The Development and Validation of Science Achievement Test

Law Hui Haw*
Faculty of Psychology and Education, University Malaysia Sabah
Jalan UMS, 88400, Kota Kinabalu, Sabah, Malaysia

Sabariah Bte Sharif
Faculty of Psychology and Education, University Malaysia Sabah
Jalan UMS, 88400, Kota Kinabalu, Sabah, Malaysia

The research is financed by Asian Development Bank. No. 2006-A171(Sponsoring information)

Abstract

Academic achievement is often regarded as a determinant of academic success. One of the most common assessment tools to evaluate the student achievement is through a well set achievement test. The main objective of the study was to develop and validate an achievement test in Science for senior secondary school students of grade 10. A Science Achievement Test (SAT) was developed by an excellent teacher in science subject based on Malaysia Science Curriculum Specification. The SAT consists of 50 multiple choice questions, which include the all 8 chapters in grade 10 science. The SAT had been validated by experts and analysed to check the difficulty index (p) and discrimination index (d), internal consistency reliability. Data were obtained from a purposive sampling of 50 students in a pilot study carried out in a secondary school in Sarawak, Malaysia. Based on the difficulty index, there are 12 easy items, 33 moderate items and 5 difficult items. 7 items found to be poor items and needed to be modified or removed due to the poor discriminating power. The reliability of SAT based on KR20 and Split half method showed the coefficient of 0.862 and 0.851 respectively. From the study, the SAT was a valid and reliable tool for measuring the students achievement in science.

Keywords: Science achievement test, validity, reliability, difficulty index, discriminant index
DOI: 10.7176/JEP/11-20-12
Publication date: July 31st 2020

Intorduction

Science is made up of huge number of fields such as Physics, Chemistry, Biology and lots more (Adeleke & Joshua, 2015). In Malaysia, science appears one of the most important subjects taught since elementary school. The goal of Malaysian Science education is to foster interest and develop student creativity through experience and investigations to acquire scientific knowledge and skills and technology as well as scientific attitudes and values (Curricular Development Department, 2018). Science education in Malaysia fosters a culture of Science and Technology by focusing on competitive individual development, dynamic, agile and resilient and capable of mastering knowledge science and technology skills. Secondary Science subjects are designed for develop a science-based, thinking-based student as well as being able to apply science knowledge, make decisions and solve problems in life real. Hence, the achievement of student in science subject needs attention and should be stressed.

According to Chawla (2016), academic achievement refers to the degree of success or achievement achieved in a particular field. In order to measure academic success, educators use different types of assessment. Assessment is a continuous process that brings some valuable information about the learning process (Linn & Gronlund, 1995). Thus far, the most commonly used approach for measuring student achievement in large-scale evaluations has been for evaluators to administer a common standardized test to students in the study (Somers, *et al.*, 2001). Sönmez and Alacapınar (2013) also stated that the tests are the often assessment tools that are used for determination of the students' gains relating to the cognitive domain within the quantitative researches of education. Through the test of achievement, an individual's mastery of a given knowledge or skill, for example science can be measured (Gay *et al*., 2009).

For the purpose of measuring student achievement in science, a set of quality science achievement test is required. In an achievement test the main emphasis is given on content coverage or course. The achievement test has the focus on the realization of objectives of teaching and learning (Sharma & Poonam, 2017). The tests used for assessing and evaluating the achievement of the student at all the stages include oral examinations, true-false tests, multiple-choice tests, matching tests, fill-in-the-blank exams, scales, short answer tests, written examinations, open ended questions, two phase testing (Kara & Celikler, 2015).

In this study, researcher wants to develop a science achievement test of which the validity and the reliability are ensured, which can be used to determine the achievements and of the senior secondary school students (ages 16-17) during the education process. Multiple choice questions (MCQs) are chosen as it appear as the most frequently used type of tests deployed on their own or in combination with other types of test tools for assessment (Baig *et al.*, 2014). Moreover, MCQs are appropriate for measuring knowledge, comprehension and could be designed to measure application and analysis (Abdel-Hameed *et al.,* 2005). MCQs are being used increasingly due to their higher reliability, validity, and ease of scoring (Case & Swanson, 2003; Tarrant & Ware, 2012).

According to Kamaruzaman (2003), item analysis needs to be done to determine whether a constructed item is good or weak. Good and weak items can be specified with a Difficulty Index (F) value. Meanwhile, discrimination index (D) is an index used to compare high performing and low performing students (Shafizan, 2013; Cohen et al., 2011; Kamaruzaman, 2003). According to Hopkins et al. (1990), if the mean value of the D value is high, then the test has high reliability.

Method

The Science Achievement Test (SAT) is a paper and pencil test designed to study student achievement levels. This test consists of 50 multiple choice questions. SAT was developed by an excellent teacher who was experienced in teaching science subject for more than 10 years. The content validity of SAT was later done by two experts in science subjects in terms of its contents and format. The SAT was drafted based on the Form Four Science Curriculum Specification (Curriculum Development Section, 2012) and the Test Specification Table (TST) according to the Bloom's Taxonomy (Bloom, 1956).

Content analysis is another very important phase in construction of an achievement test (Sharma & Sarita, 2018). The content of the form four science subject syllabus is based on five main themes, namely, 'Introduction to Science', 'Maintenance and Continuity of Life', 'Substance and Nature', 'Energy in Life' and 'Technology and Industrial Development in Society'. These five themes are further subdivided into eight topics namely 'Scientific Investigation', 'Body Coordination', 'Generation and Variation', 'Substance and Material', 'Energy and Chemical Change', 'Nuclear Energy', 'Light, Color and Vision 'and' Chemicals in the Industry '.The test was formulated based on the latest ‘Malaysian Certificate of Education’ format shown in Table 3.8. The distribution of UPS items according to the four science subject topics is as shown in Table 3.9.

Table 1: Format of Science Achievement Test (SAT)

No.	Information	Details
1	Type of Test	Multiple-choice test
2	Type of items	Multiple-choice items: Select Multiple Choice Answer Each item has four options A, B, C dan D.
3	Item number	50
4	Duration of test	1 hour 15 minutes
5	Difficulty levels	Low: Medium: High 5 : 3 : 2 (25 easy items: 15 medium items: 10 difficult items)

Table 2: Distribution of Science Achievement Test (SAT) items by Form 4 science subject

Topic	Question Order	Total items for each topic
Body coordination	1,2,3,4,5,6,7	7
Inheritence and Variation	8,9,10,11,12,13	6
Matter and Material	14,15,16,17,18,19,20,21,22,23	10
Chemical Energy and Changes	24,25,26,27,28,29,30,31,32,33	10
Nuclear Energy	34,35,36,37,38,39	6
Light, Colour and Sight	40,41,42,43,44,45,46	7
Chemical Substances in Industry	47,48,49,50	4
TOTAL		**50**

A pilot study for SAT was conducted with 50 Form 5 students at one of the national school in Limbang district. The duration of the test was one hour and fifteen minutes. The students result was used to determine its reliability and validity by analysing the item analysis, difficulty index, discriminant index and Kuder-Richardson 20 Formula.

There are differing opinions on the acceptability of difficulty index value. For example, Macinstosh and Morrison (1969) consider good F values to be between 0.4 to 0.6 while Hanna and Dettmer (2004) consider F values to be good in the range of 0.3 to 0.6. The value of difficulty index (F) can be calculated by the formula:

$$F=\frac{Ru+Rl}{Nu+Nl}$$

Note:
F = Difficulty index
Ru = the number of students in the upper group who respond correctly
Rl = the number of students in the lower group who respond correctly
Nu = the total number of students in the upper group
Nl = the total number of students in the lower group

In this study, the difficulty indices were analysed using the Henning (1987) guidelines as shown in the following table:

Table 3: Difficulty value based on Henning's suggestion

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

Besides, a good item should be able to separate or discriminate between high scores and low scores on an entire test. Thus, the discriminant index for each item was also calculated. The value of D can be obtained by using the formula:

$$D=\frac{Nu+Nl}{\frac{1}{2}N}$$

Note:
D = Discriminant index
Nu = the total number of students in the upper group
Nl = the total number of students in the lower group
N = total number of students in upper and lower groups

The discriminant indices were analysed by referring to Ebel's (1979) suggestion as follow:

Table 4: Ebel's Parameters for interpreting D values

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

From Ebel (1979, pg 267)

In addition, the reliability of SAT was also done by analysing Kuder- Richardson 20 Formula. The Kuder-Richardson Formula 20 (KR-20) first published by Kuder, G.F. & Richardson, M.W. in 1937 is a measure of internal consistency (reliability) for measures with dichotomous choices. Kuder Richardson formula has two versions (KR-20 and KR-21) for achievement and psychological test items respectively. In this study, KR-20 was used since the SAT comprises different level of difficulties. The formula for estimating reliability is given by:

$$\text{KR20}=\frac{n}{n-1}\left(\frac{SD^2-\sum PQ}{SD^2}\right)$$

Note:
K = Number of items
SD^2 = variance of scores on the test (square of the SD (standard deviation))
P = proportion of those who responded correctly
Q = proportion of those who responded incorrectly

Besides, the internal consistency reliability also tested by using Split half method. The usual method of Split half test is dividing the items into two equivalent halves is to take odd items in one half and all even items in the other half to calculate the reliability. And lastly find the correlation coefficient of the two halves.

Results and Discussion

Researcher have conducted item analysis to explain the Difficulty Index (F) and Discrimination Index (D) based on the results of the pilot test. This method is used to ensure that the selected item actually meets its requirements, the level of difficulty and reliability of the item is free of unnecessary information and irrelevant reflections (Cohen *et al.*, 2011; Linn, 1993).

The table below shows the difficulty index (F) of each item according to Henning (1987).

Table 5: Difficulty Indices of Items of the Science Achievement Test

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	1,4,5,6,9,12,14,15,17,18,19,22,23,24, 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, there were 12 items showed the difficulty index ≥ 0.67, which were item 2, 3, 10, 11, 13, 16, 21, 22, 28, 32, 34, 38, 39 and 46. These items were considered as easy item or low difficulty level. More than 50% of the items showed the medium difficulty level, that were 1, 4, 5, 6, 9, 12, 14, 15, 17, 18, 19, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 35, 36, 37, 38, 40, 42, 43, 44, 45, 47, 49 and 50. The total items in this range was 33 items. 5 items were categorized as difficult item with the difficulty index ≤ 0.33.The difficult items were 7, 8, 20, 41 and 48.

Generally, items of moderate difficulty are to be preferred to those which are much easier or much harder (Boopathiraj & Chellamani, 2013). However, Vincent and Lajium (2014) believe that good items show an F value of between 0.30 and 0.80. There was some items which had the F value less than 0.3, which were too difficult for students. Item 8, 20, 41 and 48 were too difficult and there was a need to modify or remove it. There was only a single item considered as too easy, which was item 32, with the difficulty index 0.79 nearly to 0.80.

On the other hand, there were 27 items showed the discriminant index between 0.40 and 1.0. Item 1,3,4,5,8,9,14,15,16,17,18,19,21,26, 29,32,34,35, 36,37, 38, 42,43,44,45,46,49 had a good discriminant power, which can distinguish the students ability well. Item 2,6,11,13,20,22 27,28,30,31, 39,40,47,50 were reasonably good items with the value D 0.30 - 0.39. From the analysis, there were two marginal items (item 12 and 25) and 7 poor items (item 7,10,23,24,33,41 and 48). According to Ebel (1979), the items with discrimant index less than 0.19, there was a need to reject or improve by revision. The summary of the difficulty index and discriminant index were shown in Table 7.

Table 6: Difficulty Indices of Items of the Science Achievement Test

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 and discriminant index

Item	Difficulty Index (F)	Discrimination Index (D)	Suggestion
1	0.57	0.43	Retained
2	0.68	0.36	Retained
3	0.71	0.57	Retained
4	0.54	0.79	Retained
5	0.50	0.93	Retained
6	0.50	0.36	Retained
7	0.32	0.07*	Modified/removed
8	0.29*	0.43	Modified/removed
9	0.61	0.50	Retained
10	0.68	0.07*	Modified/removed
11	0.75	0.36	Retained
12	0.46	0.21	Retained
13	0.68	0.36	Retained
14	0.61	0.79	Retained
15	0.43	0.71	Retained
16	0.68	0.50	Retained
17	0.36	0.43	Retained
18	0.39	0.64	Retained

Item	Difficulty Index (F)	Discrimination Index (D)	Suggestion
19	0.61	0.79	Retained
20	0.25*	0.36	Modified/removed
21	0.71	0.57	Retained
22	0.68	0.36	Retained
23	0.46	0.07*	Modified/removed
24	0.50	0.14*	Modified/removed
25	0.54	0.21	Retained
26	0.57	0.43	Retained
27	0.39	0.36	Retained
28	0.61	0.36	Retained
29	0.57	0.71	Retained
30	0.39	0.36	Retained
31	0.46	0.36	Retained
32	0.79	0.43	Retained
33	0.57	0.14*	Modified/removed
34	0.75	0.50	Retained
35	0.61	0.50	Retained
36	0.46	0.50	Retained
37	0.50	0.43	Retained
38	0.64	0.43	Retained
39	0.75	0.36	Retained
40	0.46	0.36	Retained
41	0.29*	0.00*	Modified/removed
42	0.57	0.71	Retained
43	0.50	0.57	Retained
44	0.46	0.50	Retained
45	0.64	0.57	Retained
46	0.68	0.50	Retained
47	0.46	0.36	Retained
48	0.25*	0.07*	Modified/removed
49	0.36	0.43	Retained
50	0.38	0.36	Retained

In order to determine the reliability of SAT, KR20 was employed was employed to evaluate the performance of the test as a whole. KR 20 coefficient was found to be 0.862. According to Fraenkel and Wallen (2008), one should attempt to generate a KR20 reliability coefficient of .70 and above to acquire reliable score. This value of KR20 presents as reliable thus revealing that this SAT is a reasonably reliable instrument. From the Split Half Reliability method, the scores of two halves were correlated. The Equal-Length Spearman-Brown correlation coefficient for reliability is 0.851. This shows again the SAT was reliable.

Conclusion

SAT is an achievement test developed to test the level and performance in science. SAT showed a high internal consistency reliability with the KR20 coefficient 0.862. Another test of reliability, the test of split half reliability coefficient was 0.851. Hence, SAT was a valid and reliable achievement test. According to the difficulty indices, there are 12 easy items, 33 moderate items and 5 difficult items. The analysis of discriminant indices showed that 7 items needed to be removed and modified. The good items may be stored in the questions bank for the future reference. The researchers, who aim to study the achievement of science subject, especially in Malaysia context, the SAT could be a good reference. The methods can be referred to those who wants to develope the achievement tests on either certain topics or different educational levels.

Other than analysing difficulty and discriminant indices, the future researchers are suggested to do the distractor analysis as well. In distractor analysis, we know how the distractors were able to function effectively by drawing the test takers away from the correct answer (Crocker & Algina, 1986). Distractors selected by students due to their misconcep-tions can inform the instructor about which skills need to improve in order to eliminate those misconceptions (Gierl *et al.,* 2017). It could be a very interesting analysis for the test developers and increase the test quality.

To conclude, it was a challenging and complicated process in consideration of developing a quality achievement test. Hence, the test developers, espescially the educators should acquire the skills in analysing the achievement test. A proper method in analysing the difficulty indices, discriminant indices as well as the distractor analysis, may improve the quality, validity and reliability of an achievement test. And thus, the students can be assessed more effectively, by using a more appropriate and valid tool.

References

Abdel-Hameed, A. A., Al-Faris, E. A., Alorainy, I. A. & Al-Rukban, M. O. (2005). The criteria and analysis of good multiple choice questions in a health professional setting. *Saudi Medical Journal. 26*. 1505-1510.

Adeleke1, A. A. & Joshua, E.O. (2015). Development and Validation of Scientific Literacy Achievement Test to Assess Senior Secondary School Students' Literacy Acquisition in Physics. *Journal of Education and Practice Vol. 6*, No.7. 28-42.

Baig, M., Ali, S.K., Ali, S. & Huda, N. (2014). Evaluation of Multiple Choice and Short Essay Question items in Basic Medical Sciences. Pak J Med Sci 2014;30(1):3-6.

Bloom, B. S. (1956). *Taxonomy of Educational Objectives, Handbook: The Cognitive Domain*. David McKay, New York.

Boopathiraj, C. & Chellamani, K. (2013).Analysis of test items on difficulty level and discrimination index in the test for research in education. *International Journal Of Social Science & Interdisciplinary Research*, ISSN 2277 3630, 2 (2), 189-193. Online available at indianresearchjournals.com

Case S &Swanson D. (2003). Constructing written test questions for the basic and clinical sciences. 3rd ed. Philadelphia: National Board of Medical Examiners, 2003.

Crocker, L., &Algina, J. (1986). Introduction to classical andmodern test theory. New York: Holt, Rinehart and Winston.

Curricular Development Department. (2018). *Form 4 Specific Science Curriculum.*Malaysian Education Ministry.

Chawla, J. (2016). Achievement in Chemistry od IX Grades in Relation to Study Habits. *International Education & Research Journal (IERJ),2(1)*,15-18.

Cohen, L., Manion, L. & Marrison, K. (2011). Research Methods in Education. 7th ed. Third Avenue, New York: Routledge.

Ebel, R. L. (1979). Essentials of educational measurement (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Fraenkel, J. R., & Wallen, N. E. (2006). *How to Design and Evaluate Research in Education (6th Ed.).* New York, NY McGraw-Hill.

Gay, L.R., Mills, G.E. and Airasian, P. (2009) Educational Research Competencies for Analysis and Applications. Pearson, Columbus.

Gierl, M., Bulut, O., Guo, Q. & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. Review of Educational Research. 87. 0034654317726529. 10.3102/0034654317726529.

Hanna, G.S. & Dettmer, P.A. (2004). Assessment for Affective Teaching: Using Context-adaptive Planning. Boston: Pearson.

Henning, G. (1987). A Guide to Language Testing: Development- Evaluation- Research. Rowley. Massachusetts: Newbury House. vii. + 198 pp.

Hopkins, K. D., Stanley, J.C. & Hopkins, B. R. (1990). *Educational and Psychological Measurement and Evaluation*. 7th ed. Massachusetts: Allyn and Bacon.

Kamaruzaman Moidunny. (2003). *Keberkesanan Program Kelayakan Profesional Kepengetuaan Kebangsaan (NPQH).* Tesis Doktor Falsafah. Universiti Kebangsaan Malaysia.

Kara, F. & Celikler, D. (2015). Development of Achievement Test: Validity and Reliability Study for Achievement Test on Matter Changing. Journal of Education and Practice. Vol.6, No.24, 2015

Linn, R. L. (1993). Educational Measurement. 3rd ed. Phonez, AZ: American Council on Education and the Oryx Press.

Linn, R.L. & Gronlund, N.E. (1995). Measurement and Evaluation in Teaching, (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Macintosh, H. G. & Morrisson, R. B. (1969). *Objective testing.* London: University of London. Press Ltd.

Shafizan Sabri. (2013). 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*, 1(12): 4-13.

Sharma, H.L.& Poonam (2017).Construction and Standardization of an achievement test in English Grammar. *International Journal of Advanced Educational Research, 2* (5), 230-235.

Sharma, H. L. & Sarita, S. (2018). Construction and standardization of an achievement test in Science.

International Journal of Research and Analytical Reviews, UGC Approved -43602. 05. 1037-1043.

Sönmez, V. & Alacapınar, F. G.. (2013). Örneklendirilmiş Bilimsel Araştırma Yöntemleri, Ankara: Anı Yayıncılık, 2nd edition, 384 p.

Somers, M-A., Zhu, P. & Wong, E. (2011). *Whether and How to Use State Tests to Measure Student*

Achievement in a Multi-State Randomized Experiment: An Empirical Assessment Based on Four Recent Evaluations. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences (IES). U. S. Department of Education.

Tarrant, M. & Ware, J. A. (2012). Framework for improving the quality of multiple-choice Assessments. Nurse Educator. 37 (3) : 98-104. doi: 10.1097/NNE.0b013e31825041d0

Vincent, P. & Lajium, D. A. (2014). *Penilaian dalam Pendidikan*. Kota Kinabalu, Sabah: Universiti Sabah.