# Analyzing the science achievement test: Perspective of classical test theory and Rasch analysis

**Law Hui Haw[1], Sabariah Bte Sharif[2], Crispina Gregory K. Han[2]**
[1]Department of Science, Institute of Teacher Education Temenggong Ibrahim Campus, Johor Bahru, Malaysia
[2]Department of Education, Faculty of Psychology and Education, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia

**Article Info**

***Article history:***

Received Jun 12, 2021
Revised Jul 21, 2022
Accepted Aug 19, 2022

***Keywords:***

Difficulty index
Discriminant index
Rasch analysis
Reliability
Science achievement test
Validity

**ABSTRACT**

Assessment in teaching and learning allows both instructor and students to evaluate the achievement, performance and improve upon it. The achievement test appears as one of the most commonly used assessment tools to determine students' gains in the cognitive domain. The study analysed science achievement test (SAT), a multiple-choice test for grade 10 senior secondary school students. The 50 SAT items were developed in compliance with the Malaysian science curriculum specification. The face and content validity of SAT were validated by experts in the science field. The SAT was administered with 50 students in a pilot study. The collected data was analysed using item analysis based on classical test theory (CTT) and Rasch analysis. In terms of difficulty index (p), item analysis showed that there were five difficult items, 33 moderately difficult items, and 12 easy items. Findings also reported eight items presented as poor items due to the poor discriminating power. The SAT showed a good coefficient of 0.862 for Kuder-Richardson 20 (KR-20) and 0.851 for the split half method. While the Rasch analysis showed a reliable and good separation for both item and person. Besides, the Rasch analysis displayed that there are two items should be refreshed or removed. The study revealed that the SAT was a solid, valid, and reliable tool, which is suitable to be used to measure the student's achievement in science.

***Corresponding Author:***

Law Hui Haw
Department of Science, Institute of Teacher Education Temenggong Ibrahim Campus
Taman Dato Onn Jaffar, 81200 Johor Bahru, Johor, Malaysia
Email: jamie_lawhh@yahoo.com

## 1. INTRODUCTION

The standards-based curriculum for secondary schools (KSSM) of Malaysian science education aims to strengthen the interest and develop student creativity through experience and investigation to acquire knowledge in science, scientific skills, thinking skills, scientific attitudes, and values [1]. In Malaysia, secondary science subjects are structured to develop students' science literacy, high order thinking skills, and the ability to apply scientific knowledge in making decisions and solving problems in real life. The student's achievement in the subject of science, therefore, needs attention and should be emphasized.

Academic achievement refers to the educational outcome, that indicates the extent to which an individual has accomplished particular educational goals [2]. The science achievement is not only providing the information about the performance and knowledge of students, but also a determinant for the future of youths and a nation, especially in promoting the industry revolution 4.0 (IR 4.0). Science seems to be one of the most important subjects taught in Malaysia, right from elementary school [3]. To date, low achievement in Science is still an issue in the country. For instance, in the recent results of third international mathematics and science study (TIMSS) in 2019 [4] the average score of science even dropped about 11 points to 460 compared to 2015. According to Awopeju and Afolabi [5], the nature of the test items and the learners' characteristics, are prominent among these factors affecting the achievement of students. The examinee's ability and the item's features could predict or explain the success of an examinee. The Malaysian education system now faces the challenge of developing suitable and meaningful ways to evaluate the extent to which students are achieving the standards [6]. The development of high quality, accessible achievement tests require substantial knowledge of a content area and the design of test items or tasks that are fair and valid measures of important knowledge and skills at a given content area [7].

Generally, there are two common theories supported the development of measurement tests, namely classical test theory (CTT) and item response theory (IRT). The CTT is always used to assess the reliability and validity of an achievement test. It introduces three concepts such as test score, true score, and error score [8], [9]. The test score (X), which is an observed score is linked to the unobserved true score (T) and error score (E) by the formula (1):

$$X = T + E \tag{1}$$

Since the true score is not easily observable, hence, there are some major assumptions made in CTT: i) The average error score of the examinees is zero; ii) True scores and error scores are uncorrelated; and iii) Error scores on the parallel tests are uncorrelated [10]. The number of correct is always accounted as examinees ability. To evaluate the tests, analysis based on CTT had been used traditionally [11]. The basic measures of item analysis were item discrimination and item facility, which was also known as item difficulty [12]. Item difficulty was analysed to see which items were more difficult than other items based on the value of the item difficulty index (p) of the items [6]. The higher the index value, the easier the item was, and the lower index value indicated a more difficult item. On the other hand, the high and low performing students were compared with the discrimination index (D) [13]–[15]. The test had high reliability if the mean value of the D value was high [16].

However, CCT had several limitations. One of the major weaknesses was essentially sample-based descriptive statistic[17]. Consequently, the generalisation of its estimators was more difficult especially when the population of the examinees was diversed in their abilities [10]. This means if the test was administered to a group of weak students, the value of (p) and (D) would be low whereas the index value would be high if administered to good students.

By comparing CTT and IRT, IRT focused on the form of examinees' responses. An assumption of IRT was that the probability of a student answering an item correctly was also the function of the item difficulty and student ability [18]. The item responses in IRT were considered the outcome (dependent variables), whereas the examinee's ability and the characteristics of the item were the latent predictors (independent variables) [19]. Rasch model appeared as one of the main models of IRT. It was a probabilistic unidimensional model that emphasised that: i) The easier the question, the more likely the student responded correctly to it; and ii) The higher the student's ability, the more likely he or she would pass the question compared to a lower ability student. Therefore, those items that did not meet the assumptions of this model were discriminated [11].

In Rasch analysis, several diagnoses are employed to check the relevance of the test construct. The reliabitity of a test can be analysed by separation diagnosis. Item separation is defined either as the number of distinct item difficulty strata or the level of the test takers' performance. It is also a range of item difficulties which are included in the test and can be ensured by examining the item strata [20]. To test the validity of the test, fit statistics are used. Analysis of fit statistics was conducted to assess the quality of the items and guide the process of measurement to detect the lack of fit and too good fit [21]. The fit statistics are normalized mean square (MNSQ) residuals (across items for each person or across persons for each item). Another diagnosis in Rasch analysis, item polarity is the PTMEA CORR between the ability measure of the respondents who took the item and the item score [20].

However, up to this point many Malaysian educators are still applying the CTT in doing analysis on test items [6]. CTT is still widely used since it is a more economical and practical method of developing quality test items [22]. After several considerations, the researchers are motivated to analyse the achievement test by considering both CTT and Rasch model of IRT. In this study, the researchers' goal was to develop a valid and reliable achievement test, to assess the science achievement of senior secondary school students (ages 16-17). Besides, the study also tried to identify the inappropriate items in the test, for the improvement of items later.

## 2. RESEARCH METHOD

### 2.1. General background

According to Sharma and Poonam [23], the achievement test focused on the realisation of teaching and learning objectives. Basically, the questions format was classified into closed-ended and open-ended. The common examples of closed-ended formats included multiple-choice, matching, and true-false; while open-ended formats are short answer, free-response, and essay writing [24]. Each format had respective strengths and limitations and therefore the instructors had to determine the convenient ones for the evaluation purposes [25]. The practical constraints, such as limited assessment time, ease of administration and grading resources were taken into consideration when choosing the question format (i.e., item type) [26].

A multiple choices question (MCQ) test paper was designed according to the curriculum specification of science subject in grade 10 level. MCQs were selected because they appeared to be the most commonly used type of test, used alone or in combination with other types of assessment test instruments [27]. They were also suitable for assessing knowledge and comprehension, and they could be designed to assess application and analysis [28]. Rodriguez [29] added that MCQ is efficient to administer, easy to score objectively, and they can be used to sample a wide range of content domains in a relatively short time using a single test administration. Furthermore, due to the higher reliability, validity, and ease of scoring, MCQ is becoming more common to be used [30], [31].

The pilot test was conducted with 50 students randomly. From the responses gathered, the researchers analysed the item analysis, difficulty index and discriminant index. The reliability in terms of Kuder-Richardson 20 (KR 20) and split half were examined by using statistical package for the social sciences (SPSS) version 22.0. Rasch analysis was also conducted on the reading subtest data using Winstep v3.73.3 to analyse item separation, Wright map, item polaritiy and item fit.

### 2.2. Sample or participants

A pilot study for science achievement test (SAT) was conducted with 50 grade 10 students at one of the government secondary schools in Limbang district. The sample was appropriate according to the suggestion of previous research [32]. The samples were randomly picked by the researchers. Researchers administrated the test as in the real examination situation, including the arrangement of sitting and time allocated. The duration of the test was one hour and fifteen minutes. The respondents will answer the paper in the answer sheet provided.

### 2.3. Instrument and procedures

The SAT is an assessment designed to evaluate the achievement level of students in science subject. The test consisted of 50 multiple choice questions, which were developed by an excellent teacher with over ten years of experience teaching science subjects. The face validity and content validity were validated by a team of experts, comprised of an associate professor, a senior lecturer and two experienced science teachers who had been teaching science for more than 20 and 30 years. The SAT was developed in accordance with the curriculum specification of grade 10 science [33] using the test specification table (TST) based on Bloom's Taxonomy [34].

According to Sharma and Sarita [35], the analysis of content is another crucial stage in the development of an achievement test. The researchers designed the SAT grounded on the ‘Malaysian certificate of education’ format as in Table 1. All the seven chapters in grade 10 science, based on four main themes, namely ‘maintenance and sustainability of life’, ‘matter in nature’, ‘energy and life’ and ‘the development of industrial technology in society’ are focused in developing the SAT.

Table 1. The design of SAT format

| No. | Information | Details |
|---|---|---|
| 1 | Item type | Multiple-choice questions: Each item has four choices<br>(One answer and three distractors)<br>A, B, C, and D. |
| 2 | Total of item | 50 |
| 3 | Test duration | 1 $^1/_4$ hour |
| 4 | Level of difficulty | Low: Medium: High<br>5: 3: 2<br>(25 easy items: 15 moderately difficult items: 10 difficult items) |

### 2.4. Data analysis

The experts had varying views on the appropriateness of difficulty index value. For instance, According to Macinstosh and Morrison [36], good p values range from 0.4 to 0.6. Whereas, Hanna and Dettmer [37] perceive good p values to do well between 0.3 and 0.6. The (2) was used to obtain the value of difficulty index (p).

$$p\frac{Ru+R1}{Nu+N1} \tag{2}$$

Where: p=Difficulty index; Ru=The number of students in the upper group who correctly respond; Rl=The number of students in the lower group who correctly respond; Nu=The total number of upper group students; Nl=The total number of lower group students.

Henning's suggestion was referred to as guidelines in this study to analyse the difficulty indices. Henning [38] suggested the item is easy, moderately difficult, and difficult if the p value is ≥0.67, 0.34-0.66 and ≤0.33 respectively. The details are shown in Table 2.

Table 2. Difficulty level according to Henning's recommendation

| Difficulty index (p) | Difficulty level |
|---|---|
| ≥0.67 | Low (easy) |
| 0.34-0.66 | Medium |
| ≤0.33 | High (difficult) |

Moreover, it is a must in assessment that an item must have a good discriminant power. To differentiate or discriminate between the low scores and high scores on an entire test and be differentiated or discriminated if an item has a well discrimination power. The analysis of the discriminant index for each item is hence important. The D value was obtained by using (3):

$$D\frac{Nu+N1}{\frac{1}{2}N} \tag{3}$$

Where: D=Discriminant index; Nu=The total number of upper group students; Nl=The total number of lower group students; N=Total number of students in upper and lower groups.

The discriminant index was analyzed with reference to the suggestion of Ebel [39] as shown in Table 3. Later, the K-20 formula was furthered to obtain the reliability of SAT. Generally, there are two versions of the formula (KR-20 and KR-21) for psychological and achievement test items. KR-20 was chosen in the study since the SAT developed comprises a different level of difficulties. Another method, split half was also used to ensure the internal consistency reliability of SAT. In this method, the SAT items were divided into two equal halves, with odd items in one part and all even items in the other. The correlation of the two halves was lastly analysed.

Table 3. Interpretation of D values based on Ebel's parameters

| Discriminant index (D) | Recommendation |
|---|---|
| 0.40-1.0 | Very good items |
| 0.30-0.39 | Reasonably good, but possibly subject to improvement |
| 0.20-0.29 | Marginal items, usually needing and being subject to improvement |
| Below 0.19 | Poor items, to be rejected or improved by revision |

In addition, Rasch analysis was used to increase the reliability and validity of SAT. Separation diagnosis was analysed for both item and person to verify the item hierarchy, and it provided the indication of how well the items were separated by a sample of persons who were giving responses to the instrument [40] and classified or separated people according to their abilities [41]. High item separation and reliability which was more than 3.0 and 0.9 respectively indicated that the person sample was enough to confirm the item difficulty hierarchy [21]. The person separation and reliability with the value greater than 2.0 and 0.8 showed that the instrument was sensitive enough to distinguish between high and low performers [21].

Item polarity was analysed to test the extent to which the construction of constructs achieved its goal. Here, the researchers tested the PTMEA CORR of SAT. If the value contained in the PTMEA CORR was the positive (+), it showed the item measured the constructs [42]. If the value was negative (-), this showed that the item might be measuring the construct in the opposite direction. The follow-up action for the negative item was then either improving or removing as the item might not lead and focus to the question, or perhaps difficult to be answered.

Furthermore, the question of model fit was one of utmost importance [12] which included the average fit (measure square and standardised) of the items and persons, showed how well the item difficulty or student ability contributes to the underlying construct of the test [43]. Fit was expressed as 'infit' (inlier-pattern-sensitive fit statistics) and 'outfit' (outlier-sensitive fit statistics). According to Green and Frantom [44], infit is weighted by distance between the item difficulty and the person position, so it is more sensitive to the unexpected patterns of observations by persons on items that are roughly targeted on them, and vice versa. Meanwhile outfit is an unweighted measure, so it is more sensitive to the unexpected observations by persons on items (such as careless mistakes or lucky guesses) that are relatively very hard or very easy for them. The details of the fit statistics are shown in Table 4.

Table 4. Rasch analysis: Item misfit diagnosis

| | Misfit Statistic too low | Predicted fit statistic | Misfit statistic too high |
|---|---|---|---|
| Standardized fit statistic | Z<–2.0 | –2.0<Z<+2.0 | Z>+2.0 |
| MNSQ fit statistic | MNSQ<0.5 | 0.5<MNSQ<1.5 | MNSQ>1.5 |
| Variation | Less than modeled | Modeled | Larger than modeled |
| Misfit type | Overfit | Good fit | Underfit |
| Response pattern | Deterministic | Stochastic | Erratic |
| Item or person performances | Muted | Productive for measurement | Noisy |
| Outcomes | Too good to be true | Expected | Unexpected |
| Interpretation | Guttman | Rasch | Unpredictable |

## 3. RESULTS AND DISUCSSION

### 3.1. Item analysis

From on the data collected in the pilot test, researchers conducted item analysis based on CTT to explain the difficulty index (p) and discrimination index (D). This method was used to ensure that the selected item had met its requirements, the level of difficulty and reliability of the item was free of unnecessary information and irrelevant reflections [14], [45]. The difficulty index (p) of each item according to Henning was shown in the Table 5.

Table 5. Difficulty indices of SAT items

| Difficulty index | Difficulty level | Items | Total |
|---|---|---|---|
| ≥0.67 | Low (easy) | 2, 3, 10, 11, 13, 16, 21, 22, 32, 34, 39, 46 | 12 |
| 0.34-0.66 | Medium | 1, 4, 5, 6, 9, 12, 14, 15, 17, 18, 19, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 35, 36, 37, 38, 40, 42, 43, 44, 45, 47, 49, 50 | 33 |
| ≤0.33 | High (difficult) | 7, 8, 20, 41, 48 | 5 |
| | | Total | 50 |

From the analysis, item 2, 3, 10, 11, 13, 16, 21, 22, 32, 34, 39, and 46 showed the difficulty index ≥0.67, showing that the items were easy items or low difficulty level. More than 60% of the items showed the medium difficulty level. The 33 items were item 1, 4, 5, 6, 9, 12, 14, 15, 18, 19, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 35, 36, 37, 38, 40, 42, 43, 44, 45, 47, and 50. And there were 5 items which considered as difficult item with the difficulty index ≤0.33. The difficult items were 7, 8, 20, 41 and 48.

As claimed by Boopathiraj and Chellamani [46] moderate difficulty items are preferred compared to those easier and more difficult. Vincent and Lajium [47] on the other hand, state that good items have a p value between 0.3 and 0.8. From the analysis, item 8, 20, 41 and 48 showed the p value less than 0.3. This means the items were categorised as difficult items and should be revised or removed. The analysis also showed that only item 32 had the p value nearly to 0.8, with the value 0.79 and was considered as too easy.

Contrarily, total of 27 items showed a very good discriminant power with the discriminant index between 0.40 and 1.0. These items could distinguish the student ability well. The items mentioned were item 1, 3, 4, 5, 8, 9, 14, 15, 16, 17, 18, 19, 21, 26, 29, 32, 34, 35, 36, 37, 38, 42, 43, 44, 45, 46, and 49. The reasonably good items were item 2, 6, 11, 13, 20, 22, 27, 28, 30, 31, 39, 40, 47, and 50, with the value D 0.30-0.39. The details are shown in Table 6. The analysis revealed two marginal items (items 12 and 25) and seven poor items (items 7, 10, 23, 24, 33, 41, and 48). Ebel [39] stated that there was a need to reject or improve an item by revision if the item has the value of discrimant index less than 0.19. The summary of the difficulty index and discriminant index identified in the analysis were shown in Table 7.

Table 6. Difficulty indices of items of the SAT

| Discriminant index | Remarks | Items | Total |
|---|---|---|---|
| 0.40-1.0 | Very good | 1, 3, 4, 5, 8, 9, 14, 15, 16, 17, 18, 19, 21, 26, 29, 32, 34, 35, 36, 37, 38, 42, 43, 44, 45, 46, 49 | 27 |
| 0.30-0.39 | Reasonably good | 2, 6, 11, 13, 20, 22, 27, 28, 30, 31, 39, 40, 47, 50 | 14 |
| 0.20-0.29 | Marginal items | 12, 25 | 2 |
| <0.19 | Poor items | 7, 10, 23, 24, 33, 41, 48 | 7 |
| | | Total | 50 |

Table 7. Summary of the difficulty index (p) and discriminant index (D)

| Item | Difficulty index (p) | Discrimination index (D) | Suggestion | Item | Difficulty index (p) | Discrimination index (D) | Suggestion |
|---|---|---|---|---|---|---|---|
| 1 | 0.57 | 0.43 | Retained | 26 | 0.57 | 0.43 | Retained |
| 2 | 0.68 | 0.36 | Retained | 27 | 0.39 | 0.36 | Retained |
| 3 | 0.71 | 0.57 | Retained | 28 | 0.61 | 0.36 | Retained |
| 4 | 0.54 | 0.79 | Retained | 29 | 0.57 | 0.71 | Retained |
| 5 | 0.50 | 0.93 | Retained | 30 | 0.39 | 0.36 | Retained |
| 6 | 0.50 | 0.36 | Retained | 31 | 0.46 | 0.36 | Retained |
| 7 | 0.32* | 0.07* | Modified/removed | 32 | 0.79 | 0.43 | Retained |
| 8 | 0.29* | 0.43 | Modified/removed | 33 | 0.57 | 0.14* | Modified/removed |
| 9 | 0.61 | 0.50 | Retained | 34 | 0.75 | 0.50 | Retained |
| 10 | 0.68 | 0.07* | Modified/removed | 35 | 0.61 | 0.50 | Retained |
| 11 | 0.75 | 0.36 | Retained | 36 | 0.46 | 0.50 | Retained |
| 12 | 0.46 | 0.21 | Retained | 37 | 0.50 | 0.43 | Retained |
| 13 | 0.68 | 0.36 | Retained | 38 | 0.64 | 0.43 | Retained |
| 14 | 0.61 | 0.79 | Retained | 39 | 0.75 | 0.36 | Retained |
| 15 | 0.43 | 0.71 | Retained | 40 | 0.46 | 0.36 | Retained |
| 16 | 0.68 | 0.50 | Retained | 41 | 0.29* | 0.00* | Modified/removed |
| 17 | 0.36 | 0.43 | Retained | 42 | 0.57 | 0.71 | Retained |
| 18 | 0.39 | 0.64 | Retained | 43 | 0.50 | 0.57 | Retained |
| 19 | 0.61 | 0.79 | Retained | 44 | 0.46 | 0.50 | Retained |
| 20 | 0.25* | 0.36 | Modified/removed | 45 | 0.64 | 0.57 | Retained |
| 21 | 0.71 | 0.57 | Retained | 46 | 0.68 | 0.50 | Retained |
| 22 | 0.68 | 0.36 | Retained | 47 | 0.46 | 0.36 | Retained |
| 23 | 0.46 | 0.07* | Modified/removed | 48 | 0.25* | 0.07* | Modified/removed |
| 24 | 0.50 | 0.14* | Modified/removed | 49 | 0.36 | 0.43 | Retained |
| 25 | 0.54 | 0.21 | Retained | 50 | 0.38 | 0.36 | Retained |

### 3.2. Reliability

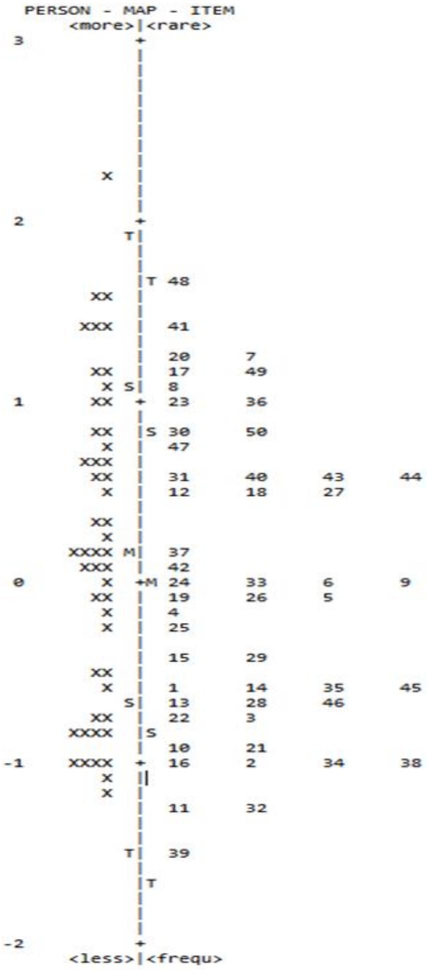
With the aim to ensure the reliability of SAT, researchers analysed KR-20 to evaluate the performance of the test. The KR-20 reliability coefficient of 0.70 and above was considered a reliable score [48]. The KR-20 coefficient from the pilot test was found to be 0.862. This value was found to be reliable, indicating that SAT is a reliable tool. Besides, the scores of two halves were correlated, using the split half reliability method. The equal-length Spearman-Brown correlation coefficient from the reliability analysis was 0.851. Once again, this proved the SAT was reliable and a valid instrument.

### 3.3. Rasch analysis

Thereafter the item analysis, Rasch analysis is also conducted using Winstep v3.73.3. From the analysis of separation diagnosis, the item separation for SAT is 2.25 with the realiability of 0.83. In term of person separation, the value index is 2.35 while the reliability is 0.85. The reliability of both item and person indicated a good and acceptable value according to Bond and Fox [42]. Even though the value of item separation was less than 3.0, Arasinah *et al.* [49] stated that the value of person and item separation which exceeds the value of 2.0, range from 2.0 to 3.0, and more than 5.0 is considered good, moderately good, and excellent respectively. Hence, the separation of SAT was good and reliable tool where the items could separate a sample of persons who were giving responses [40]. The summary of separation diagnosis and Wright map are as presented in Table 8 and Figure 1.

Table 8. Reliability and separation item and person

| | Item | Person |
|---|---|---|
| Separation | 2.25 | 2.35 |
| Reliability | 0.83 | 0.85 |



```
        PERSON - MAP - ITEM
            <more>|<rare>
 3                +
                  |
                  |
                  |
                  |
                  |
                  |
                  |
                  |
             X    |
                  |
                  |
 2                +
                 T|
                  |
                  |
                  |T 48
            XX    |
           XXX    |   41
                  |
                  |   20      7
            XX    |   17      49
             X   S|   8
 1          XX    +   23      36
                  |
            XX    |S  30      50
             X    |   47
           XXX    |
            XX    |   31      40      43      44
             X    |   12      18      27
                  |
            XX    |
             X    |
          XXXX   M|   37
           XXX    |   42
 0           X    +M  24      33      6       9
            XX    |   19      26      5
             X    |   4
             X    |   25
                  |
                  |   15      29
            XX    |
             X    |   1       14      35      45
                 S|   13      28      46
            XX    |   22      3
          XXXX    |S
                  |   10      21
-1        XXXX    +   16      2       34      38
             X    ||
             X    |
                  |   11      32
                  |
                  |
                 T|   39
                  |
                  |T
                  |
                  |
                  |
-2                +
            <less>|<frequ>
```

Figure 1. The Wright map

The Wright map in Figure 1 shows the level of person or student ability (left side) and the items difficulty (right side). On the right side, the items at the top of the scale were those more difficult and become easier to perform when further down the scale. While the person measure refers to the ability of each person to perform with the items. In short, Figure 1 was a mapping of the item difficulty distribution with the item answering ability distribution along the same continuum. It was indicated that item 48 appeared as the most difficult item since it was located at the very top of the scale. Vice versa, item 39 which was located at the very bottom showed that it was the easiest item.

To test the construct validity, item polarity and fit statistic were analysed. From Table 9, it was found that all PTMEA CORR values were positive values which ranged from 0.08 to 0.65, except for item 48. Item 48 had a negative value, which was -0.01. A positive PTMEA CORR value indicated the item was moving in one direction and measures the idea to be measured [41]. This means item 48 might test constructs in opposite directions. Therefore, item 48 should be taken into consideration either be refreshed or removed.

Summary of fit statistics was displayed in Table 10. The findings showed that the infit and outfit MNSQ values for all items were in the range recommended by Linacre [50] of 0.5 to 1.5, except for item 48 which had an outfit MNSQ value of 1.6. Next, the statistical values of the standard match Z-standard (Z-STD) infit and outfit were also analyzed. The accepted Z-STD values were in the range between -2.0 to +2.0. All items had an accepted Z-STD value except item 5 had an infit and out fit value of -2.4. Thus, item 5 was considered ‘muted’ where it was overfit and indicated the existence of an error trend [50].

Table 9. Polarity item of SAT

| Item | Measurement score | PTMEA CORR | Item | Measurement score | PTMEA CORR |
|---|---|---|---|---|---|
| 48 | 1.65 | -0.01 | 49 | 1.17 | 0.39 |
| 7 | 1.28 | 0.08 | 26 | -0.09 | 0.37 |
| 41 | 1.40 | 0.12 | 45 | -0.58 | 0.35 |
| 23 | 0.96 | 0.09 | 1 | -0.58 | 0.35 |
| 10 | -0.90 | 0.08 | 17 | 1.17 | 0.39 |
| 40 | 0.57 | 0.17 | 46 | -0.68 | 0.41 |
| 33 | 0.00 | 0.17 | 42 | 0.10 | 0.48 |
| 20 | 1.28 | 0.16 | 15 | -0.38 | 0.49 |
| 12 | 0.47 | 0.21 | 11 | -1.26 | 0.41 |
| 24 | 0.00 | 0.19 | 39 | -1.54 | 0.39 |
| 50 | 0.86 | 0.24 | 35 | -0.58 | 0.47 |
| 25 | -0.28 | 0.25 | 44 | 0.57 | 0.49 |
| 27 | 0.47 | 0.27 | 32 | -1.26 | 0.46 |
| 30 | 0.86 | 0.29 | 18 | 0.47 | 0.50 |
| 6 | 0.00 | 0.30 | 38 | -1.02 | 0.45 |
| 37 | 0.19 | 0.29 | 16 | -1.02 | 0.47 |
| 47 | 0.76 | 0.32 | 21 | -0.90 | 0.50 |
| 9 | 0.00 | 0.32 | 34 | -1.02 | 0.50 |
| 2 | -1.02 | 0.36 | 43 | 0.57 | 0.54 |
| 8 | 1.07 | 0.32 | 3 | -0.79 | 0.55 |
| 22 | -0.79 | 0.30 | 29 | -0.38 | 0.56 |
| 31 | 0.57 | 0.35 | 14 | -0.58 | 0.56 |
| 28 | -0.68 | 0.37 | 19 | -0.09 | 0.58 |
| 36 | 0.96 | 0.40 | 4 | -0.19 | 0.59 |
| 13 | -0.68 | 0.35 | 5 | -0.09 | 0.65 |

Table 10. Analysis of item misfit

| Item | Measurement | Infit | | Outfit | | Item | Measurement | Infit | | Outfit | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | score | MNSQ | ZSTD | MNSQ | ZSTD | | score | MNSQ | ZSTD | MNSQ | ZSTD |
| 48 | 1.65 | 1.26 | 1.3 | 1.60 | 1.8 | 49 | 1.17 | .95 | -0.3 | 1.01 | 0.10 |
| 7 | 1.28 | 1.19 | 1.2 | 1.46 | 1.8 | 26 | -0.09 | 1.00 | 0.0 | 1.01 | 0.10 |
| 41 | 1.40 | 1.15 | 0.9 | -0.09 | 0.65 | 45 | -0.58 | 1.00 | 0.1 | 0.99 | 0.0 |
| 23 | 0.96 | 1.21 | 1.6 | 1.40 | 2.0 | 1 | -0.58 | 1.00 | 0.10 | 0.99 | 0.00 |
| 10 | -0.90 | 1.21 | 1.4 | 1.37 | 1.5 | 17 | 1.17 | .96 | -0.2 | 0.98 | 0.00 |
| 40 | 0.57 | 1.18 | 1.6 | 1.24 | 1.6 | 46 | -0.68 | .95 | -0.3 | 0.89 | -0.50 |
| 33 | 0.00 | 1.19 | 1.7 | 1.21 | 1.5 | 42 | 0.10 | .89 | -0.1 | 0.95 | -0.30 |
| 20 | 1.28 | 1.18 | 1.1 | 1.2 | 0.9 | 15 | -0.38 | .86 | -1.2 | 0.94 | -0.30 |
| 12 | 0.47 | 1.14 | 1.3 | 1.19 | 1.4 | 11 | -1.26 | .92 | -0.4 | 0.80 | -0.60 |
| 24 | 0.00 | 1.17 | 1.5 | 1.19 | 1.4 | 39 | -1.54 | .91 | -0.3 | 0.83 | -0.40 |
| 50 | 0.86 | 1.11 | 0.9 | 1.14 | 0.8 | 35 | -0.58 | .90 | -0.7 | 0.85 | -0.80 |
| 25 | -0.28 | 1.11 | 1.0 | 1.10 | 0.7 | 44 | 0.57 | .90 | -0.90 | 0.88 | -0.80 |
| 27 | 0.47 | 1.11 | 1.0 | 1.10 | 0.8 | 32 | -1.26 | .90 | -0.5 | 0.70 | 1.0 |
| 30 | 0.86 | 1.06 | 0.5 | 1.11 | 0.7 | 18 | 0.47 | .90 | -0.9 | 0.86 | -1.00 |
| 6 | 0.00 | 1.06 | 0.6 | 1.10 | 0.7 | 38 | -1.02 | .89 | -0.7 | 0.86 | -0.50 |
| 37 | 0.19 | 1.08 | 0.8 | 1.09 | 0.7 | 16 | -1.02 | .88 | -0.7 | 0.80 | -0.80 |
| 47 | 0.76 | 1.03 | 0.3 | 1.08 | 0.5 | 21 | -0.90 | .87 | -0.80 | 0.75 | -1.10 |
| 9 | 0.00 | 1.05 | 0.5 | 1.06 | 0.5 | 34 | -1.02 | .86 | -0.80 | 0.73 | -1.10 |
| 2 | -1.02 | .94 | -0.3 | 1.06 | 0.3 | 43 | 0.57 | .85 | -1.40 | 0.83 | 1.20 |
| 8 | 1.07 | 1.05 | 0.4 | 1.01 | 0.1 | 3 | -0.79 | .84 | -1.20 | 0.70 | -1.40 |
| 22 | -0.79 | 1.05 | 0.4 | 1.02 | 0.2 | 29 | -0.38 | .82 | -1.50 | 0.80 | -1.20 |
| 31 | 0.57 | 1.02 | 0.2 | 1.04 | 0.3 | 14 | -0.58 | .82 | -1.40 | 0.74 | 1.50 |
| 28 | -0.68 | .97 | -0.2 | 1.02 | 0.2 | 19 | -0.09 | .82 | -1.70 | 0.77 | -1.80 |
| 36 | 0.96 | .95 | -0.3 | 1.02 | 0.2 | 4 | -0.19 | .81 | -1.80 | 0.77 | 1.70 |
| 13 | -0.68 | 1.01 | 0.1 | 0.96 | -0.1 | 5 | -0.09 | .75 | -2.40 | 0.70 | -2.40 |

By comparing the findings of item analysis using CTT and Rasch analysis of IRT, researcher found that there were several similarities. Firstly, item 48 from the Rasch analysis presented as the most difficult item. This was supported by the item analysis based on CTT, where item 48 also showed the p value of 0.25 which was one of the most difficult items. Besides, item 7, 8, 20, and 41 were also categorised as difficult items in the Wright map. All these difficult items showed the low p value, which indicated their level of difficulty. From the Wright map too, item 39 appeared as the easiest item and this was supported by the difficulty index of 0.75. Although items 11, 21, 32 and 34 also showed same difficulty index as item 39, they were placed in the range of easy item in the map too. Hence, findings of both analyses should be considered.

In contrast, Rasch analysis provided an extra information about the dimensionality and item misfit. For example, item 48 measured the negative direction and as 'noisy' item, which had the outfit MNSQ value. The outcome of this item was unexpected–likely due to poor item or required special knowledge and guessing. In short, there were similarities from the analysis using the Rasch model with the item analysis based on CTT. For example, item 48 was found necessary to be improved or dropped. Thus, the result of item statistics obtained from the two measuring frameworks were relatively close and similar.

Consideration of both CTT and Rasch analysis had provided a more comprehensive view about the test items. The implementation of both methods had increased the reliability and validity of an evaluation test. It might also increase the confidence of test makers or researchers before the items were gathered into the item bank and test widely. As a result of the reliability and validity of SAT, a total of 10 items were needed to be modified, namely items 5, 7, 8, 10, 20, 23, 24, 33, 41, and 48.

## 4. CONCLUSION

The SAT was an achievement test designed to test students' cognitive outcomes in science. The validity and reliability of the developed SAT were analysed by employing item analysis using both CTT and Rasch analysis. Overall, there were five difficult items, 33 moderate items, and 12 easy items based on the difficulty indices. There were seven items, according to the discriminant indices, needed to be removed or refreshed. Besides, from the Rasch analysis, there were two items needed to be revised. Overall, the study showed that SAT was a valid and reliable assessment tool.

Anyway, this study was doing analysis based on the quantitative data. The data was obtained from the students through answering the SAT. Qualitative analysis could be employed as well by implementing interviews with students, teachers, and experts. The qualitative input might provide diverse perspectives from different angles on individual items, thus giving a better description of the quality of the items.

More to add, future studies are suggested to conduct distractor analysis in addition to difficulty and discriminant indices analysis. In distractor analysis, we knew how the distractors were able to function effectively by drawing the examinee away from the correct answer. This might indicate the skills needed to be strengthened in order to eradicate those misconceptions in examinees based on the distractors chosen. This analysis could also be very interesting, and it could aid in improving test quality.

In short, developing a real good quality of achievement test was a challenging process and time-consuming task. However, it was still doable provided with the skills of analysis for achievement test. Therefore, the skills in analysing the achievement test should be acquired by the test developers, especially the educators. The SAT had been developed through systematic steps and procedures. The uniqueness of the study, which used both CTT and Rasch analysis, could be a good reference for the future researchers, or educators in designing an achievement test. A proper method in analysing the test, by using CTT and Rasch analysis are suggested way to ensure and improve the quality, reliability, and validity of an achievement test. Thus, students can be evaluated more effectively by administrating a more appropriate, solid, and valid tool.

## REFERENCES

[1] Curricular Development Department, "Form 4 specific science curriculum." Ministry of Education Malaysia, Putrajaya, 2019.

[2] T. Matingwina, "Health, academic achievement and school-based interventions," in *Health and Academic Achievement*, London: InTech, 2018. doi: 10.5772/intechopen.76431.

[3] L. H. Haw and S. B. Sharif, "The development and validation of science achievement test," *Journal of Education and Practice*, vol. 11, no. 20, pp. 103–109, Jul. 2020, doi: 10.7176/JEP/11-20-12.

[4] Ministry of Education Malaysia, "National report TIMSS 2019 - trends in international mathematics and science study," *Perpustakaan Negara Malaysia Data*. Ministry of Education Malaysia, Putrajaya, 2020. [Online]. Available: https://www.moe.gov.my/muat-turun/penerbitan-dan-jurnal/rujukan-akademik/3918-buku-laporan-timss-2019.

[5] O. A. Awopeju and E. R. I. Afolabi, "Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination," *European Scientific Journal (ESJ)*, vol. 12, no. 28, p. 263, Oct. 2016, doi: 10.19044/esj.2016.v12n28p263.

[6] A. B. Abd Latif, I. J. Yusof, N. F. Mohd Amin, W. H. Libunao, and S. S. Yusri, "Multiple-choice items analysis using classical test theory and rasch measurement model," *Man in India*, vol. 96, no. 1–2, pp. 173–181, 2016.

[7] S. N. Elliott, "Achievement tests," in *Reference Module in Neuroscience and Biobehavioral Psychology*, Amsterdam: Elsevier, 2017, pp. 41–46. doi: 10.1016/B978-0-12-809324-5.05457-2.

[8] R. K. Hambleton and R. W. Jones, "Comparison of classical test theory and item response theory and their applications to test development," *Educational Measurement Issues and Practice*, vol. 12, no. 3, pp. 38–47, 1993.

[9] T. Kline, *Psychological testing: A practical approach to design and evaluation*. California: SAGE Publications, Inc., 2005. doi: 10.4135/9781483385693.

[10] A. A. Bichi, R. Talib, R. Embong, H. B. Mohamed, M. S. Ismail, and A. Ibrahim, "Rasch-based objective standard setting for university placement test," *Eurasian Journal of Educational Research*, vol. 19, no. 84, pp. 1–14, Dec. 2019, doi: 10.14689/ejer.2019.84.3.

[11] C. Magno, "Demonstrating the difference between classical test theory and item response theory using derived test data," *The International Journal of Educational and Psychological Assessment*, vol. 1, no. 1, pp. 1–11, 2009.

[12] G. Janssen, V. Meier, and J. Trace, "Classical test theory and item response theory: Two understandings of one high-stakes performance exam," *Colombian Applied Linguistics Journal*, vol. 16, no. 2, pp. 167–184, Sep. 2014, doi: 10.14483/udistrital.jour.calj.2014.2.a03.

[13] S. Sabri, “Item analysis of student comprehensive test for research in teaching beginner string ensemble using model based teaching among music students in public universities,” *International Journal of Education and Research*, vol. 1, no. 12, pp. 1–14, 2013.

[14] L. Cohen, L. Manion, and K. Morisson, *Research methods in* education, 7th Ed. New York: Routledge, 2011.

[15] K. Moidunny, “The effectiveness of the national leadership professional eligibility program (NPQH).” Universiti Kebangsaan Malaysia, Bangi, 2003.

[16] B. R. Hopkins, K. D. Stanley, and J. C. Hopkins, *Educational and psychological measurement and evaluation*, 7th ed. Massachusetts: Allyn and Bacon, 1990.

[17] L. F. Bachman, *Statistical analyses for language assessment*. New York: Cambridge University Press, 2004. doi: 10.1017/cbo9780511667350.

[18] M. Tavakol and R. Dennick, “Psychometric evaluation of a knowledge based examination using Rasch analysis: An illustrative guide: AMEE Guide No. 72,” *Medical Teacher*, vol. 35, no. 1, pp. e838–e848, Jan. 2013, doi: 10.3109/0142159X.2012.737488.

[19] D.-T. Le, “Applying item response theory modeling in educational research,” Ph.D. Dissertation, Iowa State University, United States, 2013.

[20] E. W. Wolfe and E. V. Smith, “Instrument development tools and activities for measure validation using rasch models: Part I-instrument development tools,” *Journal of Applied Measurement*, vol. 8, no. 1, pp. 97–123, 2007.

[21] J. M. Linacre, “Is JMLE really that bad? No, it’s actually rather good!” *Rasch Measurement Transactions*, vol. 29, no. 4, pp. 1548–1549, 2016.

[22] S. K. S. Shanmugam, V. Wong, and M. Rajoo, “Examining the quality of english test items using psychometric and linguistic characteristics among grade six pupils,” *Malaysian Journal of Learning and Instruction*, vol. 17, no. 2, pp. 63–101, Jul. 2020, doi: 10.32890/mjli2020.17.2.3.

[23] H. L. Sharma and P. Gupta, “Construction and standardization of an achievement test in english grammar,” *International Journal of Advanced Educational Research*, vol. 2, no. 5, pp. 230–235, 2017.

[24] C. E. Brassil and B. A. Couch, “Multiple-true-false questions reveal more thoroughly the complexity of student thinking than multiple-choice questions: A Bayesian item response model comparison,” *International Journal of STEM Education*, vol. 6, no. 1, pp. 1–17, Dec. 2019, doi: 10.1186/s40594-019-0169-0.

[25] J. K. Hubbard, M. A. Potts, and B. A. Couch, “How question types reveal student thinking: An experimental comparison of multiple-true-false and free-response formats,” *CBE—Life Sciences Education*, vol. 16, no. 2, pp. 1–13, Jun. 2017, doi: 10.1187/cbe.16-12-0339.

[26] J. Myers, “Assessment and evaluation in social studies classrooms: A question of balance,” in *Challenges & Prospects in Canadian Social Studies*, Vancouver: Pacific Educational Press, 2004, pp. 290–301. [Online]. Available: http://www.learnalberta.ca/content/ssass/html/assessmentandevaluation/index.html

[27] M. Baig, S. K. Ali, S. Ali, and N. Huda, “Evaluation of multiple choice and short essay question items in basic medical sciences,” *Pakistan Journal of Medical Sciences*, vol. 30, no. 1, pp. 3–6, Dec. 1969, doi: 10.12669/pjms.301.4458.

[28] A. A. Abdel-Hameed, E. A. Al-Faris, I. A. Alorainy, and M. O. Al-Rukban, “The criteria and analysis of good multiple choice questions in a health professional setting,” *Saudi Medical Journal*, vol. 26, no. 10, pp. 1505–1510, 2005.

[29] M. C. Rodriguez, “Selected-response item development,” in *Handbook of test development*, New York: Routledge, 2015, pp. 259–273.

[30] S. M. Case and D. B. Swanson, *Constructing written test questions for the basic and clinical sciences*, vol. 27, no. 21. National Board of Medical Examiners, 2002. [Online]. Available: http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf.

[31] M. Tarrant and J. Ware, “A Framework for improving the quality of multiple-choice assessments,” *Nurse Educator*, vol. 37, no. 3, pp. 98–104, May 2012, doi: 10.1097/NNE.0b013e31825041d0.

[32] A. L. Whitehead, S. A. Julious, C. L. Cooper, and M. J. Campbell, “Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable,” *Statistical Methods in Medical Research*, vol. 25, no. 3, pp. 1057–1073, Jun. 2016, doi: 10.1177/0962280215588241.

[33] Curricular Development Department, “Form 4 specific science curriculum.” Ministry of Education Malaysia, Putrajaya, 2012.

[34] D. R. Krathwohl, B. S. Bloom, and B. B. Masia, *Taxonomy of educational objectives, the classification of educational goals, handbook II: Affective domain*, New York: David McKay Co., Inc., 1964.

[35] H. L. Sharma and S. Sansanwal, “Construction and standardization of an achievement test in science construction and standardization of an achievement test in science,” *International Journal of Research and Analytical Reviews*, vol. 5, no. 3, pp. 1037–1043, 2018.

[36] H. G. Macintosh and R. B. Morrison, *Objective testing*. London: Hodder & Stoughton Ltd, 1969.

[37] G. S. Hanna and P. Dettmer, *Assessment for affective teaching: Using context-adaptive planning*. London: Pearson, 2004.

[38] G. Henning, *A guide to language testing: Development evaluation research*. Cambridge: Newberry House Publishers, 1987.

[39] R. L. Ebel, *Essentials of educational measurement*, 3rd Ed. New Jersey: Prentice Hall, 1979.

[40] C. T. Sherron, “Psychometric development of the adaptive leadership competency profile,” Ph.D. Dissertation, University of North Texas, United States, 2000.

[41] T. G. Bond and C. M. Fox, *Appling the Rasch model: Fundamental measurement in the human sciences*, 3rd ed. New York: Routledge Taylor and Francis Publications, 2015.

[42] T. G. Bond and C. M. Fox, *Applying the rasch model: Fundamental measurement in the human sciences: Second edition*. New Jersey: Lawrence Erlbaum Inc. Publishers, 2007. doi: 10.4324/9781410614575.

[43] J. M. Linacre, “What do infit and outfit, mean-square and standardized mean,” *Rasch Measurement Transactions*, vol. 16, no. 2, p. 878, 2002.

[44] K. E. Green and C. G. Frantom, “Survey development and validation with the Rasch model,” in *International Conference on Questionnaire Development, Evaluation, and Testing*, 2002, pp. 14–17.

[45] R. L. Linn, *Educational measurement*, 3rd ed. Phoenix: Oryx Press, 1993.

[46] C. Boopathiraj and K. Chellamani, “Analysis of test items on difficulty level and discrimination index in the test for research in education,” *International Journal of Social Science & Interdisciplinary Research*, vol. 2, no. 2, pp. 189–193, 2013.

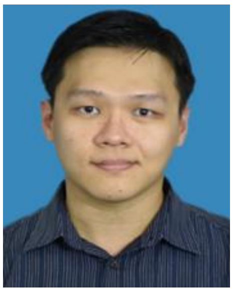
[47] V. Pang and D. Lajium, *Assessment in education*. Universiti Malaysia Sabah, 2AD.

[48] J. R. Fraenkel and N. E. Wallen, *How to design and evaluate research in education (6th ed.)*, no. 0. New York: McGraw-Hill, 2006.

[49] K. Arasinah, A. R. Bakar, H. Ramlah, A. Soaib, and H. Zaliza, “Using rasch model and confirmatory factor analysis to assess instrument for clothing fashion design competency,” *International Journal of Social Science and Humanity*, vol. 5, no. 5, pp. 418–421, 2015, doi: 10.7763/ijssh.2015.v5.492.

[50] J. M. Linacre, *A user’s guide to WINSTEPS MINISTEP: Rasch-model computer programs*. Chicago: Mesa Press, 2012.

## BIOGRAPHIES OF AUTHORS



**Law Hui Haw** iD 🎓 SC P received his PhD in Science Education from Universiti Malaysia Sabah (UMS). He is currently a lecturer at the Science Department of the Institute of Teacher Education Temenggong Ibrahim Campus, Johor Bahru, Johor, Malaysia. His research interests include science education and educational evaluation. He can be contacted at email: jamie_lawhh@yahoo.com.

**Sabariah Bte Sharif** iD 🎓 SC P obtained her PhD in Education from University of Bath, Somerset, United Kingdom. Her expertises are Educational Technology and Science Education. She is now the Dean of Faculty of Psychology and Education in University Malaysia Sabah (UMS). She can be contacted at email: sabariah@ums.edu.my.

**Crispina Gregory K. Han** iD 🎓 SC P earned her PhD from Universiti Malaysia Sabah (UMS). Her expertises are Science education and Evaluation in Education. She is now a senior lecturer in Universiti Malaysia Sabah (UMS). Currently, she has been appointed as the Head of Science with Education Programme and the coordinator for master and PhD Evaluation in Education Programme. She can be contacted at email: crispina@ums.edu.my.