MODESTUM
OPEN ACCESS

EURASIA Journal of Mathematics, Science and Technology Education, 2020, 16(7), em1863
ISSN:1305-8223 (online)
Research Paper
https://doi.org/10.29333/ejmste/8259

A Calibrated Item Bank for Computerized Adaptive Testing in Measuring Science TIMSS Performance

Mohd Ali Samsudin [1*], Thodsaphorn Som Chut [1], Mohd Erfy Ismail [2], Nur Jahan Ahmad [1]

[1] School of Educational Studies, Universiti Sains Malaysia, MALAYSIA
[2] Faculty of Technical and Vocational Education, Universiti Tun Hussein Onn, MALAYSIA

Received 25 May 2019 ▪ Accepted 23 April 2020

Abstract

The current assessment is demanding for a more personalised and less-time consuming testing environment. Computer Adaptive Testing (CAT) is seemed as a more effective alternative testing method in comparison to conventional test in meeting the current standard of assessment. This research reports on the calibration of the released Grade 8 Science objective items in Trends in International Mathematics and Science Study (TIMSS) 2003-2015 based on Rasch model framework to be used in CAT as an alternative testing tool in a low stakes test. Concurrent common item equating method was used in linking and equating sets of items. Five test sets were produced consisting of 20 unique items and 10 common items in each set. The unique items were available only in a single set while common items were available in the two-consecutive set of items. The sets were administered through Paper and Pencil test to Form 2 (Grade 8) students who had been selected through a purposive sampling method from secondary schools in the northern part of Malaysia. The fit analysis, polarity analysis, unidimensionality analysis, item measure and Person-Map-Item were conducted. The analysis produced 122 calibrated items which meet the Rasch's requirements and were suitable to be used in CAT.

Keywords: calibrated bank, Computer Adaptive Testing (CAT), TIMSS

INTRODUCTION

The implementation of technology in assessment field has raised the standard of assessment that demands a more effective and a less time-consuming testing environment. It is found that the intensive use of computers in assessment can create a positive environment, reduce the testing time and produce a more precise measurement of an individual's ability (Davey, 2011; Wang, 2010). These standards have limited the function of the conventional method which is the Paper and Pencil test in today's assessment. According to Mansoor Al-A'li (2007), although the Paper and Pencil test is suitable in assessing student's overall achievement, due to its characteristics, the instrument has a limited ability in analysing the response given by students to each item administered in the test.

Generally, Classical Test Theory is often used in constructing items for Paper and Pencil Test. The item difficulty is based on the probability value. The more people who correctly answered an item, the higher the probability value, and that item is considered easy; thus, the item difficulty parameter depends on the respondents so the obtained item difficulty parameter could not be generalised to every population (Bichi et al., 2015). Furthermore, most of the items used have an average difficulty level which is more suitable in testing students with average ability. Students with low ability would find the items used are difficult while students with high ability would find the items are very easy. According to Weiss (2011), the usage of item difficulty level which is not at the same level as a person's ability would decrease the measurement precision of the obtained score from a test. Due to this limitation, the Paper and Pencil Test provides low measurement precision in estimating the achievement of students with low and high ability thus it is not so efficient in reporting an individual's ability precisely.

In addition, in terms of marking process, a huge compilation of test papers has burdened the teachers so

✉ alisamsudin@usm.my **(*Correspondence)** ✉ saphorn.mk@gmail.com ✉ erfy@uthm.edu.my ✉ jahan@usm.my

Contribution to the literature

- This study presents the results on the calibration of Grade 8 Trends in International Mathematics and Science Study (TIMSS) released Science objective items to produce a calibrated item bank for Computerized Adaptive Testing (CAT).
- This study demonstrates the items equating technique for testlet using a concurrent common item equating method which is analysed through Winsteps based on the Rasch model.
- Calibrated item bank through Rasch Model enables two different ability parameters to be compared on one linear scale even though each student receives unique set of items in CAT.

the marking process might take few days to be completed (Mansoor Al-A'ali, 2007) and this has made the instrument seems not so practical (McMillan & Lawson, 2001). Students might forget the items and they might have low motivation in revising the items after few days they completed the test. This is why an instant score reporting after the test is completed is important because it could affect the students' thinking and emotion (Conole & Warburton, 2005). In addition, the teachers also need to rush meeting the marking deadline, thus errors might occur while marking the papers (Masrom & Abd. Rahman, 2009). Also, cheating problem might arise too as every student receives a similar set of items through a Paper and Pencil Test and this migh affect the precision of measurement (Chuesathuchon & Waugh, 2010).

Due to these characteristics of the conventional test, it often does not meet most of the current standards of assessment, thus computerized testing is becoming a popular alternative instrument (Csapó et. al., 2012 as cited in Magyar, 2015). However, linear computerised testing instrument contains similar characteristics as a Paper and Pencil Test in terms of test development in which it provides fixed items, but it is admistered through the computer. Therefore, linear computerized test provides low precision on high and low individual's ability measurement (Barker, 2008). The development of measurement theory from Classical Test Theory to Item Response Theory (IRT) plus the advanced usage of technology with extensive use of computer programming has provided a realistic path to Computerized Adaptive Testing (CAT) (Linacre, 2000; Linden & Glas, 2010).

CAT is believed to be a viable solution in improving the measurement accuracy of individuals as an alternative testing tool to linear testing which is suitable to be used in various stakes of testing environments (Linacre, 2000; Morphew et al., 2018). In the CAT, the computer starts the test by selecting the first item randomly as there is no any information about the student's ability yet. An item with an average difficulty level usually is selected as the first item (Linacre, 2000). The student's response towards the first item is analysed and followed by the measuremet of the student's ability referred as theta value. After that, the CAT will select the next item to be administered. Generally, if the student answered wrongly, an easier item will be given as the next item and vice versa. The item selection process will continue until the test meets its stopping rule and the student's ability level, refered to as theta value is reported in a logit unit (Oppl et al., 2017).

According to Way et al. (2010), CAT meets current assessment's requirement which focuses on personalised individual assessment. The adaptivity feature of CAT in selecting the item difficulty level based on the current individual's ability has enabled the usage of very difficult items or very easy items to be avoided thus every level of individual's ability whether low, average or high ability can be measured accurately. In addition, it is obvious that every examinee receives unique items based on their own current ability thus it minimised the testing time (Linacre, 2000) and the cheating problem can be avoided (Chuesathuchon & Waugh, 2010). As CAT is administered by computer, an instant score reporting right after the test has ended is possible to be done, making the instrument to be more practical (Linden & Glas, 2010).

There are few researches about CAT in Malaysia such as conducted by Md. Desa and Abdul Latif (2007) who suggested CAT as an alternative instrument to Paper and Pencil Testing by explaining its adaptivity feature without implementing the instrument in the real situation. In another study by Md. Noor and Atan (2008), the researchers configured CAT using C++ and administered CAT in assessing student's ability in programming subject. This study has found that Malaysian students showed positive motivation in using CAT in assessment as an alternative instrument to conventional method. However, none of these researchers use CAT for measuring Science subject in Trends in International Mathematics and Science Study (TIMSS) performance. Even though the research found that students showed positive acceptance in using technology in assessment, Raman and Yamat (2014) and Umar and Hassan (2015) found that the integration of ICT in educational aspect is considered poor in Malaysia.

Malaysia has participated in TIMSS since 1999 involving Grade 8 students only which is equivalent to Form 2 students based on the scope of Science syllabus stated by TIMSS. TIMSS reported achievement at four levels namely Advanced International Benchmark, High International Benchmark, Intermediate International

Benchmark and Low International Benchmark. Science TIMSS 2015 reported that only 7% of 8th grade students from all participating countries achieved advanced benchmark while 84% of the students obtained the low benchmark indicating that most of the students showed some basic knowledge in Science and were not so capable in applying the Science knowledge in abstract or in experimental contexts from various situations. It is reported that Malaysia obtained a score below the average benchmark set by TIMSS in Grade 8 Science TIMSS 2015 (Martin et al., 2016). This result has indicated that most of the Grade 8 students including Malaysian students need to improve basic Science knowledge. Thus, by having a more precise ability measurement such as using CAT, an effective action could be taken to improve the students' abilities in Science knowledge.

A calibrated item bank plays a vital role in CAT (Wise & Kingsbury, 2000). A calibrated item bank is a collection of a well organized psychometrically tested items in which the items contain data such as the subject matter objective, item difficulty level and other item's psychometric traits. Two scores measured from two different set of test items can be compared as all the items are originated from the same calibrated item bank (Choppin, 1976). For a high-stake test, 1000 items and above is needed to be used in CAT (Wise & Kingsbury, 2000) while for a low-stake test, 100 items and below is enough to be used in CAT (Gershon, 2005).

Test linking and equating needs to be done when there are a lot of items need to be calibrated. Linking is the process of pairing two scores with no strong evidence that the two scores have equivalent meaning while equating is a procedure carried out to produce a comparable score with a strong evidence that the scores have the same meaning (Ryan & Brockmann, 2018). The most common used equating design is the anchor test design or common item equating design in which two sets of tests contain a set of same items which is referred to as common items. Two criteria need to be concerned in this design is the anchor item representation and its location in a test set. A set of anchor items must represent a mini version of a test in which it contains items of every topics involved in a one test set. In terms of the anchor items location, there are three types of anchor items arrangement which are internal, embedded anchor items; internal, appended anchor items and external anchor items. In practice, the anchor items can be located anywhere in a test set but if all anchor items are positioned at the end section of a test set, the examinee's performance on the anchor items might be problematic due to declining motivation. The advantage of using anchor test design is it involve the non-equivalent group of respondents to produce basis for linking and equating the two test sets as long as the respondents have the same characteristic (López-cuadrado et al., 2008; Ryan & Brockmann, 2018). For example, Test Set A is assigned to Grade 8 students from School A while Test Set B is assigned to Grade 8 students from School B.

After the process of test linking and equating and the collection of data, the item bank should be calibrated using a suitable probabilistic model to enable the computer to select the best item based on the current ability of a student. There are three probabilistics models for CAT which are: (i) Item Response Theory (IRT), (ii) Bayesian; and (iii) neural networks (Aleksander & Morton, 1995; Culbertson, 2015; Linacre, 2000). The two most common probabilitics model used for CAT are Bayesian networks and IRT. The main difference between IRT and Bayesian networks in CAT systems is that IRT computes the probability of a correct answer to a question depending on the student's knowledge level and the answers to previous questions, whereas Bayesian networks computes the probability of a correct answer to a question taking into account the probability of the answers to previous questions.

In comparison to the three probabilitics model, IRT is considered as the most established probabilistic model for CAT as it considers both students' ability and item difficulty; thus, the Rasch model which based on IRT is seemed to calibrate the item bank for CAT. Through this model, the items' difficulty level and students' ability are classified and ordered in the same linear scale. Hence, any two different sets of items can be measured on the same scale as the items are taken from the same item bank which has been calibrated on one linear scale (Bond & Fox, 2007; Linacre, 2000). Also, the issue of test fairness will be solved when comparing students' abilities although each student answered a different set of items (Linacre, 2000). Furthermore, the Rasch model meets measurement objectivity as listed by Mok and Wright (2004), as cited in Sumintono (2016), which is the ability of the model to create one linear scale, more precise prediction, ability of the model to identify problematic items and any missing items, and the ability of the model to enable replication in measurement.

As the TIMSS items were well established (Mullis & Martin, 2017), the usage of TIMSS items in CAT might produce a more precise ability measurement of students in Science subject. Therefore, this research was carried out by focusing on the calibration of the released objective Grade 8 Science items in the Science Trends in International Mathematics and Science Study (TIMSS) using the Rasch model for a low stakes CAT in order to provide a more precise students' abilities to improve their performance in Science subject. According to Suah and Ong (2012), there are few high-quality constructed items because Malaysian teachers have limited skill in test construction. This might be because most of the teachers do not follow a proper procedure and they are more likely to depend on a textbook, reference book and national exam questions to construct items. This is why the procedures in test construction using IRT should be

Table 1. Item Specification Table of TIMSS Items for Biology

Subject	Topics	Skill Levels	Total Items
Biology	Characteristics and life processes of organisms	Knowing	10
		Applying	1
	Cells and their functions	Knowing	6
		Applying	1
	Life cycles, reproduction and heredity	Knowing	7
		Applying	1
	Diversity, adaptation and natural selection	Knowing	7
	Ecosystems	Knowing	3
		Applying	3
	Human Health	Knowing	6
Total			45

focused on teacher professional devlopment as the measurement theory is more flexible.

This study was carried out to answer the questions of research related to the process of calibration of the item bank for Computerized Adaptive Testing (CAT). The research questions were as follows:

i. Does each set of administered tests conform to the fit item analysis requirements based on the Rasch model?
ii. Does each item in each set of test meets the item's parallel condition?
iii. Is the test linking and equating process able to produce a calibrated item bank that conform to fit item analysis requirements?
iv. Is the test linking and equating process able to produce a calibrated item bank that meet Rasch Principal Component Analysis (RPCA) requirements?
v. Is the test linking and equating process able to produce a calibrated item bank that meet the requirements of the item's parallel condition?
vi. What is the distribution value of the items in the calibrated item bank acquired through Rasch model analysis?

RESEARCH METHODOLOGY

The research was a quantitative research and implemented a cross-sectional survey approach in which the instrument was given to the respondents in a specific point in time (Fraenkel & Wallen, 2009). The instrument involved in this study was a Science TIMSS test administered through a Paper and Pencil test. The respondents involved in this test were selected through purposive sampling method (Idris, 2013) from secondary schools in the northern part of Malaysia. The construction of the item bank for CAT was the focus in this study, and the methodology used for constructing item bank was based on Bjorner et al. (2007). There were four procedures involved in the process of calibration of item bank in this study which were: (i) the adoption of items, (ii) the test equating and linking, (iii) the testing of items and, (iv) the item analysis.

Adoption of Items

In this first stage, the Grade 8 Science TIMSS topics content were reviewed and compared with the Malaysia KSSM Form 1 and Form 2 Science topics before the items were adopted from the TIMSS 2003-2015 released items to ensure that the items adopted fit with the local Science curriculum content. The released TIMSS items consist of objective and subjective items (Mullis & Martin, 2017) but only objective items were selected for ease in marking.

The items adopted to be used in this research test knowing and applying skills because most of released objective TIMSS items are focusing on testing of knowing and several applying skills, while most subjective items are focusing more on testing of applying skills and reasoning skills. The number of items chosen for each topic were based on the availability of the released items that fit with the local Science curriculum so the number of items adopted may vary in quantities for each topic and skill tested. Table 1 shows the Item Specification Table for Biology and the level of testing for 45 items.

Table 2 shows the Item Specification Table and the level of testing for Chemistry with 28 items.

Table 3 shows the Item Specification Table and the level of testing for Physics with 30 items.

Table 4 shows the Item Specification Table for Earth Science and the level of testing with 22 items.

In this study, a total of 125 multiple-choice questions were adopted during the early stage of building item bank. According to Gershon (2005), an amount of 100 items and below for an item bank is enough to be used in a low stakes testing environment.

Table 2. Item Specification Table of TIMSS Items for Chemistry

Subject	Topics	Skill Levels	Total Items
Chemistry	Composition of matter	Knowing	10
		Applying	1
	Properties of matter	Knowing	7
		Applying	1
	Chemical change	Knowing	9
Total			28

Table 3. Item Specification Table of TIMSS Items for Physics

Subject	Topics	Skill Levels	Total Items
Physic	Physical states and changes in matter	Knowing	4
		Applying	1
	Energy transformation and transfer	Knowing	5
		Applying	3
	Light and sound	Knowing	4
		Applying	4
	Electricity and magnetism	Knowing	2
		Applying	1
	Forces and motion	Knowing	2
		Applying	4
Total			30

Table 4. Item Specification Table of TIMSS Items for Earth Science

Subject	Topics	Skill Levels	Total Items
Earth Science	Earth's structure and physical features	Knowing	2
	Earth's processes, cycles and history	Knowing	7
		Applying	1
	Earth's resources, their use and conservation	Knowing	9
	Earth in the solar system and the universe	Knowing	3
Total			22

Test Equating and Linking

Based on the IRT, the difficulty level of the items is determined based on the students' responses towards the items, in order to produce a calibrated item bank (Fisher & Molenaar, 1995). As there was a total of 125 adopted items, it was impossible to administer all the items to a single student, so the items need to be divided into several sets of items that were suitable to be administered on a single student.

In the second stage, the items linking and equating process was carried out when more than one set of items needed to be measured. A linking and equating procedure enabled the difficulty level of all items in every set to be measured by the Rasch model using one linear scale (Bond & Fox, 2001). This study used concurrent common item equating method in the linking and equating process because it is the best method based on the items characteristics (Linacre, 2012b). By using this method, the selection of items to be used as common items is important because the measurement from these common items will generate the one main linear scale that will be used to measure all other items (López-cuadrado et al., 2008). Common items selected from the adopted TIMSS items were based on the following criteria (Ryan & Brockmann, 2018): (1) common items should represent all the tested topics, (2) very difficult or very easy common items should be avoided, (3) a minimum of five common items should be used to link the two consecutive sets, and (4) different sets of common items that link the two consecutive sets should be used to reduce item exposure level.

From 125 adopted TIMSS items, the items were divided into five sets containing 30 items per set. Each test set consisted of 20 unique items and 10 common items. Unique items mean that the items are only available in one test set, not in another test set. Common items are the same items that appear in two consecutive sets. Table 5 shows the item ordering pattern in the linking and equating process using the concurrent common item equating technique for unique items and common items, based on the suggestion from Linacre (2012b). In total, there were 100 unique items and 25 common items.

Table 5. Item Ordering Pattern in Each Set of Questions

Sets	Item Ordering Pattern
1	5 items are common to set 5 + 20 unique items + 5 items are common to set 2
2	5 items are common to set 1 + 20 unique items + 5 items are common to set 3
3	5 items are common to set 2 + 20 unique items + 5 items are common to set 4
4	5 items are common to set 3 + 20 unique items + 5 items are common to set 5
5	5 items are common to set 4 + 20 unique items + 5 items are common to set 1

Set 1	C	U	C										
Set 2			C	U	C								
Set 3					C	U	C						
Set 4							C	U	C				
Set 5	C								C	U			

Figure 1. Visual Overview of Item Ordering Patterns

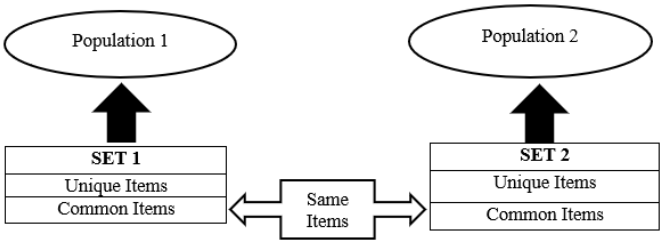


Figure 2. Anchor Test Design

Figure 1 illustrates the visual order of unique items labelled as "U" while common items labelled as "C" based on the pattern shown in Table 5. By referring to Figure 1, there were five (5) common items that linked the two consecutive sets that makes a total of 25 common items for five (5) sets. Common items are not so appropriate to be at the end of the set because the motivation to answer the item at the end might be reduced thus providing a problematic response which can affect the measurement (Ryan & Brockmann, 2018). Thus, the common items were positioned in the first part and the last part of each set as suggested by Linacre (2012b).

Testing of Items

In the third stage, all the five sets were administered to Form 2 students through the Paper and Pencil test. Computerized testing was not involved in this stage due to the limitation of the facilities that make it difficult to administer the test properly for a larger sample size concurrently. The purpose in conducting this linear conventional test was to collect students' response towards items so that an item analysis could be made to obtain the item difficulty parameter (Fisher & Molenaar, 1995). Through anchor test design (Ryan & Brockmann, 2018), the five test sets were distributed to five Dual Language Programme secondary schools in northern part of Malaysia in which each school received one test set as shown in Figure 2.

In each school, four (4) Form 2 classrooms were purposively selected consisting of advanced, high, intermediate and low achieving students. Approximately 30 Form 2 students per class involved; thus total respondents involved per school were approximately 120 Form 2 students for one test set. The minimum sample size requirement is 30 respondents for statistical stability in the analysis using the Rasch model (Linacre, 2012b). According to the Rules of Thumb, the bigger the sample size, the higher the measurement precision (Van, 2002). Therefore, a larger sample size would provide better measurement. Then, the teachers from each school were responsible for administrating this Paper and Pencil test concurrently and the students were given one hour to complete the test.

Item Analysis

After the completion of the Paper and Pencil test, all the responses given by students towards each item in five test sets were recorded in SPSS in which a score of 1 was given for every correct answer and a score 0 was given for every wrong answer. After that, the raw data from SPSS was converted into Winsteps's control and data files which then to be analysed by Winsteps software version 3.74.0 based on Rasch model. In Winteps, the fit analysis of a single test set was carried out separately followed by the fit analysis of all the sets simultaneously in assessing the psychometric characteristics of the item bank to answer the research's questions.

RESEARCH RESULTS

Item fit analysis was conducted in every set separately to detect misfiting items in each set, because items that did not meet the fit criteria within a set also did not meet the fit criteria in the overall set analysis (Linacre, 2012a). The INFIT/OUTFIT MNSQ value within the range 0.5 to 1.5 is considered useful for measurement. Items with MNSQ value within this range are considered to fit the Rasch measurement while items with MNSQ value out of the range are considered as misfiting items (Wright & Linacre, 1994 as cited in Boone et al., 2014).

Table 6. Items That Did Not Meet 'Item-Fit' Requirements in Each Set

Set	Item with INFIT/OUTFIT MNSQ value out of the range ($0.5 \leq MNSQ \leq 1.5$)	Total Item
1	26	1
2	5	1
3	-	0
4	-	0
5	19	1
Total		3

Table 7. Point Measure Correlation Value of Items in Each Set

PTMEA Corr. Value	Set 1	Set 2	Set 3	Set 4	Set 5
.00 to .20	5	4	1	7	2
.21 to .30	8	8	3	9	2
.31 to .40	8	8	9	10	10
.40 to .50	6	5	14	3	8
.51 to .60	1	2	3	1	5
.61 to .70	0	1	0	0	2
Total Item	28	28	30	30	29

Table 8. Unique and Common Items That Did Not Meet 'Item-Fit' Requirements in MFORMS

Original Set	Unique Item (U)	Common Item (C)	Total Items
1	-	-	0
2	-	-	0
3	-	-	0
4	-	-	0
5	U508	-	1
Total			1

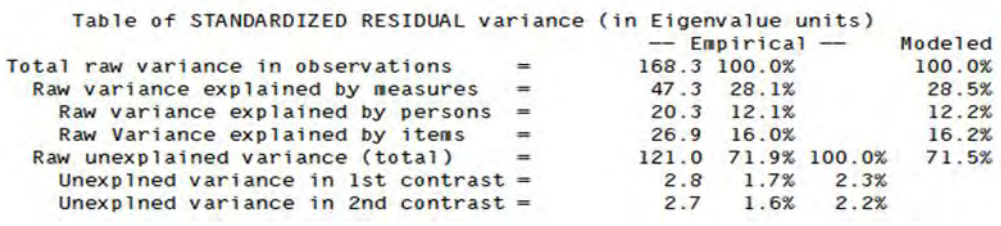
Table 6 shows the number of items which did not meet the fit-item criteria in each set of questions. According to Table 6, there were three misfitting items with MNSQ value of less than 0.5. Item 26 and Item 5 were common items found in both Set 1 and Set 2 while Item 19 was a unique item. According to Wright and Linacre (1994), as cited in Boone et al. (2014), item with MNSQ value less than 0.5 might exhibit misleading item properties thus the three misfiting items were eliminated from the the respective sets. The elimination of these three items did not affect the skills being measured as there were other items which measured similar skills after been checked by the experts.

Then, the item's parallel analysis was carried out and the remaining items showed a positive Point Measure Correlation (PTMEA Corr.) value, which meant that all the items exhibited a one-way characteristic and tested a similar construct, thus met the item's parallel criteria (Linacre, 2019a). Table 7 shows the distribution of the PTMEA Corr. values and the number of items. Based on Table 7, it was found that most of the items in each set obtained PTMEA Corr. value greater than .20. Items with PTMEA Corr. reading value smaller than .2 indicated misleading of the items, but the removal of these items were not carried out because the obtained PTMEA Corr. values were influenced by the reliability of the data, the item target on the individual sample, and the individual sample distribution (Linacre, 2019a).

The fit-item analysis of all the sets simultaneously was then executed by using the MFORMS= function that linked and equalized each common items and unique items from multiple sets into a single set of items (Linacre, 2012b). Table 8 shows unique and common items that did not meet 'item-fit' requirements after the MFORMS= analysis. Unique items labelled as "U" were found in one set only while common items labelled as "C" were similar items found in two consecutive sets.

Based on the Table 8, there was one misfiting item, U508 (item 8 in Set 5) with MNSQ value of 1.51. According to Wright and Linacre (1994), as cited in Boone et al. (2014), an MNSQ value more than 1.5 but less than 2.0 is considered unproductive for measurement but it does not distort the measurement. Item U508 had MNSQ value just slightly a little higher than 1.5 plus the TIMSS items adopted were well established; thus this misfiting item was not eliminated from the item bank. A total of 122 items met the overall fit criteria.

Figure 3 shows the Rasch Principal Component Analysis (RPCA) of Science TIMSS item bank. The construct is considered unidimensional if the Eigenvalue in the unexplained variance of the 1st contrast is equal or less than 2.0 (Linacre, 2013). From this analysis, it was found that the Eigenvalue of the unexplained variance in the 1st contrast was 2.8 units. Further analysis had been done in Winsteps and it shows that there were 9 clustered items separated from the rest of items (Appendix 1). According to Linacre (2019b), different item content such as subtraction, addition, division and multiplication in Mathematics subject could produce a second dimension statistically. However, those items are needed in measuring a student's Mathematics ability, so the construct is considered unidimensional. These 9 items in this analysis were testing basic Science knowledge namely Biology, Physic and Chemistry without the involvement of testing any numerical aspect.



```
    Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                          -- Empirical --     Modeled
Total raw variance in observations     =      168.3 100.0%        100.0%
  Raw variance explained by measures   =       47.3  28.1%         28.5%
    Raw variance explained by persons  =       20.3  12.1%         12.2%
    Raw Variance explained by items    =       26.9  16.0%         16.2%
  Raw unexplained variance (total)     =      121.0  71.9% 100.0%   71.5%
    Unexplned variance in 1st contrast =        2.8   1.7%   2.3%
    Unexplned variance in 2nd contrast =        2.7   1.6%   2.2%
```

Figure 3. Rasch Principal Component Analysis of Science TIMSS Item Bank

Table 9. Correlation Analysis on Person Measure for 122 Items and 113 Items

		Person Measure for 113 Items	Person Measure for 122 Items
Person Measure for 113 Items	Pearson correlation	1	0.983**
	Sig. (2-tailed)		< .0001
	N	523	523
Person Measure for 122 Items	Pearson correlation	0.983**	1
	Sig. (2-tailed)	< .0001	
	N	523	523

**. Correlation is significant at the 0.01 level (2-tailed).

Table 10. Item Reliability and Separation Index of Calibrated Science TIMSS Item Bank

	Total Score	Count	Measure	Model Error	Infit		Outfit	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	79.4	122.8	.00	.24	1.00	.1	.97	-.1
S.D.	38.2	41.8	1.17	.07	.11	1.3	.20	1.2
Max.	190.0	222.0	3.15	.72	1.36	4.6	1.51	4.0
Min.	18.0	78.0	-3.26	.15	.73	-2.8	.52	-2.6
Real RMSE		.26	True SD	1.14	Separation	4.39	Item Reliability	.95
Model RMSE		.25	True SD	1.14	Separation	4.39	Item Reliability	.95
S.E. of Item		Mean	= .11					

Table 11. Point Measure Correlation Value of Items in Science TIMSS Item Bank

PTMEA CORR. Value	Total Item
.00 to .20	11
.21 to .30	26
.31 to .40	46
.41 to .50	26
.51 to .60	12
.61 to .70	1
Total Item	122

Table 12. Number of Items by Difficulty Level in Calibrated Science TIMSS Item Bank

Item difficulty level (Logit)	Total item
Above 3.00	1
2.50 to 2.99	0
2.00 to 2.49	4
1.50 to 1.99	7
1.00 to 1.49	13
0.50 to 0.99	14
0.00 to 0.49	20
-0.01 to -0.50	24
-0.51 to -1.00	14
-1.01 to -1.50	13
-1.51 to -2.00	6
-2.01 to 2.50	2
-2.51 to -3.00	2
Below -3.00	2
Total	122

Then, the correlation analysis on the person measure (students' ability) had been done between the whole item bank and the item bank without those 9 items (Table 9). The analysis shows that there was a strong correlation of person measure for the two item banks. This also shows that the administration of the 9 items were measuring the same construct with the other items thus the item bank was considered unidimensional.

Table 10 shows the item reliability of .95 and separation index of 4.39 for the Science TIMSS item bank. The reliability index and separation index are considered high if the values exceed .9 and 3.0 respectively. High item reliability and separation index indicate that the total respondents involved was enough to validate the tested construct (Linacre, 2012 as cited in Boone et al., 2014).

The item's parallel condition via MFORMS= analysis was also checked. Table 11 shows the distribution of PTMEA Corr. values and number of items after being analysed. All the items obtained positive PTMEA Corr. value indicated that all items were one-way thus confirming the item's parallel criteria (Linacre, 2019a).

Table 12 shows the distribution of the number of items with the respective level of difficulty. Result from the MFORMS= analysis showed that the difficulty level for 122 items were in the range from -5.30 logit to +3.15

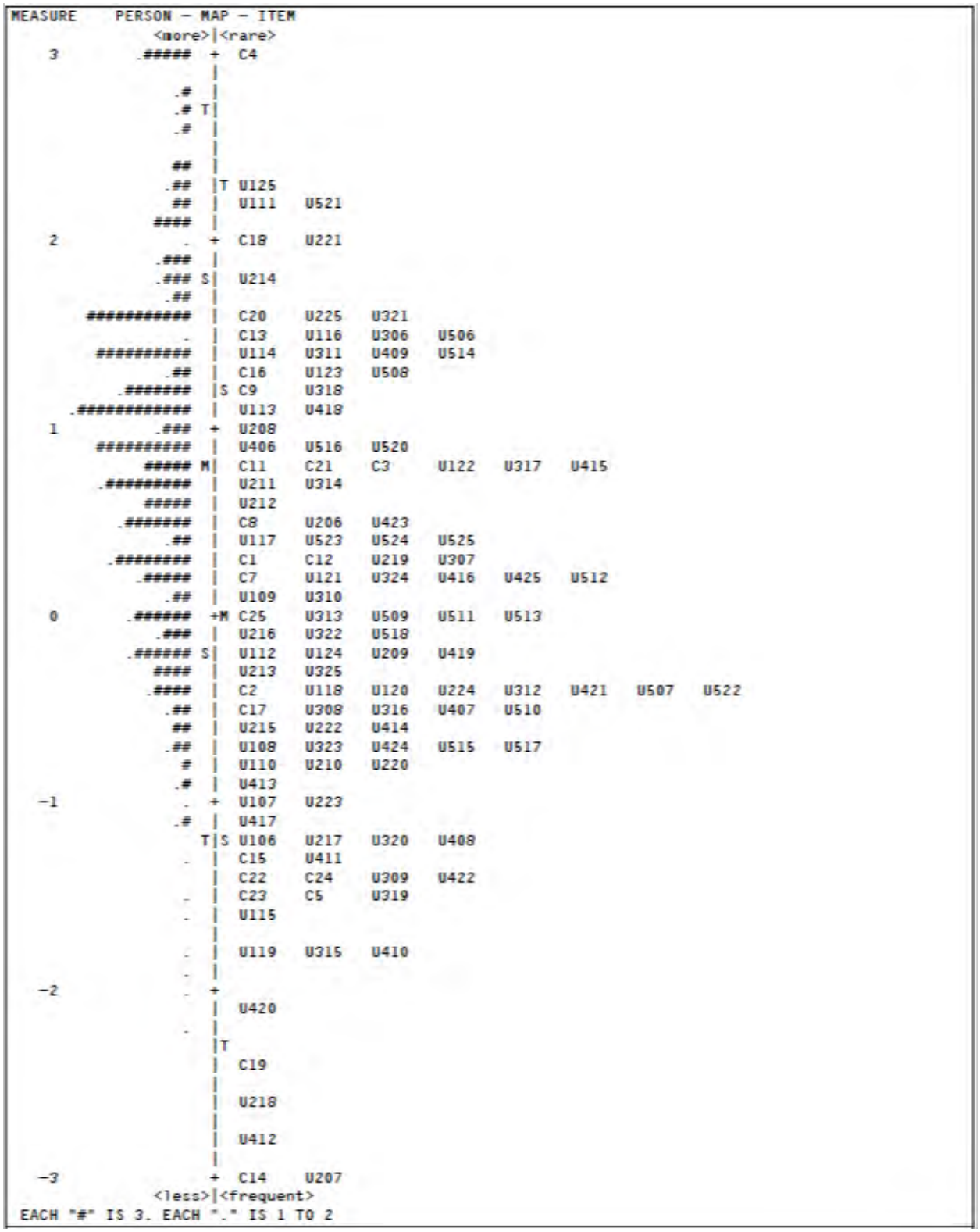


Figure 4. Map of Calibrated Science TIMSS Item Bank Individual-Item

logit. Most of the items were located at the middle scale of difficulty level ranging from -1.50 logit to +1.49 logit.

Figure 4 shows Person-Map-Item which display the distribution of person ability on the left side of map and item difficulty level on the right side of map in one linear scale measured in logit unit. Based on Figure 4, the mean (*M*) person measure was higher than the mean (*M*) item measure which was nearly one logit higher. This indicates that students' performance was better than the items' performance. Item C14 was the easiest item while item C4 was the most difficult item. In addition, there was a wide gap between difficult item C4 and U125. According to the Rasch model, this wide gap indicated construct under representation (Baghaei, 2008). In future research, another TIMSS items from the next TIMSS

cycle should be added in this gap to ensure all constructs are measured (Boone et al., 2014).

Generally, the tests performed according to the target group, as most items were scattered on the middle scale of difficulty levels between -1.50 logit to +1.49 logit. This position indicated that most items were at average level of difficulty. However, there were several items with lower difficulty levels corresponding to low ability level students and items at higher level for high achieving students. Along the scale, most of the items' difficulty level matched the majority of students' abilities. Thus, the test was relatively not so easy or so difficult for the sample respondents involved. Based on the map in Figure 4, the calibrated items were suitable to be used by every Form 2 student. This is because the item's difficulty parameter was constant across the study population, as the item bank was calibrated using IRT (Bichi et al., 2015).

DISCUSSION

The calibrated item bank is a crucial element in CAT. A large collection of items for a calibrated item bank can increase the test integrity as the students or the teachers have difficulty guessing an item that will be used (Chuesathuchon & Waugh, 2010). In order to produce a calibrated item bank, an item's psychometric properties need to be examined. The research adopted a total of 125 released objective Grade 8 Science TIMSS 2003-2015 items. Through the concurrent common item equating method, the adopted items were administered to secondary (eighth grade) school students. The response of students to every item was recorded and analysed by using Winsteps software according to Rasch model.

The psychometric analysis was carried out involving a fit analysis based on the MNSQ value range between 0.5 to 1.5, the item's polarity analysis through PTMEA Corr. value, unidimensionality analysis by referring to unexplained variance in the 1st contrast, items measure analysis, and Person-Map-Item analysis. In the fit analysis, as the released TIMSS items were limited, an MNSQ value range 0.5 to 1.5 was used to secure as many items as possible because this range is considered acceptable (Boone et al., 2014). For dichotomous items such as objective items, an MNSQ value range 0.7 to 1.3 for fit analysis supposedly needs to be used to reduce the noise in the data (Aziz et al., 2013). Although, an MNSQ value range between 0.5 to 1.5 is acceptable, the range would produce more noise than an MNSQ value range 0.7 to 1.3. In a future study, fit analysis should use an MNSQ value range 0.7 to 1.3 if the number of dichotomous items is not limited.

Fit analysis reported in this study shows that there were three (3) misfiting items. The TIMSS items' high quality in reliability and validity are undeniable as the items had been well tested by the TIMSS experts. The findings of this research also show that the adopted TIMSS items achieved high item's reliability (Table 7). There were reasons that caused the item to be misfitted. Based on the research by Chuesathuchon and Waugh (2010) who also used Rasch model to calibrate an item bank, they proposed that the misfiting items might have been due to the wording problems, translation mistakes and/or careless mistakes from the students. However, in this study, the TIMSS items were adopted without changing any wording structure or translating to other language. So, these misfiting items might be due to conflicting content of Malaysia Science curriculum and the TIMSS Science curriculum. Although the main topics for both curricula are generally the same, Malaysian teachers might teach certain topics in different depth slightly different from the TIMSS curriculum. Therefore, students might give different types of responses that cause the items to be misfited according to the Rasch model. In future research, the curriculum used should be revised by the teachers who have more than five years of teaching experience to ensure that the students learn according to the curriculum.

The number of misfiting items in this study met the expectation as the items were well developed. In the research by Chuesathuchon and Waugh (2010), more than half of the items produced in the first stage needed to be eliminated because those items did not meet the fit criteria analysis. According to Wise and Kingsbury (2000) and Bjorner et al. (2007), it is recommended to construct a large item bank in the first step before testing the items. This is because, by using a large item bank, the remaining fit item bank is considered large enough after the elimination of misfiting items; thus there are still enough items to measure all the required skills.

After the elimination of three (3) misfiting items in this study, the rest of the items had positive PTMEA Corr. values indicating that the items measure the same construct thus meeting the item's parallel condition. Next, in the unidimensional analysis, the result shows an Eigenvalue of 2.8 which was higher than the standard value 2.0. According to Linacre (2013), if the Eigenvalue displayed is higher than the standard value, further analysis in Winsteps must be carried out to check the content of the clustered items to decide whether the construct contains second strands or second dimensions (Appendix 1). For example, different operation in Mathematics such as addition, substraction, division and multiplication might produce a second dimension statistically. However, practically those operations are needed in measuring general Mathematics knowledge, thus the construct contains second strands and is said to be unidimensional. If the content of the clustered items measures a significantly different property, the construct does not meet the unidimensional criteria. In this research, there was no significant difference in terms of content in the items used because all the items measure basic Science knowledge without the involvement of any numerical testing aspect. As the analysis shows there

was a high correlation of person measure between 122 items and 113 items, thus the construct of calibrated Science TIMSS item bank was considered unidimensional. The analysis also shows that all the items measured contained a high item reliability and separation indexes.

The difficulty level of items was then checked. From the MFORMS= analysis, it was found that most of the items were located at the medium difficulty level. Since all the Science items adopted measured knowing and application skills, it makes sense that most of the items have average difficulty level. However, there were several easier items and more difficult items available for low ability and high ability students. By referring to the Person-Map-Item in this research (Figure 4), the mean of person measure was higher than the mean of item measure by nearly one logit, indicating that the person's performance was better than the items. According to Boone et al. (2014), a well targeted items on respondents are revealed when the mean item measure and the mean person measure are on the same level on the scale. However, most of the persons' abilities were paired with the respective items' difficulties level. Therefore, the items tested were considered appropriate for the sample of respondents involved.

There was a huge gap between the two most difficult items in the map indicating construct underrepresentation (Baghaei, 2008) hence additional items need to be added in this calibrated item bank to fill in the gap so that the students' ability between this gap can be measured more accurately. Based on the map, the gap was between two difficult items in which common item was the most difficult item. This analysis proved that the use of very difficult common items is not appropriate as stated from Ryan and Brockmann (2018). Very difficult common item might generate significant differences in achievement between low ability students and high ability students which then affect the scale measurement. This study chose common items based on the cognitive level of the item's information. In future study, a pilot test should be done on the items in order to choose the most suitable items to be used as common items statistically. As this research had adopted as much as possible TIMSS items from TIMSS 2003 to 2015, in future research, additional Science TIMSS's objective items could be added to improve this calibrated item bank.

Rasch analysis has shown the quality of the 122 adopted Science TIMSS items used were good and met the Rasch's assumptions. According to Gershon (2005), 100 items and below are enough to be used in a low stakes test. The use of the Rasch model in producing the calibrated item bank enables the CAT to tailor the test level with person's ability; hence, different sets of items can be measured using the same scale avoiding the test fairness issue (Eggen, 2007). It is found that the calibrated item bank is suitable to be used in CAT in testing student's ability in Science subject. The use of CAT as a testing mode for TIMSS is aligned with the students' growth model criteria. O'Malley et al. (2011) described three characteristics of "student growth models" which are: the obtained scores can be mathematically compared from one occasion to another; they can be connected for the same students over two or more occasions; and they can indicate trait changes. TIMSS, which is in testing mode of CAT, is able to measure the growth of students' performance in answering TIMSS questions. This is because of the existence of the item bank which is able to produce different set of TIMSS questions which are comparable on two different occasions. Nevertheless, those sets are comparable psychometrically as they are linked and equated by using common TIMSS items, which is the fundamental principle of the development of TIMSS item bank for CAT.

Students' performance on TIMSS implies change over time. This progress covers different levels of students' abilities, whether the students who had a problem answering a difficult TIMSS question, or the students belonged to high achieving group in answering TIMSS items. TIMSS in CAT mode begins with a set of test TIMSS items that are calibrated with different levels of difficulty to be adapted with the current students' abilities. A CAT is administered to a student and at a later date, another CAT is administered from the same set of TIMSS items, but TIMSS items previously administered to that student are not used. This process is repeated at later points in time. Simulatenously, an individual profile of change is obtained with a minimum number of TIMSS items administered for each student. When measured change is identified by this procedure, the data can also provide information on when the change occurred for each student, thus identifying the points in the instruction by teachers that had an impact on a student's measured levels on the TIMSS performance. Therefore, teachers also can do some self-evaluation on their teaching performances after giving some interventions to improve their students' performances in TIMSS. Measurement data of CAT on TIMSS performance at each point in time can also be aggregated across students to track group progress. Therefore, policymakers, researchers, curriculum developers, and educators at all levels could use TIMSS data and findings to learn about the kinds of curriculum and instructional practices that were associated with the highest levels of TIMSS performance measured by CAT.

This study has few limitations; for instance, as the mode of testing informs CAT, the choices of TIMSS questions asked were limited to multiple choice format. Due to nonexistence of questions in free response form, the measurement of CAT did not involve traits related to complex reasoning which need written explanations. Nevertheless, the results of Rasch item person map (Figure 4) had indicated that the difficulty of the TIMSS

in CAT were spread widely along a continuum from the most difficult item to the least difficult item. Moreover, the item separation index, which is greater than 4.0, indicated that the TIMSS items in CAT were able to discriminate the TIMSS students well. Although TIMSS in CAT did not measure complex reasoning, the information obtained from CAT could be used by the teachers to identify the students' performance level in answering TIMSS question in a multiple-choice format. It is argued that in order to enable students to achieve higher level of performance in TIMSS assessment test, it should start from the basic understanding of science content such as knowledge which is tested through multiple choice questions form. TIMSS items in CAT itself has its own difficulty as it contains a mixture of various science topics which challenge the students to identify a relevant science concept in order to answer each TIMSS item in CAT. This is alligned with Fensham's (1998) findings which found that students have a problem in identifying suitable science topics that they have learned and applied their knowledge to answer TIMSS questions which are presented in a different context.

CONCLUSION

The measurement quality of CAT depends on the integrity of its calibrated item bank. By using IRT such as Rasch model in assessing psychometric properties of items, the measurement by CAT becomes more flexible. Students' abilities could be compared to each other even though each student receives unique set of items. The concurrent common item equating method used in the linking and equating procedure in this study was the best method based on the items' characteristics. Using this method, unique items and common items in every sets were ordered as such based on the recommendation by Linacre (2012b) for ease in analysing through Winsteps software. At the end of this study, 122 adopted TIMSS items met all the Rasch's analysis stated in the research questions. It is hope that, the demonstration of this item bank calibration for multiple-choice questions in this study could be implemented for further research and to be used in CAT.

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

Aleksander, I., & Morton, H. (1995). *An introduction to neural computing: Information systems.* International Thomson Computer Press.

Aziz, A. A., Masodi, M. S., & Zaharim, A. (2013). *Asas model pengukuran Rasch: Pembentukan skala & struktur pengukuran (Fundamental of Rasch 's measurement model: Scale formation & measurement structure).* Universiti Kebangsaan Malaysia.

Baghaei, P. (2008). The Rasch model as a construct validity tool. *Rasch Measurement Transactions, 22*, 1145-1146. Retrieved from https://www.researchgate.net/publication/267330326_The_Rasch_model_as_a_construct_validity_tool/citations

Barker, T. (2008). Computer-adaptive testing in higher education: The validity and reliability of the approach. In F. Khandia (Ed.), *12th CAA International Computer Assisted Assessment Conference* (pp. 25-40). Lougborough University. Retrieved from http://caaconference.co.uk/pastConferences/2008/proceedings/Barker_T_final_j1_formatted.pdf

Bichi, A. A., Embong, R., Mamat, M., & Maiwada, D. A. (2015). Comparison of classical test theory and item response theory: A review of empirical studies. *Australian Journal of Basic and Applied Sciences, 9*(7), 549-556.

Bjorner, J. B., Chang, C.-H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: Item banking and computerized adaptive assessment. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *16*(1), 95-108. http://doi.org/10.1007/s11136-007-9168-6

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model fundamental measurement in the human sciences.* Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410600127

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model fundamental measurement in the human sciences* (2nd ed.). Lawrence Erlbaum Associates.

Boone, J. W., Staver, R. J., & Yale, S. M. (2014). *Rasch analysis in the human sciences.* Springer. https://doi.org/10.1007/978-94-007-6857-4

Choppin, B. (1976). *Developments in item banking. Monitoring national standard of attainment in schools*, 216-234. Retrieved from http://www.rasch.org/memo76.pdf

Chuesathuchon, C., & Waugh, R. F. (2010). Item banking and computerized adaptive testing with Rasch measurement: An example for primary mathematics in Thailand. In R. F. Waugh, (Ed.), *Applications of Rasch Measurement in Education* (pp. 1-36). Nova Science.

Conole, G., & Warburton, B. (2005). A review of computer-assisted assessment. *Research in Learning Technology*, *13*(1), 17-31. http://doi.org/10.1080/0968776042000339772

Culbertson, M. J. (2015). Bayesian networks in education assessment: The state of the field. *Applied Psychological Measurement, 40*(1), 3-21. https://doi.org/10.1177/0146621615590401

Davey, T. (2011). *A guide to computer adaptive testing systems.* Council of Chief School Officers. Retrieved from https://www.semanticscholar.org/paper/A-Guide-to-Computer-Adaptive-Testing-Systems.Davey/71d9d258a7b161db2a3848b85afaedc105ef11fa

Eggen, T. (2007). Choices in CAT models in the context of educational testing. In J. E. Hartig, E. Klieme, & D. Leutner (Eds.), *Proceeding of the GMAC Conference on Computerized Adaptive Testing* (pp. 199-217). Hogrefe & Huber Publisher. Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Fensham, P. J. (1998). Student response to the TIMSS test. *Research in Science Education, 28*(4), 481-489. https://doi.org/10.1007/BF02461511

Fisher, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models foundations, recent developments and applications*. Springer-Verlag New York Incorporation.

Fraenkel, J. R., & Wallen, N. E. (2009). *How to design and evaluate research in education* (7th Ed.). McGraw-Hill.

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement, 6*(1), 109-127. Retrieved from https://www.scholars.northwestern.edu/en/publications/computer-adaptive-testing-2

Idris, N. (2013). *Penyelidikan dalam pendidikan (Research in education)* (2nd Ed.). McGraw-Hill Education.

Linacre, J. M. (2000). Computer-adaptive testing: A methodology whose time has come. In S. Chea, U. Kang, & J. M. Linacre (Eds.), *Development of Computerized Middle School Achievement Test.* Komesa Press. Retrieved from https://www.rasch.org/memo69.pdf

Linacre, J. M. (2012a, June). *Winsteps Rasch tutorial 3*. Retrieved from http://www.winsteps.com/a/winsteps-tutorial-3.pdf

Linacre, J. M. (2012b, December). *Some question on linking and equating method.* Old Rasch Forum- Rasch on the Run: 2012. Retrieved from https://www.rasch.org/forum2012.htm

Linacre, J. M. (2013, March). *Unidimensionality with dichotomous data.* Retrieved from https://www.rasch.org/forum2013a.htm

Linacre, J. M. (2019a). *Correlations: Point-biserial, point-measure, residual.* Retrieved from http://www.winsteps.com/winman/correlations.htm

Linacre, J. M. (2019b). *Dimensionality investigation- an example.* Retrieved from https://www.winsteps.com/winman/multidimensionality.htm

Ling, S. S., Lan, O. L., Suah, S. L., & Ong, S. L. (2012). Investigating Assessment Practices of In-service Teachers. *International Online Journal of Educational Science, 4*(1), 91–106. Retrieved from http://www.iojes.net/index.jsp?mod=makale_ing_ozet&makale_id=41233

López-cuadrado, J., Armendariz, A., Pérez, T. A. & Arruabarrena, R. (2008). Helping tools for item bank calibration and development of computerized adaptive test. In E. L. G. Chova, D. M. Belenguer, & I. C. Torres (Eds.), *Proceeding of International Technology, Education and Development Conference (INTED'08), valensia* (pp. 1-9). International Association of Technology, Education and Development. Retrieved from http://www.sc.ehu.es/jiwarsar/Research-Papers/INTED08HELPINGToolsForItemBank-LopezCuadrado.pdf

Magyar, A. (2015). *Comparing measurement effectiveness of computer-based linear and adaptive test* (Doctoral Dissertation). Retrieved from http://doktori.bibl.uszeged.hu/2633/3/PhD%20theses%20MA.pdf

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

Masrom, S., & Abd. Rahman, A. S. (2009). An adaptation of agent-based computer-assissted assessment into e-learning environment. *International Journal of Education and Information Technologies, 3*(3), 163-170. Retrieved from http://www.naun.org/multimedia/NAUN/educationinformation/19-110.pdf

McMillan, J. H., & Lawson, S. R. (2001, January). *Secondary science teachers' classroom assessment and grading practices.* Metropolitan Education Research Consortium. https://pdfs.semanticscholar.org/fdd0/f7942e12f5855f55f6e726d18fd20ef48a85.pdf?_ga=2.62664684.1301240433.1582787694-554234081.1547123017

Md. Desa, Z. N. D., & Abdul Latif, A. (2007). Computerized adaptive testing: An alternative assessment method. In M. Z. Kamsah, M. N. Hassan, K. I. Abdullah, & J. H. Harun (Eds.), *Simposium Pengajaran dan Pembelajaran Universiti Teknologi Malaysia (Symposium Proceeding)* (pp. 78-85). Centre for Teaching and Learning.

Md. Noor, N., & Atan, N. A. (2008, November 5-7). *Tahap kesediaan dan keyakinan pelajar terhadap penggunaan ujian adaptif dalam mempelajari konsep pengaturcaraan komputer (Students' level of readiness and confidence in the use of adaptive tests in learning computer programming concepts)*[Paper presentation]. 2nd International Malaysian Educational Technology Convention, Pahang, Malaysia. Retrieved from https://www.academia.edu/25363636/Tahap_Ke

sediaan_Dan_Keyakinan_Pelajar_Terhadap_Penggunaan_Ujian_Adaptif_Dalam_Mempelajari_Konsep_Pengaturcaraan_Komputer

Morphew, W. J., Mestre, P. J., Kang, H. A., Chang, H.-H., & Fabry, G. (2018). Using computer adaptive testing to assess physics proficiency and improve exam performance in an introductory physics course. *Physical Review Physics Education Research, 14*(2), 1-16. https://doi.org/10.1103/PhysRevPhysEducRes.14.020110

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in Science*. Boston College, TIMSS & PIRLS International Study Center. Retrieved from http://timssandpirls.bc.edu/timss2015/international-results/

Mullis, I. V. S. & Martin, M. O. (Eds.). (2017). *TIMSS 2019 assessment framework.* Boston College, TIMSS & PIRLS International Study Center. Retrieved from http://timssandpirls.bc.edu/timss2019/frameworks/

O'Malley, K. J., Murphy, S., McClarty, K. L., Murphy, D., & McBride, Y. (2011). *Overview of student growth models* [White Paper]. Pearson. Retrieved from http://images.pearsonassessments.com/images/tmrs/Student_Growth_WP_083111_FINAL.pdf

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), 2. https://doi.org/10.1186/s41239-017-0039-0

Özdemir, B. (2016). Comparison of different unidimensional-CAT algorithms measuring students' language abilities: Post-hoc simulation study. *The European Proceedings of Social & Behavioural Sciences.* https://doi.org/10.15405/epsbs.2016.11.42

Raman, K., & Yamat, H. (2014). Barriers teachers face in intergrating ICT during English lesson: A case study. *The Malaysia Online Journal of Educational Technology. 2*(3), 11-19. Retrieved from https://eric.ed.gov/?id=EJ1086402

Ryan, J., & Brockmann, F. (2018). *A practitioner's introduction to equating with primers on Classical Test Theory and Item Response Theory* (Rev. ed.). Council of Chief State School Officers.

Suah, S. L., & Ong, S. L. (2012). Investigating Assessment Practices of In-service Teachers. *International Online Journal of Educational Sciences*, *4*(1).

Sumintono, B. (2016, September 3). *Penilaian keterampilan berpikir tingkat tinggi: Aplikasi pemodelan Rasch pada asesmen pendidikan (Assessment of higher-order thinking skills: Application of Rasch modeling in educational assessments)* [Paper presentation]. Seminar Nasional Pendidikan IPA, FKIP Jurusan PMIPA, Universitas Lambun Mangkurat, Banjarmasin. Retrieved from https://drive.google.com/file/d/0B7f_9LcFjbMTUtDUG5aQzg5a0k/view

Thompson, N. A., & Prometric, T. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research and Evaluation*, *12*(1), 1–13. Retrieved from https://pareonline.net/getvn.asp?v=12&n=1

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. https://doi.org/10.1016/j.sbspro.2015.07.586

van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing.* Springer. https://doi.org/10.1007/978-0-387-85461-8

Wang, H. (2010). Comparability of computerized adaptive and paper-pencil tests. *Test, Measurements and Research Services Bulletin*, *13*(1), 1–7. Retrieved from https://pdfs.semanticscholar.org/6f1e/;bcee90a66fcc969c4caba1362bedb39d505a.pdf

Way, W. D., Twing, J. S., Camara, W., Sweeney, K., Lazer, S., & Maeo, J. (2010). *Some considerations related to the use of adaptive testing for the common core assessments.* Educational Testing Service. Retrieved from http://www.ets.org/s/commonassessments/pdf/AdaptiveTesting.pdf

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences, 2*(1), 1-27. Retrieved from https://journals.uair.arizona.edu/index.php/jmmss/article/view/12351

Wise, S. L., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicologica*, *21*, 135-155. Retrieved from https://www.uv.es/psicologica/articulos1y2.00/wise.pdf

APPENDIX 1

Table of Standardized Standard Variance