



EVALUATING COMPUTERIZED ADAPTIVE TESTING EFFICIENCY IN MEASURING STUDENTS' PERFORMANCE IN SCIENCE TIMSS

M. A. Samsudin¹, T. SomChut*², M. E. Ismail³

¹School of Educational Studies, Universiti Sains Malaysia

²School of Educational Studies, Universiti Sains Malaysia

³Faculty of Technical and Vocational Education, Universiti Tun Hussein Onn Malaysia

DOI: 10.15294/jpii.v8i4.19417

Accepted: August 6th, 2018. Approved: December 27th, 2019. Published: December 31st, 2019

ABSTRACT

The current standards of assessment are demanding for high level of precision with less time-consuming and personalized opportunities, thus restrict the function of Paper and Pencil Test which has dominated the assessment field for a very long time. The Computerized Adaptive Testing (CAT) is viewed as an alternative testing tool to Paper and Pencil Test as it has adaptive feature which enables it to meet the current standards of assessment. This research is focusing on the evaluation of students' ability in Grade 8 Science Trend in International Mathematics and Science Study (TIMSS) using Computerized Adaptive Testing (CAT) as an alternative instrument to Paper and Pencil Test to investigate whether the implementation of CAT can produce high level of precision with fewer items administered as well as differentiate different academic level among groups of students. CAT was configured in Concerto and was administered on Form 2 and Form 4 students selected through purposive sampling method from secondary schools in northern part of Malaysia. Students' performance was analysed and compared in terms of score (theta value), SEM value and their response toward the items selected in CAT using SPSS. Finding shows that the administration of 20 objective items in fixed length CAT produced SEM ≤ 0.50 indicated that the implemented CAT increased the efficiency of assessment with fewer item administration. The t test showed that there was a significance difference between the two groups' scores in CAT in which Form 4 students had higher ability level than Form 2 students proving that CAT's configuration had been done correctly in Concerto and the test was more suitable in challenging Form 2 students' Science knowledge thus the instrument fulfilled the known-group validity.

© 2019 Science Education Study Program FMIPA UNNES Semarang

Keywords: ability, Computerized Adaptive Testing (CAT), evaluation, TIMSS

INTRODUCTION

In the conventional of conducting assessment, Paper and Pencil Test is used widely as a testing tool in educational field. The items used in this linear instrument are more suitable in testing average ability students (Chuesathuchon & Waugh, 2010). According to Mansoor Al-A'ali

(2007), this instrument is suitable in measuring student's whole performance in a large-scale assessment but it could not provide a precise evidence about individual's ability as the administration of the items is not adaptive thus a student might have to answer very easy or very difficult items which could not provide enough information on the student's ability. Therefore, Computerized Adaptive Testing (CAT) could be an alterna-

*Correspondence Address

E-mail: saphorn.mk@gmail.com

tive testing instrument to Paper and Pencil Test in current assessment practices (Bakker, 2014) as it is an instrument which focuses on measurement of an individual's ability. The adaptivity feature in CAT enables it to select the item's difficulty parameter based on the examinee's level of ability thus producing more precise individual measurement, shorter testing time and faster score reporting (Linden & Glas, 2010). Therefore, this study is done in order to test the mechanism of CAT to ensure it can work well to discriminate students with different level of ability but at the same time, it works in adaptive manner by intelligently sequence the items based on their difficulty levels. This is done by examining the theta values which is the unit of measurement for Science TIMSS achievement produced by CAT. It is hypothesized that if CAT is able to produce range of theta values for students taking CAT, therefore, it can be inferred that CAT works well under the principle of adaptive testing principle. A theoretical assumption is being speculated that students with older age will obtain higher theta values compared to the students with younger age.

Generally, in the beginning of the CAT, the computer would select the first item randomly with medium difficulty parameter as the examinee's ability could not be estimated yet. The examinee's response towards the first item is analysed and his or her current ability is estimated. Maximum Likelihood Estimator (MLE) is the most frequently used method in estimating examinee's ability in CAT because it provides low level of bias towards the ability measurement (Linden & Glas, 2010). After that, CAT will repeat these next two steps. According to Davey (2011), in the first step, an item is chosen according to examinee's current ability through various methods such as Maximum Fisher Information, Kullback-Leibler divergence or random item selection method. In the second step, the examinee's response towards the item is analysed and his or her ability is estimated. CAT will repeat these two steps until it meets the test stopping rule when all the items have been used, the measurement's precision level has been achieved or the test has reached its maximum testing time (Oppl et al., 2017). During the test, if the examinee answers wrongly, an easier item will be selected by the computer as the next item and if the examinee answers correctly, a more difficult item will be selected. The examinee's ability increased linearly with the increase of the item difficulty parameter. The repetition of these two steps produces a converging pattern of estimated ability towards a stable point which thus produces more

accurate measurement of an individual's ability. At the end of the testing session in which the test has met its stopping rule, the final estimated ability will be reported. Examinee's ability in CAT is referred as theta value estimated in logit unit. The higher the theta value, the higher the examinee's ability (Davey, 2011; Linden & Glas, 2010).

CAT can be administered by configuring it in a testing platform. The web-based platform enables the CAT to be administered online. According to Oppl et al. (2017), there are three criteria in choosing the best testing platform for CAT. Those criteria are (1) the platform offers various testing strategies; (2) the platform offers various item selection methods; and (3) the platform offers various test stopping rules. The web-based open-source platform, Concerto, meets all those criteria. This is because Concerto provides several Item Response Theory testing strategies, offers several item selection methods, and contains three types of test stopping rules. Simple coding in R language using catR-library in Concerto (Magis & Raïche as cited in Magis & Barrada, 2017) makes the test development process become easier, plus the test developer is flexible in designing the test process to meet own's standard (Aybek as cited in Aybek & Demirtasli, 2017). Therefore, Concerto becomes the ideal platform to be used by the CAT test developer (Scalise & Allen, 2015).

The main advantage of using CAT in assessment is more precise individual ability measurement can be made with shorter testing time and instant score reporting. Research by Lilley & Barker (2003) showed that the administration of linear computerized testing and CAT on similar students produced the same score. Kalender (2012) compared the score between CAT and Paper and Pencil Test in Science subtest and found that there was a high correlation of scores between two instruments. CAT provided higher reliability on ability estimation using half the number of items given in Paper and Pencil format. Besides that, research by Martin & Lazendic (2018) agreed with the finding from Davey (2011). They stated that the items selected by CAT matched well with the student's ability as the selected items were in the examinee's learning zone leading to a less measurement error thus produced better precision. Furthermore, research by Mizumoto et al. (2017) which used the Concerto as CAT testing platform found that CAT can reduce the number of items used and testing time without affecting the testing precision. Other advantages of using CAT are cheating problem can be avoided as each examinee receives unique set of items (Chuesat-

huchon & Waugh, 2010) and CAT can improve examinee's motivation towards the test because all the test takers' abilities are not being compared to each other (Betz & Turner, 2011). Due to the advantages offered by CAT, this instrument has been widely used in high stake or low stake testing environments. In educational field, CAT has been used in Adaptive English Proficiency Test for a Web, Alberta Computer Adaptive Assessment System, Japanese Computerized Adaptive Test, Measures of Academic Progress and Graduate Record Examination (IACAT, 2016), Danish National Test (Beuchert & Nandrup, 2015), School Graduation Exams in Georgia (Bakker, 2014), RISE (Utah State Board Education, 2018) and RoSA which use Concerto platform (Psychometric Unit University of Cambridge, 2018).

There was not much research conducted about the implementation of CAT in Malaysia. Desa & Latif (2007) proposed CAT as an alternative instrument to Paper and Pencil Test and explained its criteria and adaptivity feature in providing more precise ability measurement without conducting a real assessment using this instrument. Norah and Nor Azean (2008) configured CAT using C++ and implemented CAT in assessing student's ability in programming subject. They found that Malaysian students showed a positive attitude in using CAT in assessment and the students believed that there was no difference of scores between linear test and adaptive test. Although it was reported that the readiness of students in using more advance mode of testing with the integration of ICT is at the par level (Kiran et. al (2012), the integration of ICT into education in Malaysia is considered to be not at the satisfying level especially in the specific field of educational assessment and evaluation (Umar & Hassan, 2015). Therefore, the integration of ICT in education has been one major focus in Malaysian Educational Development Plan 2013-2025 to increase the quality of educational practice and to meet the current technology development (Ministry of Education, 2013). More research needs to be done about CAT in Malaysia and at the same time fulfilling the Malaysian Educational Development Plan 2013-2025. The configuration of CAT in this research is using R language which is more powerful and simpler than C++ (Venables et al., 2018).

METHODS

This research was a quantitative research which analysed numeric data (Noraini, 2013). Cross-sectional survey method was implemented

in which the instrument was given to the respondents in a specific point in time (Fraenkel & Wallen, 2009). The respondents were selected through purposive sampling method (Noraini, 2013) from secondary schools in the northern part of Malaysia. There were three stages in conducting the research: CAT configuration, CAT administration, and data analysis. In the first stage, CAT was configured in Concerto using R language. Concerto is a web-based open sourced platform that enables the linear test as well as adaptive test to be configured (Magis & Raïche, 2012). The platform consists of three types of modules. The testing module enables the planning of the test's process using R language to be done, HTML template module enables the designation of the test template, and the table module is used to record all the information, responses given by examinees as well as to record all the outputs by R (Aybek & Dermirtasli, 2017). Five nodes were selected to produce a complete test flow in which the selected nodes consist of these three modules. The selected five nodes were create_template_definition node, info node, form node, save_data node, and CAT node. The setting for five criteria of CAT namely calibrated item bank, first item selection method, item selection method, ability estimation method, and test stopping rule had been done in the CAT node. The released multiple-choice questions (MCQ) Grade 8 Science TIMSS 2003-2015 items were adopted from <https://nces.ed.gov/timss/educators.asp>. The adopted TIMSS item bank was calibrated using anchor test design (Ryan & Brockmann, 2018), and a total of 122 items was uploaded into the CAT node with the respective items' difficulty parameter measured in logit unit by Rasch model. The most difficult item had the highest difficulty parameter of 3.15 logit while the easiest item had difficulty parameter of -5.30 logit. The first item was set to be selected by CAT randomly with average difficulty level, Maximum Fisher Information had been used as the item selection method, Maximum Likelihood Estimator was used to estimate the students' ability, and 20 items fixed length test was set as stopping rule.

In the second stage, the configured CAT was administered to a class of Form 2 and Form 4 students concurrently in the beginning of the year selected through purposive sampling method. Form 2 students were selected as the Grade 8 items were used. It was assumed that older students would get better achievement thus Form 4 students were selected to compare these two groups of students' achievements in identifying the capability of CAT to function properly based on its adaptive feature. A total of 30 Form 2 stu-

dents and 35 Form 4 students were involved. The test was administered online, and the students needed to answer 20 MCQ Grade 8 Science TIMSS items within 30 minutes. As CAT is an adaptive testing, thus the arrangement of items administered depended on the responses given by the students. Therefore, every student would get a unique set of 20 items. During the test session, the first item selected by CAT was the item with average difficulty. For a given correct response, score 1 will be given while for a wrong response, score 0 will be given. After that, the student's ability (theta value) was measured in logit unit through the Maximum Likelihood Estimation. Through the Maximum Fisher Information method, the next new item with difficulty parameter near to the measured theta would be selected. Generally, easier items with lower difficulty parameter than the previous item would be selected by CAT if the student answered wrongly on the previous item and vice versa. The item selection process would continue until the test stopping rule was fulfilled and then the final theta was measured and reported.

In the third stage, the student's ability which referred to the theta value estimated by CAT, measurement error value (SEM), and students' responses towards the selected items were analysed and compared using SPSS software. The ability measured by CAT is referred as theta and it is measured in logit unit. In CAT, the theta value will be estimated for every item administered. The theta measured using Maximum Likelihood Estimation (MLE) method is the theta value that maximises the log likelihood function considering the responses given from the examinee towards the items administered (Özdemir, 2016). The higher the theta value, the higher someone's ability. The measurement error (SEM) shows how accurate the measured theta. Therefore, the lower the SEM value, the higher the accuracy of the measured theta by CAT (Barnard, 2018). In addition, the t test was used to investigate any significance different of achievement between the two groups of students to test the known-group validity.

RESULTS AND DISCUSSION

Form 2 Students' Ability in Science TIMSS Computerized Adaptive Testing (CAT)

Table 1 shows Form 2 students' ability (theta) in Science TIMSS (CAT) according to every student's ID. According to Table 1, a total of 33 Form 2 students were involved in this testing process. The highest theta value was 2.570 logit

while the lowest theta value was -3.135 logit. A total of 25 Form 2 students obtained positive theta values while the rest of the students obtained negative theta values. All the students obtained theta value with Standard Error of Measurement (SEM) less than 0.50 after the administration of 20 items which is equivalent to the reliability in the order of 0.75 (Barnard, 2018). The estimated theta value had high reliability.

Table 1. Form 2 Students' Ability in Science TIMSS CAT

ID	Theta (Logit)	SEM
i1408	2.570	0.449
i1417	2.570	0.449
i1426	2.570	0.449
i1446	2.327	0.435
i1412	2.100	0.426
i1440	2.100	0.426
i1422	2.100	0.426
i1436	1.881	0.421
i1448	1.881	0.421
i1449	1.881	0.421
i1425	1.846	0.423
i1396	1.763	0.452
i1428	1.667	0.419
i1431	1.452	0.421
i1444	1.452	0.421
i1415	1.452	0.421
i1430	1.413	0.419
i1434	1.402	0.422
i1435	1.218	0.427
i1424	1.189	0.427
i1421	0.831	0.424
i1521	0.451	0.455
i1420	0.189	0.481
i1518	0.179	0.424
i1445	0.062	0.430
i1439	-0.387	0.440
i1414	-0.439	0.433
i1419	-1.010	0.439
i1416	-1.400	0.426
i1411	-2.313	0.425
i1447	-2.626	0.454
i1432	-2.829	0.472
i1520	-3.135	0.496

Next, Table 2 shows the result of descriptive statistical analysis of Form 2 students' ability in Science TIMSS CAT. According to Table 2, the mean score for 33 Form 2 students was 0.740 logit with standard deviation of 1.657 logit and the median value was 1.413 logit. Estimated theta value of 1.452 logit had the highest frequency. Table 2 also reports that minimum theta value obtained by Form 2 students was -3.135 logit while the maximum theta value obtained was 2.570 logit producing range of 5.705 logit. Based on Table 2, percentiles analysis shows that 25% of Form

2 students obtained theta value less than -0.162 logit. Half of the students or 50% of the students obtained theta value less than 1.413 logit while the other half of the students obtained theta value more than 1.413 logit. A total of 75% of Form 2 students obtained theta value less than 1.881 logit while the rest of the students obtained theta value more than 1.881 logit. This analysis shows that Form 2 students had medium ability level. The higher the theta value, the higher the ability. Therefore, Form 2 students need an increase in their understanding of this basic science knowledge.

Table 2. Form 2 Students' Ability in Science TIMSS CAT Statistical Data

N	33
Mean	0.740
Median	1.413
Mode	1.452
Std. Deviation	1.657
Skewness	-1.093
Range	5.705
Minimum	-3.135
Maximum	2.570
Percentiles:	
25	-0.162
50	1.413
75	1.881

Furthermore, item pattern and students' responses were also being analysed. Table 3 shows the list of items with the frequency that has been selected by Science TIMSS CAT for Form 2 students. From Table 3, the total items used was 97 items out of 122 calibrated Grade 8 released Science TIMSS items. Item 15 had the highest frequency with 33 times indicated that every Form 2 student had answered this item in Science TIMSS CAT. The frequency of items used decreased gradually. Form 2 students' responses affected the item selected by the CAT and these selected items covered a wide range of item difficulty parameter from high, medium, and low level.

Table 4. The First Six Selected Items for a Form 2 Student with High Estimated Theta

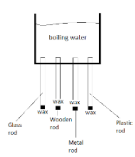
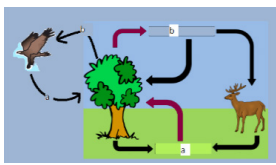
Item Number	Item	Item's Difficulty Parameter (Logit)	Student's Response	Theta (Logit)	Explanation
15	 <p>The diagram shows four identical size rods each of a different material sealed into the bottom of a container.</p>	-0.02	1	4.00	The first item selected was Item 15 with average difficulty in which the item was testing Physics content domain.

Table 3. Selected Items used by Form 2 Students in Science TIMSS CAT

Frequency	Item's Number	Total Item
33	15	1
22	11, 108	2
21	8, 26, 61, 110	4
20	1, 3, 4, 18, 60, 57, 28, 112	8
18	40	1
17	10, 34, 38	3
14	103	1
12	36, 102	2
10	52	1
9	29	1
8	119	1
7	5, 9, 32, 58, 75, 96	6
6	7, 13, 20, 45, 78, 79, 82, 85	8
5	31, 49, 56, 86, 89	5
4	48	1
3	17, 35, 59, 63, 70, 97	6
2	21, 22, 33, 41, 43, 47, 54, 61, 64, 67, 72, 74, 76, 77, 83, 87, 90, 94, 104, 107, 111	21
1	2, 12, 14, 16, 19, 23, 27, 42, 44, 51, 53, 65, 66, 69, 73, 91, 92, 95, 99, 109, 115, 117, 118, 120, 121	25
Total		97

Item Selection Process By CAT

The item selection process for a high theta and low theta students were explained in this subtopic for the first six items selected by CAT. Table 4 shows the items selected by CAT for a student with high estimated theta while Table 5 shows the items selected by CAT for a student with low estimated theta. The item selection process was explained for every item administered.

	The same amount of wax placed on the end of each rod and then the container is filled with boiling water. On which rod will wax melt first? A. Glass rod B. Wooden rod C. Metal rod				This student correctly answered the item (score 1) thus produced measured theta 4.00 logit which indicated that the current student's ability was at 4.00 logit.
11	Which of the following properties of a substance is conserved during thermal expansion? A. Mass B. Volume C. Shape D. Distance between particles	3.15	1	4.00	The second selected item was Item 11 testing Physics content domain with the highest difficulty level, 3.15 logit as this difficulty level was nearer to the previous estimated theta. The selected item's difficulty level was below the student's previous estimated theta, thus this student had high chances to answer correctly. This student answered the second item correctly thus the measured theta was 4.00 logit.
8	The diagram below shows an example of interdependence among organisms. During the day the organisms either use up or give off (a) or (b) as shown by the arrows.	2.25	1	4.00	Item 8 tested Biology content domain with difficulty level 2.25 logit was selected as the third item. Logically, the selected third item should have difficulty level more than the previous item because the student answered the second item correctly thus more difficult item should be presented. However, Item 11 had the highest difficulty level thus Item 8 was selected as the third item because the difficulty level of Item 8 was positioned just below the difficulty level of Item 11 in the calibrated bank item list and its difficulty level was the nearest to the previous estimated theta. Item 8 was answered correctly thus produced current estimated theta 4.00 logit.
					
	Choose the right answer for (a) and (b) from the alternatives given. A. (a) is carbon dioxide and (b) is nitrogen B. (a) is oxygen and (b) is carbon dioxide C. (a) is carbon dioxide and (b) is water vapour D. (a) is carbon dioxide and (b) is Oxygen				
4	Which of the following animals have been on Earth for the longest period of time? A. Human B. Birds C. Fish D. Reptiles	2.23	0	3.32	Item 4 tested Biology content domain which was selected as the fourth item because its difficulty level was the nearest to the previous estimated theta and it was positioned just below the Item 8 in the bank item list. The student answered wrongly (score 0) thus the estimated theta was decreased from the previous one. As the student failed to answer the fourth item, an easier item with difficulty level less than 2.23 logit will be selected as the next item.
60	Light travels fastest through which of the following? A. Air B. Glass C. Water D. A vacuum	2.16	1	3.62	Item 60 which tested Physics content domain was selected as the fifth item with difficulty level less than the previous item and its difficulty level was positioned just below Item 4. The student answered correctly thus the estimated theta increased 3.62 logit.

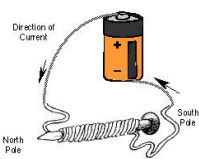
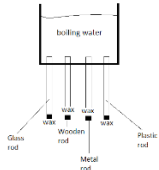
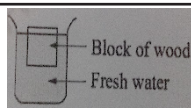
18	 <p>The figure shows an iron nail with an insulated wire coiled around it. The wire is connected to a battery. What will happen to the nail when current flows through the wire?</p> <p>A. The nail will melt B. Electric current will flow through the nail C. The nail become a magnet D. Nothing will happen to the nail</p>	2.04	2	2.83	Item 18 which tested Physics content domain was selected as the sixth item. The student answered correctly in the previous item thus the next item with difficulty level more than 2.16 logit should be selected. However, Item 18 with difficulty level 2.04 logit was selected as its difficulty level was the nearest to the previous estimated theta. The student answered wrongly thus the estimated theta decreased to 2.83 logit.
----	--	------	---	------	--

Table 5. The First Six Selected Items for a Form 2 Student with Low Estimated Theta

Item Number	Item	Item's Difficulty Parameter (Logit)	Student's Response	Theta (Logit)	Explanation
15	 <p>The diagram shows four identical size rods each of a different material sealed into the bottom of a container. The same amount of wax placed on the end of each rod and then the container is filled with boiling water. On which rod will wax melt first?</p> <p>A. Glass rod B. Wooden rod C. Metal rod</p>	-0.02	0	-4.00	The first item selected was Item 15 with average difficulty in which the item was testing Physics content domain. This student answered wrongly for this item (score 0) thus produced measured theta -4.00 logit which indicated that the current student's ability was at -4.00 logit.
102	<p>Which of the following is formed immediately after fertilization?</p> <p>A. Egg B. Sperm C. Zygote D. Embryo</p>	-3.26	0	-4.00	Item 102 which tested Biology content domain was selected as the second item because its difficulty was the nearest to the previous estimated theta. The student failed to answer the item correctly thus the estimated theta was -4.00 logit.
79	<p>The picture shows a block of wood floating in fresh water. If this block were placed in salt water from the ocean, which picture shows what would happen?</p>	-2.82	1	-3.13	Item 79 which tested Chemistry content domain was selected as the third item. This item was positioned just above Item 102 in calibrated bank item list. The student answered correctly for this item thus the estimated theta was increased to -3.13 logit.



- A. Salt Water
- B. Salt Water
- C. Salt Water
- D. Salt Water

75	Most underground caves are formed by the action of water on A. Granite B. Limestone C. Sandstone D. Shale	-2.57	1	-2.31	Item 75 which tested Earth Science content domain was selected as the fourth item. This item' difficulty level was positioned just above Item 79. The student answered correctly thus the estimated theta increased to -2.31 logit. A more difficult next item with difficulty level higher than -2.57 logit will be selected.
32	A wet towel will dry when it is left in the Sun. Which process occurs to make this happen? A. Melting B. Boiling C. Condensation D. Evaporation	-2.40	1	-1.82	Item 32 which tested Physics content domain was selected as the fifth item as its difficulty level was the nearest to the previous estimated theta. The item was more difficult than the previous item and the student answered correctly so the estimated theta increased to -1.82 logit.
5	How does the average body temperature of people living in hot climates compare to the average body temperature of people living in cold climates? A. Higher in hot climates B. Lower in hot climates C. The same in both climates	-1.81	1	-1.41	Item 5 which tested Biology content domain was selected as the sixth item with difficulty level higher than the previous item as the student answered correctly on the previous item. In addition, the selected item's difficulty level was the nearest to the previous estimated theta. the student answered correctly thus the estimated theta increased to -1.41 logit.

This selected item pattern made sense because the Science TIMSS CAT used Maximum Fisher Information in item selection method in which the difficulty level of the next item is selected closer to the estimated theta (Özdemir, 2016). Easier item will be given for a previous wrongly answered item and vice versa. Generally, this item selection procedure was similar to every student based on the response given to the previous item. In addition, Item 11 (Figure 1) with the highest difficulty level had 19 times wrongly answered by the Form 2 students from 22 times of selection. The correct answer for this item is

option A. Out of 19 students who answered item 11 wrongly, 12 students selected option D, 4 students selected option C, and 3 students selected option B. These students might have blurry understanding on the word "conserved". In conclusion, Form 2 students had medium ability level in Science TIMSS CAT. Different students' abilities affected the selection of different item difficulty parameter which covers a wide range of difficulty levels. Based on the ability estimated, Form 2 students need to improve their knowledge on basic Science especially those with the negative theta values.

Which of the following properties of a substance is conserved during thermal expansion?			
A. Mass	B. Volume	C. Shape	D. Distance between particle

Figure 1. Item 11

Form 4 Students' Ability in Science TIMSS Computerized Adaptive Testing (CAT)

Table 6. Form 4 Students' Ability in Science TIMSS CAT

ID	Theta (Logit)	SEM
i1458	4.000	0.606
i1471	4.000	0.606
i1503	4.000	0.606
i1454	4.000	0.606
i1500	4.000	0.606
i1502	4.000	0.606
i1499	3.988	0.604
i1451	3.507	0.539
i1462	3.507	0.539
i1465	3.507	0.539
i1455	3.141	0.498
i1459	2.837	0.469
i1468	2.837	0.469
i1494	2.837	0.469
i1530	2.837	0.469
i1463	2.837	0.469
i1491	2.837	0.469
i1498	2.837	0.469
i1470	2.570	0.449
i1473	2.570	0.449
i1506	2.327	0.436
i1497	2.327	0.436
i1505	2.327	0.436
i1464	2.327	0.436
i1490	2.100	0.426
i1495	1.881	0.421
i1531	1.881	0.421
i1472	1.667	0.420
i1504	1.452	0.421
i1479	1.362	0.424
i1478	1.220	0.427
i1533	-0.484	0.420
i1535	-1.122	0.419
i1460	-1.148	0.430
i1496	-1.891	0.427

Table 6 shows Form 4 students' ability (theta) in Science TIMSS CAT according to ID. By referring to Table 6, 35 Form 4 students were involved in the testing session. The highest theta value was 4.000 logit while the lowest theta value was -1.891 logit. Overall, there were 31 students who obtained positive theta value and 4 students with negative theta value. Moreover, there were ten students who obtained SEM value higher than 0.50 with theta value more than 3.500 logit. This might be due to the usage of Maximum

Likelihood Estimation method which could not produce precise measurement if nearly all items were answered correctly or wrongly (Song as cited in Özdemir, 2016). Therefore, these Form 4 students needed more difficult items for the Maximum Likelihood Estimation method to measure properly. The rest of the students who obtained SEM value less than 0.50 showed that the obtained theta had high reliability in the order of 0.75 (Barnard, 2018).

Table 7 shows statistical data of Form 4 students' ability in Science TIMSS CAT. According to Table 7, the mean score for 35 Form 4 students was 2.368 logit with standard deviation of 1.531 logit. The median value was 2.837 logit with mode value of 2.837 logit. Table 7 also reports that minimum theta value obtained by Form 4 students was -1.891 logit while the maximum theta value obtained was 4.000 logit producing range of 5.891 logit. According to Table 7, percentiles analysis shows that 25% of Form 4 students obtained theta value less than 1.881 logit. Half of the students or 50% of the students obtained theta value less than 2.837 logit while the other half of the students obtained theta value more than 2.837 logit. A total of 75% of Form 4 students obtained theta value less than 3.507 logit while the rest of the students obtained theta value more than 3.507 logit. This analysis shows that majority of Form 4 students had high ability level. The analysis made sense because Form 4 students had learnt the Grade 8 Science knowledge and currently, they are learning pure Science in the upper secondary level.

Table 7. Form 4 Students' Ability in Science TIMSS CAT Statistical Data

N	35
Mean	2.368
Median	2.837
Mode	2.837
Std. Deviation	1.531
Skewness	-1.340
Range	5.891
Minimum	-1.891
Maximum	4.000
Percentiles:	
25	1.881
50	2.837
75	3.507

After that, item pattern and students' responses were also analysed. Table 8 shows the list of items with the frequency that has been selected by Science TIMSS CAT for Form 4 students. According to Table 8, the total of items used was 75 items out of 122 calibrated Grade 8 released Science TIMSS items. Item 15 had the highest frequency with 35 times indicated that every Form 4 student had answered this item in Science TIMSS CAT. There was a huge gap between the existing item's frequency. A total of 20 items had frequency of 30 and above. Majority of Form 4 students answered items with frequency of 30 and above which comprises of these 20 items. These 20 items consisted of 19 items with high difficulty level and 1 item with negative difficulty level. These 19 items had difficulty level ranging from 1.33 logit to 3.15 logit and these items were the most difficult items in the item's difficulty level list. The remaining 1 item was item 15 with difficulty level of -0.02 logit. These frequencies of selected items indicated that CAT selected difficult items for Form 4 students as these students have higher ability in answering difficult questions.

Table 8. Selected Items used by Form 4 Students in Science TIMSS CAT

Frequency	Item	Total Item
35	15	1
31	1, 10, 11, 28, 38	5
30	3, 4, 8, 18, 26, 34, 40, 57,	14
4	60, 61, 103, 108, 110, 112	4
3	36, 52, 56, 102 20, 45, 119	3
2	5, 9, 13, 29, 31, 32, 49, 58, 62, 75, 78, 82, 85, 89, 96, 98	16
1	7, 12, 17, 23, 30, 35, 37, 43, 44, 47, 48, 50, 55, 59, 63, 65, 70, 71, 73, 74, 76, 79, 81, 83, 90, 94, 95, 109, 114, 117, 120, 122	32
Total		75

Further analysis was made on the items' pattern and the students' responses. It shows that the first item used in Science TIMSS CAT was Item 15. If correctly answered, Item 11 will be given. If wrongly answered, Item 102 will be given. This item selection pattern was similar to the item selection for Form 2 students. Selection of the next item was based on the student's response towards the previous item. From this further analysis also, 6 students with theta value 4.000 logit had answered all 19 or 20 items correctly.

As they had passed the Grade 8 Science's syllabus, their obtained theta values were considered suitable with their educational level. Therefore, these students need more difficult items to obtain more precise ability measurement.

Table 9 shows the comparison of statistical data between Form 2 students' ability and Form 4 students' ability in Grade 8 Science TIMSS CAT. Overall, the mean, median, mode, and the percentiles of Form 4 students were higher than the Form 2 students. The obtained scores for these two groups of students were appropriate with their academic levels which were taken in the beginning of the schooling year. Majority of Form 4 students obtained high theta value and they answered most items with highest difficulty parameter while Form 2 students obtained medium and slightly low theta values indicating that these Form 2 students need to learn more Science knowledge. Moreover, majority of the students obtained the estimated theta with SEM value less than or equal to 0.50 for 20 items administered. The SEM result was parallel with the CAT simulation research conducted by Barnard (2018) using Maximum Fisher Information and Maximum Likelihood Estimation method. The result indicates that this configured CAT was capable in producing precise ability estimation with fewer items. Therefore, this instrument is suitable to be used in assessing students' basic Science TIMSS knowledge as the estimated theta had high reliability.

Table 9. Comparison of Form 2 Students' Ability and Form 4 Students' Ability in Grade 8 Science TIMSS CAT

	Form 2	Form 4
Mean	0.740	2.368
Median	1.413	2.837
Mode	1.452	2.837
Std. Deviation	1.657	1.531
Minimum	-3.135	-1.891
Maximum	2.570	4.000
Percentiles:		
25	-0.162	1.881
50	1.413	2.837
75	1.881	3.507

Known-Group Validity

Known-group validity is used to test how good an instrument can differentiate between the two known groups. The validity can be conducted by administering an instrument simultaneously to two different known groups of people. The criteria for known-group validity is met when there is a statistically significance different of scores

between the two known groups and can be analysed using t test (Devellis as cited in NSSE, 2009). In this study, calibrated Grade 8 TIMSS item bank was used in CAT, thus the target group for this instrument was Form 2 students. Therefore, known-group validity was tested using t test by comparing the mean score of Form 2 students' ability with Form 4 students' ability in Science TIMSS CAT. Table 10 shows t test for these two groups of students. By referring to Table 10, the Sig.(2-tailed) shows the value of 0.000. This value was less than 0.05 thus there was a significance difference between the mean scores of Form 2 and Form 4 students in Science TIMSS CAT (Pallant, 2011).

Table 10. Independent-Samples T-Test for Two Known-Group

Equal Variances Assumed	
Lavene's Test for Equality of Variances:	
F	0.976
Sig.	0.327
t-test for Equality of Means:	
t	-4.211
df	66
Sig.(2-tailed)	0.000
Mean Difference	-1.628
Std. Error Difference	0.387
95% Confidence Interval of the Difference:	
Lower	-2.400
Upper	-0.856

The Eta squared or magnitude of differences was calculated using the following formula (1):

$$\text{Eta squared} = \frac{t^2}{t^2 + (N1 + N2 - 2)}$$

where, t = t value in t-test for Equality of Means; $N1$ = total group 1 response; and $N2$ = total group 2 response

The calculated eta squared was 0.212. According to Cohen (1988) in Pallant (2011), Eta squared value of 0.01 is considered small, 0.06 is considered medium, and 0.14 is considered large. The obtained Eta squared in this study was higher than 0.14 thus there was a big different in magnitude between the two mean scores. As there was a significance difference between mean scores of these two groups of students and the previous analysis showed that Science TIMSS CAT was more challenging to Form 2 students, this instrument met the known-group validity.

Science TIMSS' curriculum framework comprising of intended curriculum, implemented curriculum, and attained curriculum were developed based on the similarity of the Science curriculum among the participating countries. Generally, the items used in Grade 8 Science TIMSS consist of multiple-choice questions and subjective questions. Science TIMSS tests four content domains: 35% Biology items, 25% Physics items, 20% Chemistry items, and 20% Earth Science items. Furthermore, the tested items consist of three cognitive domains: 35% knowing items, 35% applying items, and 30% reasoning items. Items with knowing cognitive domain assess factual knowledge, concept, and relationship. Items in the applying cognitive domain assess the student's ability in making comparison, interpreting the information, and explaining scientifically, while items in the reasoning cognitive domain require students to analyse data and evaluate it, make a generalisation and justification (Mullis & Martin, 2017). The Science TIMSS calibrated bank item used in this research involved only multiple-choice questions with the items covered knowing and applying cognitive domains because items with reasoning cognitive domain mostly are subjective items.

According to Piaget, there are four stages of cognitive development with specific range of children's age. Children at the age between 0-2 years are experiencing sensorimotor stage in which the cognitive developments are based on the interaction of the five senses with the environment. Children at the age between 2-6 years are experiencing preoperational stage in which they use symbol to represent the image or word without having the ability to give reason logically. Children at the age between 7-12 years are experiencing concrete operation in which they can think logically based on the concrete event only, while children at the age of 12 years upwards are experiencing a formal operational stage involving abstract and logic thinking ability (Lazarus, 2010).

Based on the classification of ages from Piaget's theory, Form 2 and Form 4 students are experiencing formal operational stage. Form 2 students (Grade 8) are the beginner to this formal operational stage, and they are starting to develop the ability to think in the abstract ways. They are in the process of developing a deep abstract thinking about a concrete event and provide a systematic logic reason to the event which mainly involve the applying and reasoning cognitive domains, so they have not yet mastered these higher cognitive skills properly. Form 2 students are in

the process of engaging themselves in the problem-solving method. Therefore, majority of these students are not so capable in answering more difficult Science items. As the selected items by CAT relied on the given response to the previous item as well as the previous estimated theta, it made sense that the selected items by CAT to Form 2 students covered a wide range of item's difficulty level in the previous analysis and used almost 80% of items from the whole item bank.

Form 4 students are more mature than Form 2 students because they have been experiencing formal operational stage in a longer period than the Form 2 students. These Form 4 students have mastered more abstract thinking skills, so they can think logically in applying and relating various scientific knowledge and analyse the data better thus they are capable in learning more abstract scientific knowledge. Currently, these students are learning pure Science subjects comprising Biology, Chemistry, and Physics. Therefore, Form 4 students had more Science knowledge with higher order thinking skills than Form 2 students, so they are capable in answering more difficult Science items involving applying and reasoning cognitive domains. Due to this situation, the previous analysis showed that most of the 20 items selected by CAT to all Form 4 students were the items with high difficulty level positioned at the most upper part in the item's difficulty list. The higher the item's difficulty, the harder the item to be answered correctly. As the students were capable in answering difficult items correctly, their estimated theta increased showed that their ability also increased.

Overall analysis shows Form 4 students obtained better achievement in CAT than the Form 2 students. Form 4 students generally have mastered the Grade 8 Science TIMSS knowing and applying items more than the Form 2 students as they were capable in answering more difficult Science items correctly. The analysis indicates that the instrument was configured correctly and capable in differentiating two levels of knowledge using its adaptive feature by correctly selecting item's difficulty level with the estimate theta based on the selection method used, thus producing high ability precision using fewer items.

CONCLUSION

CAT was configured in Concerto. Concerto is a platform containing the computer algorithm which enables the testing to operate adaptively. The measurement of CAT is based on the theta values produced by the calculation

of the alignment between the student's ability level and item difficulty level. The CAT was then administered to Form 2 and Form 4 students simultaneously in the beginning of the schooling year in which Form 2 students were still in the process of learning lower secondary Science syllabus which is equivalent to Grade 8 level, while Form 4 students had learnt the lower secondary Science syllabus and now are learning upper secondary Science syllabus. Their ability towards the test was evaluated as well as the analysis of the item selection pattern and the responses given by the students. Form 2 students used overall 97 items with wide range of item difficulties while Form 4 students used a total of 75 items and most of the items had high item difficulty level. The Science TIMSS CAT met the known-group validity with Form 4 students obtained higher ability than Form 2 students indicating that the instrument is more suitable in challenging lower secondary Form 2 students. This analysis shows that the configuration of CAT in Concerto has been done correctly because it can differentiate two levels of knowledge significantly. Moreover, this configured CAT proves that it can increase the test measurement accuracy while using fewer items.

ACKNOWLEDGEMENTS

This research has been funded under the collaboration between the Research Fund Vot. E15501 from the Research Management Center (RMC) Universiti Tun Hussein Onn Malaysia (UTHM) and Fundamental Research Grant Scheme (FRGS), Ministry of Education Malaysia (203 / PGURU / 6711486).

REFERENCES

- Aybek, E.C. & Demirtasli, R.N. (2017). Computerized Adaptive Test (CAT) Applications and Item Response Theory Models for Polytomous Items. *International Journal of Research in Education and Science (IJRES)*, 3(2), 475-487.
- Bakker, S. (2014). The Introduction of Large-Scale Computer Adaptive Testing in Georgia. *Russia Education Aid for Development*. Retrieved from <http://siteresources.worldbank.org/INTREAD/Resources/BakkerIntroductionto-CATGeorgiaforREAD.pdf>
- Barnard, J. J. (2018). From Simulation to Implementation: Two CAT Case Studies. *Practical Assessment, Research & Evaluation*, 23(14), 1-8. Retrieved from <http://paronline.net/getvn.asp?v=23&n=14>
- Betz, N. E., & Turner, B. M. (2011). Using Item Re-

- response Theory and Adaptive Testing in Online Career Assessment. *Journal of Career Assessment*, 19(3), 274–286.
- Beuchert, L. V. & Nandrup, A. B. (2015). *The Danish National Test- A Practical Guide*. Economic Working Papers. Retrieved from <http://econ.au.dk/fileadmin/sitefiles/fileroeonomi/WorkingPapers/Economics/2014/wp1425.pdf>
- Chuesathuchon, C., & Waugh, R. F. (2010). Item Banking With Rasch Measurement : an Example for Primary Mathematics in Thailand. In *EDU-COM 2008 International Conference. Sustainability in Higher Education: Directions for Change* (pp. 19–21). Perth Western Australia. Retrieved from <http://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1007&context=ceducom>
- Davey, T. (2011). *A Guide to Computer Adaptive Testing Systems*. Washington, DC: Council of Chief School Officers. Retrieved from <https://www.semanticscholar.org/paper/A-Guide-to-Computer-Adaptive-Testing-Systems-Davey/71d9d258a7b161db2a3848b85afaedc105ef11fa>
- Desa, Z. N. D. & Abdul Latif, A. (2007) Computerized Adaptive Testing: An Alternative Assessment Method. In: *Symposium Pengajaran dan Pembelajaran UTM*, Johor Bahru.
- Fraenkel, J. R. & Wallen, N. E. (2009). *How to Design and Evaluate Research in Education (7th Ed.)*. New York: McGraw-Hill.
- IACAT. (2016, July 12). *Operational CAT Programmes*. Retrieved from <http://www.iacat.org/content/operational-cat-programs>
- Kalender, İ. (2012). Computerized adaptive testing for student selection to higher education. *Yükseköğretim Dergisi*, 2(1), 13-19.
- Kiran, S., Kiani, A., Kayani, S., Shahzadi, E., Sehar, A., Akhtar, K., Sohail, A. & Kayani, S. (2012). An Exploratory Study of Student's Attitudes towards Online Assessment. *International Journal of Humanities and Social Science*, 2(4), 304 - 309. Retrieved from <http://www.pearsonassessments.com/research>
- Lazarus, S. (2010). *Educational Psychology: In Social Context (4th edition)*. Cape Town: Oxford University Press.
- Lilley, M. & Barker, T. (2003). Comparison between Computer Adaptive Testing and other Assessment Methods: An Empirical Study. *Proceedings of 10th International Conference of the Association for Learning Technology (ALT-C)*, University of Sheffield, United Kingdom.
- Linden, W. J. & Glas, C. A. W. (Eds.). (2010). *Elements of Adaptive Testing*. (Statistics for Social and Behavioral Sciences Series). New York: Springer.
- Magis, D., & Barrada, J. R. (2017). Computerized Adaptive Testing with R: Recent Updates of the Package catR. *Journal of Statistical Software*, 76(1), 1-19.
- Magis, D. & Raïche, G. (2012). Random Generation of Response Patterns under Computerized Adaptive Testing with the R Package catR. *Journal of Statistical Software*, 48(8), 1–31.
- Mansoor Al-A'ali. (2007). Implementation of an Improved Adaptive Testing Theory. *Journal of Educational Technology and Society*, 10(4), 80-94. Retrieved from https://www.researchgate.net/publication/220374538_Implementation_of_an_Improved_Adaptive_Testing_Theory
- Martin, A. J., & Lazendic, G. (2018). Computer-Adaptive Testing: Implications for Students' Achievement, Motivation, Engagement, and Subjective Test Experience. *Journal of educational psychology*, 110(1), 27-45.
- Ministry of Education (MOE) (2013). *Pelan Pembangunan Pendidikan Malaysia 2013-2025*. Retrieved from http://www.moe.gov.my/cms/upload_files/articlefile/2013/articlefile_file_003107.pdf
- Mizumoto, A., Sasao, Y. & Webb, S. A. (2017). Developing and Evaluating A Computerized Adaptive Testing Version of the Word Part Levels Test. *Language Testing*, 36(1), 101-123.
- Mullis, I. V. S. & Martin, M. O. (Eds.). (2017). TIMSS 2019 Assessment Framework. TIMSS & PIRLS International Study Center Lynch School of Education. Retrieved from <http://timssandpirls.bc.edu/timss2019/frameworks/>
- National Survey of Student Engagement (NSSE). (2009). *Validity – 2009 Known Groups Validation*. Retrieved from <http://nsse.indiana.edu/html/validity.cfm>
- Norah Md. Noor, & Noor AzeanAtan. (2008). Tahap Kesiediaan dan Keyakinan Pelajar terhadap Penggunaan Ujian Adaptif dalam Mempelajari Konsep Pengaturcaraan Komputer. *Proceedings of the 2nd International Malaysian Educational Technology Convention* (pp. 271–278). Kuala Lumpur: META.
- Noraini Idris. (2013). *Penyelidikan dalam Pendidikan (2nd Ed.)*. Malaysia: McGraw-Hill Education (Malaysia) Sdn. Bhd.
- Oppl, S., Reisinger, F., Eckmaier, A. & Helm, C. (2017). A Flexible Online Platform for Computerized Adaptive Testing. *International Journal of Educational Technology in Higher Education*, 14(2).
- Özdemir, B. (2016). Comparison of Different Unidimensional-CAT Algorithms Measuring Students' Language Abilities: Post-hoc Simulation Study. *The European Proceedings of Social & Behavioural Sciences*. Retrieved from <http://dx.doi.org/10.15405/epsbs.2016.11.42>
- Pallant, J. (2011). *SPSS Survival Manual: A Step by Step Guide to Data Analysis using SPSS (4th ed.)*. Australia: Allen & Unwin.
- Psychometric Unit, University of Cambridge. (2018). *Open Source Online R-based Adaptive Testing Platform*. Retrieved from <http://www.psychometric.cam.ac.uk/newconcerto>
- Ryan, J. & Brockmann, F. (2018). *A Practitioner's Introduction to Equating*. Washington, DC: Council of Chief State School Officers.
- Scalise, K. & Allen, D. D. (2015). Use of Open-Source

- Software for Adaptive Measurement: Concerto as an R-based Computer Adaptive Development and Delivery Platform. *British Journal of Mathematical and Statistical Psychology*, 68(3), 478-496. Retrieved from <https://doi.org/10.1111/bmsp.12057>
- Umar, N. I. & Hassan, S. A. (2015). Malaysia Teachers Levels of ICT Integration and its Perceived Impact on Teaching and Learning. *Procedia-Social and Behavioral Sciences*, 197, 2015-2021.
- Utah State Board of Education. (2018). *RISE: Readiness Improvement Success Empowerment*. Retrieved from <https://www.schools.utah.gov/>
- Venables, W. N., Smith, D. M. & The R Core Team. (2018, December 20). *An Introduction to R*. Retrieved from <https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf>