Asian Social Science; Vol. 9, No. 16; 2013
ISSN 1911-2017 E-ISSN 1911-2025
Published by Canadian Center of Science and Education

Chemistry Test Items Development: Assessing Conceptual Understanding among Malaysian Students

Nur Suhaidah Sukor[1], Kamisah Osman[2] & Tuan Mastura Tuan Soh[2]

[1] Sandakan Middle School, Sandakan, Sabah, Malaysia

[2] Faculty of Education, Universiti Kebangsaan Malaysia, Selangor, Malaysia

Correspondence: Nur Suhaidah Sukor, Sandakan Middle School, P.O. Box 1458, 90716 Sandakan, Sabah, Malaysia. Tel: 60-16-828-3105. E-mail: suhaidah_sukor@yahoo.com.my

Received: August 16, 2013 Accepted: October 14, 2013 Online Published: November 28, 2013

doi:10.5539/ass.v9n16p126 URL: http://dx.doi.org/10.5539/ass.v9n16p126

Abstract

TIMSS has reported that, Malaysian students' achievement in science has exceeded the international average; however, it was still far behind its neighbouring country Singapore which is on the top three ranking. This paperwork aims to develop and validate instruments to assess the level of empowerment for chemistry-content knowledge. The instruments were adapted from previous researches on Conceptual Alternative in chemistry content knowledge. The test was administered to a group of 134 Form Five students who took chemistry subject in school. Development of a good research tool for Chemistry Content Test (CCT) needs analyses of the 31 items to find the reliability and validity values. This article discussed the steps in the item development process and a summary of how Rasch Modelling were applied to analyze the items. The implication of this research is that teachers can use the CCT as a guideline to develop questions that are needed for teaching and learning.

Keywords: assessment, validity, Rasch Modeling, chemistry

1. Introduction

Data collected from TIMSS reported that, Malaysian students scored 510 in a science subject, on average, which exceeded the international average of 474 (Mullis, Martin, Gonzalez & Chrostowski, 2004). This report also mentioned that, in comparison to other countries, Malaysia was outperformed by 19 of the 44 participating countries. The top three were Singapore, Chinese-Taipei and Republic of Korea. Thus, it is important to construct test items that must be fair and suitable in assessing all Malaysian students for those who came from high socioeconomic status as well as students with low socioeconomic status.

The instruments were adapted from previous researches on Conceptual Alternative in chemistry content knowledge. Multiple-choice items make use of common misconceptions as distracters, which allows researchers to simultaneously test for students' correct chemistry ideas and assessed their content knowledge. Thus, the development of a good research tool needs analyses of the 31 items to find the reliability and validity values. For the purpose of developing a research tool that provides an accurate assessment of the students' ability in Chemistry subjects, this paper applies Statistical Package for Social Science (SPSS) software version 18 and Rasch measurement model with Winsteps 3.69.0 software (Linacre, 2009). Rasch modelling is used to estimate and compare the items' difficulty and the popularity of each answer choice for students of differing ability (Linacre, 2007).

2. Purpose of Study

Mainly the purpose of this study is to develop a test that will assess chemistry knowledge. Specifically the objectives are: i) to gain an empirical evidence for reliability and validity of items in Chemistry Content Test (CCT), ii) to determine item difficulty separation values, and iii) to identify types of items that are difficult to answer.

3. Methodology

The preliminary version of CCT was pilot tested and administered to a group of 134 Form Five students who took the chemistry subject in school. Linacre (1994) explained that 30 examinees are enough for well-designed pilot studies. The test was administered at the beginning of the school year with the assumption that the level of

students' knowledge was still the same as form four students' level of knowledge at the end of the school year. These students have learned all the seven chapters of form four chemistry syllabuses. The students were given 35 minutes to complete the test. The data collected were used to investigate the tests' reliability and validity.

3.1 Design Instrument

The researcher, with the help of chemistry experts has developed a collection of objective test items that met the criterion of face validity. There were 31 multiple choice items constructed altogether for the Chemistry Content Test (CCT). Test items covered seven topics in form four chemistry syllabuses. The topics covered are: (i) Atomic Structure; (ii) Chemical Equation and Formula; (iii) Chemistry Periodic Table of Elements; (iv) Chemical Bonding; (v) Electrochemistry; (vi) Acids and Bases; and (vii) Salts.

In order to minimize errors, the developments of the test items were based on input of expert and teacher's opinion to serve the purpose of exhibiting the face validity. To make the test free of the confounding factors of reading ability, researcher provides a visual illustration such as diagrams, tables, and figures that serve to assist the clarification of ideas. In addition, items were referred to the Integrated Secondary School Curriculum for Chemistry Form Four Syllabus.

4. Data Collection

Since an appraisal of whether the data fit the model reasonably well is required. The item fit index was used to show how well the items function in the reflection of the traits. Point Measure Correlation (PTMEA CORR.), Outfit Mean Square values (MNSQ), and Standardized Fit Statistics (Z-std) were used in analyzing the CCT items for the purpose of determining whether items are goods and fit for the test. For fit statistics, Bond and Fox (2007) proposed that acceptable fit values fall between -2.0 and +2.0 with a sample size between 30 and 300. This pair of researchers also mentions that negative values indicate less variation than modelled: The response string is closer to Guttman-style response string in which all easy items correct then all difficult items incorrect. Positive values indicate more variations than modelled: The response string is more haphazard than expected.

It is known that separation thought of as the number of levels into which the items and persons can be separated. As Green and Frantom (2002) suggested that, for an instrument to be useful, the separation should exceed 1.0. Besides, these researchers also mention that the separation determines the reliability. Consequently, higher separation will yields to higher reliability. The reliability for the test was determined by the use of index coefficient of Cronbach Alpha.

From the same PTMEA CORR. Index values, the Discrimination Index values of items can be determined. Ong Saw Lan, Zurida, and Foong Soon Sok (2006) stressed that item value above 0.30 was considered to have satisfactory power of discrimination. Rasch analysis also has the mapping facility to allow us to see the distribution of each item in CCT together with the persons along a continuum. This graph intends to give information about person position along with the item position. The item index and the mapping facility were examined for the purpose of item revision.

5. Results

5.1 Reliability of CCT

In the analysis, the separation values are reported together with the reliability values. Items and person reliability are assessed based on person reliability and item reliability coefficient, to which are equivalent to Kuder-Richardson (KR-20).

Table 1. Summary statistics (Person Reliability)

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	8.8	31.0	-1.08	.44	1.00	.0	1.02	.0
	S.D.	2.9	.1	.54	.07	.13	.8	.24
	MAX.	18.0	31.0	.36	1.03	1.30	1.4	1.85
	MIN.	1.0	30.0	-3.65	.38	.72	-2.3	.61

REAL RMSE .45 ADJ.SD .30 SEPARATION .66 person RELIABILITY .30

MODEL RMSE .44 ADJ.SD .32 SEPARATION .72 person RELIABILITY .34

S.E. OF person MEAN = .05

person RAW SCORE-TO-MEASURE CORRELATION = 1.00

CRONBACH ALPHA (KR-20) person RAW SCORE RELIABILITY = .34

From the analysis in Table 1, it is found that the person reliability is low at 0.34 which is suggesting that similar ordering of person placement cannot be expected if this sample of people were given another set of item measuring the same construct CCT. This low reliability occurs due to the low separation value (0.66) which is supposed to be more than 1.0.

5.2 Item Fit

Attributes were checked on the Point Measure Correlation with acceptable parameters; PMC = x, $0.4 < x < 0.8$. In order to determine the item as 'problematic', Rasch requires further verification by looking at the Outfit column for Mean Square value, MNSQ = y, $0.5 < y < 1.5$ and Z-std value where Z-std = z, $-2 < z < +2$. The output table for item measure is exploited to determine misfitting item which is shown below.

Table 2. Item measures

INPUT: 134 persons, 31 items		MEASURED: 134 persons, 31 items,								
ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	ERROR	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PTMEA CORR.	items
10	24	134	.53	.23	1.09	.7	1.20	1.1	-.02	s10
11	13	134	1.26	.30	1.06	.3	1.23	.9	-.01	s11
16	23	134	.58	.23	1.07	.5	1.17	1.0	.03	s16
19	45	134	-.35	.19	1.11	1.5	1.13	1.5	.03	s19
22	22	134	.64	.24	1.05	.4	1.17	.9	.04	s22
31	34	134	.06	.20	1.07	.8	1.10	.8	.08	s31
23	10	134	1.55	.33	1.02	.2	1.06	.3	.08	s23
30	46	134	-.39	.19	1.08	1.2	1.10	1.2	.08	s30
27	20	134	.76	.25	1.04	.3	1.08	.4	.10	s27
15	16	134	1.02	.27	1.02	.2	1.10	.5	.10	s15
1	23	134	.58	.23	1.05	.4	1.04	.3	.10	s1
12	46	134	-.39	.19	1.07	1.1	1.08	.9	.10	s12
21	19	134	.82	.25	1.03	.2	1.01	.1	.12	s21
24	36	134	-.02	.20	1.04	.5	1.05	.5	.13	s24
3	63	134	-.95	.18	1.05	1.3	1.05	1.0	.15	s3
5	20	134	.76	.25	1.01	.1	1.00	.1	.16	s5
17	29	134	.28	.21	1.01	.1	1.05	.4	.17	s17
26	27	134	.38	.22	.99	.0	.98	-.1	.21	S26
18	38	134	-.09	.20	1.01	.1	1.00	.0	.21	s18
9	52	134	-.59	.18	.99	-.2	.99	-.1	.26	S9
7	38	134	-.09	.20	.97	-.3	.97	-.2	.27	S7
29	44	134	-.32	.19	.97	-.4	.95	-.5	.29	S29
20	48	134	-.46	.19	.97	-.5	.95	-.6	.30	S20
25	28	134	.33	.22	.95	-.4	.89	-.7	.31	S25
28	57	134	-.76	.18	.93	-1.5	1.03	.5	.34	S28
2	89	134	-1.80	.19	.92	-1.1	.89	-1.3	.39	S2
13	60	134	-.85	.18	.91	-2.0	.90	-1.8	.40	S13
14	55	134	-.69	.18	.91	-1.8	.90	-1.7	.40	S14
8	59	134	-.82	.18	.91	-2.0	.90	-1.9	.41	S8
4	45	134	-.35	.19	.88	-1.8	.84	-1.8	.45	S4
6	53	134	-.63	.18	.85	-3.0	.83	-2.7	.51	S6
MEAN	38.	134.	.00	.21	1.00	-.2	1.02	.0		
S.D.	18.	0.	.74	.04	.07	1.1	.10	1.0		

Based on the Table 2, analysis of the three attributes PMC, MNSQ and Z-std values show that none of the items were a misfit. The items are considered as a misfit only when all the three controls cannot be met. Hence, all items are acceptable for further analysis. The researcher continues on viewing the mapping facility in the Rasch model as a method of determining items distribution with the persons along a continuum. The map analysis is shown in Figure 1.

5.3 Item Person-Map

Rasch analysis also has the mapping facility to allow us to see the distribution of each item in CCT together with the persons along a continuum. According to Herrmann-Abell, DeBoer, and Roseman (2009) when item difficulty and person ability match, the person has a 50% chance of answering the item correctly. The item index and the mapping facility were examined for the purpose of item revision. Figure 1, below shows the item person-map for CCT.

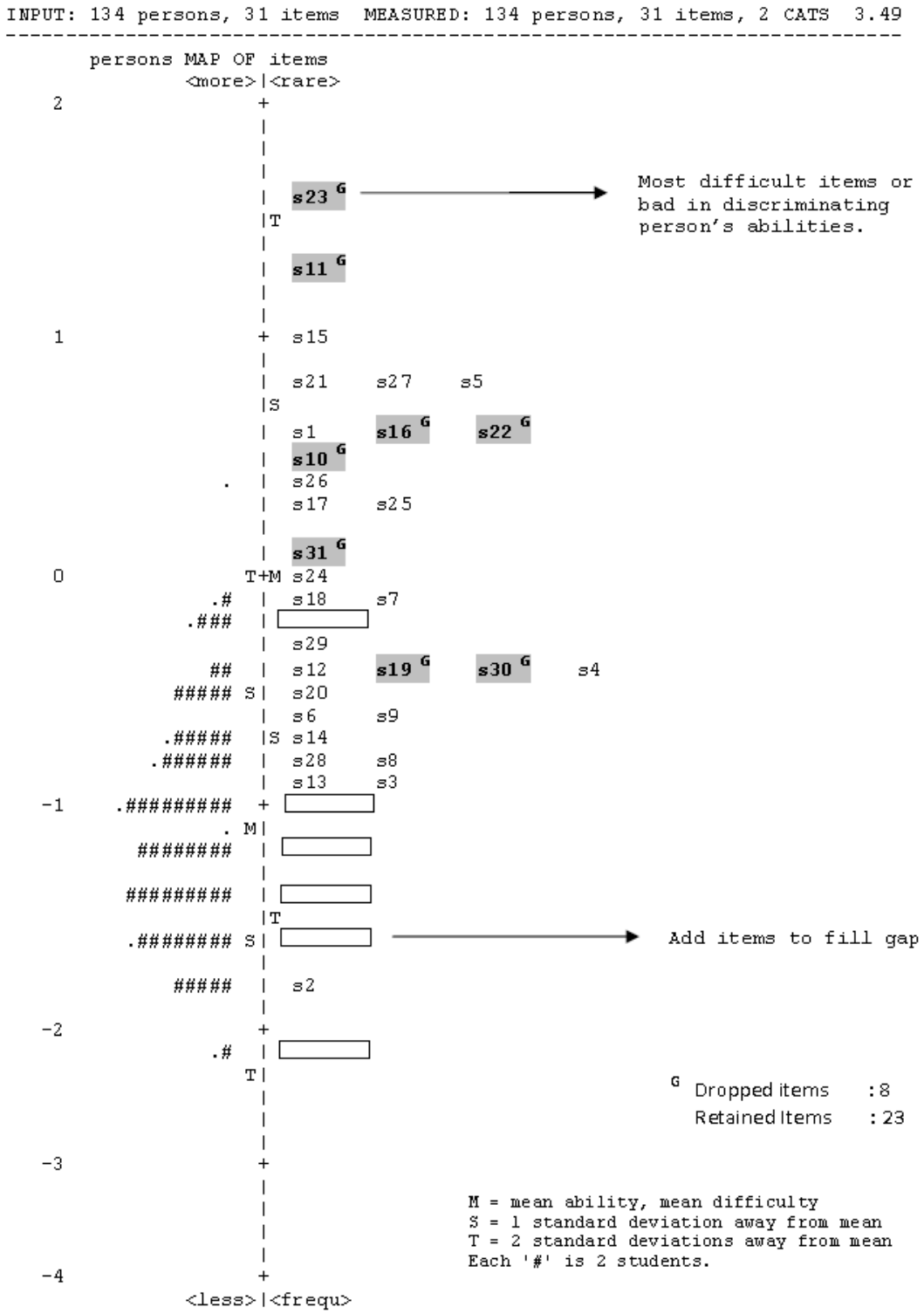

Figure 1. Item Person-Map analysis

The distribution of persons' positions is located on the left side of the vertical line and items on the right. From Figure 1, the items cover the range of -1.85 to +1.65 logits in difficulty, which is broader than the range of about -2.24 to 0.62 for persons. It also can be seen that the numerous item's position is beyond the person's capabilities. These indicate that items on the upper end of the graph are very difficult for students and as a result, the items have dropped from the item list. Meanwhile, lowest position in the graph is item s2. This item is the easiest questions, but still in persons capabilities to answer it correctly. At one point on the scale, there are 4 items at the same position. The researcher has to consider either dropping or revising one or two of them as redundant.

Although Table 2 showed that all item a fit with the model, based on the Discrimination Index, some of the items were identified as 'problematic' and needed to be revised or rejected. Before the action is taken in improving the instrument, map analysis in Figure 1 also contributes in giving information for improving quality of the items. Thus based on the data that have been collected from Table 2 and Figure 1, researcher categorized the items into four criteria for further action to be taken in developing the instrument. The categories are as follows: retain, rephrasing, simplified, and rejected.

Table 3. Categories for items and action to be taken

Item number	Total items	Action taken
18, 7, 29, 12, 4, 20, 6, 9, 14, 28, 8, 13, 3, 2	14	Retained
26, 17, 25, 24	4	Rephrased
15, 21, 27, 5, 1	5	Simplified
23, 11, 10, 16, 22, 31, 19, 30	8	Rejected and dropped from the item list

Table 3 shows the total of 31 items. A total of 14 retained items were items that have value within acceptable range, as discussed earlier. The table also shows that 4 items need to be rephrased to help students understand the questions in CCT. There are another 5 items which are located at the upper end of graph need to be simplified and matched with the students' capabilities. Lastly, 8 items that are located at the far upper end of the graph were taken out of the item list. This is because these items have negative value that really poor in discriminating students. Students with high achievement fail to answer correctly, whereas students with low achievement accidentally able to answer questions successfully. Some of the items were also taken out due to redundancy.

4. Conclusions and Implications

The objective of this paper is to develop an instrument to measure students' Chemistry Content Knowledge with referenced to the Malaysian Secondary Science Curriculum. Applying Rasch Modeling in test item development can be a powerful tool for evaluation, and refinement of items. Thus, it results in precise, valid, and relatively brief instruments that minimize response burden. The findings from the analysis showed that improvement is still needed for some of the test items to ensure that the instrument is reliable and useful. After some improvements, the test items will be distributed to sample research.

References

Bond, T. G., & Fox, C. M. (2007). *Applying The Rasch Model: Fundamental Measurement in the Human Sciences* (2nd ed.). Lawrence Erlbaum Associates. New Jersey, Landon.

Green, K. E., & Frantom, C. G. (2002). *Survey developments and validation with Rasch Model. International conference on questionnaire development, evaluation, and testing.* Charleston. Retrieved April 7, 2010, from http://www.jpsm.umd.edu.html

Herrmann-Abell, C. F., DeBoer, G. E., & Roseman, J. E. (2009). Using Rasch Modeling to Analyze Standards-Based Assessment Items Aligned to Middle School Chemistry Ideas. DR-K12 PI Meeting. AAAS Project 2061.

Lan, O. S., & Sok, F. S. (2006). *Development and Validation of Test for Integrated Science Processes.* Paper presented on 13-15 February at 3rd International Conference on Measurement and Evaluation in Education, Penang, Malaysia. Retrieved from http://www.eprints.usm.my

Linacre, J. (2007). *A User's Guide to WINSTEPS Rasch-Model Computer Programs.* Chicago: MESA Press.

Linacre, J. M. (1994). *Sample size and item calibration stability.* Rasch Measurement Transactions. Retrieved from http://www.rasch.org/memo

Linacre, J. M. (2009). *Winsteps® (Version 3.69.0) [Computer Software]*. Beaverton, Oregon: Winsteps.com. Retrieved August 1, 2009, from http://www.winsteps.com/

Mullis, V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 International Mathematics Report.* TIMSS & PIRLS International Study Centre, Lynch School of Education: Boston College.