International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume 6 Issue 21 (June 2024) PP. 131-139
DOI: 10.35631/IJMOE.621010

INTERNATIONAL JOURNAL OF MODERN EDUCATION (IJMOE)
www.ijmoe.com

GLOBAL ACADEMIC EXCELLENCE

VALIDITY AND RELIABILITY OF MULTIPLE-CHOICE BIOLOGY TEST: A RASCH ANALYSIS PERSPECTIVE

Avianna Jowinis[1], Nyet Moi Siew[2*]

[1] Fakulti Psikologi dan Pendidikan, Universiti Malaysia Sabah, Malaysia
Email: aviannajowinis@gmail.com
[2] Fakulti Psikologi dan Pendidikan, Universiti Malaysia Sabah, Malaysia
Email: sopiah@ums.edu.my
[*] Corresponding Author

Article Info:

Article history:
Received date: 07.03.2024
Revised date: 23.04.2024
Accepted date: 22.05.2024
Published date: 20.06.2024

To cite this document:

Jowinis, A., & Siew, N. M. (2024). Validity And Reliability Of Multiple-Choice Biology Test: A Rasch Analysis Perspective. *International Journal of Modern Education, 6* (21), 131-139.

DOI: 10.35631/IJMOE.621010

Abstract:

Concerns about the accuracy of multiple-choice tests for assessing biology knowledge have led to an increased emphasis on validating these tests using sophisticated analytical techniques like Rasch analysis. A 22-item multiple-choice cell division test was administered to 35 Form Four students in Kota Kinabalu, Sabah and validated through Rasch analysis, with MNSQ item values falling within the acceptable range of 0.79 to 1.24 logit. Additionally, the ZSTD components of Infit and Outfit showed favourable logit values aligning positively with the intended measures. Overall, the results indicated that item fits were within an acceptable range and PTMEA-CORR values measured what was intended to assess. Meanwhile, the Cronbach's alpha was a good value at 0.65, with person reliability of 0.72 and item reliability of 0.76. The person separation was observed at a value of 1.62 while the item separation considered acceptable at a value of 1.80; indicating good measurement capability which supports its use in evaluating students' comprehension of biological concepts. Applying the Rasch model demonstrated consistent and reliable test enabling assessment of students' achievement in biology; contributing towards continuous improvement in teaching and learning biology.

Keywords:

Multiple- Choice Questions, Rasch Analysis, Validity and Reliability

Introduction

In biology classes, multiple-choice exams are frequently used to assess students' comprehension of biological concepts. According to Butler (2018), many studies have been conducted to identify the most effective methods for utilizing multiple-choice exams to gauge

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 21 (June 2024) PP. 131-139
DOI: 10.35631/IJMOE.621010

student learning. As stated by Zhu *et al.* (2019), multiple choice questions (MCQs) are still one of the most widely used types of assessment questions in standardized tests. Many secondary and higher education exams use multiple choice questions (MCQs) because they are simple and accurate to score, which saves a significant amount of manpower and time (Jia *et al*., 2020). These benefits make MCQs a popular choice for educators across different levels of education systems around the world.

However, despite their widespread use, there are challenges in developing high-quality multiple-choice questions in biology. One of the primary challenges is ensuring the validity and reliability of these assessments. The objective of this research is to evaluate a set of multiple-choice questions for cell division test for validity and reliability using Rasch analysis. Rasch analysis provides a valuable tool for addressing these challenges. Rasch analysis application allows researchers to assess the psychometric aspects of multiple-choice biology examinations, such as their validity and reliability.

Validity and Reliability in Biology Education

In the Malaysian education system, biology is mainly taught for Form four and Form five students who have chosen science as an elective subject. Azzeme and Yusri (2018) stated that many students believe that biology is a challenging field of study. Güngör and Özkan (2017) found that cell division continues to be a challenging topic for many students. Efforts to improve understanding have not always been successful. The study conducted by Fauzi and Mitalistiani (2018) found that Indonesian students viewed the topics of genetics, metabolism and cell division as challenging. Meanwhile, Basri and Abdullah (2020) noted that Sabah's students continue to struggle to understand basic genetic ideas, particularly those related to genes and chromosomes. The outcome of the test offers a clear indication of how poorly young people understand basic concepts. They are consequently unable to understand the relationships and connections between these concepts and concepts such as cell division and genetics. Subsequently, Salleh *et al*. (2021) found that cell division, cell structure and organization, and the chemical composition of cells are difficult topics for assessing teacher and student judgments on complicated biological concepts.

According to Susongko (2016), measuring student achievement consists of providing numbers that are intended to show students' competence in a subject. A multiple-choice test created by experts is used to assess how well students performed on the cell division test. Concerns about the reliability and validity of multiple-choice exams for assessing students' knowledge of biology have increased in recent years. It is important to ensure that these tests accurately measure knowledge and are both valid and reliable. Heale and Twycross (2015) emphasize the importance for researchers to assess their measurements for reliability and validity. Gardner (2000) emphasized that assessments are only valuable if they are recognized as valid and reliable, as the quality of the data generated depends on this. As defined by Ghauri and Gronhaug (2005), validity refers to how well the data collected reflects the actual research topic. Ary *et al.* (2006), on the other hand, defined reliability as the ability of the instrument to consistently measure the things it is intended to assess. The validity and reliability of multiple-choice biology tests were assessed using Rasch analysis.

Integrating Rasch analysis into biology education also serves to address the challenges associated with developing and implementing high-quality assessment tools. Particularly when it comes to multiple-choice tests in biology, educators often face hurdles in ensuring the validity and reliability of these assessments. Factors such as the content of the question, the

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 21 (June 2024) PP. 131-139
DOI: 10.35631/IJMOE.621010

clarity of the options, and the test's ability to distinguish between levels of understanding must be carefully considered to ensure the accuracy of the assessment. Additionally, incorporating statistical measurements and utilizing Rasch analysis can lead to a more comprehensive understanding of test attributes, biases, and its alignment with intended learning outcomes. It helps to identify inappropriate elements and areas for improvement in the assessment tool, thereby improving its validity and reliability.

Rasch Analysis

Rasch theory is a mathematical framework for examining various assessment kinds. It is also referred to as the Rasch measurement model. Georg Rasch created this theory at the start of the 20th century, and it serves as a foundation for a probabilistic understanding of the connection between test item difficulty and individual abilities.

According to Yasin *et al.* (2015), content validation guarantees that predetermined objectives are satisfied, while a high reliability score suggests consistent instrument validity. Bond and Fox (2013) highlight that the Rasch measurement model assists researchers in determining how effectively an instrument represents the concept or latent characteristic being studied. Basran and Lajium (2020) argue that the Rasch model should bring social science measures closer to traditional physical measurements.

Azizah and Wahyuningsih (2020) described Winstep software, which employs a Rasch model computer to assess test-generated results and determine factors like as MNSQ outfit, point measure correlation, item reliability, and others. As stated by Chan *et al.* (2014), Cronbach's alpha provides researchers with insight into the reliability of the instrument during testing, whereas person reliability refers to the reproducibility of every individual's series of instructions when asked an alternate set of questions reviewing the same construct.

Findings from Rasch analysis offer guidance on how to enhance validity and reliability by reevaluating targeted cognitive processes, adjusting difficulty levels, or revising scoring criteria. Sumintono and Widhiarso (2015) categorized Cronbach's Alpha (KR-20) values as low (lower than 0.5), medium (0.5 to 0.6), good (0.6 to 0.7), high (0.7 to 0.8), and very high (more than 0.8). In addition, Sumintono and Widhiarso (2015) assessed item and person reliability based on the following criteria: less than 0.67 as low, 0.67 to 0.80 as sufficient, 0.81 to 0.90 as good, 0 .91 to 0.94 as very good and more than 0.94 as excellent.

Furthermore, fit statistics were also used to find items that were inconsistent with participants' responses to assess the reliability of the scale. Rasch's item fit analysis illustrates how well a student's proficiency or item difficulty relates to the test's underlying construct. The fit of an object can indicate whether it is suitable for assessing what it is intended to measure and whether it is functioning normally. Rasch analysis helps identify misfit items that do not align with intended cognitive levels or discriminate effectively between high and low-performing students. Azizah and Wahyuningsih (2020) recommended that the point-measure correlation (PTMEA-CORR) be between 0.4-0.85, the acceptable mean square value (MNSQ) be between 0.5 and 1.5, and the z-standardized values (ZSTD) be between −2 and 2.

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 21 (June 2024) PP. 131-139
DOI: 10.35631/IJMOE.621010

Method

Participant

Purposive sampling was used to choose study participants. The participants were 35 Form Four students that enrolled in science stream classes in Kota Kinabalu, Sabah. All students had their cell division topic lesson in the same year.

Instrument

The cell division test included 22 multiple-choice questions (MCQ). This was a criterion-referenced test based on Bloom's revised taxonomy (Anderson, 2001). This criterion-referenced test allowed the researcher to assess an individual's competence or skill level with respect to a specific body of knowledge that the test covered. In a new version of Bloom (Anderson *et. al.,* 2001), the original taxonomy is divided into subcategories and the various category titles are changed to their active verb counterparts: remember, understand, apply, analyze, evaluate, and create.

In order to create the MCQ items, questions from biology reference books and textbooks were modified using the Form 4 biology curriculum as a guide. It covered the topic of cell division such as meiosis, mitosis and cell cycle. Participants had 35 minutes to answer all questions. Table of Specification with the subtopic and their corresponding cognitive objectives listed as in the Table 1.

Table 1: Table of Specification in Cell Division Test

TOPIC	Remember	Understand	Apply	Analyze	Evaluate	Create	Number of Item	%
Introduction to cell division	1	1	1				3	14
Cell cycle and mitosis	1	1	2	2	1	1	8	36
Meiosis	1	1	1	2	1	2	8	36
Issues in cell division on human health			1	1	1		3	14
TOTAL	3	3	5	5	3	3	22	
Total Percentage Weight								100

Procedure

The Biology Curriculum for Form Four KSSM served as a guide for the researchers to create the multiple-choice questions from biology reference books and textbooks. After expert evaluation, a pilot study with multiple-choice cell division test was conducted. The results of the multiple-choice test taken by the Form Four students were entered into the analysis using Microsoft Excel 2010. The response patterns determined were quantitatively analyzed using

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 21 (June 2024) PP. 131-139
DOI: 10.35631/IJMOE.621010

WINSTEPS® Version 3.73. The test items were assessed for validity and reliability using Rasch analysis. Parameters measured in the Rasch analysis included Cronbach's alpha value, item fit, item and person reliability/separation, and point measure correlation (PTMEA-CORR). These parameter values are tested within a certain range. Figure 1 shows the flowchart of the validation and reliability testing procedure for the cell division test.

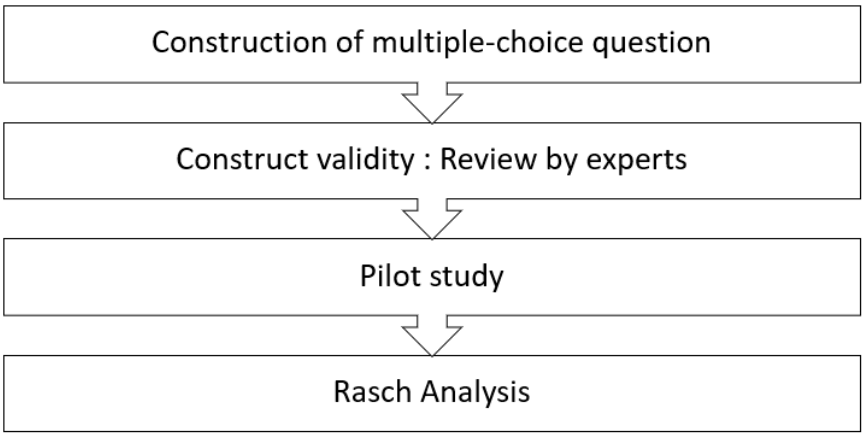

Figure 1: The Procedure Of Validation And Reliability Testing For The Cell Division Test

Findings and Discussion

Nugroho *et al.* (2022) suggested that Point measure correlation, outfit z-standard, and outfit mean-square value are three criteria that determine the level of item fit. According to Boone *et al.* (2014), the suitability of the items can be determined using the infit and outfit ranges for MNSQ items, which are between 0.5 to 1.5, and the infit and outfit ZSTD ranges, which are between -2 and 2. Linacre (2010) claims that the outfit z standard value index (ZSTD) can be ignored as long as the outfit mean square value (MNSQ) values are appropriate for outfit and fitness.

Azizah and Wahyuningsih (2020) agreed that test item is considered to be fit or valid if they fulfilled one or both of the criteria for Outfit of MNSQ and ZSTD, and PTMEA-CORR. Higher values of PTMEA-CORR are considered “mismatched” items. As suggested by Camilia *et al.* (2023), if the questions do not meet these three requirements, they are probably not good enough and should be corrected or changed. Table 1 shows that the infit and outfit values of the MNSQ item range between 0.79 and 1.24 logit. It was also found that all ZSTD components of Infit and Outfit had logit values between -1 and 1.8. In addition, the point-measure correlation (PTMEA-CORR) values or item alignments for all items were positive (>0), indicating that the items can measure what they are intended to assess. This enabled the test instrument to include all 22 items that met the valid criteria.

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume 6 Issue 21 (June 2024) PP. 131-139
DOI: 10.35631/IJMOE.621010

Table 2: Rasch Analysis

Item	Infit		Outfit		PT- MEASURE
	MNSQ	ZSTD	MNSQ	ZSTD	CORR
Q1	.84	-.5	.79	-.4	.44
Q2	.97	-.3	.96	-.2	.36
Q3	.84	-.9	.89	-.4	.47
Q4	1.02	.2	1.00	.1	.30
Q5	1.10	.9	1.09	.6	.18
Q6	.95	-.3	.85	-.7	.38
Q7	1.09	.9	1.08	.6	.20
Q8	1.07	.6	1.05	.4	.24
Q9	1.10	.7	1.07	.4	.17
Q10	1.21	1.8	1.20	1.2	.03
Q11	.89	-1.0	.86	-1.0	.46
Q12	.99	0	1.01	.1	.32
Q13	.95	-.5	.91	-.7	.40
Q14	.92	-.7	.88	-.8	.44
Q15	1.04	.3	1.07	.3	.20
Q16	.93	-.7	.92	-.5	.41
Q17	.99	.0	.94	-.2	.32
Q18	.83	-1.6	.79	-1.6	.55
Q19	.88	-.8	.86	-.6	.44
Q20	1.15	1.3	1.15	1.0	.12
Q21	1.04	.4	1.20	1.4	.21
Q22	1.24	1.1	1.24	1.1	.06

Table 3: Cronbach's Alpha (KR-20), Person-Item Reliability and Separation

Analysis	Value
Cronbach's Alpha (KR-20)	0.65
Person Reliability	0.72
Item Reliability	0.76
Person Separation	1.62
Item Separation	1.80

Separation shows how well a group of people can identify the test items, while person separation shows how well a set of items can distinguish the people being tested (Boone and Noltemeyer, 2017). According to Duncan *et al.* (2003), a person with a separation index of 1.5 is considered acceptable, 2 is considered exceptional, and 3 is considered excellent. As shown in Table 2, the value of Cronbach's alpha was 0.65 and considered as good. It was also indicated that person reliability value was 0.72, while item reliability value was 0.76. These findings indicated that the instrument exhibited sufficient reliability. The separation value can be used to determine how persons and/or items are grouped. In this study, person separation was found to be 1.62 and item separation value is 1.8, which considered as acceptable value for person separation.

Conclusion

This study aims to evaluate a series of multiple-choice questions for the cell division test for validity and reliability using Rasch analysis. All 22 multiple-choice test items in the biology assessment tool were accepted and validated by Rasch analysis as they meet the criteria mentioned by Azizah and Wahyuningsih (2020). MNSQ item scores were within an acceptable

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 21 (June 2024) PP. 131-139
DOI: 10.35631/IJMOE.621010

range, ranging from 0.79 to 1.24 logit. Furthermore, the ZSTD components of Infit and Outfit showed favorable logit values that were positively consistent with the intended measures. The results showed that the item fits were within the acceptable range and PTMEA-CORR (point-measure correlation) values can measure what they intended to assess.

Meanwhile, Cronbach's alpha was 0.65, the person reliability value was calculated as 0.72, while the item reliability value was calculated as 0.76. The person distance is observed at a value of 1.62, while the object distance is considered acceptable at a value of 1.80. The value of Cronbach's alpha is good, the person's item reliability and separation are acceptable and sufficient. This result demonstrated that the cell division assessment tool can be used to assess students' understanding of biological concepts. The results can be used to develop a standardized test that assesses students' knowledge of cell division in biology classes.

Overall, the use of Rasch analysis in the evaluation of biology tests has enormous potential for improving the accuracy and validity of students' performance in biology, not only in the area of cell division, but also in other areas such as the chemical composition of a cell and the respiratory system. Additionally, it allows for the identification of potential biases or limitations in the test tasks and provides a basis for making necessary adjustments or improvements. By continually applying and refining Rasch analysis in educational assessment, educators are better able to create fair and reliable assessments that accurately measure student abilities and effectively influence instructional practice. This comprehensive approach to exam assessment ultimately contributes to the continuous improvement of learning in biology classes.

Acknowledgement
The researchers would like to express their appreciation to the University of Malaysia Sabah, Sabah, Malaysia, which has funded the publication of this research under the Cluster Fund Research Grant, DKP0005 Phase 1/2023.

References

Anderson, L. W., Krathwohl, D. R., and Bloom, B. S. (2001). A Taxonomy for Learning, Teaching, and Assessing a Revision of Bloom’s Taxonomy of Educational Objectives. New York, NY: Longman.

Ary, D., Jacobs, L. C., Razavieh, A., & Sorensen, C. (2006). Introduction to research in education (7th Ed.): California: Thomson Wadsworth.

Azzeme, A. M., & Yusri, I. W. (2018). Digital Board Game in Teaching and Learning of Biochemistry. Proceedings of The International University Carnival on e-Learning (IUCEL) 2018. 470-472.

Azizah, A., & Wahyuningsih, S. (2020). Use of the Rasch model for analysis of test instruments in actuarial mathematics courses. Journal of Mathematics Education, 3(1), 45-50.

Basran, A., & Lajium, D. (2020). Aplikasi Model Rasch DalamPengujian Intrumen Inventori Konsep Daya. International Journal of Modern Education, 2(6), 14-27.

Basri, S., & Abdullah, M. S. (2020). Level of Understanding and Alternative Frameworks in Genetics Fundamental Concepts among Form Four Biology Students in Sabah, Malaysia. International Journal of Academic Research in Business and Social Sciences. 10(9), 522-541.

Bond, T. G., & Fox, C. M. (2013). Applying the Rasch model: Fundamental measurement in the human sciences. Psychology Press

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume 6 Issue 21 (June 2024) PP. 131-139
DOI: 10.35631/IJMOE.621010
Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Rasch Analysis in the Human Sciences. Dordrecht, Netherlands: Springer.

Boone, W. J. & Noltemeyer, A. (2017) Rasch analysis: A primer for school psychology researchers and practitioners. Cogent Education, 4(1), 1–13.

Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? Journal of Applied Research in Memory and Cognition, 7(3), 323–331.

Camila, C. G., Afandi, A., Tenriawaru, A. B., Artika, W., & Siregar, N. (2023). Development of higher order thinking skill questions using Stahl and Murphy's taxonomy on excretion system topic. Assimilation: Indonesian Journal of Biology Education, 6(2), 97-108.

Chan, S. W. & Zaleha, I. 2014. A Technology-Based Statistical Reasoning Assessment Tool in Descriptive Statistics for Secondary School Students. The Turkish Online Journal of Educational Technology, 13(1), 29–46.

Duncan, P. W., Bode, R., Lai, S. M., & Perera, S. (2003) Rasch analysis of a new stroke-specify outcome scale: The stroke impact scale. Archives of Physical Medicine and Rehabilitation, 84 (7), 950-963.

Fauzi, A., & Mitalistiani, M. (2018). High school Biology topics that perceived difficult by undergraduate students. Didaktika Biologi: Jurnal Penelitian Pendidikan Biologi, 2(2), 73-84.

Gardner, R.C. (2000) Correlation, causation, motivation, and second language acquisition. Canadian Psychology, 41, 10-24.

Ghauri, P. & Gronhaug, K. (2005). Research Methods in Business Studies. Harlow, FT/Prentice Hall.

Gungor, S. N., & Ozkan, M. (2017). Evaluation of the concepts and subjects in Biology perceived to be difficult to learn and teach by the pre-service teachers registered in the pedagogical formation program. European Journal of Educational Research, 6(4), 495-508.

Heale, R. & Twycross, A. (2015). Validity and Reliability in Quantitative Studies. Evid Based Nurs, 18(3), 66–67.

Jia, B., He, D., & Zhu, Z. (2020). Quality and feature of multiple-choice questions in education. Problems of Education in the 21st Century, 78(4), 576-594.

Linacre, J. M. (2010). Predicting responses from Rasch measures. Journal of Applied Measurement, 11(1), 1–10

Nugroho, A. T., Karyanto, P., & Sugiharto, B. (2022). Profile of critical thinking ability of high school students on animalia material during hybrid learning in the pandemic era. Journal of Science Education Research, 8(6), 2635-2640.

Salleh, W.N.W.M, Ahmad, C.N.C., Setyaningish, E. (2021). Difficult topics in biology from the view point of students and teachers based on KBSM implementation. EDUCATUM Journal of Science, Mathematics and Technology, 8(1), 49-56.

Sumintono, B. & Widhiarso, W. (2015) Aplikasi Pemodelan Rasch Pada Assessment Pendidikan. Cimahi: Trim Komunikata Publishing House

Susongko, P. (2016). Validation of Science Achievement Test with The Rasch Model. *Jurnal Pendidikan IPA Indonesia, 5*(2), 268-277.

Yasin, R. M., Yunus, F. A. N., Rus, R. C., Ahmad, A., & Rahim, M. B. (2015) Validity and Reliability Learning Transfer Item Using Rasch Measurement Model. Procedia - Social and Behavioral Sciences, 204, 212–217.

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 21 (June 2024) PP. 131-139
DOI: 10.35631/IJMOE.621010

Zhu, Z., Wang, C., & Tao, J. (2019). A Two-Parameter Logistic Extension Model: An efficient variant of the Three-Parameter Logistic Model. Applied Psychological Measurement, 43(6), 449-463.