Psychometric Properties of Chemistry Multiple-Choice Paper for Science Students: An Extensive Analysis By Rasch Model

Rose Enne Emellia Mohamed Razali[1*] ,Ahmad Adnan Mohd Shukri[2] ,Harris Shah Abd. Hamid[3]

[1]School of Inspectorate Kuala Lumpur, Tingkat 24, Blok Utama Menara Takaful Malaysia, No. 4 Jalan Sultan Sulaiman, 50000 Kuala Lumpur, Malaysia

[2]School of Educational Studies, Universiti Sains Malaysia, 11800 USM Pulau Pinang, Malaysia

[3]Department of Education and Psychology, Faculty of Management, Education and Humanities, University College MAIWP International, 50480 Kuala Lumpur, Malaysia

[*]Corresponding author

DOI: https://doi.org/10.47772/IJRISS.2026.10100588

Received: 30 January 2026; Accepted: 05 February 2026; Published: 19 February 2026

ABSTRACT

An achievement test is an essential element in the teaching and learning process as its primary purpose is to measure student performance. However, high-quality test items require extensive time and effort to be produced, particularly multiple-choice questions. Teachers are known as content experts and test developers however some of them may lack knowledge in test development. Therefore, some of the constructed items may be flawed, biased, or unreliable to measure the students' performance. The present study aims to determine the psychometric properties of newly developed multiple-choice questions for the Chemistry test paper using the Rasch Model. A study was conducted among 435 respondents from four randomly selected secondary schools in Klang Valley, Malaysia. Data were analyzed using a software called Winsteps. From the point of unidimensionality, Principal Component Analysis explained the test dimension of the instrument was moderate and acceptable, with 27.8% of raw variance measured. The reliability estimates for items were 0.99 while for person reliability is 0.87. multiple-choice paper. The aforementioned findings provide a dimension of validity of the test. In conclusion, the Rasch model has proven that the Chemistry test paper is a valid and reliable unidimensional instrument in measuring students' ability and item difficulty.

Keywords: Rasch model, chemistry education, unidimensionality, reliability estimates, separation index

INTRODUCTION

For decades, many students have found Chemistry as a mundane subject due to the abstract concepts and unfamiliar language setting (Langitasari et al., 2024). Apart from that, students tend to put off this subject when they cannot relate their learning to real-life phenomena. Abstract concepts in Chemistry are ideas generated by scientists' creative imaginations that are constrained by observations of natural phenomena, for instance, atoms, molecules, and electron transitions, while the unfamiliar language setting typically refers to chemical equations, scientific terms, and symbols. In a world of science and technology, it is crucial to understand the fundamental chemical and scientific ideas because they significantly impact society, especially on the quality of life. This understanding is essential for students to address the problems and issues of their daily lives. Therefore, teachers play a crucial role in assessing student performance to ensure their understanding reaches the targeted outcomes by conducting tests and examinations.

Many researchers are intrigued with the chemistry achievement test due to the predominant question in chemistry is how to trigger students into what will be a new way of seeing and thinking (Jegstad, 2023). An achievement test is one of the measuring tools for teachers and students to analyze the success rate of the learning process in the specific content areas (Shukri et al., 2020). The analysis is indispensable to strengthening the process of teaching and learning.

Since a comprehensive assessment is required to assess student performance, all types of evaluation need to meet the essential requirements of validity, reliability, and usability. An ideal measurement can be used for various purposes and is accurate, particularly by using the test scores. It has a stable frame of reference for comparing various students and offers a linear measure that can give significance to scores and detect misfits. Hence, a valid instrument is crucial in measurement as it can provide reliable data for meaningful analysis and generate useful information, notably used in decision-making. Despite various measurement models, the Rasch model has been proved as a suitable approach for examining and validating the educational instrument (Darmana et al., 2021)

Measurement experts have used the Rasch model in their studies to analyze the instrument's psychometric properties due to its objectivity and comprehensive analysis output. For instance, a study conducted by Suryadi et al. (2025) applied an open-ended response format Scientific Inquiry Competence (SIC) instrument to demonstrate the Rasch approach's strength in conducting psychometric analysis as an instrument. The authors stated that the raw data is not assumed linear in Rasch computation compared to traditional analysis. Therefore, the data wouldn't be flawed. The Rasch model works perfectly well to validate the multiple-choice questions because it rejects the concept of raw score and on the other hand, provides person and item estimates by placing them on the same interval scale as well as provide validity evidence for the test (Darmana et al., 2021). Yet, the Rasch model is able to illuminate the weaknesses of the problematic items in details through the statistical analysis such as item fit and distractor analysis (Bakytbekovich et al., 2023). In the nutshell, Rasch analysis assists researchers in improving the quality of the instrument by allowing them to optimize the instrument.

Assessment is among the important factors in education discipline because it's able to measure students' performance. Nonetheless, teachers have insufficient or absence of knowledge on test development, substantially item analysis. It is vital to validate the test items to ensure the test paper is able to assess the students' performance accurately. In this study, the instrument used to measure the students' performance in Chemistry was a standardized test paper constructed by a researcher and content-validated by few content-expertise as well as sanctioned by the District Education Department. The advantage of using this test paper is it has been constructed according to the table of specification (TOS) and national examination format developed by Malaysia Examination Syndicate (MES). The constructed test paper also has been through the test development process. The novelty of this study is item validation of multiple-choice questions in standardized Chemistry test using Rasch model. The study's contribution is to assist teachers in improving their instructions by designing suitable and effective instruction according to the students' ability and increase their level of understanding in learning Chemistry. Furthermore, teachers are also able to enhance the existing assessment practice in school through laying out the importance of item analysis. This study aims to determine the psychometric properties of the Chemistry multiple-choice test paper. As a result, this study addresses the questions: (a) To what extent does the data set of Chemistry test fit the Rasch model? (b) What are the student reliability and item reliability of the Chemistry test paper?; (c) What are the item validity of the Chemistry test paper?; (d) What are the appropriateness between item difficulty and students' ability?

METHOD

The present study was conducted using a quantitative research design because this design focuses on describing and explaining (Thomas & Zubkov, 2023). Thus, it is very suitable for the current situation as this study aims to establish the relationship between students' ability and item difficulties quantitatively. The sample of this study was 435 students that were retrieved from four randomly selected secondary schools in Klang Valley, consisting of 245 (56.32%) male students and 190 (43.68%) female students. All of them were Form Four Pure Science students at the age of 16 and requisite to learn Chemistry, one of the elective Science subjects offered in Science Stream. The sample size was adequate to establish 99% confidence that the estimated item difficulty is within a definitive of its stable value. The expected sample size as small as 30 respondents would be sufficient for a Rasch model with dichotomous items, in terms of item difficulty calibration to be within one logit of a stable value with 95% confidence (Tesio, et al., 2024). Hence, 435 respondents are ample for a stable item (50 items) and person measures. The Rasch Model is a psychometric technique designed to improve the accuracy of a designed instrument, track the consistency of an instrument, and measure respondents' performance (Boone, 2016). The item complexity and person capability in the Rasch model are calculated in a logit scale (Runnels, 2012). Despite its sophisticated mechanism, the Rasch model

was used to analyze the data by utilizing the student's raw test scores to compute their performance on a linear scale.

The Rasch mathematical equation used for dichotomous data is as the following:

$B_n - D_i = \text{In} (P_{ni} / 1 - P_{ni})$

B_n is the student's ability along the variable; D_i is the difficulty of a test item; P_{ni} is the probability of the student answering a test item correctly; and $1 - P_{ni}$ is the probability of a student answering a test item wrongly.

A computer software called WINSTEPS 4.5.5; a Rasch-based item analysis program, was used due to the simplicity of data handling, flexibility, and exhaustive detailed understandable documentation. This software was able to scrutinize whether test items fulfilled the basic assumptions of the Rasch model.

The Chemistry test paper was developed by researcher in collaborating with several experience Chemistry teachers and District Education Department. The items were based on a matrix known as Table of Specification (TOS) of Chemistry Test as shown in Table 1.

Table 1. Table of specification (TOS) based on the number of ítems according to cognitive level for each topic in chemistry test

Learning Area	Knowledge		Comprehension		Application
A. Introduction Chemistry					
A1. Introduction to Chemistry					
B. Matter Around Us					
B1. The Structure of the Atom	4L, 6L, 34L		16M, 43M, 47M		21L
B2. Chemical Formulae and Equations	2L, 7L		19L, 37H, 40M, 46M		29H, 35M, 36H, 50H
B3. Periodic Table of Elements	32L, 39M		8L, 12M, 30M		38H
B4. Chemical Bonds			9L, 22M		17H, 22M
C. Interaction Between Chemicals					
C1. Electrochemistry	1L, 3L		11L, 25H		14H, 42H
C2. Acids and Bases	5L, 10L, 31L		15M, 26M, 44M, 48M		41H
C3. Salts	13L		27M, 33H, 45M		20H, 24H
D. Production and Management of Manufactured Chemicals					
D1. Manufactured Substances in Industry			18L, 23L		
D2. Chemicals for Consumers			28M		
Total	20		16		14
Answers	A	B	C	D	
Total Number	9	15	13	13	
Weightage	25L	15M	10H		
Total	50				

L : Lower level item

M : Moderate level item

H : High level item

Item 1 – 20 is to analyze knowledge construct according to topic and level

Item 21-35 is to analyze comprehension construct according to topic and level

Item 36-50 is to analyze application construct according to topic and level

TOS is widely used for content validation because it helps teachers frame the decision-making process of test creation and strengthen the validity of teacher assessments based on tests constructed (Danushka & Gamage,

2024). Typically, TOS provides a precise and relevant outline to the teachers or test developers in writing test items. It describes the topics to be measured and the number of items that are associated with each topic. In the present study, the TOS prepared was equivalent to the actual Malaysian Certificate of Education standard of Chemistry multiple-choice test paper format. There are 50 multiple-choice questions with different degrees of difficulty to be answered within one hour and fifteen minutes. These questions are used to assess students' knowledge, skills and abilities in a specific content area or subject area after they have received instruction over a set period of time (Illene et al., 2023). The difficulty level of each item was determined by the professional judgement since the panels are content experts. All the items developed need to be pretested and have empirical evidence that meets the standards before setting an instrument. The test scores obtained are used as a piece of explicit evidence to infer that the student has learned and vice versa.

RESULTS AND DISCUSSION

Unidimensionality requires the measurement to aim only one latent trait at a time and it can be distinguished through observed raw variance. Ahmad and Siew (2021) recommended that an observed raw variance measure should be able to explain more than 20% of the variance to substantiate the unidimensionality assumption. Rosli et al. (2020) contended that the raw variance explained by measure of 20% or more is acceptable. Table 2 shows that the observed raw variance measures were 27.7%, exceeding 20% of the variance for all items This data shows the Rasch dimension only explains 27.7% of the variance. Comparison between the observed raw variance measure for the data and the raw variance of the expectation model shows that both variances were quite equivalent, which was 27.8%. In addition, the unexplained variance of the eigenvalue for the first contrast is 2.1, less than the recommended cut-off value, 3.0 (Abdellatif, 2023). This signifies that the test dimension of the Chemistry test paper was moderate and acceptable. Periphrastically, it also means all Chemistry items were intelligible and not confusing. The small percentage of explained raw variance could be due to narrow ranges of ability in students or the difficulty level of some items. In other words, similar abilities among the Pure Science students and equal difficulty of the Chemistry test items could have caused the small percentage of explained raw variance.

Table 2. Summary of principal component analysis (PCA)

Standardized Residual Variance (In Eigenvalue Unit)		Observed (%)		Expected (%)
Total raw variance in observations	69.2	100		100
Raw variance explained by measures	19.2	27.7		**27.8**
Raw variance explained by persons	6.9	9.9		9.9
Raw variance explained by items	12.3	17.8		17.8
Raw unexplained variance (total)	50	72.3	100	
Unexplained variance in first contrast	2.1	3.0	4.2	
Unexplained variance in second contrast	1.8	2.6	3.6	
Unexplained variance in third contrast	1.6	2.4	3.3	

Factor loadings of all Chemistry test items ranged from -0.01 to 0.45. However, Table 3 exhibits items 40 and 27 had positive factor loadings that exceeded the factor loading benchmark of 0.40. The grouping of these two items is significant because it recommends that both items have a common meaning that differs from the Rasch measurement standard (Sigudla & Maritz, 2023). Therefore, it can be summarized that a secondary dimension exists in this instrument with only a small influence. Sigudla and Maritz (2023) stressed out that any item with factor loading ≥0.40 should be examined.

Table 3. Factor loading of Chemistry test items that signify multidimensionality

Contrast	Loading	Measure	Infit MnSq	Outfit MnSq	Item
1	**0.45**	0.55	1.24	1.24	**40**
1	**0.41**	− 0.39	1.2	1.30	**27**
1	0.39	1.57	1.25	1.36	45
1	0.35	0.68	1.09	1.11	20

1	0.32	0.35	1.24	1.29	18
1	0.30	0.27	1.22	1.36	39
1	0.29	0.25	1.15	1.17	28
1	0.19	− 0.81	1.05	1.03	2
1	0.19	0.85	1.12	1.19	4
1	0.17	1.78	1.17	1.40	16
1	0.15	1.33	1.03	1.07	44
1	0.12	2.12	1.09	1.31	38

Notes: Factor loading ≥ 0.40 are in boldface. The information presented is an excerpt from the complete table.

The separation index supports the notion of a logit interval scale in segregating items and persons. The item separation index can be used as a construct validity index while the person separation index represents criterion validity. A high separation index indicates that an item or person is subject to adequate discrimination. In contrast, a low separation index signifies that several items are redundant and low variability of person on the trait. Data are considered widely spread in terms of range if the separation index is greater than one (1) (Bakytbekovich et al., 2023).

In Table 4, summary of statistical analysis established that the person separation index was 2.54. This index implies that students' ability was aesthetic and the Chemistry test paper reliably separated Pure Science students into at least three statistically different ability groups. The separation index represents how good the test can distinguish students according to their ability (Fitrah et al., 2024). The item separation index was 8.95, indicating very reliable item difficulty estimation and good variability. This index also denotes that the Chemistry test paper items could be separated into nine groups based on the students' answer. Deng et al. (2023) note that this separation pattern is considered well dispersed. The items' position has high reliability when the separation index exceeds the minimum value of two (2). Thus, a higher separation index produces a higher quality instrument.

As for the item reliability, the index was 0.99, indicating that the Chemistry test items were reasonably well-distributed along the logit interval scale. This index also adverting an adequate breath of position on the linear continuum from students with insufficient knowledge to students with sufficient knowledge in Chemistry. The high-reliability index signifies a high level of confidence in replicating items' placement within the measurement error. A reliability index greater than 0.94 is considered excellent (Hlynsson et al., 2025). Hence, it can be concluded that all items in the Chemistry test paper were in the acceptable range of 0.6 to 1.4 and were considered excellent.

Table 4. Analysis of reliability and separation index

	Person		Item
N	435		50
Measures			
Mean	1.00		1.00
Standard deviation	0.14		0.11
Standard error	0.05		0.15
Outfit mean square			
Mean	0.98		0.98
Standard deviation	0.25		0.19
Separation	2.54		8.95
Reliability	0.87		0.99
Cronbach's alpha		0.87	
Chi-square		23075.57	
Unidimensionality		19.20%	

An achievement test is considered ideal when the difficulty level is set up in accordance with the abilities of students (Sahin et al., 2023). In other words, the test represents the full range of the abilities of all students. A

standardized multiple-choice achievement test such as Chemistry test paper consisting of 50 questions was constructed according to the item ratio principle of 5:3:2 that referred to the different constructs in the Bloom's Taxonomy. This ratio represents 25 items on the knowledge construct, 15 items on the understanding construct and 10 items on the application construct (Malaysian Examination Council, 2023). In TOS, each item was classified into difficult, moderate and easy based on the professional judgement by panels members. However, in Rasch statistical analysis, items were clustered into various difficulty level by utilizing logit unit of the standard deviation.

Table 5 shows the analysis of Chemistry test items according to the different construct based on the students' answer, 8 items (16%) are on knowledge construct, 34 items (68%) are on understanding construct and 8 items (16%) are on application construct. Ratio comparison with the standardized multiple-choice achievement test format indicates that the ratio for the constructed test paper was 2:8.5:2. Therefore, it can be concluded that the Chemistry test measures the abilities of students in certain constructs only, which is merely on their understanding.

Table 5. Item difficulty level

Item Difficulty Level			
	Difficult (1.22 -2.21)	Moderate (0.91 – (-0.96))	Easy (-1.22 - (-2.41))
Item no.	35, 38, 16, 45, 48, 30, 44, 9	11, 25, 47, 4, 10, 22, 20, 46, 40, 42, 18, 39, 28, 24, 26, 50, 29, 8, 23, 12,33, 1, 36 ,27, 13, 37, 41, 14, 6, 31, 2, 3, 49, 43	21, 32, 17, 19, 15, 7, 5, 34
Total (Percentage)	8 (16%)	34 (68%)	8 (16%)
Ratio	2	8.5	2

The person-item map (Wright map) depicted in Figure 1 is a significant feature of the Rasch model. It is a graphical representation that is able to exhibit the relationship of person and item. This map demonstrates the distribution of persons estimates and items difficulty on a common logit scale. The person-item map illustrates a clear picture of the linear continuum of the students' ability in comparison to the Chemistry test items. The distribution of students' ability level, indicated by "#s", is displayed on the left side of the map, from highest to the lowest, from top to the bottom of the scale. The higher logit values of the person measure indicate a higher students' ability level in Chemistry test and better test performance while the lower logit values of the person measure signify the low student's ability and substandard test performance. The upper left quadrant represents students with sufficient knowledge in Chemistry whereas the lower left quadrant signifies students who have insufficient knowledge. On the right side of the map, the difficulty level of items is distributed in descending order from the most difficult item to the easiest item. More difficult items are located towards the top of the person-item map and easier items are located towards the base of the map.

The letter M denotes the mean of the students' ability ("M" on the left side of the map) and item difficulty level ("M" on the right side of the map). The mean item difficulty is usually set to 0 logit. If a student is plotted at the same level as an item, a student has a 50% chance of responding to that item correctly (Danushka & Gamage, 2024). As the items tend to be difficult, the odd of success is reduced means that less chance to answer correctly due to those items are estimated beyond the student's ability. The mean of students' ability is compared with the mean of item difficulty to establish how well the Chemistry test items have distributed according to the level of students' ability. For this data set, mean items is 0 while mean for students is 0.35 logit which are very close to each other. This indicates that the test items for the students are well-targeted (Danushka & Gamage, 2024). It also means the difficulty level of test items is appropriate for the science students.

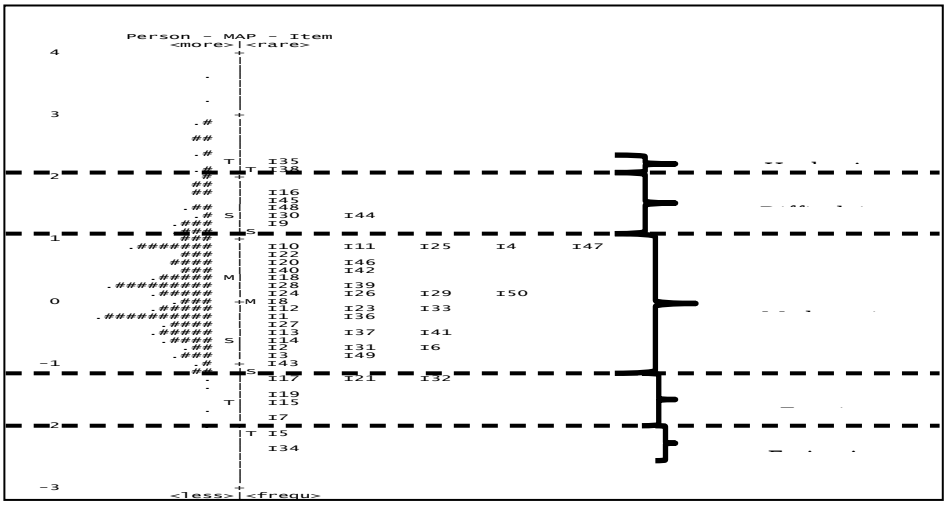
The maximum level of item measurement was 2.21 logit while the maximum measure of a student was 3.66 logit. The broader distribution of students' ability compared to item difficulty span exhibits that only certain

items are able to cover the range of the measurement traits (Bakytbekovich et al., 2023). Most of the Chemistry test items have difficulties level near the mid-range of the logit scale which is within one standard deviation. Students with high ability can answer the difficult items, while the easy item can be answered easily by students with high and low ability. The most difficult item answered by students is Item 35 with 74 correct responses out of 435. This item is about molar mass of a compound. In contrast, Item I34 is the easiest item with 400 correct responses out of 435. The easy item is about the naphthalene graph.

The person-item map posits a normal distribution of items and students in the logit interval scale continuum and falls into the scale's mid-range zone. The distribution of students' ability should be coordinated with the distribution of item difficulty when norm reference interpretations are requested (Tesio et al., 2024). Nonetheless, there are some gaps exist in the item location distribution (Item 9 and Item 10, Item 38 and Item 16) in the map that indicate students in the middle and upper level were not reasonably targeted by the Chemistry test items due to the content aspects for the constructs under study may be lack of some representation and potentially compromising the validity of the test. The gap between two consecutive items in the mid-range of the person-item map and lacking appropriate items with the higher ability students at the upper level of logit scale indicate that some important aspects have not been measured by the instrument (Samsudin et al., 2020).

Eight items (Item 17, Item 21, Item 32, Item 19, Item 15, Item 7, Item 5, Item 34) fall below students' ability. Despite fitting the model, they do not contribute to the measurement precision. Hence, these items may be discarded from the test. On the contrary, a few students with high ability are located above the logit 2.21. If there were many students at the high end of the difficulty range, then more items may be required to ensure measurement of all levels of ability was being covered, but unlike the aforementioned items, the difficult items should not be removed from the instrument in order to prevent ceiling effect (Noroozi & Karami, 2022). Precision of measurement would be useless if students' ability beyond the demand of the test.

Figure 1. Person-item map of chemistry multiple-choice test paper



In this study, psychometric analysis was conducted by contemplating four indicators; index value of infit MNSQ, item measure in logit unit, item polarity index and distractor analysis to determine the quality of each item and ensure they meet the standards. The test quality is assessed by students' response against each test item. *Table 6 and Table 7 are examples of psychometric analysis of several Chemistry test items using Winsteps software.*

Table 6. Analysis of Item 1

Infit MNSQ	0.90			
Item Measure	-0.28			
Option	Data		Average Ability	PTMea Corr
	Count	Percentage		

A	113	26	-0.13	-0.29
B	15	3	-0.56	-0.18
C	33	8	-0.31	-0.20
*D	271	62	0.7	0.46

In Table 6, the infit MNSQ value for item 1 is 0.90, which was within the acceptance range. The point-biserial measure correlation of item 1 is 0.46, showing excellent discrimination index and the difficulty level of the item is moderate. Option D is the key answer; thus it has a positive average ability and point-biserial correlation. A positive average ability, 0.7 showed that majority students responds to option D. Distractor A is a good distractor because it has a negative point-biserial measure correlation value and able to attract more than 5% of the students. However, distractor A has the potential to be the right answer due to the number of students who choose this distractor is high. Distractor C also a good distractor due to a negative point-biserial measure correlation value and the percentage of students who select this distractor is exceeds 5%. Different from distractor A and C, distractor B managed to attract only 3% of the students although the PTMEA Corr value has a negative value. Therefore, in distractor analysis, it can be concluded that distractor B is not a good distractor and needs to be examined and modified. Overall, based on the infit MNSQ value, point-biserial measure correlation, item measure and distractor analysis, it can be concluded that item 1 is an excellent item.

Table 7. Analysis of Item 7

Infit MNSQ	0.91			
Item Measure	-1.90			
Option	Data		Average Ability	PTMea Corr
	Count	Percentage		
A	9	2	-0.65	-0.15
B	14	3	-0.48	-0.16
*C	381	88	0.47	0.34
D	27	6	-0.36	-0.19

Table 7 shows the infit MNSQ value for item 7 is 0.91 which was within the acceptance range. The point-biserial measure correlation of item 7 is 0.34, indicates that this item has a good discrimination index, and this item is an easy item. Option C is the key answer because it has a positive average ability and point-biserial measure correlation. Due to the majority of students choosing this option, there is a possibility that low ability students also select the key answer by making either educated guess or blind guess. Distractor D is a good distractor compared to distractor B and A. However, distractor D managed to attract only 6% of students which is 1% exceeding the benchmark. The small percentage difference indicated that the distractor D might be confusing or the language used is inapproriate. Distractors B and A are considered not functioning properly due to the frequency of distractor selection is less than 5%. The implausible distractors such as distractor B and A can make the key answer obvious and reduced the effectiveness of the item. In a nut shell, it can be concluded that item 7 is a good item but lacks of quality. As a recommendation, this item need to be modified by revising in term of language used, the plausibility of the distractors and the order of the distractors.

IMPLICATIONS

The present study is conducted to determine the psychometric properties of the Chemistry multiple-choice test paper using the Rasch Model. The validity of the Chemistry test items was established based on the separation index analysis. Based on the result of the study, all test items are valid and reliable as they were in the acceptable range. On that account, no item was modified or discarded from the instrument. There are practical and methodological implications gained from this study.

For practical implication, this study significantly affects test developers and teachers in reviewing the multiple-choice items and the assessment standards primarily on item development. The data analysis from this study reveals that the Chemistry test paper is unidimensional and has good psychometric properties. Psychometric

data are important because it shows that the instrument only measures the intended construct (Swan et al., 2023; Ahmad & Siew, 2021). The psychometric evidence dictates the validity of interpretation from the test scores. Therefore, collection and reporting validity and reliability of the evidence are the important aspect (Hlynsson et al., 2025).

A valid instrument can increase the confidence level of teachers in using the instrument for measuring the knowledge of students and their level of understanding. The findings of this study serve as an indicator of the state of Chemistry measurement. A previous study conducted on the psychometric properties showed that teachers must be able to utilize validated instruments as self-assessment tools in identifying their strengths and weaknesses (Abdellatif, 2023; Bakytbekovich et al., 2023; Darmana et al., 2021; Hlynsson et al., 2025). Furthermore, the assessment of learning such as multiple-choice items are able to evaluate students' progress in the learning process and provide guidance for creating chemical learning strategies and recognizing students' understanding of chemical material (Al-Kafawein & Al-Hilal, 2025).

For methodological implication, numerous measurement experts have proclaimed that Rasch Model to be the "gold standard" approach for psychometric studies, as it is solely measurement model which have the properties of invariance for objective measurement that overcomes the limitation of the traditional statistical models (Hadzibajramovic et al., 2020; Murray et al., 2024; Salzberger et al., 2021; Sandoval et al., 2021). The present study offers a comprehensive psychometric validation using this state-of-the-art measurement model since the psychometric evidence gathered is viewed as a collective activity and reported at the level of detail (Luperdi-Roman et al., 2025). In general, this study provides a useful tool for test developers and teachers to measure how well students understand what has been taught in the classrooms. It is crucial to have insights into what exactly students understand, as conflicting views with teachers may potentially result in inaccurate methods of instruction and interpretation.

CONCLUSION

The psychometric properties of the test items are essential elements for assessing the quality of the test items. An accurate and reliable test outcome provides useful information on the students' progress, the pedagogical method's effectiveness, and valid prediction of the students' achievement. A far-reaching analysis of items using the Rasch model provides accurate empirical information on the psychometric properties of items rather than raw scores that determine items' quality. This valuable information benefits the teachers and the test developers to determine the functional and non-functional items in constructing a high-quality test.

Appertaining to the measurement theory, the presence of even a few flawed items are able to reduce the reliability and validity of the test. The unreliable and invalid test could not measure the students' understanding and ability of the subject's content. Therefore, these flawed items must be identified to ensure the tests result are meaningful. Regardless of reducing the reliability and validity of the test, flawed items also confuse students during the test-taking process. Therefore, removing the problematic items is able to enhance the quality of the instrument.

Psychometric interpretation enables teachers to improvise and modify their instructions according to the students' ability besides ensuring the appropriate use of a test as a tool of assessment. The result from the extensive analysis designates the Chemistry test paper possesses good psychometric properties and is capable of yielding valid and reliable scores in measuring the cognitive domain of students. Despite the target of the item particularly well on students' ability, there are no suitable items to assess students with the highest ability. The Chemistry test paper measures mostly the students' understanding. Teachers or test developers could refine the test by eliminating items with low difficulty, lessening the number of items with the same difficulty and adding items with a higher level of difficulty to create a better instrument. Furthermore, more difficult items should be added to the instrument to measure the students with the highest ability.

For a comprehensive evaluation of the Chemistry test paper, future studies could be conducted on Chemistry test paper consisting of subjective and essay questions by using different approach of Rasch Model such as multi-facet, partial-credit model, rating scale model or graded response theory based on the type of data collected.

ACKNOWLEDGEMENTS

Authors wish to acknowledge the support of Ministry of Education, Malaysia, District Education Departments, University of Malaya, and teachers who gave permission to collect the data for this study. This research has been approved (KPM.600-3/2/3-eras (5143) by Education Planning and Research Department (EPRD), Ministry of Education however no grants were accepted by the researchers in this study.

Declaration Of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

REFERENCES

1. Abdellatif, H. (2023). Test results with and without blueprinting: Psychometric analysis using the Rasch model. Educacion Medica, 24(3), 1-14. https://doi.org/10.1016/j.edumed.2023.100802
2. Ahmad, J., & Siew, N. M. (2021). Curiosity towards STEM education: A questionnaire for primary school students. Journal of Baltic Science Education, 20(2), 289-304. https://doi.org/10.33225/jbse/21.20.289
3. Al-Kafawein, J., & Al-Hilal, M. (2025). Assessing students' progress in chemistry: Using multiple-choice questions and performance-based assessments. Journal of Curriculum & Teaching, 14(1), 174-183. https://doi.org/10.5430/jct.v14n1p174
4. Bakytbekovich, O. N., Mohammed, A., Alghurabi, A. M. K.,...& Afif, A. N. S. (2023). Distractor analysis in multiple-choice items using the Rasch model. International Journal of Language Testing, 13 (Special Issue), 69-78. https://doi.org/10.22034/ijlt.2023.387942.1236
5. Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? CBE-Life Sciences Education, 15(4), rm4. https://doi.org/10.1187/cbe.16-04-0148
6. Danushka, N. A. S., & Gamage, P. S. Y. (2024). Assurance of test authenticity: Power of table of specification (TOS). University of Vocational Technology.
7. Darmana, A., Sutiani, A., Nasution, H., Ismanisa, I., & Nurhaswinda, N. (2021). Analysis of Rasch model for the validation of chemistry national exam instruments. Jurnal Pendidikan Sains Indonesia, 9(3), 329-345, https://doi.org/10.24815/jpsi.v9i3.19618
8. Deng, T., Sousa, L. M., Garg, V., & Bradley, M. S. A. (2023). Segregation of formulated powders in direct compression process and evaluations by small bench-scale testers. International Journal of Pharmaceutics, 647, 1-14. https://doi.org/10.1016/j.ijpharm.2023.123544
9. Fitrah, M., Sofroniou, A., Ofianto, Judijanto, L., & Widihastuti. (2024). Reliability and separation index analysis of mathematics questions integrated with the cultural architecture framework using the Rasch model. Journal of Education & e-Learning Research, 11(3), 499-509. https://doi.org/10.20448/jeelr.v11i3.5861
10. Hadzibajramovic, E., Schaufeli, W., & Witte, H. D. (2020). A Rasch analysis of the Burnout Assessment Tool (BAT). PLOS One, 15(11), Article e0242241. https://doi.org/10.1371/journal.pone.0242241
11. Hlynsson, J. I., Sjoberg, A., Strom, L., & Carlbring, P. (2025). Evaluating the reliability and validity of the questionnaire on well-being: A validation study for a clinically informed measurement of subjective well-being. Cognitive Behaviour Therapy, 54(2), 208-230. https://doi.org/10.1080/16506073.2024.2402992
12. Illene, S., Feranie, S., & Siahaan, P. (2023). Create multiple-choice tests based on experimental activities to assess students' 21st century skills in heat and heat transfer topic. Journal of Education and Learning (EduLearn), 17(1), 44-57. https://doi.org/10.11591/edulearn.v17i1.20540
13. Jegstad, K. M. (2023). Inquiry-based chemistry education: A systematic review. Studies in Science Education, 60(2), 251-313. https://doi.org/10.1080/03057267.2023.2248436

14. Krishnan, S., & Idris, N. (2014). Investigating reliability and validity for the construct of inferential statistics. International Journal of Learning, Teaching and Educational Research, 4(1), 51-60.
15. Langitasari, I., Aisyah, R. S. S., Parmandhana, N., & Nursaadah, E. (2024). Enhancing students' conceptual understanding of chemistry in a SiMaYang learning environment. KnE Social Sciences, 9(13), 191-200. https://doi.org/10.18502/kss.v9i13.15919
16. Luperdi-Roman, C. J. M., Goni-Cruz, F. F., & Deroncele-Acosta, A. (2025). Design and psychometric validation of the research competency scale for university students in Peru. International Journal of Evaluation & Research in Education, 14(6), 4887-4902. https://doi.org/10.11591/ijere.v14i6.35752
17. Malaysian Examination Council. (2023). Laporan Peperiksaan Kimia SPM. Pelangi Sdn. Bhd.
18. Murray, A. L., King, J., Xiao, Z., Ribeaud, D., & Eisner, M. (2024). Psychometric evaluation of a bried measure to capture general population-level variation in ADHD symptoms from childhood through the transition to adulthood. International Journal of Behavioral Development, 49(1), 12-25. https://doi.org/10.1177/01650254241268865
19. Noroozi, S., & Karami, H. A. (2022). A scrutiny of the relationship between cognitive load and difficulty estimates of language test items. Language Testing in Asia, 12(13), 1-19. https://doi.org/10.1186/s40468-022-00163-8
20. Rosli, R., Abdullah, M., Siregar, N. C., Hamid, N. S. A., Abdullah, S., Beng, G. K., Halim, L., Daud, N. M., Bahari, S. A., Majid, R. A., & Bais, B. (2020). Student awareness of space science: Rasch model analysis for validity and reliability. World Journal of Education, 10(3), 170-177. https://doi.org/10.5430/wje.v10n3p170
21. Runnels, J. (2012). Using the Rasch model to validate a multiple-choice english achievement test. International Journal of Language Studies, 6(4), 141-153.
22. Sahin, M. G., Yildirim, Y., & Ozturk, N. B. (2023). Examining the achievement test development process in the educational studies. Participatory Educational Research, 10(1), 251-274. http://dx.doi.org/10.17275/per.23.14.10.1
23. Salzberger, T., Cano, S., Abetz-Webb, L., Afolalu, E., Chrea, C., Weitkunat, R., & Rose, J. (2021). Addressing traceability of self-reported dependence measurement through the use of crosswalks. Measurement, 181, Article 109593. https://doi.org/10.1016/j.measurement.2021.109593
24. Samsudin, M. A., Chut, T. S., Ismail, M. E., & Ahmad, N. J. (2020). A calibrated item bank for computerized adaptive testing in measuring science TIMSS performance. EURASIA Journal of Mathematics, Science & Technology Education, 16(7), 1-15. https://doi.org/10.29333/ejmste/8259
25. Sandoval, I., Gilar-Corbi, R., Veas, A., & Castejon, J-L. (2021). Promoting equality in higher education: Development and internal validity of a selection test for science university degrees in Ecuador. Psychological Test & Assessment Modeling, 63(2), 191-204.
26. Shukri, A. A. M., Ahmad, C. N. C., & Daud, N. (2020). Integrated STEM-based module: relationship between students' creative thinking and science achievement. Jurnal Pendidikan Biologi Indonesia, 6(2), 173-180. https://doi.org/10.22219/jpbi.v6i2.12236
27. Sigudla, J., & Maritz, J. E. (2023). Exploratory factor analysis of constructs used for investigating research uptake for public healthcare practice and policy in a resource limited setting, South Africa. BMC Health Services Research, 23. https://doi.org/10.1186/s12913-023-10165-8
28. Suryadi, N. R. S., Nurmegawati, L., Mu'aziyah, S. E. S., Perdani, A. S. (2025). Rasch model for analysis of scientific attitude instruments in the context of secondary school science education. Equator Science Journal, 3(2), 98-106.
29. Swan, K., Speyer, R., Scharitzer, M., Farneti, D., Brown, T., Woisard, V., & Cordier, R. (2023). Measuring what matters in healthcare: A practical guide to psychometric principles and instrument development. Frontiers in Psychology, 14, 1-18. https://doi.org/10.3389/fpsyg.2023.1225850
30. Tesio, L., Caronni, A., Kumbhare, D., & Scarano, S. (2024). Interpreting results from Rasch analysis 1. The "most likely" measures coming from the model. Disability & Rehabilitation, 46(3), 591-603. https://doi.org/10.1080/09638288.2023.2169771
31. Thomas, D., & Zubkov, P. (2023). Quantitative research designs. In Quantitative research for practical theology (pp. 103-114). Andrews University Press.