarity with CAA. Improved performance in subsequent diagnostic tests is more likely to reveal improved understanding, provided that the subsequent tests comprise validated and different items of similar difficulty levels. In this case study, the fact that learners were subsequently able to attempt and score from harder diagnostic items that they had hitherto never done, could have indicated improvements in conceptual understanding, besides the improvements due to learning the test.

Bearing in mind the common enthusiasm learners have to use computers, the freer test atmosphere, instantaneous marking, and the diagnostic nature of test items also supported self-diagnosis and led learners to request for more testing, all of which increased competitiveness and excitement, especially when learners improved. Anything that can motivate learners to self-assessment is desirable and contributes towards self-diagnosis. According to Stiggins (2002) self-assessment *builds confidence* of learners to take responsibility for own learning, and lays foundations for lifelong learning. Similarly, Croft *et al.* (2001: 62) reported positive attitudes towards CAA, and argued that enthusiasm contributed to harder and consistent study*,* and to more knowledge and better performance. However, in this case, one teacher associated learner scores with her own success but failed to mention the possibility that learners could have worked harder.

The capacity of teachers to diagnose using CAA

Teachers alluded to diagnosis indirectly in their answers. For example, T1 said that data could show learner's problems and thoughts (so teachers can focus on learners with problems), and Teacher 2 replied that CAA was 'okay' when commenting on its diagnostic value. T2 indicated that CAA could show problems, errors, and mistakes. Another diagnostic suggestion was that CAA helps in revision and can test large numbers of learners. Besides that these teachers had little time and no reason to attend to this study, their actions and responses did not show capacity to diagnose learners' problems. They were reluctant to interrogate the data that QM Reporter produced such as facility, and discrimination. Teachers were worried about marks, but not the quality of responses their learners chose.

Yet remediation following diagnosis required the teachers to reflect on assessment data and to establish what could have gone wrong in teaching (Bright, 1987: 81). In this case, QM Reporter revealed distracters that were plausible to learners as alternative constructs. Remediation required the use of such distracters in subsequent tests (Tamir, 1996: 97, 107; Maloney, 1987: 510-513), to *inform the moment-to-moment instructional decisions* during the learning process (Stiggins, 1999). A comparison of the frequency of a particular misconception in subsequent tests can be used to measure the success of diagnosis and remediation (see for example, item analysis, question 10 – Test 3). Therefore, the diagnostic value the teachers attached to CAA was compromised by the teachers' reluctance or by incapacity to use data from QM Reporter. Thus, Mann's (1999) suggestion of minding other factors to accurately determine the effects of technology is relevant in this case. It is reasonable to speculate that these teachers required further training on diagnosis, part of which I tried by setting diagnostic items with them.

Diagnosis of learners' problems

Diagnosing learners' problems by these teachers should include dialogue with learners as they did the test. This implies a change in classroom practices towards freer practical oriented constructivist lessons.

In the interview, a statement such as … *'it was put in a difficult manner so that we can not exactly understand the question'* (L14) can indicate problems with language or a higher level of thinking required in answering a question than what the learner was used to. L4 said that 'n*eeded to know your facts'.* Learners who obtained low marks rushed through the test (see the List Report; e.g. L1 & L 2). This implied lack of careful thinking before choosing answers, and showed that processing information was rare. Possibly, learners dealt with tasks as right or wrong facts (Scott *et al.,* 1987: 19), even those that were weighed differently, and did well in recall items than in challenging questions (see item analysis report; e.g., Question 4 (80%), Question 10 (40%)). Learners found the test difficult because previous assessments had not been challenging (Test 1). Thus, although on the one hand the Subject Advisor and I sought to improve test validity by covering science processes well, on the other hand, the introduction of higher-level items reduced the validity of that test to learners (i.e., the test was difficult). My suspicion is that teachers approached science as factual.

Assessment, diagnosis, and remediation proceeded simultaneously in a natural way (Cunningham, 1991: 15) only when learners invited each of us for consultation (Hannafin, *et al.,* 2004: 13). It would have been better if there were opportunities for testing learners' constructs practically.

Diagnosing teaching

From the above, it appears that data can provide insight on the teaching style and problems. The results from the tests indicate that there were problems with teaching. Setting diagnostic tasks as given for example, by Tamir (1996: 97), appeared difficult for teachers, *inter alia* because teachers lacked the necessary depth of conceptual understanding to set such tasks (see Test 1). However, the plausibility of distracters catered for some learners' conception, since each alternative was chosen, but at the same time reflected upon the teaching they received.

Further to the arguments above, 93% of learners believed that the strength of current influences resistance (Question 6, Test 2). Teachers could have been the source of this popular faulty belief (Pollitt, 1990: 885).

Diagnosing, remedying, and learning theories

Empiricists or positivists, as well as constructivists would argue for practical work to avoid errors that learners made in Question 6. Nonetheless, The New Zealand Council for Educational Research (2001) advises teachers to identify the point at which learning faltered, but this might differ with the learning theory adopted in class, since assessment and learning are co-incidental. The first dilemma is that classroom practices and so assessment, diagnosis and remediation are rarely located in a single learning theory. As an example, the questions in Test 2 can be positioned in behavioural, cognitive or constructivist camps for which diagnosis and remedy might be different. A starting point is to see how each learning theory would deal with diagnosis and remediation.

For behavioural tasks (Fuchs, 1995: 2) discrete and single concepts appeared in items such as Question 4 (see item analysis report). Learners did these well to the satisfaction of Tamir (1996) that multiple-choice tests can reveal faulty memories. Behavioural remediation was achieved since learners improved every time they re-did the test - i.e., they eventually identified the correct answers through drill and practice.

Cognitive and constructivist items required teacher involvement with learners at the time they were solving problems (Hein & Lee, 2000: 7). It is unusual for learners to be allowed to talk during a test, and one can argue that these constructivist/cognitive items, which learners considered thought-provoking with the opportunities of instant marking that CAA provided, helped in constructivist diagnosis and remediation.

However, difficulties in diagnosis and remediation emerge with Wood's (as cited in Gipps, 1996: 255) advice to identify levels of ‘how well’ one understands. Although Test 3 provided multiple distracters that represented some of the learners' levels of understanding (Hannafin *et al.,* 2004: 13), there was an inherent assumption that thinking is systematic and has a road map where the distance translates into different marks or presumably different levels of understanding. The ways I personally think is certainly not linear, and is abstract and haphazard that it requires research; it is hard for anyone to tap into and discover accurately where my thoughts could be stuck. Hence, the way such distracters or personal schema (Tamir, 1996; Piaget cited in Mwamwenda, 1993: 71) were assessed might domesticate constructivism (Greening, 1998: 23-24). It thus raises questions about the practicality of constructivism and the design of rubrics in the new South African science curriculum. Additionally, Harlen (1993: 28-36) raises the issue of the difficulty of demarcating science processes from each other, and therefore, the difficulty in designing rubrics as required by curricula, which assume that each process or outcome can be identified, assessed and remedied. However, I propose that, although Rubrics might be hypothetical, they offer a platform for conjectures about concepts, researching the concepts, as well as possibilities of linking MCQs, rubrics, practical work, and continuous assessment.

Diagnosing science processing

CAA can support such possibilities (Figure 1) because CAA can handle multiple measurements. Figure 1 is my hypothetical example of assessment of some of the science processes involved, for example, in understanding Ohm's Law. Figure 1 also borrows from (Kuiper, 1997) in that no level or construct is wrong. The figure combines the design principles of a rubric, and MCQ, which could also encourage a dialogue between the teacher and a learner. The question would be open-ended and present as many alternative constructs as possible typed in CAA.

Table 10: *Rubric, MCQ or Dialogue?*
Question on Ohm's Law: What do you know about the relationship between potential difference and resistance?

Competency Level	Outcomes/constructs/alternative conceptions in an MCQ	Practical evidence	Score
1	V is proportional to R	Table of readings // V vs. R	2
2	V proportional to R because charge loses more energy across higher resistance		3
3	V proportional to I if R is constant / the rate at which charge passes R is depends on V if R remains the same // simple V=IR calculation).	Table of readings – draws graph // V vs. I. Determines R	7
4	Factors that affect R affect I with V constant (Factors include, temp, state or phase, thickness, length).	Table of readings // V vs. I under different temperatures - Determines R Vs T	10
5	Conductors or resistors whose R changes (due to some factors) do not obey this relationship // gives examples	Table of readings and graph // V vs. I. Explains the graph	13
6	R's in parallel provide more passage to charge and become less effective. Too many Rs allow too many charges to pass thus increasing power consumption or even a short circuit // examples of calculations of current in parallel Rs.	Connects resistors in parallel and records V as well as I.	15

Processes include hypothesising (i.e., rationalism), collection of data (i.e., empiricism), and social negotiations (i.e., social constructivism), bearing in mind that the journey is not terminal and could culminate into research, for example, in the form of project work. CAA provides the capacity to deal with such multiple constructs in form of a rubric (Jonassen *et al.,* 2003: 229), especially in a practical exercise and a dialogue – it reduces the labour, which, as we know, teachers are complaining about when they deal with continuous assessment and rubrics in classrooms with large numbers of learners.

CONCLUSION

These teachers (and learners) confirmed findings about the worth of CAA as claimed by in Oliver (2000), Gretes & Green (2000), as well as Croft *et al.* (2001), that CAA is valuable because it instantly provided volumes data and analyses, which teachers and learners could use to diagnose and remedy problems in teaching and learning. In the course of formulating strategies for diagnostic assessment that would benefit learners and teachers (Stiggins, 1999, 2002), I propose the following action plan for the two teachers:

- Improved conceptual understanding in science, and ability to set meaningful diagnostic tasks. They have to be persuaded probably with incentives such as CAA software and certification about using them to enrol for assessment learnerships in SA.
- Teachers would need to change their teaching styles to fit diagnosis and remedial work.
- A review of the curriculum to accommodate CAA, including reductions in class loads.

- Common misconceptions on each topic in science should be researched and be incorporated into item databases.
- Learners should have access to CAA, with guidance from the teacher, any time.

Other capital problems, such as the costs of computers and of CAA in these parts of the world remain hindrances that are beyond the teachers' influence, and should be addressed by all stakeholders.

REFERENCES

Amir, R. & Tamir, P. 1994. In-depth Analysis of Misconceptions as a Basis for Developing Research-Based Remedial Instruction: The Case of Photosynthesis. *The American Biology Teacher, Volume 56, No. 2, February 1994*. pp. 94-99.

Birenbaum, M. 1996. Assessment 2000: Towards a Pluralistic Approach to Assessment. In Birenbaum, M. and Dochy, F. J. R. C. 1996 (Eds). *Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge.* London: Kluwer Academic Publishers.

Bojic, P. 1995. What is Computer-Based Assessment? [Online] Available: http://www.warwic.ac.za/ETS/interactions/Vol2no3/links.htm. [6th October 1999].

Bright, G. W. 1987. *Microcomputer Applications in the Elementary Classroom. A Guide for Teachers.* Boston: Allyn and Bacon, Inc.

Croft, A. C., Danson, M., Dawson, B. R. & Ward, J. P. 2001. Experiences of Using Computer Assisted Assessment in Engineering Mathematics. *Computers & Education 37 (53-66).*

Cunningham, D. J. 1991. Assessing Constructions and Constructing Assessments: A Dialogue. *Educational Technology. May 1991. Volume 31 (5).*

Department of Education, South African Government. 23rd December 1998. Vol. 402, No. 19640. No. 6397 No. R. 1718. Government Notice, Department of Education. National Education Policy Act, 1996 (ACT NO. 27 OF 1996). Assessment Policy in the General Education and Training Band, Grades R To 9 and Abet. [Online] Available: http://www.polity.org.za/govdocs/regulations/1998/reg98-1718.html [18th July 2000].

Department of Education, South African Government, 30 November 2005, *National Curriculum Statement Grades 10-12 (General) Learning Programme Guidelines Physical Sciences*

Dietel, R. J., Herman, J. L., and Knuth, R. .A. 1991. *NCREL, What Does Research Say About Assessment?* [Online] Available: http://www.ncrel.org/sdrs/areas/stw_esys/4assess.htm [2000, July 31]

Fraser, W. J. 1991. Basic Considerations in the Construction of Classroom Tests. In Dreckmeyr, M. & Fraser, W. J. (Eds.). *Classroom Testing in Biology and Physical Science.* Bloemfontein. HAUM Tertiary.

Fuchs, L. S., 1995. *Connecting Performance Assessment to Instruction: A Comparison of Behavioral Assessment, Mastery Learning, Curriculum-Based Measurement, and Performance Assessment*. [Online] Available: http://ericec.org/digests/e530.htm [2001, March 23].

Gay, L. R. & Airasian, P. 2000. *Educational Research. Competencies for Analysis and Application. Sixth Edition.* Columbus, Ohio: Merrill.

Gipps, C. 1996. Assessment for Learning. In Little, A. & Wolf, A. (Eds.). *Assessment in Transition. Learning, Monitoring, and Selection in International Perspective*. Oxford. Pergamon.

Greening, T. 1998. Building the Constructivist Toolbox: An Exploration of Cognitive Technologies. *Educational Technology / March-April 1998, (23-35).*

Gretes, J. A. & Green, M. 2000. Improving Undergraduate Learning with Computer-Assisted Assessment. *Journal of Research on Computing in Education.* Volume 33. Number 1. Fall 2000.

Haladyna, T. M. 2002. Supporting Documentation: Assuring More Valid Test Score Interpretation and Uses. In Tindal, G. & Haladyna, T. (Eds.). 2002. *Large-Scale Assessment Programs For All Students. Validity, Technical Adequacy, and Implementation.* Mahwah, New Jersey. Lawrence Erlbaum Associates, Publishers. (Pages 89-108).

Hannafin, M. J., Kim, M. C., & Kim, H. 2004. Reconciling Research, theory, and Practice in Web-Based Teaching and Learning: the Case for Grounded Design. *Journal of Computing in Higher Education. Spring 2004, Vol. 15(2), (30-49).*

Harlen, W. 1993. *The Teaching of Science*. London. BPCC Wheaton Ltd.

Harlen, W. 2000. Assessment in the Inquiry Classroom. [Online] Available: http://www.nsf.gov/pubs/2000/nsf99148/lcd/ch_11.htm [25th April 2001].

Hein, G. E. and Lee, S. 2000. Assessment of Science Inquiry. [Online] Available: http://www.nsf.gov/pubs/2000/nsf99148/lcd/ch_12.htm [25th February 2000].

Heinecke, W. F., Blasi, L., Milman, N. & Washington, L. 1999. New Directions in the Evaluation of the Effectiveness of Educational Technology. Paper given at Papergiven at *The Secretary's Conference on Educational Technology-1999.* [Online] Available: http://www.ed.gov/Technology/TechConf/1999/whitepapers/paper8.html [30th October 2002].

Hickey, D. T., Horwitz, P., D. T., Kindfield, A. C. H., & Christie, M. A. T. 2003. Integrating Curriculum, Instruction, Assessment, and Evaluation in a technology-Supported Genetics Learning Environment. *American Educational Research Journal. Summer 2003, Vol. 40, No. 2, (495-538).*

Jonassen, D. H., Howland, J. L., Moore, J. L., & Marra, R. M. 2003. *Learning to Solve Problems with Technology. A Constructivist Perspective. Second Edition.* Upper Saddle River: Merrill Prentice Hall.

Knight, M. & Brown, A. 2000. Computer based Assessment. [Online] Available: http://ctiweb.cf.ac.uk/HABITAT/HABITAT4/compass.html [14th December 2001].

Kuiper, J. 1997. Quirks and *Quarks*: Changing Paradigms in Educational Research. *Meeting of the Association for Research in Mathematics, Science and Technology Education. 22-26 January 1997. University of Witwatersrand, Johannesburg. (530-534).*

Lawton, D. & Gordon, P. 1996. *Dictionary of Education.* Second edition. London: Hodder & Stoughton.

Linn, R. L. 2002. Validation of the Uses and Interpretations of Results of State Assessment and Accountability Systems. In Tindal, G. & Haladyna, T. (Eds.). 2002. *Large-Scale Assessment Programs For All Students. Validity, Technical Adequacy, and Implementation.* Mahwah, New Jersey. Lawrence Erlbaum Associates, Publishers, (27-66).

Little, A. & Wolf, A. (Eds.) 1996. *Assessment in Transition. Learning, Monitoring, and Selection in International Perspective*. Oxford. Pergamon.

Madaus, G. F. & Kellaghan, T. 1992. Curriculum Evaluation and Assessment. In Jackson, P. W. (Ed.) 1992. *Handbook of Research on Curriculum.* (119-154). New York. Macmillan Publishing Company.

Maloney, D. P. 1987. Ranking tasks. A New Type of Test Item. *Journal of College Science Teaching*, May 1987, (510 – 515).

Manzini, S. 2000. Learners' Attitudes Towards the Teaching of Indigenous African Science as Part of the School Science Curriculum. *Journal of the Southern African Association for Research in Mathematics, Technology and Science Education. Volume 4, Number 1, (19-32).*

McComas, W., Clough, M., & Almazroa, H. 1998. The Role and Character of the Nature of Science in Science Education, (3-39). In McComas, W. F. (ed.) 1998. The Nature of Science in Science Education. Rationales and Strategies. Science & Technology Education Library. London. Kluwer Academic Publishers.

Muwanga-Zake, J. W. F. 2000. Is Science Education in South Africa in a crisis? The Eastern Cape Experience. *Journal of the Southern African association for Research in Mathematics, Technology and Science Education. Vol. 4 (1), (1-11)*

Muwanga-Zake, J. W. F. December 2004. PhD Thesis. University of KwaZulu-Natal, South Africa. *Evaluation Of Educational Computer Programmes As A Change Agent In Science Classrooms.*

Mwamwenda, T. S. 1993. *Educational Psychology. An African Perspective.* Durban: Butterworth Publishers (Pty) Ltd.

New Zealand Council for Educational Research. 2001. [Online] Available: http://arb.nzcer.org.nz/nzcer3/nzcer.htm [21th February 2002].

Ogunniyi, M. B. 2000. Teachers' and Pupils' Scientific and Indigenous Knowledge of Natural Phenomena. *Journal of the Southern African Association for Research in Mathematics, Technology and Science Education. Volume 4, Number 1, (70-77).*

Ogunniyi, M. B. 1997. Multiculturalism and Science Education Research in the New South Africa. *Proceedings of the Fifth Meeting of the Southern African Association for Research in Mathematics and Science Education. 22-26 January 1997. University of Witwatersrand, Johannesburg, (50-53).*

Oliver, A. 2000. Computer Aided Assessment – the Pros and Cons. [On Line] Available: http://www.Herts.ac.uk/ltdu/learning/caa_procon.htm [14th December 2000].

Pachler, N. & Byrom, K. 1999. Assessment of and through ICT. In Leask, M. and Pachler, N. (1999). (Eds.) *Learning to Teach Using ICT in the Secondary School*. London: Routledge.

Pollitt, A. 1990. Diagnostic Assessment Through Item Banking. In Entwistle, N. (Ed.) 1990. *Handbook of Educational ideas and Practices.* London and New York: Routledge.

Question Mark. [Online] Available: http://www.qmark.com/ [9th September 1999].

Ryan, J. M. & DeMark, S. 2002. Variation in Achievement Scores Related to Gender, Item Format, and Content Area Tested. In Tindal, G. & Haladyna, T. (Eds.). 2002. *Large-Scale Assessment Programs For All Students. Validity, Technical Adequacy, and Implementation.* Mahwah, New Jersey. Lawrence Erlbaum Associates, Publishers. (Pages 67-88).

Sandals, L. H. 1992. An Overview of the Uses of Computer-Based Assessment and Diagnosis. *Canadian Journal of Educational Communication, Vol. 21, No. 1 (67-78).*

Scott, P., in association with Dyson, T. & Gater, S. 1987.*A constructivist view of learning and teaching in science.* Leeds. Centre for Studies in Science and Mathematics Education, University of Leeds. Leeds LS2 9JT.

Stiggins, R. J. Summer 1999. Teams. *Journal of Staff Development*, Summer 1999 (Vol. 20, No. 3) [Available] 22nd July 2006. Online. http://www.nsdc.org/library/publications/jsd/stiggins203.cfm

Stiggins, R. J. 2001. *Student-Involved Classroom Assessment.* 3rd ed. Upper Saddle River, NJ: Prentice-Hall, Inc.

Stiggins, R. J. 2002. Assessment Crisis: The Absence Of Assessment *FOR* Learning. *Phi Delta Kappan*. http://www.pdkintl.org/kappan/k0206sti.htm [Available] 22nd July 2006.

Taiwo, A. A. 1995. *Fundamentals of Classroom Testing.* New Delhi: Vikas Publishing House PVT LTD.

Tamir, P. 1996. Science Assessment. In Birenbaum, M. and Dochy, F. J. R. C. 1996 (Eds.). *Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge.* London: Kluwer Academic Publishers.

The Department of Computer-Based Education, University of Cape Town. 2000. [Online] Available: http://www.uct.ac.za/projects/cbe/mcqman/mcqchp2.html [21st November 2001].

The Maricopa Center for Learning & Instruction, Maricopa Community Colleges, 2000. What is assessment of learning? [Online] Available: http://www.mcli.dist.maricopa.edu/ae/al_what.html [20th March 2001].

The New Zealand Council for Educational Research. 2001. [Online] Available: http://arb.nzcer.org.nz/nzcer3/nzcer.htm [2001, March 20].

Thelwall, M. 2000. Computer-Based Assessment: a Versatile Educational Tool. *Computers & Education 34 (2000)* (37-49).

Weiss, C. H. 1998. Evaluation. *Methods for Studying programs and Policies.* New Jersey: Prentice Hall.

Wise, S. L. & Plake, B. S. 1990. Computer-Based Testing in Higher Education. *Measurement and Evaluation in Counselling and Development / April 1990 / Vol 23.*

Original article at: http://ijedict.dec.uwi.edu//viewarticle.php?id=226&layout=html