JPII 10 (4) (2021) 571-584

Jurnal Pendidikan IPA Indonesia

http://journal.unnes.ac.id/index.php/jpii

EXAMINING THE STEM-SCIENCE ACHIEVEMENT TEST (SSAT) USING RASCH DICHOTOMOUS MEASUREMENT MODEL

J. Jamaludin[1], Y. F. Lay*[2], C. H. Khoo[3], A. S. Y. Leong[4]

[1,2]Universiti Malaysia Sabah, Sabah, Malaysia
[3,4]Institut Pendidikan Guru Kampus Kent, Tuaran, Malaysia

DOI: 10.15294/jpii.v10i4.29405

Accepted: March 8th 2021. Approved: December 27th 2021. Published: December 31st 2021

ABSTRACT

The main purpose of this study is to develop a valid and reliable instrument to measure the STEM-Science achievement of primary school students in Malaysia. Six Year 4 Science topics (Scientific Skills, Life Processes of Human, Properties of Materials, Measurement, Solar System, Importance of Technologies in Life) and Six Year 5 Science topics (Rules and Regulation in Science Lab, Life Processes of Plants, Acids and alkali, Electricity, Earth and Space Science, Technology and Sustainable Life) have been included in the development of the STEM-Science Achievement Test (SSAT). 226 Year 4 and 226 Year 5 primary school students in Sabah responded to the developed instrument to test their STEM-Science knowledge. The Rasch dichotomous measurement model approach was used to evaluate the validity and reliability of the SSAT. The validity assessed the Point-Measure Correlation (PTMEA CORR), Principal Component Analysis of Residuals (PCAR), as well as Infit and Outfit Mean Squares (MNSQ). In terms of reliability, Cronbach's alpha, item reliability, and item separation index were analysed. The analysis results revealed the presence of unidimensionality for both objective and subjective items. For objective items, the reported values for Cronbach's Alpha are .81 and .83; item reliability are .95 and .95; item separation are 4.21 and 4.25 for Year 4 and Year 5 students, respectively. Standardised residual correlations for Year 4 and Year 5 subjective items also showed satisfactory values. The assessment using Rasch measurement model has proven that SSAT is a valid and reliable instrument to measure Malaysian primary students' STEM-Science-related knowledge.

Keywords: validation; Rasch measurement model; science achievement test

INTRODUCTION

Validity is considered by numerous experts to be critical for any effective assessment (Bond & Fox, 2015). There have been many definitions of what validation of a test entails. For some, a valid test is simply one that measures what it was designed to measure. However, in order to evaluate the quality of these questions, a discussion of reliability is essential (Reardon et al., 2021). Reliability is the degree to which an instrument consistently measures the ability of an individual or group (Morales, 2009; Taber, 2018). Lamentably, not all countries evaluate the validity and reliability of their achievement test instrument. Hence, because of that, the use of poorly designed achievement test affects students' interest and achievement (Osadebe, 2015).

Science is considered as a core subject in school, and Malaysia is formulating its education policy to improve the system based on research findings by experts and consequently improve student achievements in science (Sumintono, 2017). According to Bhagat & Baliya (2016), an achievement test is a tool for teachers to measure the development of skills or knowledge proficiency (Anwer, 2019) of an individual into a certain topic that has been taught. Apart from that, a test design can determine the strengths and weaknes-

*Correspondence Address
E-mail: layyf@ums.edu.my

ses of students in the topic being taught (Laurens et al., 2017). Student performance is usually measured by performing several tasks such as answering tests or quizzes, final examination, and assignments (Ghulman & Mas' odi, 2009). Efficient assignments should be tailored to the same level of cognitive thinking skills (Scully, 2017) for all students in order to assess what they have learned. Well-designed, systematic tasks are either organised or constructed, which, referring to Bloom's cognitive thinking skills, should consider the students' abilities (Rahman & Manaf, 2017), thus contributing to improve students' performance (Haliza et al., 2012). Teachers assess the student's learning to determine how well they progress and understand the lesson. One way is through achievement tests. A suitable assessment tool in the teaching and learning process is required to quantify students' understanding (Parsons et al., 2019) and ability decently and fairly (Haliza et al., 2012). The current study has presented an analysis of the pilot test of a recently evolved science achievement test intended to decide how well students comprehend learning through STEM module taught by the teacher for Year 4 and Year 5 primary school students in Sabah. O'Reilly & McNamara (2007) expressed that students' science achievement is portrayed as content-based science performance, which can be estimated by some tests such as students' comprehension of a science entry, science course grade, and state science test scores (Román-González et al., 2019).

To ensure a valid and reliable measurement of students' science achievement, it is extremely crucial to determine the validity and reliability of the STEM-Science Achievement Test (SSAT). This initiative will help to develop an alternative and a more complex way of measuring students' science achievement (Boone, 2016; Planinic et al., 2019; Komarudin et al., 2021). Furthermore, with the help of the Rasch dichotomous measurement model, the likelihood of students (with different abilities) answering items (with different difficulty indices) correctly will be identified (Bond & Fox, 2015). The validity and reliability of the measurement depend on the correlation between the person's ability and the item difficulty (Susongko, 2016; Chan et al., 2021).

The study aims to develop an achievement test of which the validity and the reliability are ensured and can be used to determine Year 4 and Year 5 (aged 10-11) primary school students' science achievement. The Year 4 Science topics include 'Scientific Skills', 'Life Processes of Human', 'Properties of Materials', 'Measurement', 'Solar System', and 'Importance of Technologies in Life'. The Year 5 Science topics comprise of 'Rules and Regulation in Science Lab', 'Life Processes of Plants', 'Acid and Alkali', 'Electricity', 'Earth and Space Science', and 'Technology and Sustainable Life'. This study embarks on a research objective to validate the STEM-Science Achievement Test (SSAT) in measuring students' science achievement.

METHODS

The validity and reliability of the STEM-Science Achievement Test were determined prior to the actual study. Rasch dichotomous measurement model was used to examine the validity and reliability of the multiple-choice items, whereas partial credit model was used to examine the validity and reliability of the subjective items (Arnold et al., 2018).

A list of all primary schools in Penampang district and their relevant statistics for 2018 was obtained from the Sabah State Education Department. The samples in this study were drawn from the list using a purposive (Sharma, 2017) and random sampling routine (Rahi, 2017). The first sampling unit is the school. Once the schools have been selected, the Year 4 and Year 5 classes in these schools were made the targets of a second random sampling routine. The science teachers who have been appointed as the research assistants in the selected primary school carried out the random sampling routine. Once the classes have been chosen, the students in these classes formed the samples of this study. Due to grant project requirements, the research team adopted a purposive sampling strategy in rural and urban schools in the Penampang district. The most number of students in two urban and rural primary schools were selected to participate in this pilot study. 452 students were involved as samples in this study.

An official letter was sent to the sampled school in order to invite them to participate in the pilot achievement test. A project description, together with a copy of the permission letter from the Educational Planning and Research Division, Ministry of Education and the Sabah State Education Department, was attached. The school principals were asked to enclose the statistics on the number of Year 4 and Year 5 classes by a stream (track), as well as the number of students in each class for further sampling purpose. For practical purposes, the sampling of the classes was performed by the research assistant in the selected schools. A specific instruction was given to the research assistants to randomly select the classes and students who were involved in this study.

The STEM-Science Achievement Test (SSAT) consists of six topics in Year 4 and six topics in Year 5, with several individual items in Table 1.

Table 1. Learning Topic in SSAT

Year 4	Learning Topic	Number of Questions		Learning Outcomes
		Objective Type	Subjective Type	
	Scientific Skills	10	11	To master basic science process and manipulative skills.
	Life Processes of Human	10	15	1) To describe human breathing using nose, trachea, and various media. 2) Sketch the airway during inhale and exhale. 3) To encounter that the rate of respiration depends on the type of observing the movement of the chest.
	Properties of Materials	10	19	To examine the nature of the material in terms of its ability to absorb water.
	Measurement	10	17	1) To measure time using the right tools and using the right units in the right way. 2) Create a time measuring tool.
	Solar System	10	17	1) To list members of the Solar System. 2) To describe members in the Solar System. 3) To arrange the order of the planets in the Solar System. 4) To recognise that the planet revolves around its axis and at the same time circulates around the Sun. 5) To formulate planets in the solar system that spin on their axis and circulate the Sun in their orbits with simulations. 6) To create a creative and innovative model of the Solar System in creative form.
	Importance of Technologies in Life	10	17	1) To identify activities that the brain, senses, and limbs can and cannot do by carrying out activities. 2) To imply that humans have the capacity to perform activities. 3) To explain through examples of tools used to overcome human limitations through observation through various media. 4) To state that technology is one of the applications of science knowledge to overcome human limitations.
Total of Questions		60	96	

Year 5	Learning Topic	Number of Questions		Learning Outcomes
		Objective Type	Subjective Type	
	Rules and Regulation in Science Lab	10	9	To explain the rules and regulation in the Science Lab.
	Life Processes of Plants	10	12	1) To describe how plants disperse seeds or fruits through observation of actual plants or through various media: through water samples of lotus, coconut; by way of example, wind; through human and animal examples of buds, papaya; explosive mechanisms such as rubber, saga, cork; 2) To describe the method of exposure of plants with characteristics of seeds or fruits through observation of actual specimens or through various media.
	Acid and Alkali	10	12	1) To define acidic, alkaline, and neutral materials in terms of colour change in litmus paper. 2) To determine the properties of acidic, alkaline, and neutral materials in terms of colour-changing litmus paper, taste, and touch by testing several ingredients. 3) To give examples of acidic, alkaline, and neutral substances.
	Electricity	10	12	1) To state the complete electric circuit form. 2) To build complete electrical circuits using dry cells, bulbs, switches, and connectors. 3) To specify the function/function of the switch in the electrical circuit.
	Earth and Space Science	10	12	1) To describe the Moon not emitting light, but reflecting light from the Sun. 2) To describe the Moon rotating around its axis and at the same time circulating the Earth in terms of direction and duration by conducting simulations. 3) To use space and time relationships to describe the phases of the Moon such as moon, crescent, semicircular and full moon in a complete circulation following the Qamari calendar. 4) To describe observations through sketching, ICT, writing, or speaking.
	Technology and Sustainable Life	10	13	1) To state the importance of sustainable building value for sustainable living. 2) To create a stable model. 3) To improve the built-in-model based on the results that have been implemented.
Total of Questions		60	70	

Winsteps 4.8.0.0 software for Rasch was used to evaluate the validity and reliability of the STEM-Science Achievement Test (SSAT). Before running the Rasch analysis, the collected data were screened to ensure that they were no missing values, as well as suspicious response patterns and outliers. In this paper, the appropriate table was referred to in order to check the item value. Samples of certain tables in Winsteps that were used as reference in this article were explained briefly to get a clear picture of the analysis. In this analysis, reliability was determined by using Cronbach's alpha, item reliability, and item separation, whereas the validity of the items was examined through Point-Measure Correlation (PTMEA CORR). The interpretation of the result analysis has been explained by referring to the rating scale instrument quality criteria (Fisher, 2007), as shown in Table 2.

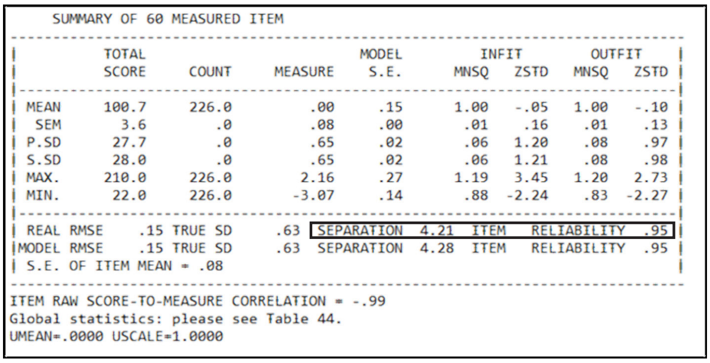
Table 2. Rating Scale Instrument Quality Criteria (Source: Fisher, 2007)

Criterion	Poor	Fair	Good	Very Good	Excellent
Targeting *	> 2 errors	1-2 errors	< 1 error	< .5 error	< .25 error
Item Model Fit Mean-Square Range Extremes	< .33 - >3.0	.34 - 2.9	.5 - 2.0	.71 - 1.4	.77 - 1.3
Person and Item Measurement Reliability	<.67	.67-.80	.81-.90	.91-.94	>.94
Person and Item Strata Separated	2 or less	2-3	3-4	4-5	>5
Ceiling Effect: % Maximum Extreme Scores	>5%	2-5%	1-2%	.5-1%	<.5%
Floor Effect: % Minimum Extreme Scores	>5%	2-5%	1-2%	.5-1%	<.5%
Variance in Data Explained by Measures**	<40%	50-60%	60-70%	70-80%	>80%

In Rasch analysis, PTMEA CORR is used as an immediate check to ensure that the response-level scoring makes sense. If the observed correlation is negative, something may have gone wrong (MCQ miskey, reversed survey item, etc.). Besides that, PTMEA CORR is also used to check the extent of the construct accuracy used in achieving the testing objectives. A positive PTMEA-CORR value indicates that the item is capable of measuring a construct, while negative PTMEA-CORR values indicate the opposite. In general, because correlations are much too difficult to interpret, we thus switched over to using mean-squares. Mean-Square (MNSQ) infit and outfit, *z*-STD infit and outfit standard fit statistics, and dimensional uniformity in Principal Component Analysis of Residuals (PCAR) were reported.

RESULTS AND DISCUSSION

Reliability refers to the extent to which assessments are consistent. The first statistics that we are referring to is called 'separation', which is the dispersion index for item positions. The table in Figure 1 is referred to in order to check the item reliability and item separation value. The table can be found in Table 3.1 in Winsteps.



```
    SUMMARY OF 60 MEASURED ITEM
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT         OUTFIT    |
|          SCORE      COUNT     MEASURE   S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|------------------------------------------------------------------------------|
| MEAN     100.7      226.0        .00    .15      1.00   -.05   1.00   -.10 |
|  SEM       3.6         .0        .08    .00       .01    .16    .01    .13 |
| P.SD      27.7         .0        .65    .02       .06   1.20    .08    .97 |
| S.SD      28.0         .0        .65    .02       .06   1.21    .08    .98 |
| MAX.     210.0      226.0       2.16    .27      1.19   3.45   1.20   2.73 |
| MIN.      22.0      226.0      -3.07    .14       .88  -2.24    .83  -2.27 |
|------------------------------------------------------------------------------|
| REAL RMSE    .15 TRUE SD      .63  SEPARATION  4.21  ITEM   RELIABILITY  .95 |
|MODEL RMSE    .15 TRUE SD      .63  SEPARATION  4.28  ITEM   RELIABILITY  .95 |
| S.E. OF ITEM MEAN = .08                                                      |
-------------------------------------------------------------------------------
ITEM RAW SCORE-TO-MEASURE CORRELATION = -.99
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000
```

Figure 1. Summary of 60 Measured Item in Objective Item for Year 4

Table 3 summarises the value of item reliability and item separation for Year 4 and Year 5. It can be seen that the item separation for objective item is very good for Year 4 and Year 5, which is 4.21 and 4.28, respectively. This separation index translates into approximately five levels of item difficulties, e.g., very easy, easy, moderate, difficult, and very difficult. Next, with the reliability index of person valued at .81 for Year 4 and .83 for Year 5 (analogous to the standard Cronbach's alpha) and the item reliability .95 for both classes, it is indicated that the items are excellent to consistently generate participant's score. In ensuring the quality of the items, the higher the level of item difficulties, the better the item is in examining the respondent's ability (Sener & Tas, 2017). "High reliability" (of persons or items) means that there is a high probability that persons (or items) are estimated with high measures than persons (or items) estimated with low measures. This data have revealed high item reliability, which indicate that the tests developed have a wide range of item difficulty and that the sample size is sufficient to establish a reproducible item difficulty hierarchy and minimise the measurement errors (Saidi & Siew, 2019).

Table 3. Reliability of Objective Item for Years 4 & 5

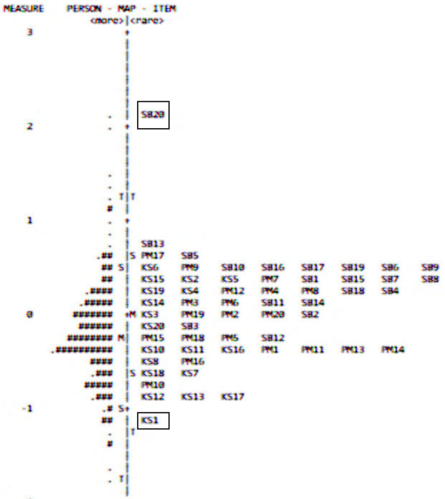
	Cronbach's Alpha	Item Reliability	Item Separation
Year 4	0.81	0.95	4.21
Year 5	0.83	0.95	4.28

In Table 4, the item separation for subjective item is shown to be good for Year 4 and Year 5, which is 3.27 and 3.56, respectively. This separation index translates to about four levels of item difficulties, e.g., easy, moderate, difficult, and very difficult. Next, with the reliability index of person valued at .78 for Year 4 and .84 for Year 5, this indicates that the items are consistently reproducing a participant's score. Parallel to this, the item reliability is .91 for Year 4 and .93 for Year 5.

Table 4. Reliability of Subjective Item for Year 4 & 5

	Cronbach's Alpha	Item Reliability	Item Separation
Year 4	0.78	0.91	3.27
Year 5	0.84	0.93	3.56

Both objective and subjective items in Year 4 and Year 5 indicate that a hierarchy of similar items across variables can be reproduced in a similar sample of population. This indicates good reliability of the items in measuring students' learning abilities. Item difficulty and person ability have been laid outside to side on the same measurement scale (vertical line with logit unit) in the Person-Item Distribution Map, as depicted in Fig. 2. The distribution map can be found in Table 16.3 in Winsteps. The results demonstrate that the mean of respondent is -0.26, and the mean of item is .0. In Year 5, the results show that the mean of respondent is .05, and the mean of item is 0.9. If the difference between the two values is less than 0.50, this test is hence within target (Linacre & Wright, 2012). On the person distribution area, both symbols: "." and "X" represent one and two test participant(s). "S" indicates a standard deviation away from the mean. The right-hand side shows the test items that are defined by the code, followed by the number of questions. Proper coding is important for item retrieving for identification.



```
MEASURE      PERSON - MAP - ITEM
                    <more>|<rare>
  3                       +
                          |
                          |
                          |
                          |
                          |
                          |
                        . |  SB20
  2                     . +
                          |
                          |
                          |
                        . |
                        . |
                        . T|T
                        # |
  1                     . +
                        . |
                        . |  SB13
                      .## |S PM17  SB5
                       ## S|  KS6   PM9   SB10  SB16  SB17  SB19  SB6   SB9
                       ## |  KS15  KS2   KS5   PM7   SB1   SB15  SB7   SB8
                    .#### |  KS19  KS4   PM12  PM4   PM8   SB18  SB4
                   .##### |  KS14  PM3   PM6   SB11  SB14
  0               ####### +M KS3   PM19  PM2   PM20  SB2
                   ###### |  KS20  SB3
                 ######## M|  PM15  PM18  PM5   SB12
             .########### |  KS10  KS11  KS16  PM1   PM11  PM13  PM14
                     #### |  KS8   PM16
                     .### |S KS18  KS7
                    ##### |  PM10
                     .### |  KS12  KS13  KS17
 -1                    .# S+
                       ## |  KS1
                        . |T
                        # |
                          |
                        . |
                        . T|
                          |
 -2                     . +
                          |
                          |
                        . |
                          |
                          |
                          |
```

Figure 2. Person-Item Distribution Map of Year 4 Objective Test

As can be seen in Figure 2, the most difficult item (SB20) is located at the top, and the easiest item (KS1) is located at the bottom. When the difficulty of the item and the respondent's ability are matched, the respondent has a 50% chance of answering the item correctly (DeBoer et al., 2009). The results show that the mean of respondents in Year 4 is .04, and the mean of item is .08. In Year 5, the results reveal that the mean of respondent is -1.7, and the mean of item is .0. If the difference between the two values is less than .50, the telling of that is that this test is within target (Linacre & Wright, 2012).

Table 5 shows the PTMEA CORR for Year 4 objective items. To examine the construct validity, item polarity (PTMEA CORR) findings have been assessed. Based on Table 5, all items show a positive PTMEA CORR value of .03 to .44, pointing that the response latencies of the items are in line with ideas (Linacre, 2003). A positive PTMEA CORR value shows that the item can measure what it needs to measure (Bond & Fox, 2015).

Table 5. Point-Measure Correlation (PTMEA CORR) for Year 4 Objective Item

Item	Measurement Score	PTMEA CORR	Item	Measurement Score	PTMEA CORR
46	54	.03	22	-.04	.28
52	-23	.11	14	.12	.28
6	.47	.13	56	.49	.28
37	.58	.14	55	.36	.29
5	.32	.14	9	-1.01	.29
32	.26	.17	48	.34	.30
4	.20	.18	35	-.21	.30
2	.41	.18	34	-.35	.30
8	-.48	.19	7	-.64	.30
49	.47	.20	44	.22	.30
3	-.02	.20	10	-.33	.31
45	.60	.20	21	-.37	.31
59	.49	.21	43	-.13	.32
60	.69	.21	29	54	.32
31	-.39	.21	42	.00	.32
16	-.35	.22	30	-.76	.32
53	.74	.24	36	-.50	.33
38	-.21	.25	24	.22	.34
28	.22	.25	51	.14	.34
40	-.02	.25	23	.08	.34
41	.30	.26	13	-.82	.34
27	.32	.26	15	.38	.35
26	.16	.26	1	-1.15	.35
19	.22	.27	25	-.21	.35
57	.54	.27	33	-.40	.35
58	.24	.27	18	-.60	.36
39	-.06	.28	20	-.13	.37
50	.43	.28	12	-.87	.39
11	-.35	.28	17	-.93	.44

Table 6 shows the PTMEA CORR for Year 5 objective items. The negative PTMEA CORR values indicate that the items are opposite to the objective of the testing. Therefore, the items need to be checked because they are either difficult to be answered, or it did not address the construct accurately.

Table 6. Point-Measure Correlation (PTMEA CORR) for Year 5 Objective Item (Before Item Adjustment)

Item	Measurement Score	PTMEA CORR	Item	Measurement Score	PTMEA CORR
52	1.28	-.08	15	-.16	.35
49	.75	-.01	18	-.26	.35
6	.28	.05	25	-.04	.36
46	.96	.07	26	.00	.37
5	-.06	.14	43	-.14	.39
35	.32	.14	32	-.24	.40
60	.21	.18	33	-.04	.41
53	.64	.18	8	-1.11	.30
47	.57	.17	48	.41	.31
40	.61	.20	12	-.47	.33
54	.61	.21	3	-1.08	.33
37	.44	.21	39	.10	.34
9	.76	.22	17	-.43	.41
58	.44	.22	13	-.22	.41
10	-.31	.24	29	.20	.43
59	.26	.24	7	-.65	.44
34	.72	.23	20	-.24	.45
31	.24	.25	19	.06	.45
11	.00	.25	28	-.24	.46
16	.40	.25	27	-.78	.46
42	.10	.27	21	-.76	.47
44	-.14	.27	22	-.43	.47
56	.40	.27	23	-.51	.51
36	.82	.27	24	-.36	.52
55	.46	.29			

Based on Table 6, it was found that items 52 and 49 have negative PTMEA CORR values of −.08 and −.01, signifying that the response latencies of the items are inversely proportional (Linacre, 2003). Therefore, items 52 and 49 were checked and re-adjusted.

Table 7. Point-Measure Correlation (PTMEA CORR) for Year 5 Objective Item (After Item Adjustment)

Item	Measurement Score	PTMEA CORR	Item	Measurement Score	PTMEA CORR
6	.31	.05	49	-.46	.31
46	1.01	.06	8	-1.08	.31
5	-.03	.14	14	-.22	.31
2	-2.54	.13	48	-.38	.32
35	.35	.14	12	-.44	.33
47	.61	.16	3	-1.06	.34
51	.68	.17	39	.13	.34
45	.44	.18	15	-.13	.35
58	2.09	.20	18	-.22	.35
40	.65	.21	25	-.01	.36

37	.48	.21	26	.03	.38
52	.65	.22	43	-.11	.39
56	.48	.22	32	-.20	.40
9	-.73	.23	33	-.01	.41
34	.77	.23	17	-.40	.41
31	.27	.24	13	-.18	.42
10	-.27	.24	29	.23	.43
57	.29	.24	7	-.62	.44
16	.44	.25	20	-.20	.45
11	.03	.26	19	-.09	.45
4	-.16	.27	28	-.20	.46
54	.44	.27	21	-.73	.46
42	.13	.27	22	-.40	.47
50	.65	.27	27	-.75	.47
30	.31	.27	23	-.48	.51
44	-.11	.28	24	-.32	.52

After adjustment, the positive PTMEA CORR value indicates that the item is moving in one direction and measures the ideas to be measured (Bond & Fox, 2015), as shown in Table 7. Based on Table 8, all 96 PTMEA CORR values have been found to be positive .00 to .50. A positive PTMEA CORR value indicates that the item is actuating in one direction and can measure what is supposed to be measured (Bond & Fox, 2015).

Table 8. Point-Measure Correlation (PTMEA CORR) for Year 4 Subjective Item

Item	Measurement Score	PTMEA CORR	Item	Measurement Score	PTMEA CORR
75	1.07	.26	83	-.05	.34
68	1.02	.24	16	-.05	.20
25	1.01	.24	20	-.05	.11
46	1.00	.20	28	-.05	.24
26	.99	.20	80	-.07	.50
41	.99	.24	34	-.09	.27
15	.99	.15	89	-.09	.30
27	.99	.24	38	-.16	.35
67	.99	.24	62	-.19	.26
94	.99	.18	19	-.19	.16
18	.96	.30	58	-.19	.15
29	.96	.18	7	-.23	.38
30	.96	.33	10	-.27	.45
24	.95	.23	55	-.33	.19
79	.94	.17	82	-.33	.31
61	.93	.42	73	-.36	.28
23	.92	.28	74	-.36	.30
48	.82	.09	8	-.37	.39
56	.82	.19	6	-.37	.33
49	.75	.10	35	-.39	.33
36	.71	.32	88	-.39	.40
70	.69	.15	17	-.48	.28

22	.62	.24	63	-.48	.26
78	.61	.30	12	-.49	.06
54	.56	.20	1	-.54	.25
42	.53	.21	14	-.54	.16
32	.39	.18	5	-.57	.15
51	.39	.12	43	-.58	.26
13	.29	.00	39	-.60	.37
50	.25	.14	65	-.61	.31
52	.25	.09	95	-.65	.45
57	.25	.16	72	-.65	.36
93	.25	.17	2	-.68	.16
44	.20	.22	59	-.78	.32
53	.20	.08	86	-.80	.42
76	.20	.30	91	-.82	.33
64	.18	.29	37	-.83	.35
45	.18	.23	31	-.86	.24
40	.16	.25	4	-.89	.14
69	.14	.25	92	-.96	.37
60	.11	.31	81	-1.00	.46
71	.11	.23	90	-1.00	.48
87	.08	.40	77	-1.02	.22
66	.08	.22	96	-1.05	.42
3	.07	.13	85	-1.12	.42
47	.07	.14	84	-1.25	.50
33	.03	.27	11	-1.38	.44
21	-.01	.17	9	-1.41	.47

Based on Table 9, all values of PTMEA CORR 70 items have been shown to be positive values of .01 to .54. A positive PTMEA CORR value stresses that the item is moving in one direction and measures the idea that it wants to measure (Bond & Fox, 2015).

Table 9. Point-Measure Correlation (PTMEA CORR) for Year 5 Subjective Item

Item	Measurement Score	PTMEA CORR	Item	Measurement Score	PTMEA CORR
26	1.30	.14	18	.03	.45
66	1.26	.33	60	.00	.20
58	1.24	.47	16	-.07	.28
37	1.22	.36	41	-.07	.26
49	1.22	.35	57	-.10	.26
35	1.20	.39	3	-.10	.23
56	1.15	.37	63	-.13	.24
4	.85	.29	20	-.18	.39
52	.66	.18	6	-.18	.26
59	.66	.16	31	-.19	.41
62	.61	.40	50	-.19	.22
64	.61	.42	54	-.21	.45
30	.51	.27	67	-.28	.29
34	.46	.06	11	-.31	.27

44	.46	.15	36	-.31	.37
39	.41	.36	53	-.31	.40
29	.38	.29	9	-.44	.23
42	.33	.13	33	-.47	.27
48	.33	.35	19	-.52	.53
27	.29	.30	2	-.54	.33
43	.29	.17	25	-.54	.54
61	.29	.01	55	-.62	.27
68	.29	.25	17	-.74	.42
24	.25	.24	47	-.78	.37
28	.25	.11	10	-.78	.31
70	.25	.20	15	-.78	.32
51	.21	.08	14	-.81	.32
65	.18	.13	22	-.85	.27
23	.14	.05	21	-.91	.43
7	.10	.34	13	-.94	.26
32	.10	.16	45	-.96	.07
38	.10	.25	69	-.98	.19
40	.10	.17	46	-1.02	.25
1	.07	.15	12	-1.18	.40
5	.03	.30	8	-1.41	.25

The Component Analysis of Residuals (PCAR) technique is used to cinch undeviating of the instrument dimensions. It ascertains the ability of the instrument to scale in a uniform dimension with the adequate level of distraction. From the analysis, it has been found that the variance explained by the measures for Year 4 objective item is 14.0%. Meanwhile, the variance not explained in the first contrast is 7.4%. For Year 5 objective item, the variance explained by measures is 15.6%, while the variance not explained in the first contrast is 5.4%, as depicted in Table 10. From the data obtained, the variance in the data described is less than the minimum that covers 40% of Rasch's requirements. However, the variance unexplained in Contrast 1, which is 5.0% and 3.7% high, is well controlled and far from the ceiling value of 15%. The strength of the Rasch dimension, 9.8 and 10.7, can be directly compared to the strength of the secondary dimension (1st contrast) 5.0 and 3.7, showing that for testing purposes, the data can be clarified unidimensional.

Table 10. PCAR for Objective Item

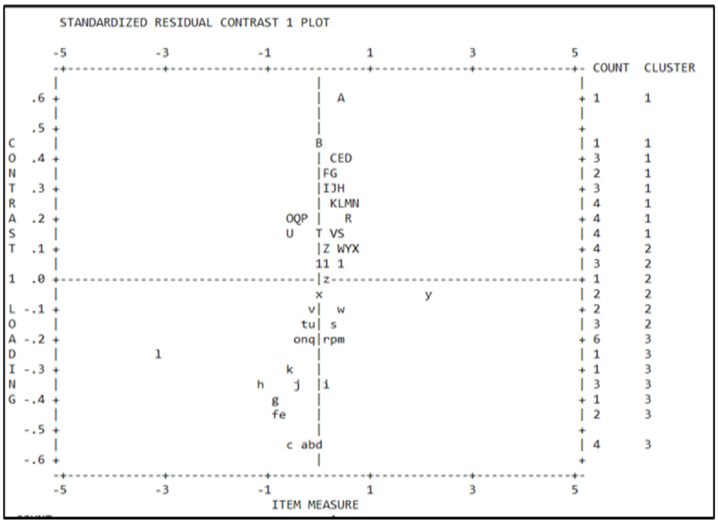
	Raw Variance Explained by Measures % (Eigenvalue)	**Variance Unexplained % (Eigenvalue)**	**Variance Unexplained in Contrast 1 % (Eigenvalue)**
Year 4	14.0% (9.8)	86.0 % (60)	7.4 % (5.1)
Year 5	15.6% (10.7)	84.4 % (58)	5.4 % (3.7)

From the analysis of the subjective item, it has been found that the variance explained by the measures for Year 4 subjective item is 19.2%. Meanwhile, the variance not explained in the first contrast is 3.7%. For Year 5 subjective item, the variance explained by measures is 16.7%, while the variance not explained in the first contrast is 3.6%, as shown in Table 11. From the data gathered, the variance in the data described is less than the minimum that covers 40% of Rasch's requirements. However, the variance unexplained in Contrast 1, which is 4.5% and 3.0% high, is well measured and far from the ceiling value of 15%. The strength of the Rasch dimension, 22.9 and 14.1, can be directly compared to the strength of the secondary dimension (1st contrast), 4.5 and 3.0, showing that for testing purposes, the data can be perceived unidimensional.

Table 11. PCAR for Subjective Item

	Raw Variance Explained by Measures % (Eigenvalue)	Variance Unexplained % (Eigenvalue)	Variance Unexplained in Contrast 1 % (Eigenvalue)
Year 4	19.2% (22.9)	80.8 % (96.0)	3.7 % (4.5)
Year 5	16.7% (14.1)	83.3 % (70.0)	3.6 % (3.0)

The next process of identification is by looking at the contrast 1 plot, as shown in Figure 3. Figure 3 is the contrast plot that shows that the position of the items is randomly scattered along the logit, but in the formation of clusters. Therefore, further examination should be carried out by looking at the value of disattenuated correlation in Table 12.



```
       STANDARDIZED RESIDUAL CONTRAST 1 PLOT

      -5            -3            -1             1             3             5
      -+-------------+-------------+-------------+-------------+-------------+- COUNT  CLUSTER
      |                                   |                                   |
   .6 +                                   |  A                                + 1      1
      |                                   |                                   |
   .5 +                                   |                                   +
C     |                                   B                                   | 1      1
O  .4 +                                   | CED                               + 3      1
N     |                                   |FG                                 | 2      1
T  .3 +                                   |IJH                                + 3      1
R     |                                   | KLMN                              | 4      1
A  .2 +                               OQP |   R                               + 4      1
S     |                               U   T VS                                | 4      1
T  .1 +                                   |Z WYX                              + 4      2
      |                                   11 1                                | 3      2
1  .0 +-----------------------------------|z----------------------------------+ 1      2
      |                                   x              y                    | 2      2
L -.1 +                                  v|  w                                + 2      2
O     |                                 tu| s                                 | 3      2
A -.2 +                                onq|rpm                                + 6      3
D     |             1                     |                                   | 1      3
I -.3 +                               k   |                                   + 1      3
N     |                           h    j  |i                                  | 3      3
G -.4 +                             g     |                                   + 1      3
      |                             fe    |                                   | 2      3
  -.5 +                                   |                                   +
      |                               c abd                                   | 4      3
  -.6 +                                   |                                   +
      -+-------------+-------------+-------------+-------------+-------------+-
      -5            -3            -1             1             3             5
                                    ITEM MEASURE
```

Figure 3. Standardised Residual Contrast 1 Plot Objective Item

The purpose of checking the disattenuated correlation coefficient according to Muchinsky (1996) is that the disattenuation tells us whether the correlation between two sets of measures is low because of measurement error or because the two sets are uncorrelated. Disattenuated correlation does not change the quality of the measures or their predictive power; not directly comparable with uncorrected correlations; not suited to statistical hypothesis testing; and not a substitute for precise measurement. Therefore, based on Table 12, the disattenuated value is more than zero and positive, which indicates an item measuring in the same direction as another item. All items, either objective or subjective, were examined on their disattenuated correlation for unidimensionality purposes. In this study, unidimensional STEM-science achievement test would be designed to measure only STEM-science achievement, but not other constructs.

Table 12. Disattenuated Correlation for Objective Item

Approximate Relationships Between the PERSON Measures			
PCA Contrast	Item Clusters	Pearson Correlation	Disattenuated Correlation
1	1-3	.0097	.1310
1	1-2	.5179	.7892
1	2-3	.4973	.7318

The item misfit is explained based on the data in Table 13. The Mean Square (MNSQ) value of 60 items is within the range suggested by Wright & Stone (1999). However, there are several items with *z*-STD infit values that are outside the range, as suggested by Linacre & Wright (2012), ranging from -2.0 to +2.0. If IN z-STD is below the set range, this means the that the model is overfit data, where it can be said that the item is in line with Rasch's model predictions. Meanwhile, IN z-STD values that exceed the range of this setting show that the model is underfit data, indicating that the items do not follow Rasch model predictions (Linacre, 2014). Therefore, these items were re-examined and later determined whether they require elimination or repairing. Outfit items indicate that the items were answered casually or based on guessing, while items that are considered as infit are more sensitive. However, it is very difficult to explain the cause. The number of items misfit in the objective and subjective test is shown in Table 13.

Table 13. Item Misfit

	Year		Underfit Model	Overfit Model
Objective Item	Year 4	Number of Items	3	2
		Item No.	46, 4, 52	20, 17
	Year 5	Number of Items	4	8
		Item No.	6, 46, 5, 35	13, 29, 7, 20, 19, 28, 21, 22
Subjective Item	Year 4	Number of Items	7	3
		Item No.	4, 5, 6, 7, 11, 12, 77	81, 84, 90
	Year 5	Number of Items	2	0
		Item No.	6, 46	-

Action has been taken to this misfit item such as reconsideration in terms of the way of questioning, re-structuring the item questions, or re-adjusting the item level of difficulty.

CONCLUSION

The assessment of science achievement item in the STEM-Science Achievement Test (SSAT) is a valid and reliable instrument to measure Malaysian primary students' achievement in Year 4 and Year 5 STEM-Science. The remaining items in the SSAT have met the minimum threshold values, as required by outer, Cronbach's alpha, item reliability, item separation, PTMEA CORR, Mean Square Fit, and Standardised fit statistics.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the research grant (SDK0026-2018) of Universiti Malaysia Sabah (UMS) and all those who have been involved, directly and indirectly, in the validation of research instruments in various phases of this study.

REFERENCES

Anwer, F. (2019). Activity-based teaching, student motivation and academic achievement. *Journal of Education and Educational Development*, *6*(1), 154-170.

Arnold, J. C., Boone, W. J., Kremer, K., & Mayer, J. (2018). Assessment of competencies in scientific inquiry through the application of Rasch measurement techniques. *Education Sciences*, *8*(4), 184.

Bhagat, P., & Baliya, A. (2016). Construction and validation of achievement test in science. *International Journal of Science and Research*, *5*(6), 2277-2280.

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

Boone, W. J. (2016). Rasch analysis for instrument development: why, when, and how?. *CBE—Life Sciences Education*, *15*(4), rm4.

Chan, S. W., Looi, C. K., & Sumintono, B. (2021). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. *Journal of Computers in Education*, *8*(2), 213-236.

DeBoer, G. E., Herrmann-Abell, C. F., Wertheim, J., & Roseman, J. E. (2009, April). Assessment linked to middle school science learning goals: A report on field test results for four middle school science topics. In *Annual Meeting of the National Association of Research in Science Teaching, held* (pp. 17-21).

Fisher, W.P. Jr. (2007). Rasch measurement transactions. *Journal of Measurement, 21*(1),1094.

Ghulman, H. A., & Mas' odi, M. S. (2009, December). Modern measurement paradigm in Engineering Education: Easier to read and better analysis using Rasch-based approach. In *2009 International Conference on Engineering Education (ICEED)* (pp. 1-6). IEEE.

Haliza, O., Izamarlina, A., Hafizah, B., Zulkifli, M. N., & Nur Arzilah, I. (2012). Application of Rasch measurement model in reliability and quality evaluation of examination paper for engineering mathematics courses. *Social and Behavioral Sciences*, *60*, 163 – 171.

Komarudin, K., Suherman, S., & Anggraini, A. (2021). Analysis of Mathematical Concept Understanding Capabilities: The Impact of Makerspae STEM Learning Approach Models and Student Learning Activities. *Journal of Innovation in Educational and Cultural Research, 2*(1), 35-43.

Laurens, T., Batlolona, F. A., Batlolona, J. R., & Leasa, M. (2017). How does realistic mathematics education (RME) improve students' mathematics cognitive achievement? *Eurasia Journal of Mathematics, Science and Technology Education*, *14*(2), 569-578.

Linacre, J. M. (2003). Dimensionality: Contrasts and Variances. Help for Winsteps Rasch Measurement Software. Available: http://www.winsteps.com/winman/principalcomponents.htm

Linacre, J. M. (2014). Infit mean square or infit z-std [From Rasch Measurement Forum discussion board]. Available: raschforumboards.netnals.

Linacre, J. M., & Wright. B. D. (2012). A User's Guide to WINSTEPS Ministeps Rasch Model Computer Programs. Chicago: Mesa Press.

Morales, R. A. (2009). Evaluation of mathematics achievement test: A comparison between CTT and IRT. *The International Journal of Educational and Psychological Assessment, 1*(1), 19-26.

Muchinsky, P. M. (1996). The correction for attenuation. *Educational & Psychological Measurement, 56* (1), 63-75.

O'Reilly, T., & McNamara, D. S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high-stakes" measures of high school students' science achievement. *American educational research journal, 44*(1), 161-196.

Parsons, S., Kruijt, A. W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378-395.

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.

Osadebe, P. U. (2015). Construction of Valid and Reliable Test for Assessment of Students. *Journal of Education and Practice*, *6*(1), 51-56.

Rahi, S. (2017). Research design and methods: A systematic review of research paradigms, sampling issues and instruments development. *International Journal of Economics & Management Sciences*, *6*(2), 1-5.

Rahman, S. A., & Manaf, N. F. A. (2017). A critical analysis of Bloom's taxonomy in teaching creative and critical thinking skills in Malaysia through English literature. *English Language Teaching*, *10*(9), 245-256.

Reardon, S. F., Kalogrides, D., & Ho, A. D. (2021). Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale. *Journal of Educational and Behavioral Statistics*, *46*(2), 138-167.

Román-González, M., Moreno-León, J., & Robles, G. (2019). Combining assessment tools for a comprehensive evaluation of computational thinking interventions. In *Computational thinking education* (pp. 79-98). Springer, Singapore.

Saidi, S. S., & Siew, N. M. (2019). Reliability and validity analysis of Statistical Reasoning Test survey instrument Using the Rasch Measurement Model. *International Electronic Journal of Mathematics Education*, *14*(3), 535-546.

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research, and Evaluation*, *22*(1), 4.

Sener, N., & Tas, E. (2017). Developing achievement test: A research for assessment of 5th grade biology subject. *Journal of Education and Learning*, *6*(2), 254-271.

Sharma, G. (2017). Pros and cons of different sampling techniques. *International Journal of Applied Research*, *3*(7), 749-752.

Susongko, P. (2016). Validation of science achievement test with the Rasch model. *Jurnal Pendidikan IPA Indonesia*, *5*(2), 268-277.

Sumintono, B. (2017). Science education in Malaysia: Challenges in the 21st. century. *Jurnal Cakrawala Pendidikan*, *36*(3), 459-471.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273-1296.

Wright, B., & Stone, M. (1999). Measurement Essentials 2nd Edition. Wilmington, Delaware: Wide Range, Inc.