Improving educational assessment: A computer-adaptive multiple choice assessment using NRET as the scoring method

Lau Sie Hoe[1], *Lau Ngee Kiong*[1], *Hong Kian Sam*[2], *Hasbee Bin Usop*[2]

(1. Faculty of Information Technology & Quantitative Science, University Technology MARA, Sarawak 94300, Malaysia;

2. Faculty of Science Cognitive and Human Development, University Malaysia, Sarawak 94300, Malaysia)

Abstract: Assessment is central to any educational process. Number Right (NR) scoring method is a conventional scoring method for multiple choice items, where students need to pick one option as the correct answer. One point is awarded for the correct response and zero for any other responses. However, it has been heavily criticized for guessing and failure to credit partial knowledge. Despite continued research in scoring methods, conventional NR scoring method remains the most popular and widely used procedure. Thus, the problem of crediting partial knowledge and guessing still exist. In addition, the current assessment practices do not take students' misconceptions into consideration. Thus, there is a need to propose a new scoring method that is able to minimize guessing, credit partial knowledge, yet able to diagnose misconceptions. This study proposed the development of a Computer-Adaptive Assessment Software (CAAS) for MC items using the new Number Right Elimination Testing (NRET) which take into consideration these three elements. This paper described the rational for using the NRET, the design and development of CAAS and the results of the pilot study carried out in three secondary schools in Malaysia.

Key words: multiple-choice; guessing; partial knowledge

1. Introduction

Assessment is central to any educational process. However, only a small fraction of the potential benefit is typically obtained (Paul, 1994). Although research has shown that multiple choice (MC) items can elicit complex thinking process (Haladyana, 1997; Hopkins, Stanley & Hopkins, 1990; Maguire, Hattie & Haig, 1994; Norman, 1997; Rothman, 1995; Sax, 1989), the way of evaluating thinking processes is limited by among others the way these items are scored (Ndalichako, 1997). Number Right (NR) scoring method is a conventional scoring method, where student need to pick one option as the correct answer. One point is awarded for the correct response and zero for any other responses.

However, it has been heavily criticized for guessing and failure to credit partial knowledge. It does not capture the full information available in the responses concerning a student's ability (Bock, 1972; Claudy, 1978;

Lau Sie Hoe, senior lecturer, Faculty of Information Technology & Quantitative Science, University Technology MARA; research field: mathematics education and assessment.

Lau Ngee Kiong, Ph.D., senior lecturer, Faculty of Information Technology & Quantitative Science, University Technology MARA; research field: mathematics education and assessment.

Hong Kian Sam, Ph.D., associate professor, Faculty of Science Cognitive and Human Development, University Malaysia; research fields: mathematics education, e-learning, computer supported learning.

Hasbee Bin Usop, Ph.D., senior lecturer, Faculty of Science Cognitive and Human Development, University Malaysia; research fields: learning science, science education.

Haladyana, 1994; Hambleton, Robert & Traub, 1970; Thissen, Steinberg & Fitzpatrick, 1989; Wainer, 1989). There is no information on how students think with respect to each alternative: complete certainty, vacillation between several alternatives, blind guess, or even more or less serious fallacy (Holmes, 2002). Thus it is impossible to understand the thinking process of students. As a result, useful diagnostic feedback that stimulates cognitive functioning leading to reflection and enabling students to review and monitor their work could not be provided.

The assessment of partial knowledge and the control of guessing behavior have been the two goals of measurement specialists since MC items were introduced in 1917 (Hirtz & Jacobs, as cited in Hutchinson, 1985). Crediting partial knowledge and controlling guessing will create a fairer assessment that lead to a more valid inference about student's ability. A valid assessment is very important because many important or irreversible decisions and policies are made based on the results of testing. Alternative scoring method such as Number Right with correction for guessing (Trow, 1923), Confidence Weighting (De Finetti, 1965), Probability Measurement (De Finite, 1965), Free Choice method or Subset Selection Testing (SST) (Dressel & Schmid, 1953), Elimination Testing (ET) methods (Coombs, 1953), Rank Ordering the options method (Aiken, 1968), Answer-Until-Correct (AUC) method (Gilman & Ferry, 1972), Option Weighting method (Nedelsky, 1954), and Item Weighting method (Chevalier, 1998) were proposed. However due to the complexity of the procedure involved in administering and obtaining test scores, the NR scoring method remains the most popular and widely used procedure (Ndalichako & Rogers, 1997). Thus the problems of crediting partial knowledge and guessing still exist. In addition, there is little or no emphasis in taking students' misconceptions into consideration in assessment.

Coombs, Millbolland and Womer (1956) describe partial knowledge as knowledge that students have which allows them to recognize some of the incorrect response options as clearly wrong in a MC item while being unable to recognize the correct option. Partial knowledge is correct information: it is just insufficient knowledge. In contrast, misconception is incorrect information. A student is having misconception if he eliminates the correct answer from among the answer options of the multiple choices (Lowry, 1979). A student believes that the correct answer is definitely not the answer. Misconception might lead to serious consequence in critical fields such as medical practices. Thus, students' misconceptions should be detected. Unfortunately, the current assessment practices do not take misconceptions into consideration. From the educational point of view, the misconceptions that students have with them are critical barriers to learning.

Thus, there is a need for a new scoring method that is able to minimize guessing, credit partial knowledge, yet able to diagnose misconceptions. Computer-Adaptive Assessment Software (CAAS) for MC items using the new Number Right Elimination Testing (NRET) was proposed as a solution in the present study. NRET is a hybrid of two existing MC scoring method: Number Right (NR) and Elimination Testing (ET). In addition, timely hints and feedback were also provided in CAAS to support continuous learning.

2. Related review of literature

According to Holmes (2002), NR scoring method is a very coarse measuring tool which lumps all personal certainties, doubts, fallacies and guesswork into just two categories: correct or incorrect. Coombs (1953) proposed ET where students must cross out all alternatives that they consider incorrect. However, according to Jaradat and Tollefson (1988), ET test instructions are confusing despite prior practice. ET also conflicted with the instructional process where students are taught to solve for the correct answer but instead assessed on their ability to identify the wrong answer. Dirkwager (1996) stated that based on partial knowledge, students can identify that some option(s) which is/are more likely to be true than others. However, under ET, students are denied the opportunity to indicate their preferred option as the possible answer. Thus, this portion of partial knowledge is not captured.

In a survey on test-taking strategies employed by students, Oh (2004) reported that out of 264,900 students who took the Fall, 2001 administration of the SAT, 60% or roughly 159,940 students first eliminated incorrect alternative before selecting one alternative as the answer. Likewise, Dirkwager (1996) also discovered that most students used this strategy. In Malaysia, it is a well-known fact that most of the students, even at primary level used this strategy while attempting MC items. This strategy has been encouraged and even taught by teachers in schools to maximize test score though there is no review to support it. This method of answering MC item has a long history as a test taking strategy (Shepard, 1982; Grant, 1989; Devine & Meagher, 1989; Shepard, 1990). Millman, Bishop and Egel (1965) regard them as deductive reasoning strategies.

Rogers and Bateson (1991a, 1991b), and Roger and Wilson (1993) suggested that effective application of these deductive reasoning strategies depends upon the partial knowledge that the students possess. No doubt students can increase their test scores by being test-wise. Nonetheless, it has been shown that effective application of test-wise strategies require partial knowledge. Thissen (1976) strongly suggested that students who do not possess the necessary knowledge to answer an item do not randomly select an answer, but rather use partial knowledge to first eliminate the incorrect and select one of the remaining option. Tomkowicz (2000) also shares his view. According to Mehrens and Lehmann (1991), logic and evidence suggest that it is informed guessing that predominates. Downing (2003) further argued that informed guessing should not be discouraged because it is a reality of life. Much of what individual does throughout lives is in fact based on incomplete knowledge.

Although frequently used by students, the strategy of first eliminating the incorrect option before selecting the answer is only regarded as a test taking strategy to maximize test score. It has not been formalize as a scoring method. This strategy is actually the hybrid of NR (choose one as the answer) and ET (eliminate all incorrect alternatives). Based on the arguments above, this test-taking strategy has the potential to be an alternative scoring method to replace the conventional NR and ET scoring in addressing the issue of guessing and crediting partial knowledge. In addition, misconceptions can be detected based on students' response patterns. Thus, Number Right Elimination Testing (NRET) which formalizes this test-taking strategy as a scoring method is proposed.

3. NRET test instructions and scoring guides

NRET scoring method is a hybrid of two existing scoring method, Number Right (NR) and Elimination Testing (ET). Under NRET, students need to eliminate alternative(s), which is/are incorrect and based on the remaining alternatives, and choose one as the answer. The calculation of item score for NRET scoring method is based on the combination of ET and NR scoring method. One point will be awarded for each incorrect answer crossed out. However, a penalty of -3 will be given if the correct answer is crossed out. This is to discourage guessing and penalizing misconceptions. One additional point will be awarded if the answer chosen is correct. Thus, the item score ranges from -3 to 4. Table 1 shows the comparison of the NR, ET and the NRET scores for all possible responses patterns.

Table 1 NRET, ET and NR item score for all possible response pattern

Level of knowledge	Response pattern	NR score	ET score	NRET score
Full knowledge	Answer correct Identify 3 options as definitely incorrect	1	3	4
Partial knowledge	Answer correct Identify 2 options as definitely incorrect	1	2	3
	Answer correct Identify 1 options as definitely incorrect	1	1	2
	Answer wrong Identify 2 options as definitely incorrect	0	2	2
	Answer wrong Identify 1 options as definitely incorrect	0	1	1
	Answer correct Identify no option as definitely incorrect	1	0	1
Absence of knowledge	Answer wrong Identify no option as definitely incorrect	0	0	0
Partial misconception	Identify answer as definitely incorrect Identify 2 options as definitely incorrect	0	-1	-1
	Identify answer as definitely incorrect Identify 1 option as definitely incorrect	0	-2	-2
Full misconception	Identify only the answer as definitely incorrect	0	-3	-3

For example, consider four students having different level of knowledge for the question below.

Question: Which of the following value is greater than 15?
(A) $\sqrt{37}$ (B) 42
(C) 16-1 (D) 10000 Correct Answer: (B)

Student V: full knowledge—very sure that (B) is definitely the answer and others definitely wrong.

Student W: partial knowledge—very sure that option (C) and (D) are definitely not the answer. Between options (A) and (B), he feels that (B) has the higher chance to be correct based on what he can recall.

Student X: partial knowledge—very sure that option (C) and (D) are definitely not the answer. Between options (A) and (B), he feels that (A) has the higher chance to be correct based on what he can recall.

Student Y: partial misconception—able to identify (C) as definitely incorrect. However, he also has misconception, thinking (B) is definitely wrong. Between options (A) and (D), he feels that (D) is more likely to be the correct answer.

Student Z: full misconception—thinks that (B) is definitely wrong and between (A), (C) and (D), he feels that (A) is most likely the answer.

Based on their knowledge level, their scores under NR, ET and NRET scoring method are as shown in Table 2.

Table 2 Example of student's NR, ET and NRET scores

Students	Knowledge level	NR score	ET score	NRET score
V	Full knowledge	1	3	4
W	Partial knowledge	1	2	3
X	Partial knowledge	0	2	2
Y	Partial misconception	0	-2	-2
Z	Full misconception	0	-3	-3

Table 1 and Table 2 clearly show that NRET gives a finer discrimination of students' knowledge level, thus giving a more valid assessment of student's knowledge state. In addition, it offers students more opportunities for them to tell what they know as compared to NR. This may also improve students' attitudes toward testing as suggested by Lukin (1989). Inflexibility in NR can lead to cynical attitudes and loss of faith in objective testing by the students (Abu-Sayf, 1979). The NRET scoring method has a total of 32 possible responses as compared to only four in NR and 16 in ET. Increase in total possible responses will increase the discrimination ability. Hence, it can differentiate better students of nearly similar abilities.

Paul (1994) says the objectives of assessment should include: establishing and revealing status ("knowing what students know"), diagnosis of weaknesses ("knowing what students don't know") and comparative assessment with larger populations (answering "where do they stand?"). According to Kellaghan, Madaus and Airsian (1982), traditional assessment methods have failed to adequately achieve the desired objectives. NRET scoring method offers a promising method for gathering more diagnostic information from MC testing procedures, thus, improving the psychometric qualities of the tests. Students' knowledge state can be identified based on the students' response pattern to each option on an MC item. Their knowledge state may range from full knowledge to full misconception as shown in Table 1 above. This additional diagnostic information may be used to improve remediation activities such as in content misconceptions.

4. Methodology

4.1 The design of Computer-Adaptive Assessment Software (CAAS)

CAAS is web-based assessment software for learning mathematics and is currently available at http://caa.bestservices.com.my. It is designed for three types of potential users: the administrators, teachers and students. For security reasons, all users have to register and their accounts activated before they can access the software. It is adaptive in nature where the selection of the successive item will depend on students' performance of the item. Thus, students with high ability will not waste time in solving average questions that do not excite, challenge, or interest them. In addition, hints and timely feedback are also given for continuous learning. Students will not only obtain their scores, but also learn about their mistakes or misconceptions instantly. Students are also given a second chance to correct their thinking if their response to an item is incorrect in their first attempt. The flow for item selection is as shown in Figure 1.

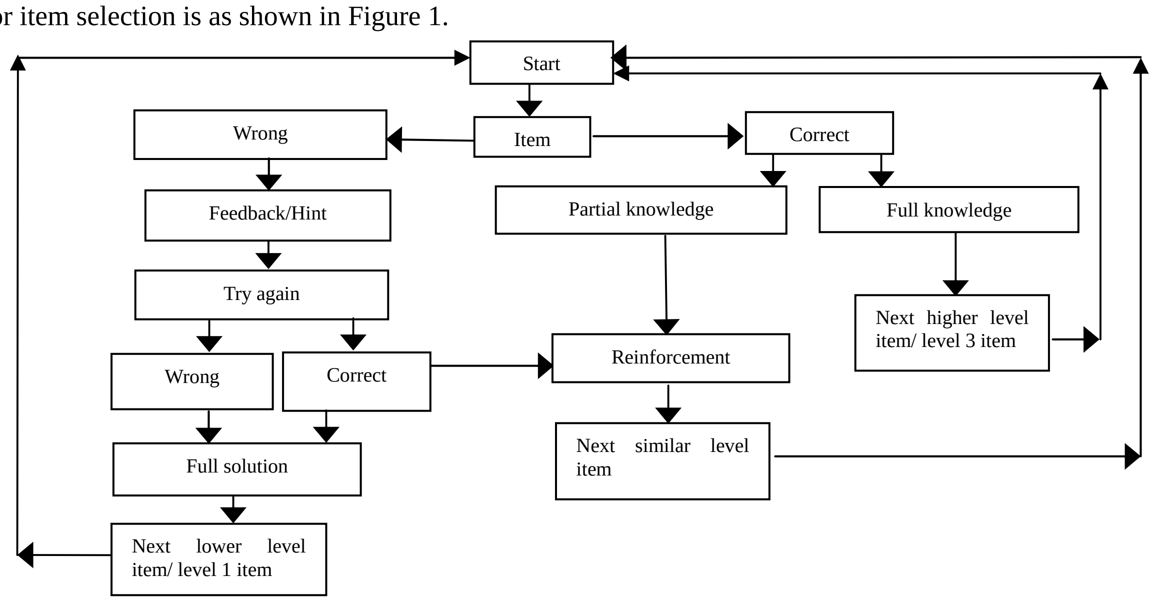


Figure 1 Flow chart for item selection

There are a total of 13 chapters in the Form Two Malaysian Mathematics syllabus in CAAS. For each chapter, there are 57 items graded into three levels of difficulty. The items are constructed by experience mathematics teachers and lecturers. Each question is followed by a general hint and a full solution. The items were pilot tested in schools to determine the difficulty levels. Based on the students' abilities, CAAS will generate 20 items for each chapter.

An item is presented on the screen, for each response option in a multiple-choice item, students should indicate on each option how they think about that response: "Correct", "Wrong" or "Not Sure". Students must choose one option as "correct". However, they have the flexibility in choosing none, one, two or three as "Wrong" or "Not Sure". Figure 2 shows the tools used by students.

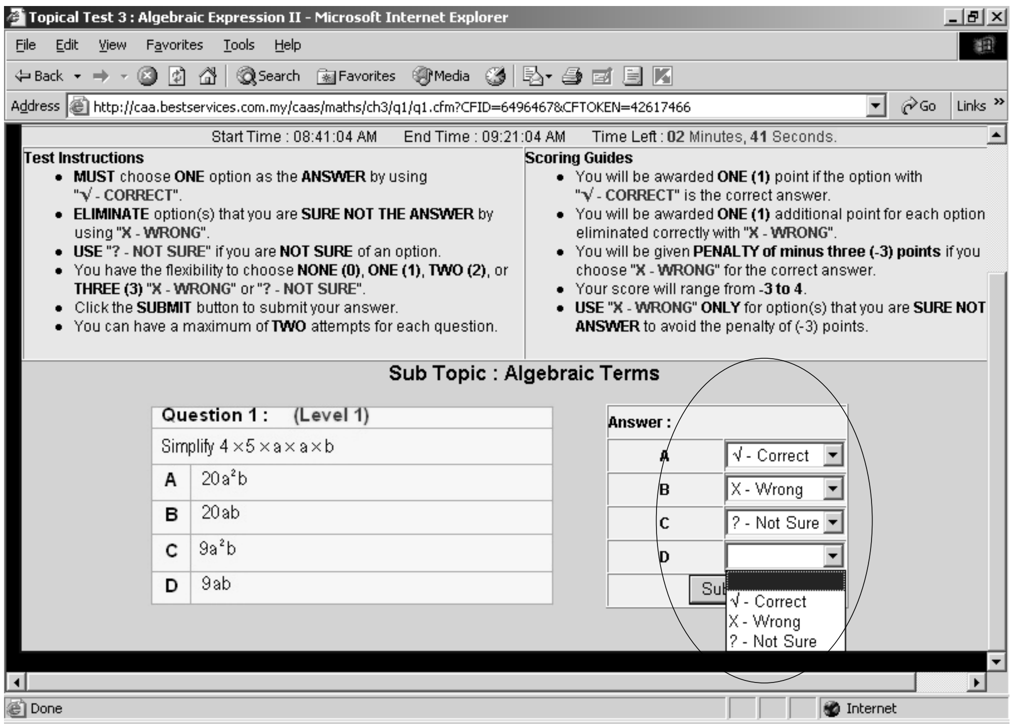
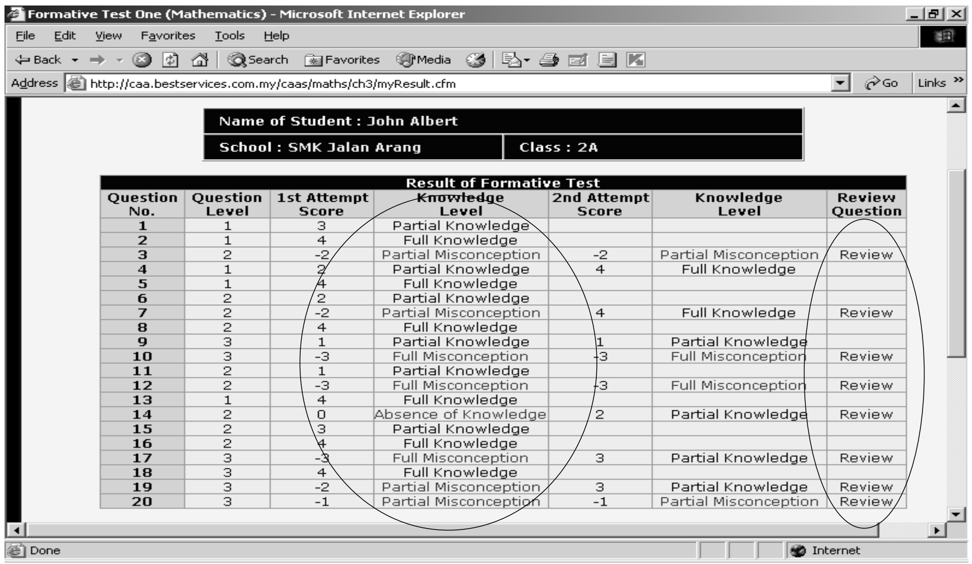


Figure 2 Tools used by the students

A reminder will appear if the students do not follow the test NRET test instructions. Based on the students' responses, CAAS will generate the item scores and classify the students' knowledge state for each item. By reviewing the knowledge state, students will be more aware of whether they overvalue or undervalue their knowledge. This will give them a more accurate self-assessment of their knowledge state. According to Barnett and Hixon (1997), and Paris and Winograd (1990), accurate self-assessment is also essential to support metacognition, which empowers students to regulate and control their own learning and for which laboratory studies show positive correlations with studying and achievement. For items having misconception or partial knowledge, students will be given opportunity to review the questions, their responses and the actual answers. This feedback feature makes it easier for teachers to give students feedback about their thinking and for them to revise their work. In addition, students can view their ranking against all participating students in their school as well as against all participating students. Figure 3 shows the detail analysis of students' performance.



Name of Student : John Albert

School : SMK Jalan Arang | Class : 2A

Result of Formative Test						
Question No.	Question Level	1st Attempt Score	Knowledge Level	2nd Attempt Score	Knowledge Level	Review Question
1	1	3	Partial Knowledge			
2	1	4	Full Knowledge			
3	2	-2	Partial Misconception	-2	Partial Misconception	Review
4	1	2	Partial Knowledge	4	Full Knowledge	
5	1	4	Full Knowledge			
6	2	2	Partial Knowledge			
7	2	-2	Partial Misconception	4	Full Knowledge	Review
8	2	4	Full Knowledge			
9	3	1	Partial Knowledge	1	Partial Knowledge	
10	3	-3	Full Misconception	-3	Full Misconception	Review
11	2	1	Partial Knowledge			
12	2	-3	Full Misconception	-3	Full Misconception	Review
13	1	4	Full Knowledge			
14	2	0	Absence of Knowledge	2	Partial Knowledge	Review
15	2	3	Partial Knowledge			
16	2	4	Full Knowledge			
17	3	-3	Full Misconception	3	Partial Knowledge	Review
18	3	4	Full Knowledge			
19	3	-2	Partial Misconception	3	Partial Knowledge	Review
20	3	-1	Partial Misconception	-1	Partial Misconception	Review

Figure 3 Analysis of students' performance

4.2 Pilot testing of CAAS

Theoretically, CAAS holds great potential in improving educational assessment. Nevertheless, it does not means that CAAS is ready to be used. Improvement should be supported by empirical evidences. Thus, a pilot study was conducted to determine students' perceptions on the suitability of CAAS as a learning tool and the usefulness of feedback on the knowledge state.

4.3 Sample for the pilot test

The samples for the pilot study consisted of 73 students from three secondary schools in Kuching and Samarahan divisions, Sarawak, Malaysia. There were 34 male and 39 female students. A total of 35 students or 47.9% students surveyed had computers at home. Out of 35 students having computers at home, 31 students spent at least one hour on learning using computer in a week. All the three schools were equipped with computers and 51 students out of 73 students spent at least an hour learning using computer at their schools per week. 31 out of 73 students surveyed stated that they had been exposed to other learning software besides CAAS.

5. Findings and discussions

5.1 Perceptions on CAAS as a learning tool

Out of 73 students surveyed, 60.3% or 55 students agreed that CAAS helped them in learning mathematics. Only one student disagreed while other 23.3% or 17 students were undecided. In addition to this, 64.4% or 57 students stated they would always use the CAAS if it was available on the Web as compared to only 2.7% or 2 students who say otherwise. The rest of the students were undecided. A further 61.6% or 45 student suggested that CAAS with NRET scoring method be extended to other subjects such as science. Thus, the majority of the students accepted CAAS as a learning tool and was willing to use it if it was made available on the Web.

These findings were generally consistent with past studies, which reported the benefits of using computer-based assessment. Fleischman (2001) and Lockwood (2001) reported that Web-based assessment promoted learning because students had more opportunities for practice, self-testing, self-evaluation and self motivation. According to Bennett (2001), Bransford, Brown and Cocking (1999), instant feedback enabled students to learn about their mistakes or misconceptions thus providing them the opportunity to readjust their thinking instantly. Laborde (1991) stresses the need of students being given a second chance to answer the same item to confirm their adjustment if the answer in the first attempt is incorrect. Paul (1994) points that a good assessment tool for learning should include informing students on what they know and do not know. In addition, students should know where they stand as compared to others. The design of CAAS took all the above factors into consideration. Thus, as shown by the findings, CAAS was regarded as an acceptable learning tool by the students.

5.2 Perceptions on the usefulness of feedback on the knowledge state

The pilot study showed that 42.5% or 31 students perceived feedback on the knowledge state as useful and helped them in learning mathematics. Only 13.7% or 10 students said otherwise while the remaining 43.8% or 32 students were undecided. 57.5% or 42 students agreed that feedback on knowledge level enabled them to identify the area they were weak in. In contrast, only 1.4% or 1 student disagreed. 47.9% or 35 students reported that these feedbacks helped them in planning what to study as compared to only 5.4% or 4 students that disagreed. The remaining 46.7% or 34 students were undecided. This finding was consistent with suggestions by Barnett and Hixon (1997), Paris and Winograd (1990), and Paul (1994) that accurate self-assessment is useful and essential to support metacognition which empowers students to regulate and control their own learning. This indicates that students value feedback on their knowledge state. By reviewing their knowledge state, students will be more aware of whether they overvalued or undervalued their knowledge and help them in planning their learning.

The majority of the students viewed feedback on knowledge state as useful. However, 50.7% or 37 students stated they were more interested in their marks than feedback on their knowledge state. For items having misconception or partial knowledge, CAAS gave opportunity for students to review the questions, their responses and the actual answer to determine where their mistakes were. Unfortunately only 35.6% or 26 students reported that they always return to check why they get the item wrong. This finding was consistent with finding by Wotjas (1998), Hounsell (1987), Crooks (1988), and Jackson (1995) who reported that students seem to be only concerned with the final grades or marks. They further found that most of the students did not read the feedback if they did not like their grades or marks. According to Ding (1998), even if students read feedback comments, they do little with them. In contrast, Brookhart (2001) found that successful students used both marks and feedback and actively self-assess, learn and to direct their future learning. Black and William (1998) suggested that in the absence of marks, students will read feedback much more carefully and use it to guide their learning.

6. Conclusions

A good assessment tool should enable students to evaluate their understanding of a topic before it is too late. CAAS using the NRET scoring method has the potential to assist students to evaluate their understanding by providing feedback on the knowledge state. Generally the majority of the students surveyed accepted CAAS as learning tool for mathematics and were willing to use it if it is available on the Web. Students perceived feedback on knowledge level as useful in learning. However, finding on what student did with feedbacks were rather discouraging and suggestion by Black and William (1998) is worth considering for effective feedback that promotes learning. This study has produced some encouraging empirical evidences to support the use of CAAS using the NRET scoring method as assessment tool.

References:

Abu-Sayf, F. K. (1979). Recent development in the scoring of multiple-choice items. *Educational Review, 31*, 269-270.

Aiken, L., Jr. (1968). Weighting and guessing on varieties of the multiple-choice test. *Educational and Psychological Measurement, 28*, 1087-1101.

Angoff, W. H. (1989). Does guessing really help? *Journal of Education Measurement, 26*, 323-336.

Barnett, J. E. & Hixon. J. E. (1997). Effect of grade level and subject on student test score prediction. *The Journal of Educational Research, 90*, 170-174.

Bennett, R. F. (2001). How the Internet will help large-scale assessment reinvent itself. *Education Policy Analysis Archives, 9*(5). Retrieved April 24, 2002, from http://epaa.asu.edu/epaa/v9n5.html.

Black, P. & William, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51

Bransford, J. D., Brown, A. L. & Cocking, R. R. (1999). *How people learn*. Washington, D.C.: National Academy Press.

Brookheart, S. M. (2001). Successful students' formative and summative uses of assessment information. *Assessment and Evaluation in Higher Education, 8*(2), 154-169.

Chevalier, S. A. (1998, April). A review of scoring algorithms for ability and aptitude tests. Paper presented at *the Annual Meeting of Southwestern Psychological Association*, New Orleans, LA.

Claudy, J. G. (1978). Biserial weights: A new approach to test item option weighting. *Applied Psychological Measurement, 2*, 25-30.

Coombs, C. H. (1953). On the use of objective examinations. *Educational and Psychological Measurement, 13*, 308-310.

Coombs, C. H., Milholland, J. E. & Womer, F. B. (1956). The assessment of partial knowledge. *Educational & Psychological Measurement, 16*, 13-37.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 129-138.

De Finetti, B. (1965). Methods for discriminating levels of partial knowledge. *British Journal of Mathematics and Statistical Psychology, 18*, 87-123.

Devine, T. G. & Meagher, L. D. (1989). *Mastering study skills: A student guide*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Ding, L. (1998, Sept. 25-26). Revisiting assessment and learning: Implications of students' perspectives on assessment feedback. Paper presented to *Scottish Educational Research Association Annual Conference*, University of Dundee.

Dirkwager, A. (1996). Testing with personal probabilities: Eleven year olds can correctly estimate their personal probabilities. *Educational and Psychological Measurement, 56*, 957-971.

Downing, S. M. (2003). On guessing corrections. *Medical Education, 38*, 113.

Dressel, P. L. & Schmid, J. (1953). Some modifications of the multiple-choice item. *Educational and Psychological Measurement, 13*, 574-595.

Fleischman, J. (2001). Approaches and strategies for assessment: Project-based learning using technology. *Converge Journal, 2*, 38-40.

Gilman, D. A. & Ferry, P. (1972). Increasing test reliability test through self-scoring procedures. *Journal of Educational Measurement, 9*, 205-207.

Grant, P. (1989). *Reading and study skills*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Haladyana, T. M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Erlbaun Associates Inc.

Haladyana, T. M. (1997). *Writing test items to evaluate higher order thinking*. Boston: Allyn and Bacon.

Hambleton, R. K., Roberts, D. M. & Traub, R. E. (1970). A comparison of reliability and validity of two methods of assessing partial knowledge on multiple-choice tests. *Journal of Educational Measurement, 7*, 75-82.

Holmes, P. (2002). *Multiple evaluations versus multiple choice as testing paradigm—Feasibility, reliability and validity in practice.* (Doctoral Dissertation, University of Twente)

Hopkin, K. D., Stanley, J. C. & Hopkin, B. R. (1990). *Educational and psychological measurement and evaluation (7th ed.).* Englewood Cliffs. NJ: Prentice Hall.

Hounsell, D. (1987). Essay writing and the quality of feedback. In: J. T. E. Richardson, M. W. Eysenck & D. Warren-Piper. (Eds.). *Student learning: Research in education and cognitive psychology.* Milton Keynes: Open University Press and Society for Research into Higher Education.

Hutchinson, T. P. (1985). Predicting performance in variants of the multiple-choice test. Paper presented at *the 4th Annual Meeting of Psychometric Society and the Classification Societies*. Cambridge. (ERIC No. ED 263 177)

Jackson, M. (1995). *Making the grade, the formative evaluation of essays*. Retrieved from http://ultibase.rmit.edu.au/Articles?jacks1.html.

Jaradat, D. & Tollefson, N. (1988). The impact of alternative scoring procedure for multiple-choice item on test reliability, validity, and grading. *Educational and Psychological Measurement, 48,* 627-635.

Kellaghan, T., Madaus, G. F. & Airasian, P. W. (1982). The effect of standardized testing. In: G. F Madaus & D. F. Stufflebeam. (Eds.). *Evaluation in education and human services*. Boston: Kluwer-Nijhoff Publishing.

Laborde, C. (1991). The computer as part of learning environment: The case of geometry. In: C. Keitel & K. Ruthven. (Eds.). *Learning from computers: Mathematics education and technology*. Berlin Heidelberg, Germany: Springer-Verlag, 48-67.

Lockwood, F. (2001). Innovation in distributed learning: Creating the environment. In: F. Lockwood & A. Gooley. (Ed.). *Innovation in open and distance learning: Successful development of online and Web-based learning*. London: Kogan Page Limeted, 1-14.

Lowry, S. R. (1979). The effect of misinformation on item discrimination indices and estimation priorities of multiple-choice test scores. Paper presented at *the Annual Meeting of the American Educational Research Association*, San Francisco, USA.

Lukin, L. E. (1989). *The psychometric properties of confidence testing as a function of computerized administration, training, and development of probabilistic reasoning*. (Unpublished doctoral dissertation, University of Nebraska)

Maguire, T. O., Hatte, J. & Haig, B. (1994). Construct validity and achievement assessment. *The Alberta Journal of Educational Research, XL*(2), 109-126.

Mehrens, W. A. & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology*. Toronto: Holt, Rinehart, and Winston, Inc

Millman, J., Bisghop, C. H. & Ebel, R. L. (1965). An analysis of test wiseness. *Educational and Psychological Measurement, 25,* 707-726.

Ndalichako, J. L. (1997). *Comparison of number right, item response, and finite state approaches to scoring multiple-choice items*. (Unpublished doctoral dissertation, University of Alberta)

Ndalichako, J. L. & Rogers, W. T. (1997). Comparison of finite state theory, classical test theory and item response theory in scoring multiple- choice items. *Educational and Psychological Measurement, 57*(4), 580-589.

Nedelsky, L. (1954). Ability to avoid gross error as a measure of achievement. *Educational and Psychological Measurement, 14*, 459-472.

Norman, G. R. (1997). Assessment in problem-based learning. In: D. Boud & G. Feletti. (Eds.). *The challenge of problem-based learning (2nd ed.)*. London: Kogan Page, 263-266.

Oh, H. J. (2004 April, 12th-16th). Reasoning test taker' guessing strategy and their understanding of formula scoring. Paper presented at *Annual Meeting of the American Educational Research Association (AERA)*, San Diego, CA.

Paris, S. G. & Winograd, P. (1990). How metacognition can promote academic learning and instruction. In: B. F. Jone & L. Idol. (Eds.). *Dimensions of thinking and cognitive instruction*. Hillsdale, NJ: Erlbaum, 15-52.

Paul, J. (1994). Improving education through computer based alternative assessment methods. In: G. Cockton, S. W. Draper & G. R. S. Weir. (Eds.). *People and computers (Vol. 9)*. Cambridge: Cambridge University Press.

Roger, W. T. & Wilson, C. (1993). *The influence of test-wiseness upon performance on high school students on Alberta education's diploma examinations*. (Alberta Education, Student Evaluation Branch, Contract No. 91-0143). Edmonton, AB: University of Alberta.

Rogers, W. T. & Bateson, D. J. (1991a). The influence of test-wiseness on performance of high school seniors on school learning examinations. *Applied Measurement in Education, 4*(2), 159-183.

Rogers, W. T. & Bateson, D. J. (1991b). Verfication of the model of test-taking behavior of high school seniors. *Journal of Experimental Education, 59*, 331-350.

Rothman, R. (1995). *Measuring up: Standard, assessment, and school reform*. San Francisco: Jossey-Bass Publishers.

Sax, G. (1989). *Principles of educational and psychological measurement and evaluation (3rd ed.)*. Belmont, CA: Wadsworth.

Shepard, J. F. (1982). *The Houghton Mifflin study skills handbook*. Boston, Massachusetts: Houghton Mifflin Company.

Shepard, J. F. (1990). *College study skills (4th ed.)*. Boston, Massachusetts: Houghton Mufflin Company.

Thissen, D. (1976). Information in wrong response to the Raven progressive matrices. *Journal of Educational Measurement, 13*, 201-214.

Thissen, D., Steinberg, L. & Fitzpatrick, A. R. (1989). Multiple choice models: The distractors are also part of item. *Journal of Educational Measurement, 26*(2), 161-177.

Tomkowiz, J. T. (2000). *Comparison of ability estimates from dichotomously and nominally scored testwise susceptible and non-susceptible items*. (Unpublished doctoral dissertation, University of Alberta)

Trow, W. C. (1923). The psychology of confidence. *Archives of psychology, 65*, 1-47.

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement, 26*(2), 191-208.

Wotjas, O. (1998). Feedback? No, just give us the answers. *Times Higher Education Supplement, 25*.

(Edited by Max and Jean)