JOURNAL OF BALTIC SCIENCE EDUCATION

ISSN 1648-3898 /Print/
ISSN 2538-7138 /Online/

ASSESSING THE VALIDITY AND RELIABILITY OF SCIENCE MULTIPLE CHOICE TEST USING RASCH DICHOTOMOUS MEASUREMENT MODEL

Najah Hazirah Mohd Dzin,
Yoon Fah Lay

Abstract. *Multiple choice tests are widely applied to assess students' knowledge in science education. This study aimed at assessing the validity and reliability of Science Multiple-choice Test in Malaysia. The items for this test were formulated by the researcher together with a panel of science teachers and the head of the science department with close reference to Secondary School Standard Curriculum (KSSM) syllabus. The test consists of 50 multiple-choice items with four options. Rasch measurement model was adopted to evaluate the quality of the Science Multiple-choice Test in terms of reliability analysis, item polarity analysis (PTMEA-CORR), item fit analysis and Principal Component Analysis of Residuals (PCAR). The reliability analysis was performed using Cronbach's Alpha, and the results of reliability and separation index respectively indicated good reliability level of the test items. In order to improve the validity of the test, two negatively worded items (Q39 and Q40) were removed. Lastly, the PCAR analysis showed the unexplained variance in the 1st contrast (5.4%) was found to be well controlled and was below the ceiling value of one-third of the variance explained by the item (18.7%). However, the positive value of the disattenuated correlations indicate no evidence of the presence of secondary dimension.*

Keywords: *multiple-choice test, Rasch measurement model, reliability and validity, science subject, secondary school standard curriculum*

Najah Hazirah Mohd Dzin,
Yoon Fah Lay
University of Malaysia Sabah, Malaysia

Introduction

Achievement is often regarded as a key driver in the field of education, especially as pacemakers for measuring the effectiveness of the teaching and learning process involving critical subjects such as science. Achievement is also an inherent element in describing student motivation in a particular domain (Bedford, 2017; Bryan et al., 2011). Students who are more motivated succeed in more difficult areas than students who are less motivated. Achievement also describes assessment of a student's skills in a particular subject. More skilled students usually obtained higher scores than less skilled ones. Apart from comparing achievement gained among the students, the scores obtained were also used to measure the extent of their mastery on the subjects.

However, even until today, there are still many constraints that need to be rectified in order to improve achievement in science. Among the possible reasons why science achievement has not been comprehensively improved are due to certain elements such as in the way students learn science or negligence of the science teachers towards the issue (Owens, 2009). Therefore, in order to remedy this, various methods were taken into consideration to ensure improvement in science achievement. One main area of focus includes determination of appropriate student performance assessment methods that focus mainly on laboratory skills test and achievement test which are frequently used to measure student performance in science.

Achievement test is one of the most widely used way for teachers to observe the development of students' knowledge or their performance in a particular domain. The tests or examinations conducted in the classroom are mainly carried out in the form of multiple-choice (Moreno et al., 2006) or subjective responses, depending on the content of learning in certain subject. Multiple choice tests are suitable for testing the effectiveness of learning in the classroom (Haladyna, 2004). In fact, tests with this format also help educators formulate strategies to guide students in understanding the content of learning, especially in science education (Chu et al., 2009).

https://doi.org/10.33225/jbse/21.20.927

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Multiple-choice (MC) testing can be categorized as one of the assessments that is still widely used in education today. The nature of multiple-choice testing is more objective and has a clearer structure in testing students' cognitive skills (Gierl et al., 2017). Although the use of MC in assessing students' cognitive skills is often debated, it is not impossible to implement this matter especially to assess students' performance at a higher level. A well-constructed item based on the most well-known Bloom Taxonomy and qualified distractor options are often a good guide for teachers to construct the items that can be used to denote students' levels more thoroughly (Scully, 2017). Matters such as verb manipulation and the use of high-quality distractors (Gierl et al., 2017) need to be taken into consideration when constructing items to test higher order thinking skills.

The construction of items for MC format went through not only rigorous process, but the item qualities were also confirmed using face validity, content validity and expert validity. Validity and reliability assessments are important procedures in ensuring that the data obtained meet the objectives of the study. Validity describes the extent to which the instrument is capable of measuring things to be measured in a study. In research, validity refers to the process of collecting and analysing evidence that support that the instrument used is appropriate, accurate, useful, and meaningful for the study (Fraenkel et al., 2012). Meanwhile instrument reliability refers to the measurement of instruments that are reliable, can be replicated and are always consistent (Ruel et al., 2015).

Two well-known approaches to define the quality of MC items are classical test theory and item response theory. Most quantitative studies perform MC item analysis using item response theory (Akhter et al., 2019; D'Sa & Visbal-Dionaldo, 2017; Jansen et al., 2015; Toksöz & Ertunç, 2017). The characteristics such as difficulty index, discrimination index and reliability index approve the quality of the MC items. Ayanwale and Adeleke (2020) concluded that the use of item response theory is however more effective than classical test theory in the development and scoring of MC items.

Recently, many researchers have shown interest in MC item testing by using the Rasch measurement model. The Rasch measurement model is a psychometric technique that aims to improve accuracy in terms of item construction, item quality as well as measuring student performance (Boone, 2016). In fact, Rasch measurement model offers different analysis techniques as compared to item responses theory. While IRT can be modified using several parameters, the measurement principle of the Rasch measurement model is more in line with the definition of a Thurstone measurement scale (Boone et al., 2014). Thus, the use of Rasch measurement model is more sample independent in MC item testing.

In Malaysian education system, teachers are responsible for designing, constructing, distributing, and evaluating tests in their respective subjects at the school level based on the National Education Philosophy and the school curriculum (Ong, 2010). The construction of science multiple-choice test items respectively for the Form Four in this study are based on the Secondary School Standard Curriculum (KSSM) which was introduced in 2020 for the first batch of Form Four students in this particular year. Moreno et al. (2006) stated that there are three main elements that need to be emphasized when constructing a test item, namely, the content of the item, the way of presenting the content of the item and the construction of appropriate distractor options. Based on these elements, Science multiple-choice test items for the Form Four students are developed based on five chapters in the national standard document namely *Dokumen Standard Kurikulum & Pentaksiran* (DSKP) for the Science subject, as well as the use of several types of multiple-choice format with a uniform number of distractor option.

Research Problem

Most previous studies that examined the level of science achievement as one of the variables did not use specific achievement test constructed by the researchers themselves (Andrew, 1998; Britner & Pajares, 2006; Lee et al., 2014). In fact, some researchers acknowledge the use of subject grades obtained through school as one of the limitations of their study (Chen & Usher, 2013). A study conducted by Joët et al. (2011) for example, used students' existing grades through early-year tests administered in schools. Although the researchers claimed that these data were controlled through the previous end-of-year test results of students and only used as pacemakers, there was no clear evidence regarding the validity and reliability of the items used in assessing student achievement sampled for this study. The use of existing grades raises the issues of validity and reliability as items in school test or examination were formulated for the purpose of obtaining grades during schooling only.

Meanwhile, Malaysia recorded inconsistent average science achievement scoring over the decade in international assessment like TIMSS and PISA (Hin, 2020; Phang et al., 2020). In fact, this scenario is linked to students' negative attitudes towards science (Lay et al., 2020; Ismail et al., 2018). In Malaysia, most students in middle school

https://doi.org/10.33225/jbse/21.20.927

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

avoid choosing science streams due to their beliefs that the pure science subjects were extremely difficult for them. In 2017, Malaysian Ministry of Education has implemented a comprehensive new curriculum starting as the shift of drastic transformation in educational system (Ministry of Education, 2013). The Standard Curriculum for Secondary School (KSSM) had been introduced for first batch of Form One in year 2017 (Curriculum Development Division, 2018). By 2020, they were already in Form Four and had learned the subjects they had chosen.

Research Focus

The selection of instruments in a study depends on the type of data needed to measure the objectives of the study (Cohen et al., 2013). In this study, the construction of the Science-multiple Choice Test (SMCT) aimed to measure the current achievement of Form Four students in Science subjects. The construction of items for this instrument was based on the latest syllabus of the Secondary School Standard Curriculum (KSSM) for Form Four Science subjects. The KSSM syllabus for Form Four Science subjects was used by schools under the Ministry of Education Malaysia starting in 2020. Since the learning process of the latest syllabus for Science subjects is still ongoing process, SMCT items were built based on the topics in the KSSM syllabus that have been studied in secondary schools. The construction of multiple-choice test items should be followed by an assessment of the validity and reliability of the items (Haladyna, 2004). In the context of this study, the researchers have formulated their own questions based on the revised Bloom's taxonomy (Anderson et al., 2001) as well as the KSSM syllabus. SMCT items were analysed using the Rasch measurement model.

Research Aim and Research Questions

In this study, the validity and reliability analysis of multiple-choice test items were performed using Rasch measurement model. The Rasch measurement model is a psychometric technique that aims to improve accuracy in terms of item construction, item quality as well as measuring student performance (Boone, 2016). Based on the purpose of the study, four main objectives are developed to assess the validity and reliability of a Science multiple choice test instrument through the Rasch measurement model namely (i) reliability analysis, (ii) polarity item analysis, (iii) item fit analysis and (iv) Principal Component Analysis of Residuals (PCAR).

Research Methodology

General Background

This study adopted quantitative approach using the survey design. The main objective for actual study is to test the mediation effect of science self-efficacy for the relationship between sources of self -efficacy and science achievement. The study was conducted in September 2020 for a month. This data collection process took place face-to-face after school closures across the country due to the COVID-19 pandemic. The first phase involved data collection for questionnaires was related to student self -efficacy. Meanwhile, science achievement tests were conducted after 2 weeks from the date of handling the questionnaire form. All respondents involved were informed that all data were for research purposes and did not involve a scoring system from the school.

Research Sample

The multi-stage sampling method was adopted in this study. This study was conducted in the west coast division of Sabah, Malaysia. Out of the seven districts in the west coast of Sabah, two district education offices were selected through random sampling method. Next, the researchers selected few secondary schools from the two districts for pilot study. Lastly, purposive sampling was used to select Form Four students who took general science subject to meet the needs of the study. Form Four students in Malaysia have similar characteristics as compared to the international Grade 10 students. They are around 16 years old and are studying in secondary schools. However, the science curriculum in Malaysia might be different from other countries.

The best sample size determination method for structural equation modelling in a study is based on power analysis with specified features (Hair et al., 2007). The minimum sample size in this study was determined through calculations based on G*Power software version 3 (Faul et al., 2007). Based on the calculations, a total of 74 mini-

https://doi.org/10.33225/jbse/21.20.927

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

mum sample numbers in referral to the structural model were proposed for this study. However, previous studies have suggested that 100 to 200 samples are a good starting point for studies related to path estimation analysis especially for structural equation model (Hoyle, 1995; Kline, 2005). Therefore, in order to satisfy this requirement, the researchers have collected data from 109 samples. The samples were considered to be homogenous as all schools in Malaysia are using the same curriculum and syllabus as stipulated by Malaysian Ministry of Education.

Instrumentation

The SMCT used in this study aims to measure the level of Science achievement of Form Four students in five chapters (Chapter 1: Safety Measures in Laboratory, Chapter 2: Emergency Help, Chapter 3: Techniques in Measuring the Parameters of Body Health, Chapter 4: Green Technology for Environmental Sustainability and Chapter 5: Genetic). The Science multiple-choice test is a paper and pencil test that contains 50 multiple choice items. The items for this test were formulated by the researcher together with a panel consisting of a Form Four Science subject teacher and the head of the Science department with close reference to the Secondary School Standard Curriculum (KSSM) syllabus. Originally, the SMCT items were drafted based on all 12 topics in the new syllabus introduced by Ministry of Education. However, there are some limitations while conducting this study. The COVID-19 pandemic that hit the whole world, including Malaysia caused the closure of industries and social institutions. In Malaysia, all schools were closed starting on March 2020 causing drastic changes to the learning system at the time. The learning process at that time was conducted online for 6 months until September but was less than satisfactory. The teaching and learning process occurs more slowly due to the drastic adaptation of the method. Based on the feedback received from the science subject teachers, the content was therefore being finalized to five topics only.

However, these changes did not affect the original purpose of the study as validation process was done to ensure the appropriateness of the question content as well as the level of question testing on the sample. Based on 5 topics selected by the researcher, the preparation of 50 questions for SMCT is according to the test specification table as shown in the Table 1.

Table 1
Test Specification Table for SMCT

Topic	Subtopic	Knowledge			Understanding			Skills			Total Item
		L	M	H	L	M	H	L	M	H	
Theme: Scientific Methodology											
1.0 Safety Measures in laboratory	1.1 Self Protection Equipment	2									2
	1.2 Disposable of Waste	1	2								3
	1.3 Fire Extinguisher		1	1	1	1		1			5
2.0 Emergency Health	2.1 Cardiopulmonary resuscitation (CPR)	1	1	1			1	1			5
	2.2 Heimlich Maneuver		1	3							4
3.0 Techniques in Measuring the Parameters Body Health	3.1 Body Temperature	1	1	1	1						4
	3.2 Pulse rate	1				1					2
	3.3 Blood Pressure					1		1			2
	3.4 Body Mass Index (BMI)								3		3
Theme: Maintenance and Continuity of Life											

https://doi.org/10.33225/jbse/21.20.927

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Topic	Subtopic	Knowledge			Understanding			Skills			Total Item
		L	M	H	L	M	H	L	M	H	
4.0 Green Technology for Environmental Sustainability	4.1 Sustainable Environment	1	1								2
	4.2 Energy Sector	1				1					2
	4.3 Waste and Waste Water Management Sector					2					2
	4.4 Agriculture and Forestry Sector			2							2
	4.5 Transportation Sector			1							1
	4.6 Green Technology and Life			1							1
5.0 Genetic	5.1 Cell Division						1				1
	5.2 Inheritance			1	1		1	1	1		5
	5.3 Mutation		1								1
	5.4 Genetic Engineering Technology						1				1
	5.5 Variation			1						1	2
	Total Item	8	8	12	3	6	4	4	4	1	50

Data Analysis

The validity and reliability of SMCT were determined through Rasch Dichotomous Measurement Model by using Winsteps 4.8.0.0 software (Linacre, 2021). Assessment of reliability (Cronbach's Alpha coefficient, item reliability, item separation) and construct validity include item polarity analysis (PTMEA-CORR), item fit analysis in terms of mean square (MNSQ) infit and outfit paired with standardized fit statistic (*z*-STD) infit and outfit continued with Principal Component Analysis of Residuals (PCAR) were reported as evidence in this paper. Table 2 specifies the acceptable criterion for the reliability and validity analysis by using Rasch measurement model that used to report the quality of SMCT.

Table 2
Validity and Reliability Criterion in Rasch Measurement Model

Criterion	Assumption
Reliability	
a) Cronbach's Alpha (α)	• 0.90 – 1.00 very good and effective with a good degree of consistency • 0.70 - 0.80 good and acceptable • 0.60 - 0.70 acceptable • < 0.60 item need to be fixed. • < 0.50 item need to be removed.
b) Item Reliability	• 0.94 excellent • 0.91 - 0.94 very good • 0.81 - 0.90 good • 0.67 - 0.80 acceptable • < 0.67 low reliability
c) Item Separation	• The greater the separation value, the better the item will be. • Separation values serve to categorize items in stratified group. • > 2.0 considerably good

https://doi.org/10.33225/jbse/21.20.927

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Criterion	Assumption
Construct Validity	
a) Item polarity (PTMEA-CORR)	• A positive PTMEA-CORR value indicates that the response intertwining for the item or individual is in the same direction with the construct. • A negative or zero PTMEA-CORR value indicates that the response linkage for the item or individual is contrary with the construct.
b) Principal Component Analysis of Residuals (PCAR)	• Raw variance explained by measures > 40% • Unexplained variance in the 1st contrast < 2.0 Eigen value. • Observed variance < 1/3 of variance explained by the item.
c) Mean square (MNSQ) infit and outfit	• 2.0, distorts or degrades the measurement system. • 1.5 < MNSQ < 2.0, unproductive for construction of measurement, but not degrading. • 0.5 < MNSQ < 1.5, productive for measurement. • < 0.5, less productive for measurement, but not degrading. May produce misleadingly good reliabilities and separation.
d) z-STD infit dan outfit	• -2.0 < MNSQ < +2.0

Research Results

Based on the criteria of the Rasch measurement model, the researchers reported a quality evaluation of a multiple-choice test instrument through (i) reliability analysis based on Cronbach's Alpha (α), reliability and separation respectively for the item and person, (ii) polarity item analysis based on PTMEA-CORR values, (iii) item fit analysis based on mean square (MNSQ) and *z*-STD values respectively for the infit and outfit bound and (iv) Principal Component Analysis of Residuals (PCAR).

Reliability Analysis

Table 3 shows the reliability obtained based on the Cronbach's Alpha value which is .72. In this case, the value of the Cronbach's Alpha coefficient for this instrument is good and generally acceptable, whereas, the values of item reliability and item separation, were .95 and 4.30, respectively. This indicates that the item reliability for this instrument is very good. Meanwhile, the item separation values are in the range of levels 4 to 5. This means that the items in this instrument were categorized into four or five hierarchy of difficulty item.

Table 3
Summary of Reliability Analysis for SMCT

Cronbach Alpha	Item Reliability	Item Separation
0.72	0.95	4.30

Figure 1 shows the Wright Map, a mapping of the item difficulty distribution with the ability to answer the item distribution along the same continuum (Planinic et al., 2019). From Figure 1, we can see that the most difficult item (Q40) is located at the top and the easiest item (Q18) is located at the bottom of the map. When the difficulty of the item and the respondent's ability are matched, the respondent has a 50% chance of answering the item correctly (Herrmann-Abell et al., 2009). In this study, the mean for respondents (0.06) and the mean of items (.15) as shown in Table 4. The difference between these two values is less than .50 which is indicating that this test is within the target (Jamaludin & Lay, 2020).

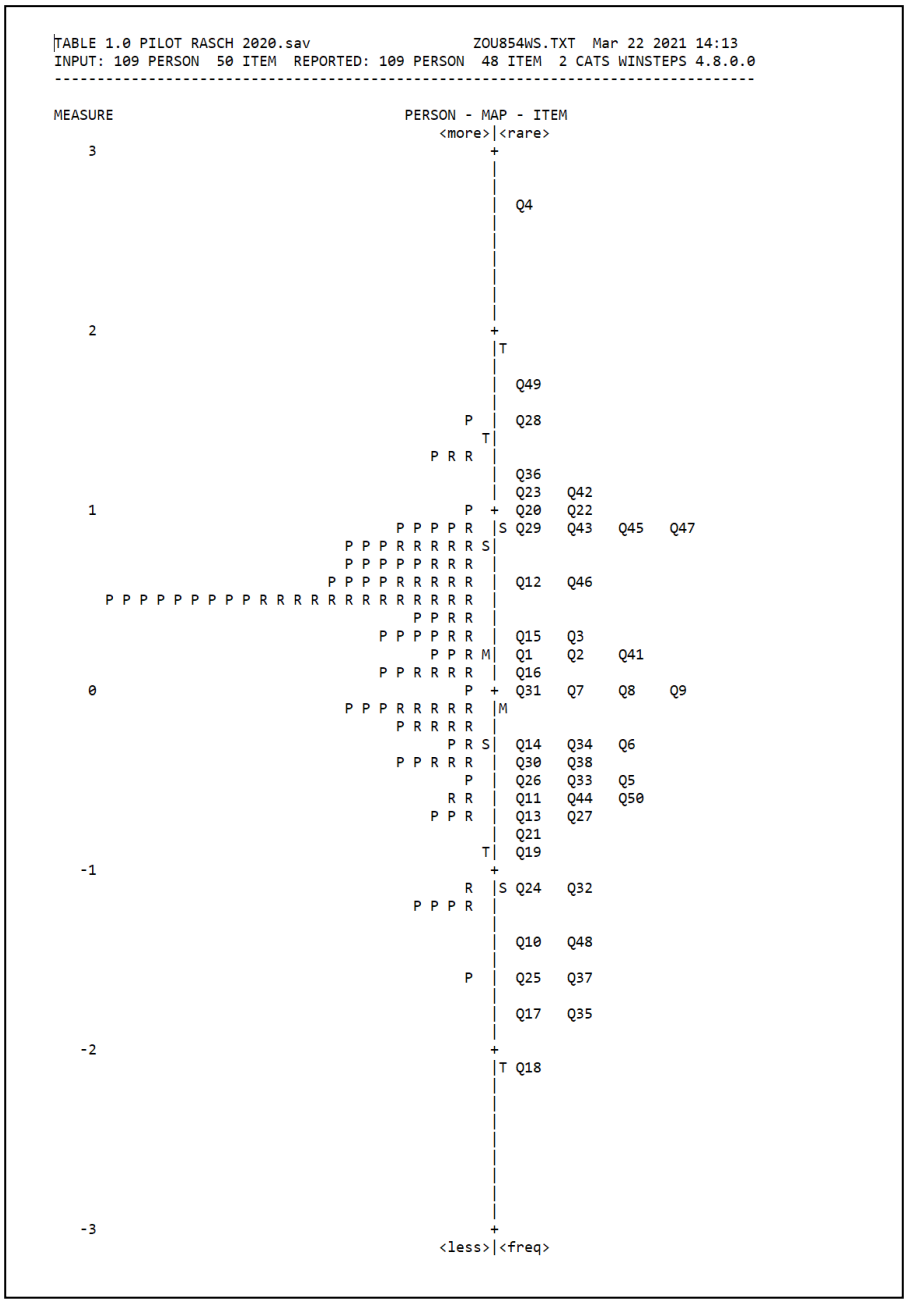
Table 4
Mean Values of Person and Item for SMCT

Person Mean	Item Mean
0.06	0.15

https://doi.org/10.33225/jbse/21.20.927

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Figure 1
Wright Map for Science Multiple-choice Test



Item Polarity Analysis (PTMEA-CORR)

Based on Table 5, items 39 and 40 clearly have negative PTMEA CORR values, which were -.20 and -.19 respectively. This suggests that the response of the items is contrary to the construct. Therefore, items 39 and 40 were removed from the SMCT item listing. Once adjustment through the item removal method was performed, all items had positive PTMEA CORR values indicating that all the items were moving in same direction and measure the construct as they should be (Bond & Fox, 2015). The list of the items with positive PTMEA-CORR values after the adjustment was performed shown in Table 6.

https://doi.org/10.33225/jbse/21.20.927

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Table 5
Point Measure Correlation (PTMEA CORR) Values of SMCT Items Before Adjustment

Item	Measurement Score	PTMEA CORR	Item	Measurement Score	PTMEA CORR
39	1.39	-0.20	15	0.27	0.26
40	3.05	-0.19	9	0.00	0.28
42	1.05	0.02	5	-0.51	0.28
24	-1.05	0.07	37	-1.56	0.29
31	-0.04	0.09	38	-0.42	0.30
28	1.54	0.10	45	0.92	0.34
4	2.66	0.11	32	-1.05	0.34
36	1.24	0.11	6	-0.29	0.35
3	0.31	0.13	10	-1.35	0.35
19	-0.94	0.13	34	-0.29	0.35
49	1.71	0.13	13	-0.69	0.35
7	0.04	0.13	12	0.59	0.36
43	0.92	0.15	50	-0.60	0.37
41	0.23	0.16	16	0.08	0.37
1	0.23	0.16	48	-1.35	0.38
30	-0.38	0.16	22	0.97	0.38
20	1.01	0.17	25	-1.64	0.41
46	0.55	0.17	17	-1.80	0.42
47	0.88	0.17	2	0.23	0.42
27	-0.74	0.18	23	1.05	0.43
33	-0.46	0.19	18	-2.08	0.43
8	0.04	0.21	11	-0.55	0.46
26	-0.46	0.21	44	-0.65	0.52
29	0.92	0.23	21	-0.84	0.53
14	-0.33	0.25	35	-1.80	0.53

Note. Q39 and Q40 have been removed.

Table 6
Point Measure Correlation (PTMEA CORR) Values of SMCT Items After Adjustment

Item	Measurement Score	PTMEA CORR	Item	Measurement Score	PTMEA CORR
42	1.16	0.01	9	0.09	0.27
24	-0.97	0.07	5	-0.42	0.28
28	1.65	0.09	37	1.48	0.29
36	1.34	0.10	38	-0.33	0.32
31	0.05	0.10	45	1.02	0.33
4	2.77	0.12	32	-0.97	0.34
19	-0.86	0.12	6	-0.20	0.35
3	0.41	0.13	34	-0.20	0.35
7	0.13	0.13	12	0.69	0.36
49	1.82	0.13	13	-0.61	0.36
47	0.98	0.15	10	-1.27	0.36
43	1.02	0.15	50	-0.51	0.37
41	0.33	0.16	48	-1.27	0.38

https://doi.org/10.33225/jbse/21.20.927

Item	Measurement Score	PTMEA CORR	Item	Measurement Score	PTMEA CORR
20	1.11	0.16	22	1.07	0.39
30	-0.29	0.17	16	0.17	0.40
1	0.33	0.18	25	-1.56	0.42
46	0.65	0.18	2	0.33	0.42
27	-0.65	0.19	17	-1.72	0.42
33	-0.37	0.20	23	1.16	0.42
26	-0.37	0.21	18	-2.01	0.44
8	0.13	0.22	11	-0.46	0.47
29	1.02	0.23	44	-0.56	0.52
14	-0.24	0.26	35	-1.72	0.53
15	0.37	0.26	21	-0.75	0.53

Item Fit Analysis

Based on Table 7, the mean squared infit values (MNSQ) for 48 items after the removal of items Q39 and Q40 were in the range recommended by Wright and Linacre (1994) of .5 to 1.5. meanwhile, the same result shown for the *z*-STD infit values which are still in the acceptable range between -2.0 to +2.0 (Liu, 2020). However, the *z*-STD value of outfit item 36 and item 42 were outside the recommended range. According to Boone et al. (2014), since the MNSQ value of all items were in the range recommended of .5 to 1.5, the *z*-STD values of respective items that out of ranges can be ignored.

Table 7
Item Misfit Analysis Based on MNSQ and z-STD Range

Item	Measurement Score	Infit		Outfit	
		MNSQ	*ZSTD*	*MNSQ*	*ZSTD*
42	1.16	1.12	1.5	1.27	2.2
24	-0.97	1.13	1.0	1.19	1.2
28	1.65	1.05	0.5	1.19	1.2
36	1.34	1.03	0.4	1.33	2.4
31	0.05	1.12	1.9	1.13	1.9
4	2.77	1.01	0.1	1.02	0.2
19	-0.86	1.10	0.8	1.13	0.9
3	0.41	1.10	1.9	1.10	1.6
7	0.13	1.10	1.6	1.13	1.9
49	1.82	1.03	0.3	1.10	0.6
47	0.98	1.04	0.7	1.16	1.6
43	1.02	1.05	0.8	1.12	1.2
41	0.33	1.08	1.5	1.10	1.5
20	1.11	1.04	0.6	1.12	1.1
30	-0.29	1.08	1.1	1.07	0.8
1	0.33	1.05	1.2	1.09	1.4
46	0.65	1.06	1.0	1.08	1.1
27	-0.65	1.04	0.4	1.13	1.1
33	-0.37	1.04	0.5	1.11	1.2
26	-0.37	1.05	0.6	1.04	0.5
8	0.13	1.04	0.8	1.05	0.8

https://doi.org/10.33225/jbse/21.20.927

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Item	Measurement Score	Infit		Outfit	
		MNSQ	*ZSTD*	*MNSQ*	*ZSTD*
29	1.02	1.01	0.1	1.02	0.3
14	-0.24	1.02	0.2	1.03	0.4
15	0.37	1.01	0.3	1.00	0.0
9	0.09	1.01	0.1	1.02	0.3
5	-0.42	1.00	0.1	1.00	0.0
37	1.48	1.01	0.1	0.86	-0.6
38	-0.33	0.98	-0.2	0.97	-0.3
45	1.02	0.96	-0.6	0.93	-0.7
32	-0.97	0.96	-0.3	0.93	-0.4
6	-0.20	0.96	-0.6	0.97	-0.3
34	-0.20	0.96	-0.5	0.94	-0.7
12	0.69	0.95	-0.9	0.92	-1.0
13	-0.61	0.95	-0.4	0.94	-0.5
10	-1.27	0.94	-0.4	0.88	-0.5
50	-0.51	0.95	-0.5	0.93	-0.6
48	-1.27	0.93	-0.4	0.85	-0.7
22	1.07	0.92	-1.1	0.87	-1.3
16	0.17	0.93	-1.2	0.91	-1.4
25	-1.56	0.87	-0.6	0.88	-0.5
2	0.33	0.92	-1.6	0.89	-1.7
17	-1.72	0.91	-0.4	0.72	-1.1
23	1.16	0.90	-1.3	0.83	-1.6
18	-2.01	0.89	-0.4	0.66	-1.1
11	-0.46	0.88	-1.4	0.87	-1.3
44	-0.56	0.84	-1.7	0.83	-1.5
35	-1.72	0.83	-0.8	0.63	-1.6
21	-0.75	0.85	-1.4	0.78	-1.8

Principal Component Analysis of Residuals (PCAR)

Principal Component Analysis of Residuals (PCAR) technique was used to ensure the unidimensionality of the instruments used in this research. The statistical analysis summary of PCAR was shown in Table 8. PCAR technique itself can detect the ability of this instrument to measure in a uniform dimension with acceptable level of distractor. Based on the PCAR analysis, it was found that the total raw variance explained by the measurement was 23.7% whereas the unexplained variance in the first contrast was 5.4% as shown in Table 7.

Table 8
Principal Component Analysis of Residuals (PCAR) Statistic for SMCT

Raw Variance Explained by Measures	Variance Unexplained	Variance Unexplained in 1st Contrast
23.7% (12.1)	76.3% (39.0)	5.4 % (2.8)

The raw variance explained by Standardized Residual Variance (in Eigenvalue units) as shown in the Figure 2 is 23.7% exceed the expectations of the model (23.5%). This indicates that Eigenvalue of 12.1 was detected in raw variance explained by measures. From the data obtained, the value of the raw variance explained by the data has reached the minimum level of instrument uniformity requirements (20%) but did not reach the minimum level involving 40% of the Rasch requirements. In addition, the value of unobserved variance in the 1st contrast has

https://doi.org/10.33225/jbse/21.20.927

reach 2.8, which is more than 2.0 Eigenvalue as required by Rasch analysis. However, the unexplained variance in the 1st contrast which is as high as 5.4% was found to be well controlled and below from the ceiling value of one-third of variance explained by the item (18.7%).

Figure 2
Standardized Residual Variance (in Eigenvalue units)

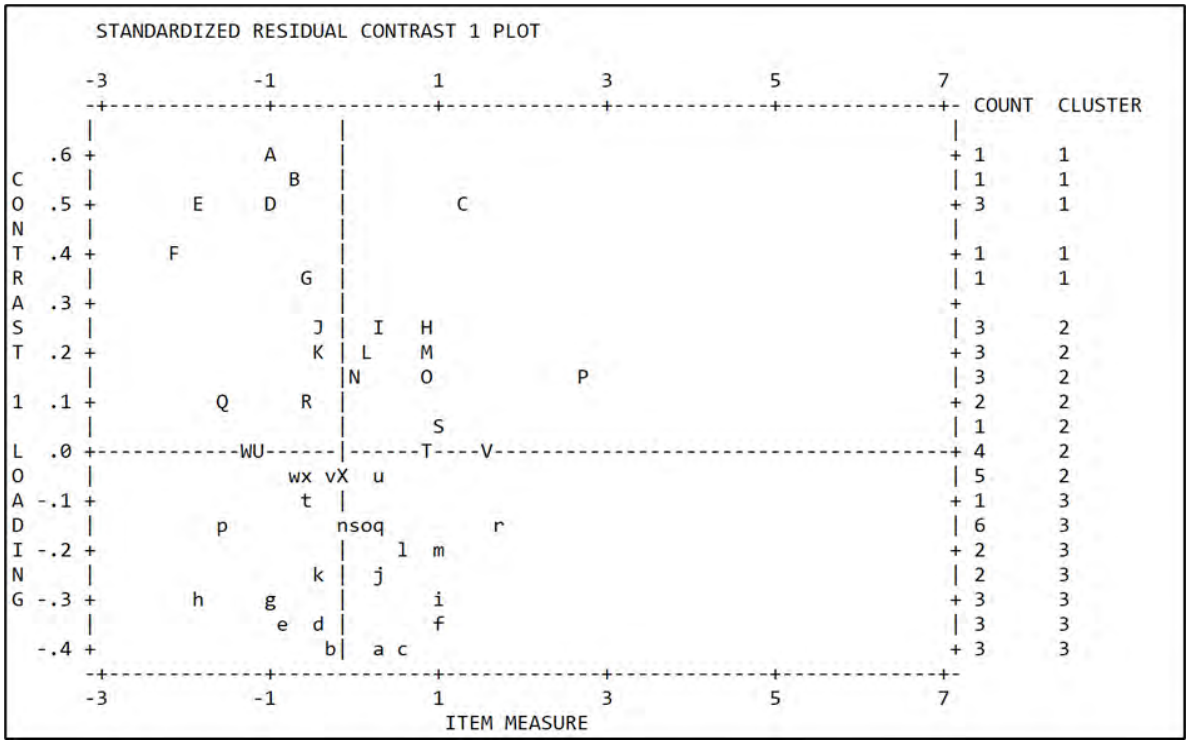
TABLE 23.0 PILOT RASCH 2020.sav ZOU509WS.TXT Mar 24 2021 13:24
INPUT: 109 PERSON 50 ITEM REPORTED: 109 PERSON 48 ITEM 2 CATS WINSTEPS 4.8.0.0

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units

	Eigenvalue	Observed		Expected
Total raw variance in observations =	51.0972	100.0%		100.0%
Raw variance explained by measures =	12.0972	23.7%		23.5%
Raw variance explained by persons =	2.5235	4.9%		4.9%
Raw Variance explained by items =	9.5737	18.7%		18.6%
Raw unexplained variance (total) =	39.0000	76.3%	100.0%	76.5%
Unexplned variance in 1st contrast =	2.7828	5.4%	7.1%	
Unexplned variance in 2nd contrast =	2.6222	5.1%	6.7%	
Unexplned variance in 3rd contrast =	2.5468	5.0%	6.5%	
Unexplned variance in 4th contrast =	2.3618	4.6%	6.1%	
Unexplned variance in 5th contrast =	2.0810	4.1%	5.3%	

Explanation of the unidimensionality of this instrument continued by looking Contrast 1 Plot table as shown in Figure 3. Based on Figure 3, there are three clusters that exist in the item distribution for the Science Multiple-choice Test. Most items normally distributed, but there are some adhesions tracked on the plot items like "nsoq" and "vX". However, to ascertain whether these three clusters measure the same thing statistically, an analysis needs to be done on the disattenuated correlation that exists between these three clusters.

Figure 3
Contrast 1 Plot for Science Multiple-choice Test Items



https://doi.org/10.33225/jbse/21.20.927

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Based on Figure 4, the disattenuated correlation value that exists between the second and third clusters is .8033, where this value is the highest value and exceeds the minimum level of .5 required to ensure that this cluster measures the same thing statistically. Meanwhile, the value of disattenuated correlation that exists between cluster 1 and cluster 3 is .31. Furthermore, the value of disattenuated correlation that exists between cluster 1 and cluster 2 is .41. While both values are below the minimum level of .5, the proof of the existence of the second dimension is remain unclear as disattenuated correlation value displayed is positive. The action of correcting an item is an appropriate step especially in terms of verse processing and the arrangement of distractor options in a particular item.

Figure 4
Disattenuated Correlation

Approximate relationships between the PERSON measures

PCA Contrast	ITEM Clusters	Pearson Correlation	Disattenuated Correlation	Pearson+Extr Correlation	Disattenuated+Extr Correlation
1	1 - 3	0.1740	0.3148		
1	1 - 2	0.2404	0.4142		
1	2 - 3	0.4289	0.8033		

The examination of item correlation was performed to identify the correlation between any of the item in SMCT through Standardized Residual Correlations analysis. Figure 5 shows the satisfactory results obtained as there were no item correlation values that exceeded the control level of 0.7. This indicates that the instrument was free from any confusion in terms of purpose and objectives of the test.

Figure 5
Standardized Residual Correlations

TABLE 23.99 PILOT RASCH 2020.sav ZOU295WS.TXT Mar 24 2021 13:58
INPUT: 109 PERSON 50 ITEM REPORTED: 109 PERSON 48 ITEM 2 CATS WINSTEPS 4.8.0.0

LARGEST STANDARDIZED RESIDUAL CORRELATIONS
USED TO IDENTIFY DEPENDENT ITEM

CORRELATION	ENTRY NUMBER	ITE	ENTRY NUMBER	ITE
.42	17	Q17	18	Q18
.38	13	Q13	19	Q19
.36	1	Q1	27	Q27
.36	19	Q19	42	Q42
.34	17	Q17	19	Q19
.34	10	Q10	34	Q34
.33	26	Q26	37	Q37
-.42	35	Q35	43	Q43
-.39	5	Q5	41	Q41
-.38	7	Q7	17	Q17
-.37	19	Q19	46	Q46
-.37	10	Q10	47	Q47
-.36	28	Q28	44	Q44
-.36	7	Q7	12	Q12
-.35	9	Q9	34	Q34
-.35	32	Q32	45	Q45
-.35	5	Q5	15	Q15
-.35	25	Q25	49	Q49
-.33	22	Q22	45	Q45
-.33	13	Q13	46	Q46

https://doi.org/10.33225/jbse/21.20.927

Discussion

In this research, SMCT was designed to assess the science achievement of the Form Four student on five topics in science subject. Since the topics tested are based on the new curriculum, the data can provide preliminary information on student performance in this subject. This information is really useful for science teachers to observe the cognitive development of students in the domain of science.

Based on reliability analysis through Rasch measurement model, Cronbach Alpha value of 0.72 indicates that SMCT is within good and acceptable criteria. Furthermore, item reliability value of 0.95 indicates that the constructed items are excel for testing the sample of this study. The separation index of 4.30 indicates the constructed item has four difficulty levels of testing despite the actual three levels in test specification table. The Wright Map analysis shows that the level of item difficulty was matched to the ability of the students being tested. In this study, the selected sample were students with a moderate background of past science achievement overall. This concludes that SMCT is a reliable as a whole-body test.

While doing item polarity analysis, some items were removed while others retained with modifications. The negative values of PTMEA-CORR indicate that the items were moving opposite direction as well as are unsuitable to be used as test items. The negative values also suggest that the responses given contrary to the expected one. However, the value of MNSQ paired with z-STD of remaining 48 items had been analysed to make sure the items statistically fit. Overall, the unidimensionality of the SMCT had been confirmed based on PCAR analysis and dissatuenated correlation analysis.

This finding is consistent with the other previous studies on validation by using the same methodological approach (Azizah et al., 2021; Jamaludin & Lay, 2020; Soeharto & Csapó, 2021; Yang et al., 2018). Overall, assessment of Rasch Dichotomous Model proves that the items in Science Multiple-choice Test are valid and reliable with removal of two negatively worded items (Q39 and Q40). However, some item modifications suggested in item fit analyses to improve the quality of instrument. Table 9 shows the summary of item misfit and action to be taken for potential items. Item Q36 and Q42 pass the threshold level of MNSQ but exceed the limit of *z*-STD values which means that both items are possibly underfit. The modification in terms of sentence structure and arrangement of distractor options as to enhances the quality of the instrument.

Conclusions and Implications

The reliability and validity analysis using Rasch measurement model ensures a reliable and valid Science Multiple-choice Test which meets the assessment needs of the Science subject at secondary school level. The results of the Rasch analysis proved that the validity and reliability of the test items were established with the removal of two items. However, some items modification had been made in terms of verse processing and arrangement of distractor options to improve the quality of the instrument.

This study has an impact on several aspects, especially methodology and practice in the field of science education. From a methodological point of view, the Rasch measurement model promises a more comprehensive validation of the instrument. In fact, reliability analysis such as separation index prove the difficulty level of constructed instrument. Moreover, information such as Wright map analysis also helps the researcher to know the suitability of the instrument to the sample studied. The validity and reliability of this instrument also proves that SCMC can be used for actual testing purposes in the classroom as well as suitable for the use against groups with characteristics equivalent to the study sample.

Some limitations of the study need to be clarified by researchers. The minimum sample with total of 109 students was used due to time as well as location constraints. Therefore, this sample size may not be suitable for generalization purposes. However, a larger sample size could be used in future studies. Moreover, the drafting of SMCT items only involved five topics in the new science syllabus. This is due to the constraints of the teaching and learning process which did not take place in accordance with the teacher's planning. The construction of science multiple choice items in the future should involve all 12 topics in Form 4 science.

References

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of educational objectives*. Longman.

Andrew, S. (1998). Self-efficacy as a predictor of academic performance in science. *Journal of Advanced Nursing*, *27*(3), 596–603.

Ayanwale, M. A., & Adeleke, J. O. (2020). Efficacy of Item Response Theory in the validation and score ranking of dichotomous response mathematics achievement test. *Bulgarian Journal of Science and Education Policy*, *14*(2), 260-285.

https://doi.org/10.33225/jbse/21.20.927

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Azizah, A., Wahyuningsih, S., Kusumasari, V., Asmianto, A., & Setiawan, D. (2021, March). Validity and reliability of mathematical instruments in online learning using the Rasch measurement model at UM lab school. *AIP Conference Proceedings,* 2330 (1), p.040024. AIP Publishing LLC.

Bedford, S. (2017). Growth mindset and motivation: A study into secondary school science learning. *Research Papers in Education*, *32*(4), 424-443. https://doi.org/10.1080/02671522.2017.1318809

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental measurement in the Human Sciences* (3rd ed.). Routledge.

Boone, W. (2016). Rasch analysis for instrument development: Why, when, and how? *Cell Biology Education*, *15*, 1-7. https://doi.org/10.1187/cbe.16-04-01488

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer Science & Business Media.

Britner, S. L., Pajares, F. (2006). Sources of science self-efficacy beliefs of middle school students. *Journal of Research in Science Teaching*, *43*(5), 485–499. https://psycnet.apa.org/doi/10.1002/tea.20131

Bryan, R. R., Glynn, S. M., & Kittleson, J. M. (2011). Motivation, achievement, and advanced placement intent of high school students learning science. *Science Education*, *95*(6), 1049–1065. https://coe.uga.edu/assets/downloads/mse/smqii-bryan-et-al-2011.pdf

Chen, J., & Usher, E. (2013). Profiles of the sources of science self-efficacy. *Learning and Individual Differences*, *24*, 11–21. https://doi.org/10.1016/j.lindif.2012.11.002

Chu, H.-E., Treagust, D., & L., C. (2009). A stratified study of students' understanding of basic optics concepts in different contexts using two-tier multiple-choice items. *Research in Science & Technological Education*, *27*, 253–265. https://doi.org/10.1080/02635140903162553

Cohen, L., Manion, L., & Morrison, K. (2013). *Research methods in education* (7th ed.). Routledge.

Curriculum Development Division (2018). *Form 4 and 5 curriculum and assessment standard documents: The Science Standard Curriculum for Secondary School (KSSM)*. Ministry of Education Malaysia.

D'Sa, J. L., & Visbal-Dionaldo, M. L. (2017). Analysis of multiple-choice questions: Item difficulty, discrimination index and distractor efficiency. *International Journal of Nursing Education*, *9*(3), 109-114.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. https://doi.org/10.3758/bf03193146

Fraenkel, J., Hyun, H., & Wallen, N. (2012). *How to design and evaluate research in education* (8th ed.). McGraw-Hill Education.

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, *87*(6), 1082-1116. https://doi.org/10.3102%2F0034654317726529

Hair, J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2016). *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. SAGE Publications.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Lawrence Erlbaum Associates.

Herrmann-Abell, C. F., DeBoer, G. E., & Roseman, J. E. (2009, November). Using Rasch modeling to analyze standards-based assessment items aligned to middle school chemistry ideas. In *Poster presented at the DR-K12 PI Meeting*.

Hin, K. K. (2020). PISA 2018 and Malaysia. *International Journal of Advanced Research in Education and Society*, *2*(3), 12-18.

Hoyle, R. H. (1995). *Structural Equation Modeling: Concepts, issues, and applications*. SAGE Publications.

Ismail, M. E., Samsudin, M. A., Amin, N. F. M., Kamarudin, N., Daud, K. A. M., & Halim, L. (2018). Contributing factors to science achievement in TIMSS Malaysia: Direct model and indirect model. *International Journal of Engineering & Technology*, *7*(4.30), 423-428.

Jansen, M., Scherer, R., & Schroeders, U. (2015). Students' self-concept and self-efficacy in the sciences: Differential relations to antecedents and educational outcomes. *Contemporary Educational Psychology*, *41*, 13-24. https://doi.org/10.1016/j.cedpsych.2014.11.002

Jamaludin, J., Fah, L. Y., Hoon, K. C., & Yee, L. S. (2020). Examining the STEM-Science Achievement Test (SSAT) using Rasch Dichotomous Measurement Model. *Solid State Technology*, *63*(1s), 648-661.

Joët, G., Usher, E. L., Bressoux, P. (2011). Sources of self-efficacy: An investigation of elementary school students in France. *Journal of Educational Psychology*, *103*(3), 649-663. https://psycnet.apa.org/doi/10.1037/a0024048

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. SAGE Publications.

Lay, Y. F., Ng, K. T., & Parahakaran, S. (2020). *Contributing factors to TIMSS 2011 eighth graders' science achievement: A comparison between Malaysia and Singapore*. Universiti Malaysia Sabah Press.

Lee, W., Lee, M. J., & Bong, M. (2014). Testing interest and self-efficacy as predictors of academic self-regulation and achievement. *Contemporary Educational Psychology*, *39*(2), 86-99. https://doi.org/10.1016/j.cedpsych.2014.02.002

Linacre, J. M. (2021). Winsteps® Rasch Measurement Computer Program. In *Beaverton, Oregon: Winsteps. com* (4.8.0). Oregon: Winsteps. com.

Liu, X. (2020). *Using and developing measurement instruments in science education: A Rasch Modeling approach* (2nd ed.). Information Age Publishing.

.Ministry of Education Malaysia (2013). *Malaysia Education Blueprint 2013-2025 (Preschool to Post-Secondary Education)*. Putrajaya: Ministry of Education Malaysia.

Moreno, R., Martínez, R. J., & Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *2*, 65–72. https://doi.org/10.1027/1614-2241.2.2.65

Ong, S. L. (2010). Assessment profile of Malaysia: High-stakes external examinations dominate. *Assessment in Education: Principles, Policy & Practice*, *17*(1), 91–103. https://doi.org/10.1080/09695940903319752

Owens, T. M. (2009). Improving science achievement through changes in education policy. *Science Educator*, *18*(2), 49-55.

Phang, F. A., Khamis, N., Nawi, N. D., & Pusppanathan, J. (2020). TIMSS 2019 science grade 8: Where is Malaysia standing? *Asean Journal of Engineering Education*, *4*(2), 37-43. http://tree.utm.my/wp-content/uploads/2021/01/AJEE-4.2.6.pdf

https://doi.org/10.33225/jbse/21.20.927

Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch analysis in Physics education research: Why measurement matters. *Physical Review Physics Education Research*, *15*(2), 020111. https://doi.org/10.1103/PhysRevPhysEducRes.15.020111

Ruel, E., Wagner, W. E., & Gillespie, B. J. (2015). *The practice of survey research*. SAGE Publications.

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research, and Evaluation*, *22*(1), 4. https://doi.org/10.7275/swgt-rj52

Soeharto, S., & Csapó, B. (2021). Evaluation and development of students' misconception using diagnostic assessment in science across school grades: A Rasch measurement approach. *Journal of Turkish Science Education*, *18*(3), 351-370. https://www.tused.org/index.php/tused/article/view/1227/702

Toksöz, S., & Ertunç, A. (2017). Item analysis of a multiple-choice exam. *Advances in Language and Literary Studies*, *8*(6), 141-146. http://dx.doi.org/10.7575/aiac.alls.v.8n.6p.141

Wright, B., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transaction*, *8*(3), 370.

Yang, Y., He, P., & Liu, X. (2018). Validation of an instrument for measuring students' understanding of interdisciplinary science in grades 4-8 over multiple semesters: A Rasch measurement study. *International Journal of Science and Mathematics Education*, *16*(4), 639-654. https://eric.ed.gov/?id=EJ1172609

Received: *May 08, 2021* Accepted: *November 30, 2021*

Cite as: Mohd Dzin, N. H., & Lay, Y. F. (2021). Assessing the validity and reliability of science multiple choice test using RASCH dichotomous measurement model. *Journal of Baltic Science Education*, *20*(6), 927-941. https://doi.org/10.33225/jbse/21.20.927

Najah Hazirah Mohd Dzin PhD Student, Faculty of Psychology and Education, University of Malaysia Sabah, Jalan UMS 88400 Kota Kinabalu, Sabah, Malaysia.
E-mail: zendoter@gmail.com

Yoon Fah Lay *(Corresponding author)* PhD, Professor of Science Education, Faculty of Psychology and Education, University of Malaysia Sabah, Jalan UMS 88400 Kota Kinabalu, Sabah, Malaysia.
E-mail: layyf@ums.edu.my
ORCID: https://orcid.org/0000-0002-5219-6696