Research and Practice in Technology Enhanced Learning
Vol. 9, No. 2 (2014) 283－300
© Asia-Pacific Society for Computers in Education

ROBUSTNESS OF NUMBER RIGHT ELIMINATION TESTING (NRET) SCORING METHOD FOR MULTIPLE-CHOICE ITEMS IN COMPUTER-ADAPTIVE ASSESSMENT SYSTEM (CAAS)

SIE-HOE LAU

Universiti Teknologi MARA, Malaysia
lausiehoe@sarawak.uitm.edu.my

KIAN-SAM HONG

Universiti Malaysia Sarawak, Malaysia
hksam@fcs.unimas.my

PAUL NGEE-KIONG LAU

Universiti Teknologi MARA, Malaysia
plnk@sarawak.uitm.edu.my

HASBEE USOP

Universiti Malaysia Sarawak, Malaysia
uhasbee@fcs.unimas.my

This paper compares the robustness of the Number Right Elimination Testing (NRET) scoring method for multiple-choice items in Computer-Adaptive Assessment System (CAAS) with two existing scoring methods: Number Right (NR) and Elimination Testing (ET). The NRET scoring method is more reflective of the reality at a workplace that credits partial knowledge and penalizes guessing and detects misconceptions. Quasi-experimental research design was employed where error due to scoring was the prime focus and the scoring method was the main manipulated variable. A total of 449 Form Two students in 19 Malaysian secondary schools participated in the study. The robustness of the NRET method was evaluated twice; one using mathematics items and another using science items. In addition, students' perceptions of NRET and CAAS were also studied and discussed. The results showed that the NRET method is more efficient in estimating students' ability with the NRET scores having higher reliability and lower Standard Error of Measurement. Furthermore, the results showed that the test length could be shortened while retaining the desired reliability if the NRET method was used. They also showed that the NRET scores were similar to the ET scores but were different from the NR scores. The findings on students' perceptions of the NRET method rated it as a practical scoring method and indications of students' willingness to use CAAS if it was available.

Keywords: Multiple-choice; scoring method; partial knowledge; misconception.

1. Introduction

Assessment is a very important component of the educational process because it measures students' learning. Over the last decade, the use of information and communication technology has grown rapidly and this offers enormous prospect for innovations in testing and assessment. Nonetheless, according to Zimmaro (2003), testing continues to be predominantly paper-and-pencil based. Buccino (2000) recommended changing the traditional mode of assessment to the alternative of Computer-Based Testing (CBT). However, the majority of the CBTs are mere replications of paper-and-pencil format to computer-based format which use conventional multiple-choice items.

Multiple-choice items are scored using the conventional Number Right (NR) method where students are instructed to choose an option as the answer and one point is awarded for each correct answer. This method encourages guessing, fails to credit partial knowledge and cannot detect misconceptions. Full knowledge and lucky guesses are lumped as correct whereas partial knowledge, absence of knowledge, and misconceptions are lumped as wrong (Lau, 2010). According to Richard and Joseph (2013), some students are simply better multiple-choice test takers than others, and this ability can translate to higher scores even in subject areas where they have little knowledge. This may skew assessment results and hide valuable information about a student's state of knowledge comprehension. Thus, a student's mental model may remain inadequately sampled, or in some cases, mis-sampled, leading to inappropriate instructional interventions (Moore, 2007).

In addition, the Number Right method of grading multiple-choice items contradicts with the reality in a workplace. Under this method, we are signalling to students that partial knowledge is not important, guessing is an acceptable practice and it is all right to have misconceptions. However in reality, we rarely have the opportunity to guess in our decision making. Furthermore, many important and irreversible decisions are mostly made based on partial knowledge because time does not permit us to delay decision making till full knowledge is attained. Errors due to misconceptions can be disastrous in critical fields such as medicine, aerospace and engineering. In reality, a person could face termination, lawsuit or jail sentence for errors made due to misconceptions.

The Elimination Testing (ET) method was proposed by Coombs (1953) as an alternative to the Number Right method. Elimination Testing requires students to cross or eliminate any option they consider incorrect. One point is awarded for each wrong option eliminated but a penalty of -3 is given if the correct option is crossed out. A study by Bradbard and Green (1986) found that ET decreases guessing, captures partial knowledge and detects misconceptions. However, the ET test instructions are confusing despite prior practice (Jaradat & Tollefson, 1988). Additionally, it is also conflicting where students are taught to solve for the correct answer but being assessed on their ability to identify the wrong answer (Lau, 2010).

To counter the weakness of the NR and ET methods, Lau, Hong, Lau, and Usop (2009) proposed the Number Right Elimination Testing (NRET) method which is a hybrid of Number Right (NR) and Elimination Testing (ET) scoring methods. Students must choose one option as the correct answer and, for the remaining options, they have the choice of crossing out as definitely wrong or mark them as unsure. One point is awarded for each wrong option eliminated and a penalty of -3 is given if the correct answer is eliminated. In addition, one point is awarded if the answer chosen is correct. No point is given for selecting unsure for an option. The test instruction for the NRET method is simple and familiar to students. Studies by Lau et al. (2009, 2011) and Lau (2010) found that the NRET method is able to minimize guessing, credit partial knowledge and detect misconceptions.

Although theoretically the NRET method seems more robust than the NR or the ET method, it does not imply that the NRET method is ready to be used. Improvement in efficiency in ability estimation must be demonstrated, and empirical evidence on its practicality is required.

2. Purposes of The Study

The purposes of this study were to develop a Computer-Adaptive Assessment System (CAAS) using the Number Right Elimination Testing (NRET) for multiple-choice items and evaluate the robustness of the NRET method as compared to the NR and the ET methods. The robustness of the NRET method was evaluated twice, namely, using secondary mathematics items and secondary science items. In addition, students' views on the NRET and CAAS were also evaluated. Specifically, this study was designed to answer the following three research questions:

RQ1: What is the efficiency of the NRET method in CAAS for estimating students' ability as compared to the NR and the ET methods?
RQ2: How similar are scores of NRET, NR, and ET?
RQ3: What are the students' perceptions of NRET and CAAS?

3. Related Review of The Literature

Oral examination was the primary means of educational testing before the mid-nineteenth century. Subsequently, written test in the form of essay questions was introduced to replace oral examination. However, research in the early part of the twentieth century clearly demonstrated that essay tests tended to be highly subjective and unreliable. These findings motivated educators to develop more objective educational measurements. One of the first reported uses of multiple-choice test occurred in 1917 for the selection and classification of military personal for the United States Army (Ebel, 1979). Today, multiple-choice tests are the most highly regarded and widely used type of objective test for measurement of knowledge, ability, or achievement (Ben-Simon, Budescu, & Nevo, 1997; Lee & Winke, 2013).

3.1. *Number Right*

Number Right (NR) is the most commonly used method for scoring multiple-choice items because of its simple test instructions. It requires students to pick only one option as the answer, and one point is awarded for the correct answer and zero point for the wrong answer. However, students have no opportunity to be credited for partial knowledge. Full knowledge and lucky guesses are lumped as correct whereas partial knowledge, absence of knowledge, and misconceptions are lumped as wrong (Lau, 2010). In addition, lucky guesses cannot be distinguished from correct answers based on knowledge (Bar-Hillel, Budescu, & Attali, 2005). The NR method assumes that all correct answers are the results of full knowledge and all wrong answers reflect an absence of knowledge. Thus, it can only distinguish between full knowledge and an absence of knowledge.

3.2. *Elimination Testing*

Elimination Testing (ET) method requires students to cross out all alternatives that they consider incorrect. One point is awarded for each wrong answer eliminated but if the correct answer is crossed out, the student receives a penalty of 1-k points, where k is the number of options. Hence, the score for a multiple choice item with four options can vary from -3 to 3. With these scores, student knowledge can be categorized into full knowledge (3), partial knowledge (1 or 2), absence of knowledge (0), partial misconception (-1 or -2), and full misconception (-3).

Coombs, Miholland, and Womer (1956) compared the ET method with the NR method and reported a higher reliability for ET scores. However, a body of similar studies done in the 1970s (Collet, 1971; Hakstian & Kansup, 1975; Traub & Fisher, 1977) indicated that reliability for the ET scores were equal to or greater than the NR scores but the improvement in reliability was not statistically significant. Bradbard and Green (1986) further suggested that ET decreases guessing, and Bradbard, Parker, and Stone (2004) reported that there was no loss of reliability compared to the NR scoring. However, the study by Bradbard et al. evidenced that guessing was reduced, partial knowledge was measured, and ET scoring provided a finer discrimination with respect to students' knowledge.

3.3. *Number Right Elimination Testing (NRET)*

Number Right Elimination Testing (NRET) method is a hybrid of Number Right (NR) and Elimination Testing (ET) scoring methods for multiple choice questions (Lau et al., 2009). Students must choose one option as the correct answer, and for the remaining options, they have the choice of crossing out as definitely wrong or mark them as unsure. One point is awarded for each wrong option eliminated and a penalty of -3 is given if the correct answer is eliminated. In addition, one point is awarded if the answer chosen is correct. No point for selecting the option as unsure. Table 1 shows the comparison of

Table 1. Comparison of item scores for NR, ET, and NRET method by level of knowledge for all possible responses patterns.

Knowledge level	Response pattern	NR	ET	NRET
Full knowledge	Answer correct Identify 3 wrong options as incorrect	1	3	4
Partial knowledge	Answer correct Identify 2 wrong options as incorrect	1	2	3
	Answer correct Identify 1 wrong option as incorrect	1	1	2
	Answer wrong Identify 2 wrong options as incorrect	0	2	2
	Answer wrong Identify 1 wrong option as incorrect	0	1	1
	Answer correct Identify none wrong option as incorrect	1	0	1
Absence of knowledge	Answer wrong Identify none wrong option as incorrect	0	0	0
Partial misconception	Identify the answer as incorrect Identify 2 wrong options as incorrect	0	-1	-1
	Identify the answer as incorrect Identify 1 wrong option as incorrect	0	-2	-2
Full misconception	Identify only the answer as incorrect	0	-3	-3

Note: NR = Number Right, ET = Elimination Testing, NRET = Number Right Elimination Testing.

item scores for the NR, the ET and the NRET methods for multiple-choice items with four options.

Based on Table 1, it is clear that the NRET method gives a better representation of students' actual knowledge. The NRET gives more flexibility to students to express their knowledge. In addition, it also credits partial knowledge, penalizes guessing and detects misconceptions, which is more reflective of the reality at a workplace.

3.4. *Computer-Based Testing (CBT)*

The emergence of computer technology in the 1980s provides another opportunity for researchers to address the problem of guessing and crediting partial knowledge in multiple-choice testing. According to Holmes (2002), one of the earliest reported experiments with computer-based testing was in 1965 by Shuford. As the learning environment continues to evolve in the digital age, there is a growing interest in the development of the CBT (Baucer & Anderson, 2000; Boettcher & Conrad, 1999). As a result, a number of innovative CBT has been proposed. Early CBT mostly adopted and modified the various existing scoring methods. For instance, Baker (1968), Dirkzwager (1975, 1993, 1996), Holmes (2002), Shuford (1965) and Sibley (1974) adopted the probability measurement method while Chambers (1990), Farrell and Leung (2004), Klinger (1997), Paul (1994), and Rippey (1986) used a modified version of the confidence weighting scoring method.

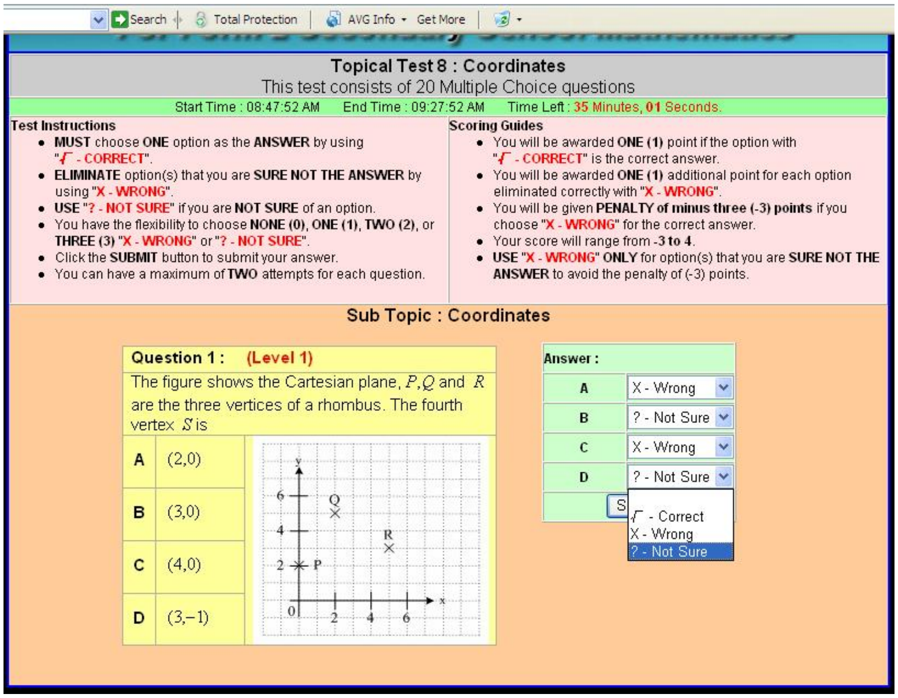
4. Research Methodology

The first part of this section describes the features of the Computer-Adaptive Assessment System (CAAS) which uses the NRET method of scoring multiple-choice items. This is followed by the second section describing the research design to answer the three research questions in this study. The last section touches on the data analyses procedures.

4.1. *Computer-Adaptive Assessment System (CAAS with NRET)*

Computer-Adaptive Assessment System (CAAS) is an online formative assessment system for multiple-choice items that uses NRET as the scoring method. For any option of a multiple choice item, students have the choice of choosing "√Correct", "X Wrong" or "? Not Sure". Students must choose one option as "correct". However, they have the flexibility in choosing none, one, two or three as "X Wrong" or "? Not Sure". Figure 1 shows the starting interface of CAAS.

CAAS ensures that students' responses conform to the NRET response mode. A reminder would appear if the students did not follow the NRET test instructions.



Topical Test 8 : Coordinates
This test consists of 20 Multiple Choice questions

Start Time : 08:47:52 AM End Time : 09:27:52 AM Time Left : 35 Minutes, 01 Seconds.

Test Instructions

- **MUST** choose **ONE** option as the **ANSWER** by using "√ - CORRECT".
- **ELIMINATE** option(s) that you are **SURE NOT THE ANSWER** by using "X - WRONG".
- **USE** "? - NOT SURE" if you are **NOT SURE** of an option.
- You have the flexibility to choose **NONE (0)**, **ONE (1)**, **TWO (2)**, or **THREE (3)** "X - WRONG" or "? - NOT SURE".
- Click the **SUBMIT** button to submit your answer.
- You can have a maximum of **TWO** attempts for each question.

Scoring Guides

- You will be awarded **ONE (1)** point if the option with "√ - CORRECT" is the correct answer.
- You will be awarded **ONE (1)** additional point for each option eliminated correctly with "X - WRONG".
- You will be given **PENALTY of minus three (-3) points** if you choose "X - WRONG" for the correct answer.
- Your score will range from **-3 to 4**.
- **USE** "X - WRONG" **ONLY** for option(s) that you are **SURE NOT THE ANSWER** to avoid the penalty of (-3) points.

Sub Topic : Coordinates

Question 1 : **(Level 1)**

The figure shows the Cartesian plane, P,Q and R are the three vertices of a rhombus. The fourth vertex S is

A	(2,0)
B	(3,0)
C	(4,0)
D	(3,−1)

Answer :

A	X - Wrong
B	? - Not Sure
C	X - Wrong
D	? - Not Sure

Figure 1. Interface of CAAS.

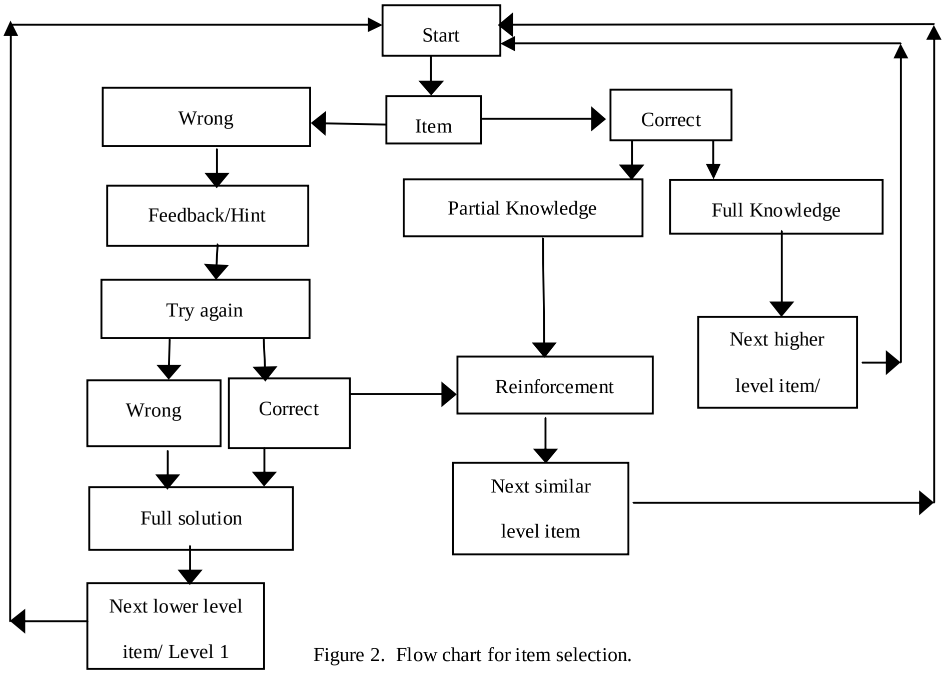


Figure 2. Flow chart for item selection.

Feedback on the NRET score is given so that students are aware of the point gained or penalty for each response pattern. In addition, CAAS is adaptive in nature where selection of the successive items depends on students' performance of the item presented. To support continuous learning, hints and timely feedback are given. The flow chart for item selection and the timing of feedback is as shown in Figure 2.

4.2. *Research design*

The quasi-experimental research design was employed in this study where error due to scoring was the prime focus. The scoring method was the main manipulated variable. Scoring methods compared in the present study were the NR, the ET, and the NRET. The main response variables were test scores reliability, standard error of measurement (SEM), the h-statistics, mean absolute difference (MAD), standard deviation (SD) and the correlations between scores from the three different scoring methods. In this study, students' responses to the same tests using the NRET test instructions were used to calculate the NR, the ET, and the NRET scores. Since the item responses for the same tests were used to calculate the NR, the ET, and the NRET method scores, other errors such as errors due to guessing, distraction in testing situation, administration, content sampling and fluctuations in the individual student's behavior were held constant. Thus, any differences in the response variables were solely due to difference in scoring. The approach of obtaining different scores from the same test and using one common test instructions had been used by past researchers such as Holmes (2002), Kansup (1973), and Ndalichako (1997). A survey questionnaire was also used to gauge students' perceptions of CAAS using NRET.

4.3. *Participants*

A total of 449 Form Two students from 19 secondary schools in Sarawak, Malaysia participated in this study. There were 255 female students and 194 male students aged between 13 and 14 years. They had gone through six years of primary education and at least one year of secondary education. The medium of instruction for mathematics and science was English.

4.4. *Research instruments*

One set of 40 mathematics multiple-choice items and another set of 40 science multiple-choice items were used to evaluate the robustness of NRET in estimating students' ability. These items were adopted and modified from the mathematics items for the eighth grade (13 years old) of the Trends in International Mathematics and Science Study (TIMSS) for the year 2003. A survey questionnaire was used to gauge students' perceptions of CAAS using NRET. The first eight questions were on perceptions toward the NRET test instructions. This was followed by four questions on the NRET scoring system and four questions on perceptions of CAAS as a learning tool. Each statement had five optional choices: "Strongly Agree", "Agree", "Neutral", "Disagree" or "Strongly Disagree".

4.5. *Data collection procedures*

Permissions for conducting the study were obtained from the Educational Planning and Research Division of the Malaysian Ministry of Education and Sarawak State Education Department before meeting the principals of the selected secondary schools to identify students who were willing to be participants of the study. Trainings were conducted before the final data were collected. The researchers visited the participating schools to conduct road shows to promote and conduct briefings on CAAS to teachers and students. Guidelines on CAAS were distributed and students registered as CAAS users, logged-in and tried the sample exercise in the school computer laboratories. Five topical exercises and two mathematics tests were uploaded to CAAS for training purposes; the trainings were carried out in the school computer laboratories under the supervision of the mathematics teachers appointed as the research assistants of the study. However, many students faced the problem of slow and unstable Internet connectivity in their schools. As an alternative, students were permitted to complete the exercises and tests online at home if they were unable to complete them in schools due to inadequate computer facilities and poor Internet connectivity. Their parents were duly informed and students were given four months to complete these exercises. The final mathematics and science tests were conducted in the computer laboratories of each school using 15 laptop computers linked to a local server. This method of data collection was necessary to avoid interruption due to poor Internet connectivity of certain participating schools. The participants were quarantined and sat for the final test in batches.

4.6. *Data analysis procedures*

The first research question focused on the efficiency of the NRET scoring method in estimating students' ability. Efficiency referred to accuracy and reliability. It was evaluated by three indices consisting of the internal consistency, the standard error of measurement (SEM) and the index of "relative information" or the h-statistics. Internal consistency was measured by the Cronbach Alpha reliability coefficient. SEM was the standard deviation of the discrepancies between a student's true score and the observed score over an infinite number of repeated testing (Crocker & Algina, 1986). The statistic, h, indicated how much the NR test would have to be increased in length in order to obtain the same reliability when the NRET or the ET method was used. The positive h value of greater than one indicates greater efficiency. For instance, if the h statistic for a new method is 1.1, it means that a 40-item test using the NR method gives similar reliability as a 36-item ($40/1.1 \cong 36$) test using the new method. Thus, efficiency is increased by decreasing the number of items used while maintaining the desired reliability.

The second research question compared the scores from the NRET, the NR, and the ET methods to determine any similarity. The scores based on the NRET, the NR and the ET scoring methods were transformed into standardized score distribution with mean 50 and standard deviation of 10 (Glass & Hopkins, 1984) before comparisons were conducted. Comparisons were done using two procedures. The first procedure was to examine if the students were ranked differently by the different methods. Correlations between scores from NR, ET, and NRET methods were computed while Pearson-product moment correlations were compared. Higher correlation coefficient would imply closer agreement among scores. The second procedure was to examine if the scores for the students were equal in absolute sense. According to Ndalichako (1997), the extent to which scores from each method provided similar or different information was determined by using the mean absolute deviation (MAD) and standard deviation (SD). Smaller values of MAD and SD mean more similarities among scores.

To ensure a more stable and valid values of the indices being compared, simulations were carried out using the actual data by varying the sub-test length. Simulations were done by selecting at random the number of items according to the length of each sub-test. There were a total of ten sub-test lengths ranging from 36-item test to 20-item test. Five simulations were conducted for each sub-test length and the average value of each index was computed for comparison.

Responses to the survey questionnaires were used to answer the third research question. Descriptive statistics such as mean and standard deviation were used. The response to each of these statements was coded 5 (strongly agree) to 1 (strongly disagree).

5. Results

The results for the three research questions are presented in this section.

Research question 1: What is the efficiency of the NRET method in CAAS for estimating students' ability as compared to the NR and the ET methods?

Table 2 shows the reliability analyses for test scores under the three different scoring methods. The average reliability of the NRET scores is the highest for all sub-test lengths for both mathematics and science tests. The NR scores have the lowest reliability.

Table 3 shows the SEM values for the mathematics and science tests. The SEM values of the NR scores are the largest for both tests. This is consistent with the results of the NR scores having the lowest reliability (see Table 2). This is followed by the NRET scores. The ET scores have the smallest SEM values. The differences in the SEM values between ET and NRET are relatively small when compared with the difference between NR and ET or NR and NRET. Thus, measurement error of the test scores would be larger if the NR method was used. Accordingly, the NR scores are less reflective of the students' true scores. The SEM values for both tests have also increased with decrease in the sub-test length. This trend is expected as the number of items decreased, the reliability of the test scores decreased. Consequently, the SEM values also increased.

Table 4 shows that the *h* statistics for ET and NRET are greater than 1.0 for both tests. This indicates that the tests could be shortened while retaining the desired reliability if the ET or the NRET method was used instead of the conventional NR method. The *h* statistics for ET and NRET are consistently in the range of 1.063 to 1.143 for both tests. This means that for a 40-item test under the NR method, the number of items can be reduced to between 38 items (40/1.063 ≅ 38) to 35 items (40/1.143 ≅ 35) under the ET or the NRET method. Further analysis revealed that average *h* statistics for the NRET method was consistently higher than the ET method. This indicates that the NRET method is more efficient than the ET method.

Table 2. Average reliability of tests scores.

Sub-test length	Average Cronbach alpha value					
	Mathematics			Science		
	NR	ET	NRET	NR	ET	NRET
36	0.907	0.913	0.914*	0.853	0.868	0.869*
35	0.903	0.909	0.911*	0.849	0.864	0.865*
33	0.899	0.905	0.906*	0.845	0.860	0.861*
31	0.893	0.899	0.900*	0.838	0.853	0.855*
29	0.887	0.893	0.895*	0.828	0.845	0.846*
27	0.877	0.886	0.887*	0.812	0.829	0.830*
25	0.868	0.877	0.879*	0.799	0.818	0.819*
23	0.852	0.861	0.864*	0.787	0.808	0.809*
21	0.847	0.855	0.857*	0.770	0.792	0.793*
20	0.842	0.854	0.854	0.757	0.775	0.778*

Note: * denotes the highest value in its respective category.

Table 3. Average standard error of measurement.

Sub-test length	Average SEM					
	Mathematics			Science		
	NR	ET	NRET	NR	ET	NRET
36	6.74	4.39#	4.66	7.35	4.74#	5.02
35	6.83	4.47#	4.71	7.46	4.81#	5.10
33	6.99	4.58#	4.86	7.65	4.95#	5.24
31	7.28	4.74#	5.04	7.92	5.12#	5.41
29	7.55	4.97#	5.24	7.98	5.18#	5.48
27	7.87	5.14#	5.45	8.67	5.64#	5.97
25	8.18	5.35#	5.65	9.06	5.86#	6.22
23	8.51	5.59#	5.88	8.80	5.73#	6.06
21	8.88	5.72#	6.08	9.74	6.26#	6.64
20	9.17	5.84#	6.24	9.71	6.29#	6.65

Note: # denotes the lowest value in its respective category.

Table 4. The *h* statistic.

Sub-test length	Index of "relative information" (*h* statistic)			
	Mathematics		Science	
	ET	NRET	ET	NRET
36	1.076	1.090	1.133	1.143
35	1.073	1.100	1.130	1.140
33	1.070	1.083	1.127	1.136
31	1.067	1.078	1.122	1.140
29	1.063	1.086	1.131	1.140
27	1.090	1.101	1.127	1.135
25	1.084	1.105	1.126	1.134
23	1.076	1.104	1.139	1.146
21	1.065	1.083	1.137	1.144
20	1.098	1.098	1.106	1.125
Average	1.076	1.093	1.128	1.138

Research question 2: How similar are scores of NRET, NR, and ET?

The results in Table 5 show that the correlations between the ET and the NRET scores are near perfect at 0.999 for both tests. In contrast, the correlations between the NR and the ET or the NRET scores range from 0.975 to 0.986. The near perfect correlation in the ET and the NRET scores is also reflected by their lower MAD (0.19 and 0.28) and lower SD (0.26 and 0.33). This indicates that the ET and the NRET scores give similar ranking to students (near perfect correlation). In addition, their scores are similar in absolute sense (lower MAD and SD). On the other hand, larger MAD (0.78 and 1.14) and SD (1.09 and 1.39) among the NR and the NRET scores indicate that the NR and NRET scores are not similar in absolute sense. The lower correlation coefficient between the NR and the NRET scores also shows that there is a slight difference in the ranking of students for the NR and the NRET scores.

Table 5. Correlation, MAD, and SD among NR, ET and NRET scores.

Scoring method	NR		ET		NRET	
	Math	Science	Math	Science	Math	Science
NR	-	-	0.986	0.975	0.981	0.984
ET	0.98 (1.35)	1.42 (1.73)	-	-	0.999	0.999
NRET	0.78 (1.09)	1.14 (1.39)	0.19 (0.26)	0.28 (0.33)	-	-

Research question 3: What are the students' perceptions of NRET and CAAS?

A considerable number of students took a neutral stand in their opinions on the statements regarding the NRET test instructions as shown in Table 6. However, the number of students who agreed (SA and A) were more than those who disagreed (D and SD) for all except the seventh statement which is a negatively worded statement. Thus, in general, the majority of the participating students that provided their opinions perceived the NRET test instructions as familiar and were actually practicing it when answering the multiple choice items.

The results of the students' perceptions of the NRET scoring system is shown in Table 7. As before, a significant number of students did not provide an opinion for or against the NRET scoring system. Nonetheless, many students agreed that they could calculate the NRET marks, the scoring system was fair, and the NRET method of marking should be used for all tests. However, most of the students still preferred the NR method of marking.

Table 6. Students' perceptions of the NRET test instructions.

	Responses						
Statements	SA	A	N	D	SD	M	s.d.
The NRET test instructions are easy to follow.	92	190	145	19	3	3.7	0.8
I am familiar with the NRET test instructions.	59	168	196	24	2	3.6	0.8
I cross out options which are wrong before choosing the answer.	76	155	151	53	14	3.5	1.0
I can find options which are "Sure Wrong" for mathematics items.	44	135	205	60	5	3.4	0.9
I can find options which are "Sure Wrong" for science items.	30	123	232	57	7	3.3	0.8
I can decide whether the option is "Sure Wrong" or "Not Sure".	29	156	214	42	8	3.4	0.8
The NRET method is not suitable for mathematics.	23	53	186	147	40	2.7	0.9
The NRET method is more suitable for factual subjects (such as Science or Geography).	46	115	217	52	19	3.3	0.9

Note: SA = Strongly Agree, A = Agree, N = Neutral, D = Disagree, and SD = Strongly Disagree, M = Mean, s.d. = Standard Deviation.

Table 7. Students' perceptions of the NRET scoring system.

	Responses						
Statements	SA	A	N	D	SD	M	s.d.
I am able to calculate NRET marks.	33	150	186	62	18	3.3	0.9
The NRET scoring system is fair.	60	169	179	34	7	3.5	0.8
I prefer NR method of marking.	37	106	231	49	26	3.2	0.9
The NRET method of marking should be used for all tests.	43	113	206	69	18	3.2	1.0

Note: SA = Strongly Agree, A = Agree, N = Neutral, D = Disagree, and SD = Strongly Disagree, M = Mean, s.d. = Standard Deviation.

Table 8. Perception of CAAS as assessment tool.

	Responses						
Statements	SA	A	N	D	SD	M	s.d.
The exercises in CAAS help me in learning mathematics.	143	219	80	5	2	4.1	0.7
I enjoy learning mathematics using CAAS.	120	172	141	13	3	3.9	0.9
I will always use CAAS if it is available.	63	205	156	18	7	3.7	0.8
I want to try similar exercises for other subject such as science.	90	216	126	12	6	3.8	0.8

Note: SA = Strongly Agree, A = Agree, N = Neutral, D = Disagree, and SD = Strongly Disagree, M = Mean, s.d. = Standard Deviation.

The results of the analyses in Table 8 indicate that the majority of the students viewed CAAS as a helpful learning tool. They enjoyed learning mathematics with CAAS and were willing to use CAAS for learning mathematics and other subjects. Thus, it could be concluded that students perceived CAAS using NRET to be a usable and practical assessment tool.

6. Discussions

The NRET method was found to be more efficient in estimating students' ability. The reliability of the NRET scores was consistently higher than the NR and ET scores. This was also reflected by their lower SEM. In addition, the positive *h* statistics value of greater than one also showed that the test length could be shortened while retaining the desired reliability if the NRET method was used instead of the NR or ET method. The results of the study were consistent with the findings of past studies. Firstly, the item score for NR was either 1 or 0. On the other hand the item scores range for NRET was from -3 to 4. Ma (2004) argued that when test items were scored dichotomously, potential useful information about an individual's level of proficiency that was contained in the other response options was lost. Thus, the precision of measurement was reduced. Frey (1989) found that an increase in scoring range would increase variability among students and, therefore, would increase test reliability. Secondly, one of the major weaknesses of the NR is guessing. The NRET method controls guessing by having penalty instructions. According to Swineford and Miller (1953), and Traub and Hambleton (1972), wild guesses were reduced under penalty instructions. Hence, the NRET method with penalty instructions yielded more precise score. Third, the NR score was less precise because the score awarded did not reflect the actual knowledge state of the students. According to Ben-Simon et al. (1997) and Chang, Lin, and Lin (2007), students' knowledge for a given item could be classified into five categories, namely, full knowledge, partial knowledge, absence of knowledge, partial misconception, and full misconception. This contrasts with the NR method which lumped students' knowledge into just two categories, namely, "correct" and "incorrect". On the other hand, students' knowledge was graded into five levels under NRET. The results from this study are consistent with past studies on other scoring methods which consistently show that the NR method is less efficient than the scoring methods that take into consideration partial knowledge and guessing (Bokhorst, 1986; Hanna, 1975).

The NRET scores were similar to the ET scores. However, the NRET scores were different from the NR scores. The scores for the NRET and ET methods were almost similar in values based on their small MAD. In addition, the high correlation indicated that the ranking of students by test scores were almost identical for both NRET and ET. The strong correlation between NRET and ET was expected. The theoretical rationale behind the NRET was similar to the ET. Both methods had penalty instructions to discourage guessing. Furthermore, there was not much difference in their score range (-3 to 3 for ET and -3 to 4 for NRET). Finally, for both the NRET and the ET methods, students' knowledge level could be classified into five levels. On the other hand, scores yielded by the NR methods were different from the NRET and the ET methods. The difference was expected. The NR scores were scored dichotomously (correct or wrong) and information on the incorrect responses was not captured. According to Bock (1972) and Thissen (1976), scoring method that took into consideration the incorrect responses (as in the case of NRET and ET) yielded nearly twice the information of dichotomous scoring (correct or wrong). Thus, the difference between the NRET, the ET and the NR scored methods were most likely due to credit for partial knowledge and penalty for guessing.

Generally, the students in this study accepted the NRET method as a practical scoring method. The majority of them viewed the NRET test instructions as easy to follow and familiar. In addition, they also felt that the NRET scoring system was fair. This finding contrasted with previous findings where teachers and students often had negative perceptions toward alternative scoring methods (Frey, 1989; Guilford, 1954; Jaradat & Tollefson, 1988). One possible explanation could be the simple and easy-to-follow NRET test instructions which resemble one of the most common tests taking strategy used by students while answering multiple choice items. Another possible reason was the flexibility for students to indicate their knowledge under the NRET. Lukin (1989) suggested that offering students more opportunity to tell what they know may also improve their attitude toward testing. Inflexibility in NR can lead to cynical attitudes and loss of faith in multiple-choice testing by the students (Abu-Sayf, 1979). The students also reported that CAAS helped them learn mathematics and also indicated their willingness to use CAAS if it was available. This was consistent with the findings by Nguyen (2002), and Galbraith and Haines (1998).

7. Conclusions

The ultimate goal of testing is to estimate a person's ability. The accuracy of these scores is important to create a fair assessment that leads to valid inferences about students' ability. The results of the study revealed that the NRET scores were more reliable with smaller SEM as compared to the NR scores. Thus, the NRET scores were more valid and accurate. The students reported that CAAS helped them learn mathematics and they also indicated that they were willing to use CAAS if it was available. Educators should seriously look at the possibility of incorporating or using CAAS or equivalent to supplement the traditional assessment method in the classrooms. Students growing up with computer technology are attracted to computer-based activities. Thus, educators should capitalise on the "pull factor" that already exists to maximise learning.

The above conclusions are based on the analyses conducted in this study. Clearly further research is warranted. Firstly, this study used mathematics and science items and it involved only Form Two students (aged 14 years old) in Malaysia. Further studies conducted across different subject area and student age group would help to clarify the generalizability of the finding of the study. Secondly, training students to be accustomed with the new NRET test instructions and Internet connectivity were among the challenges faced while carrying out this study. Further studies can be conducted to identify other problems and challenges in implementing CAAS in schools.

References

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

Baker, J. D. (1968). *The uncertain student and the understanding computer*. Papers presented at the O.T.A.N. conference (Nice, May 20-24, 1968).

Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple choice test: A case study in irrationality. *Mind and Society*, *3*, 3–12.

Baucer, J. F., & Anderson, R. S. (2000). Evaluating student's written performance in the online classroom. In R. E. Weiss, D. S. Knowlton & B. W. Speck (Eds.), *Principles of effective teaching in the online classroom: New directions for teaching and learning* (pp. 65–71). San Francisco: Jossey-Bass.

Ben-Simon, A., Budescu, D. V., & Nevo, N. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement*, *21*, 65–88.

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

Boettcher, J. V., & Conrad, R. M. (1999). *Faculty guide for moving teaching and learning to the web*. Los Angeles, CA: League for Innovation in the Community College.

Bokhorst, F. D. (1986). Confidence-weighting and the validity of achievement tests. *Psychological Reports*, *59*, 383–386.

Bradbard, D. A., & Green, S. B. (1986). Use of the Coombs elimination procedure in classroom tests. *Journal of Experimental Education*, *54*, 68–72.

Bradbard, D. A., Parker, D. F., & Stone, G. L. (2004). An alternative multiple-choice scoring procedure in a macroeconomics course. *Decision Science Journal of Innovative Education*, *2*(1), 11-26

Buccino (2000). Politics and professional belief in evaluation: The case of calculus renewal. In S. L. Ganter (Ed.), *Calculus Renewal: Issues for undergraduate mathematics education in the next decade* (pp. 121–146). NY: Kluwer Academic/Plenum Publishers.

Chambers, D. B. (1990). *An evaluation of a simplified response confidence testing method for assessing partial knowledge in computer-based tests*. Unpublished doctoral dissertation, University of California, Berkeley.

Chang, S. H., Lin, P. C., & Lin, Z. C. (2007). Measures of partial knowledge and unexpected responses in multiple-choice tests. *Educational Technology & Society*, *10*(4), 95–109.

Collet, L. S. (1971). Elimination scoring: An empirical evaluation. *Journal of Educational Measurement*, *8*, 209–214.

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

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Florida: Holt, Rinehart, & Winston.

Dirkzwager, A. (1975). Computer-based testing with automatic scoring based on subjective probabilities. In O. Lecarme & R. Lewis (Eds.), *Computer in education* (pp. 305–311). Amsterdam: North-Holland Publishing Company.

Dirkzwager, A. (1993). A computer environment to develop valid and realistic predictions and self-assessment of knowledge with personal probabilities. In D. A. Leclercq & J. E. Bruno (Eds.), *Item banking: Interactive testing and self-assessment* (pp. 66–102). Berlin: Springer Verlag.

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

Ebel, R. L. (1979). *Essential of educational measurement (3rded.)*. Englewood Cliffs, NJ: Prentice Hall.

Farrell, G., & Leung, Y. K. (2004). Innovative online assessment using confidence measurement. *Education and Information Technologies*, *9*(1), 5–19.

Frey, A. S. (1989). *A test validation study for capturing partial information in multiple-choice questions using polychotomous scoring*. Unpublished doctoral dissertation, University of San Francisco, California.

Galbraith, P., & Haines, C. (1998). Disentangling the Nexus: Attitudes to mathematics and technology in a computer learning environment. *Educational Studies in Mathematics*, *36*(3), 275–290.

Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology*. London: Allyn & Bacon.

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

Hakstian, A. R., & Kansup, W. (1975). A comparison of several methods of assessing partial knowledge in multiple-choice tests testing procedures. *Journal of Educational Measurement*, *12*, 231–239.

Hanna, G. S. (1975). Incremental reliability and validity of multiple choice tests with an Answer-until-Correct procedure. *Journal of Educational Measurement*, *12*, 175–178.

Holmes, P. (2002). *Multiple evaluations versus multiple-choice as testing paradigm: Feasibility, reliability and validity in practice*. Unpublished doctoral dissertation, University of Twente, Enschede, Holland.

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.

Kansup, W. (1973). *A comparison of several methods of assessing partial knowledge in multiple-choice tests*. Unpublished master thesis, University of Alberta, Edmonton, Canada.

Klinger, A. (1997). *Experimental validation of learning accomplishment* (Technical Report No. 970019). Pittsburgh: Frontiers in Education.

Lau, N. K., Lau, S. H., Hong, K. S., & Usop, H. (2011). Guessing, partial knowledge, and misconception in multiple-choice tests. *Journal of Education Technology & Society*, *14*(4), 99–110.

Lau, S. H. (2010). *Practicality and robustness of Number Right Elimination Testing (NRET) for multiple-choice items in paper-and-pencil testing (PPT) and computer-based testing (CBT)*. Unpublished doctoral dissertation, Universiti Malaysia Sarawak, Malaysia.

Lau, S. H., Hong, K. S., Lau, N. K., & Usop, H. (2009). Improving educational assessment: A computer-adaptive multiple choice assessment using NRET as the scoring method. *US-China Education Review*, *6*(54), 51–60.

Lee, H. S., & Winke, P. (2013). The differences among three-, four, and five-option-item formats in the context of a high-stakes English-language listening test. *Language Testing*, *30*(1), 99–123.

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, Nebraska.

Ma, X. (2004). *An investigation of alternative approaches to scoring multiple response items on a certification examination*. Unpublished doctoral dissertation. University of Massachusetts, Amherst.

Moore, D. R. (2007). Concept acquisition and confidence using a spatial probability measure instrument. *Journal of Educational Multimedia and Hypermedia*, *16*(1), 25–38.

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, Edmonton, Canada.

Nguyen, D. M. (2002). *Developing and evaluating the effects of web-based mathematics instruction and assessment on student achievement and attitude*. Unpublished doctoral dissertation, Texas A&M University, Texas.

Paul, J. (1994). Alternative assessment for software engineering education. In J. L. Diaz-Herrera (Ed.), *Software engineering education* (pp. 463–472). New York: Springer-Verlag.

Richard, B. G., & Joseph, M. L. (2013). Inherent limitation of multiple-choice testing. *Academic Radiology*, *20*(10), 1319–1321.

Rippey, R. M. (1986). A computer program for administering and scoring confidence tests. *Behavior Research Methods, Instruments, and Computers*, *18*, 59–60.

Shuford, E. H., Jr. (1965). *Cybernetic testing* (Report ESD-TR-65-467). Hanscom Field, Bedford, MA: Decision Science Laboratory, L.G.

Sibley, W. L. (1974). *An experimental implementation of computer-assisted admissible probability testing* (Research Report). Santa Monaco: Rand Corporation.

Swineford, F., & Miller, P. M. (1953). Effects of directions regarding guessing on item statistics of a multiple-choice vocabulary test. *Journal of Educational Psychology*, *44*(2), 129–139.

Thissen, D. (1976). Information in wrong response to the Raven Progressive Matrices. *Journal of Educational Measurement*, *13*, 201–214.

Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement*, *1*, 355–369.

Traub, R. E., & Hambleton, R. K. (1972). The effect of scoring instructions and degree of speedness on the validity and reliability of multiple-choice tests. *Educational and Psychological Measurement*, *32*, 737–758.

Zimmaro, D. M. (2003). *Systemic validity of unproctored online testing: Comparison of proctored paper-and-pencil and unproctored web-based mathematics placement testing*. Unpublished doctoral dissertation, Pennsylvania State University, Pennsylvania.