Volume 12. Issue 2. 265-276 DECEMBER 2023

Jurnal Ilmiah Pendidikan Fisika Al-BiRuNi
https://ejournal.radenintan.ac.id/index.php/al-biruni/index
DOI: 10.24042/jipfalbiruni.v12i2.17037

P-ISSN: 2303-1832
e-ISSN: 2503-023X

Validity and Reliability of Elasticity Multiple-Choice Items (EMCI) Using Rasch Model

Vivi Mardian[1*], Achmad Samsudin[2], Judhistira Aria Utama[3], Irma Rahma Suwarma[4], Bayram Coştu[5]

[1, 2, 3, 4] Department of Physics Education, Universitas Pendidikan Indonesia, Bandung, Indonesia
[5]Yildiz Technical University, Istanbul, Turkey

**Corresponding Address: vivimardian1111@gmail.com*

Article Info

Article history:

Received: May 31, 2023
Accepted: November 18, 2023
Published: December 30, 2023

Keywords:

Elasticity Concept;
Multiple Choice Items;
Rasch Model.

ABSTRACT

Assessment in the form of an instrument is very important to test for validity and reliability so that it can measure student learning outcomes. We implemented a measurement instrument designed to assess students' understanding of elasticity. The measurement instrument consisted of 15 items. The instrument was administered to students who were taking physics subjects in senior high school in their second year. In total, 74 students were taken from two classes in Padang, Indonesia. In this study, the students' performance was collected as quantitative data and evaluated using the Rasch model. The results showed that the data matched the Rasch model measurements. Moreover, female students answered 53% of the questions better than male students. Furthermore, the DIF plot shows that S8 is the most difficult problem, while S14 is the easiest. There are two gender bias questions, namely S12 and S14. Both questions were easily solved by female students. However, all questions can be used to measure students' abilities in elasticity and Hooke's law. This study seeks to make a contribution to the literature on EMCI evaluations by offering a case study for academics and researchers to use in assessing students' elasticity skills.

INTRODUCTION

Elasticity concepts are an essential element of the physics education program's physics subject fundamental. Learning the notion of elasticity is critical for pupils (Vázquez-Bernal & Jiménez-Pérez, 2023). Hooke's law, based on Silva et al., (2019), is an excellent technique to expose secondary school pupils to investigative procedures appropriate to PISA level 6. Elasticity is also linked to biological concerns such as DNA bending. Mierke (2020) explains that spring simulations are applied to demonstrate how a spring may operate on mechanical processes such as DNA bending and membrane fluctuations at cell surfaces.

One type of assessment to measure students' mastery of the concept of elasticity is multiple-choice questions. Multiple-choice tests are a popular final assessment tool in teaching (Dziob, 2020; Gamage et al., 2022; Podschuweit & Bernholt, 2018). They usually ask pupils to choose the best answer from a list of available options (Carotenuto et al., 2021; Kashihara & Fukaya, 2022; Taber, 2018). This usually indicates that there will be one correct answer among two, three, or four possibilities, meanwhile, variants might include choosing the best-possible answer or several potential replies ("multiple-response"). Multiple-choice questions can

How to cite Mardian, V., Samsudin, A., Utama, J.A., Suwarma, I.R., & Coştu, B. (2023). Validity and reliability of elasticity multiple-choice items (EMCI) using rasch model. *Jurnal ilmiah pendidikan fisika Al-Biruni, 12*(2), 265-276.

be employed in the preliminary (formative) and final (summative) examinations.

Multiple-choice tests are frequently utilized for educational purposes as an evaluation technique (McKenna, 2019; Susanti et al., 2018; Wammes et al., 2022). How extensively and in which circumstances this occurs cannot be determined with any degree of certainty. It is a commonly employed type of evaluation globally and ubiquitous (Khan & Krell, 2019; McKenna, 2019). Although not as useful for humanities topics, multiple-choice tests are regularly used in a variety of STEM subjects, including the area in which this study's experiment is organized, computational science, and by expert, statutory, and regulatory bodies, including those in critical fields such as health, law, and finances (Borda et al., 2020; Brassil & Couch, 2019; McKenna, 2019). They often save staff time in terms of labeling, moderating, and offering feedback because they may be marked instantly - and, in theory, objectively (Grover & Wright, 2023; McKenna, 2019). This implies that multiple-choice questions can help teachers verify students' answers faster because they only select one answer that they believe is accurate (Schwarz, 2023; Van et al., 2023). The multiple-choice questions generated in this study are utilized for assessing how well learners perform at the end of the class, which is known as summative evaluation.

The Rasch model may be used to assess the reliability and validity of multiple-choice tests. The Rasch model was developed as part of a wider set of measuring tools known as item response theory, and it has been widely utilized in educational research to assess psychometric data (Chan et al., 2021; Heritage et al., 2023; Wind et al., 2019). Rasch model analysis has proven to be a powerful approach for assessing psychometric features and removing response bias (Chan et al., 2021; Ha, 2021). The Rasch model's psychometric analysis technique might have been employed to generate test items and functioned as an important tool in the learning evaluation (Chan et al., 2021; Rodríguez-Mora et al., 2022). The results of the Rasch model using the logit ruler met Mok and Wright's five measurement standards for human research, which are as follows: (a) yielding a straight-line measure; (b) overcoming incomplete data; (c) providing an accurate estimate; (d) detecting outliers or misfits; and (e) reproducibility (Chan et al., 2021). The Rasch model is used in this investigation because it reflects a person's measures on the same scale, independent of the participants' test (Al-Owidha, 2018; Chan et al., 2021). Human and object traits were used to evaluate estimations of latent qualities. The Rasch model might be used to analyze students' exam success rates depending on the difficulty level of the substances and their level of competence (Chan et al., 2021).

Based on the studies mentioned previously, the purpose of this study is to test the validity, and reliability of a newly created elasticity multiple choice instrument (EMCI). A valid test can measure the material that students must measure (Henry et al., 2021; Larrain & Kaiser, 2022) Reliable questions are shown by consistent student test results from time to time (Pedaste et al., 2021) The instruments that have been developed need to be examined to determine whether they are valid or not and whether they are reliable or not by using Rasch modeling, the results of the analysis will be deeper and more detailed, researchers Hasanah and Purwanto (2023) tested the validity and reliability of the Rasch model on elasticity. However, researchers have not revealed whether there is gender bias or not. Previous research revealed that multiple-choice questions in physics material experienced gender bias (Gladys et al., 2023). To prove the results of this research, we formulated two research questions:

RQ1: To what extent does the data that was gathered from senior high school pupils suit the Rasch model?

RQ2: Is there a gender bias in the questions that have been developed?

METHODS

Procedure

The EMCI applied to this investigation was derived from prior work Suwono et al. (2021). The purpose of this study is to create a diagnostic test instrument for pre-service teachers' misunderstandings regarding cell biology. Suwono et al. (2021) developed the instrument based on Treagust (1988) instrument development. The EMCI development steps can be seen in Figure 1.

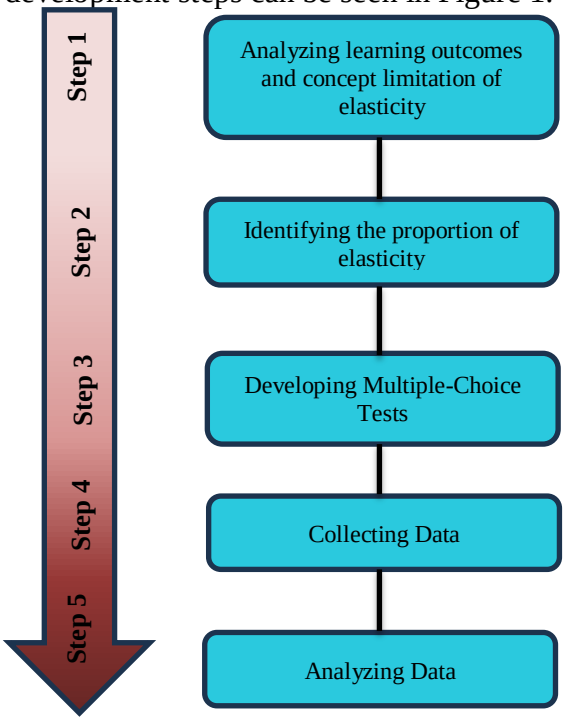

Figure 1. Procedure of development EMCI

Analyzing learning outcomes and concept

The core competency of the elasticity topic is analyzing the elastic properties of materials in everyday life. The learning objectives were developed into 31 learning objectives based on core competencies, textbooks, and the physics syllabus.

Identifying the proportion of elasticity

The learning objectives that had been developed in the initial stage were reduced to 12 items based on suggestions from the lecturer. The sub-concepts assessed are the definition of elasticity (1 item), strain and stress (1 item), Young's modulus (3 items), Hooke's law (3 items), the arrangement of springs (5 items), and the application of elasticity in human life (2 items).

Developing Elasticity Multiple-Choice Instrument (EMCI)

The measuring apparatus has 15 questions. In summary, when constructing the instrument, the researcher constructed physics-content themes incorporating the elasticity idea depending on the aim of the study. Each item addressed one of the elasticity concepts. For example, given a picture, students are asked to interpret the graph of the force relationship with the displacement of length. Students were asked to select only one choice from a list of four. The EMCI component of the instrument was scored in two ways.

The instrument developed consisted of 15 multiple-choice questions. Students have been taught elasticity material before working on the problem. The following details the questions depicted in Table 1.

Table 1. Detail of EMCI

Items	Core of question
1	Calculate the magnitude of the strain
2	Calculate the young's modulus
3	Summarize the data in the experimental results table
4	Calculate the magnitude of the spring constant
5	Determine the force graph with the increase in length that has the minimum elasticity constant
6	Sort the images with the spring replacement constant from largest to smallest
7	Complementing the unknown constant values in the figure
8	Determine the load ratio on the spring circuit so that it produces the same increase in length if it is arranged in series and parallel
9	Determine the length of the shortened spring
10	Calculate the increase in spring length from the experimental data
11	Calculate the potential energy of the spring
12	Complete the empty data in the table
13	Calculating the average spring constant from the experimental table
14	Calculate the amount of stress on the rubber
15	Calculate the magnitude of the steel strain from the table

Collecting Data

The instrument was distributed to senior high school pupils participating in the study. In total, 74 students (29 males, and 45 females) were taken from 2 classes, in Padang, Indonesia. Participant students had taken elasticity about 3 weeks before being given tests. Due to the new normal conditions, the instrument is distributed online via Google Forms. Participant students are given 90 minutes to complete all of these items.

Analyzing Data

The portion of credit The Rasch Model parameterizes the component challenges of each item's separate "steps" Masters in Park and Liu (2021), allowing researchers to accommodate the possibility of varied numbers of response possibilities for different items Bond & Fox in Park and Liu (2021). Below is a diagram of the Rasch partial credit model.

$$\ln\left({}^{P_{nik}}/_{1\_P_{nik}}\right) = B_n - D_{ik} \qquad (1)$$

For item and person measurements, the Rasch model employs the logit scale, yielding equal-interval linear data for parametric statistical testing. This linear measure corresponds better with the main hypotheses of the tests and is used to evaluate item difficulty and student ability based on the data set.

RESULTS AND DISCUSSION

In the present study, the Rasch model was used to validate the EMCI leveraging quantitative data. Rasch models, which are part of IRT, are frequently employed in educational research to assess psychometric data because they reliably represent measures on the same scale (Khine, 2020). The results of the study will show if the data matches the Rasch model measurements, explore the capability of EMCI among high school students, and decide whether to add tests whose aim varies depending on the student's gender. The consequences of these factors are discussed in the next part.

RQ1: To what extent does the data that was gathered from senior high school pupils suit the Rasch model?

The EMCI's fit validity was used to evaluate its content validity. Rasch item-fit statistics were used to assess how well an item fitted the model and suited the concept of one feature (Chan et al., 2021). Infit MNSQ (mean-square) and Outfit MNSQ (mean-square) fit indices were examined for EMCI. The Outfit MNSQ of the EMCI item mean was 1.00, as shown in Table 2, which was a reasonable value within the recommended range of 0.5 to 1.5 (Akhtar & Sumintono, 2023; Chan et al., 2021). Furthermore, the internal consistency of the students' test responses was evaluated using Cronbach's alpha (KR-20) person raw score test reliability (Chan et al., 2021). The EMCI's Cronbach's alpha was 0.64, which was deemed adequate (Chan et al., 2021). In addition, Akhtar and Sumintono (2023) calculated that the raw variance explained by measures was 34.5%, which was higher than the 20% threshold. Consequently, the items were useful for measuring and producing a reasonable forecast.

If identical items were provided to a separate sample of people with the same skills, the item dependability index assessed the repeatability of item position within an item organization as well as the measured parameter (Chan et al., 2021). Table 2 indicates that the EMCI had a very reliable item dependability value of 0.95, an item separation or dispersion of items combined with the measured variable (Chan et al., 2021). Because it was greater than three, or 4.20, the EMCI had adequate dispersion. Table 1 shows that the item's mean measure (logit) is 0.00 logit and the standard deviation is more than one logit (1.38), showing a highly large dispersion of measures in item difficulty level over the logit scale. This suggests that the instrument can assess a larger variety of pupil abilities in the setting of EMCI.

Table 2. Person and item statistics

	Person	Item
N	74	15
Logit (L)		
M	0.57	0.00
StDev	1.16	1.38
SE	0.14	0.37
Outfit Mean-Square (OMS)		
M	1.00	1.00
StDev	0.61	0.39
Separation (S)	1.42	4.20
Reliability (R)	0.67	0.95
Cronbach's Alpha (CA)	0.64	
Measures explain raw variance	34.5%	

The person dependability index, on the contrary, was an assessment of the replication of person placement that might be expected if a particular group of individuals were given an additional set of relevant items assessing an identical construct (Chan et al., 2021). In Table 2, the person dependability for EMCI was 0.67, which was weak (Chan et al., 2021). A person separation index calculates the separation or dispersion among individuals on a given variable (Chan et al., 2021). For EMCI, the value of person separation was larger than one, i.e., 1.42, exhibiting that the samples were sufficient to identify a person's ability (Chan et al., 2021). The person logit mean was +0.57 logit, indicating that every participant performed better than average (better than item mean) on the EMCI. It has a standard deviation of 1.16, indicating a significant level of ability dispersion across the students.

To assess item compatibility, the values of Outfit MNSQ, Outfit ZSTD, and Pt-Measure Corr were used for each item. A range of 0.5 to 1.5 for the item and person's Outfit in the MNSQ showed a satisfactory fit of the data to the model. The Outfit ZSTD value should be in the -2.0 to 2.0 range to show that the components were somewhat predictable. To test whether all of the components were performing properly, the Pt-measure corr was used. To show that the items were somewhat predictable, the Pt-measure corr number should be between -4.0 and 0.85.

Despite this, Q11 was consistently beyond the range of Outfit MNSQ in Table 3, although Outfit ZTSD and Pt-Measure Corr remained within acceptable ranges. The Outfit ZSTD and Outfit MNSQ values for items Q5 and Q6 were out of range, while the Pt-Measure Corr values were in range. As a consequence, all of these objects were saved and did not need to be discarded. When three standards (Outfit MNSQ, Outfit ZSTD, and Pt-Measure) are not met, the item is deemed unfit. However, if only one or two criteria are not met, the object can still be used for evaluation. The test had fifteen items in all. In summary, the Rasch model measurement was appropriate for the entire data set collected from senior high school students.

Rasch model also has the advantage of providing useful information on item fit quality. Item fit indicates whether or not the item performs measurements normally. Table 3 contains the findings of the item fit analysis.

Table 3. Item-fit analysis

Ques Tions	Logit	Infit MNSQ	Outfit MNSQ	Outfit ZSTD	Pre-Measure Corr
S1	0.72	0.90	0.83	-0.91	0.48
22	1.11	1.02	0.95	-0.15	0.36
S3	-1.71	1.02	1.12	0.41	0.44
S4	-1.57	0.66	0.52	-1.30	0.69
S5	0.72	1.12	1.93	3.94	0.25
S6	0.91	1.20	1.51	2.23	0.20
S7	1.31	0.99	1.16	0.68	0.35
S8	2.78	0.90	0.75	-0.34	0.33
S9	0.52	1.16	1.18	1.00	0.29
S10	-0.86	1.05	1.03	0.19	0.43
S11	-2.46	0.71	0.32	-1.24	0.63
S12	0.59	0.93	0.85	-0.79	0.47
S13	-0.10	0.99	1.26	1.35	0.43
S14	-1.71	0.87	0.59	-0.95	0.58
S15	-0.25	0.99	1.00	0.05	0.45

On the same logit scale, the Wright Map in Fig. 1 displays the distribution of item difficulty and student competencies. The item complexity was displayed on the right side of the Wright Map, while the pupils' abilities were displayed on the left. The greater the logit, the harder the objects and the pupils' abilities. Lower logits provide simpler challenges for students with lower

abilities. Figure 1. Item S8 was the most difficult to indicate, while item S11 was the easiest to demonstrate. Although, a student had the greatest ability: 44. Student 01 had the lowest ability level among the 74 students. Six items, namely S1, S2, S5, S6, S7, and S8, were tough for the pupils since their difficulty levels were higher than the person's average. There were three pairing of items with the same difficulty, namely S1 with S5, S9 with S12. It signifies that the chances of successfully answering these questions are less than 50%.

Figure 1. illustrates the results of the Rasch model investigation as a Wright Map. Students can be classified into three groups based on the Wright Map: high, medium, and low. Students in the high group are those who can answer tough questions or have a large logit score of +2. There were two pupils (S44 and S69) in the group of kids with high talents who were able to work on the most challenging S8 questions. Students in the medium group are those who can answer problems of moderate difficulty or who have a logit value greater than -2 but less than +2. Students with moderate abilities account for up to 92% of all students. Students in the low category can only answer questions categorized as very easy to answer or have a logit value less than -2. Four pupils make up the group with the lowest category.

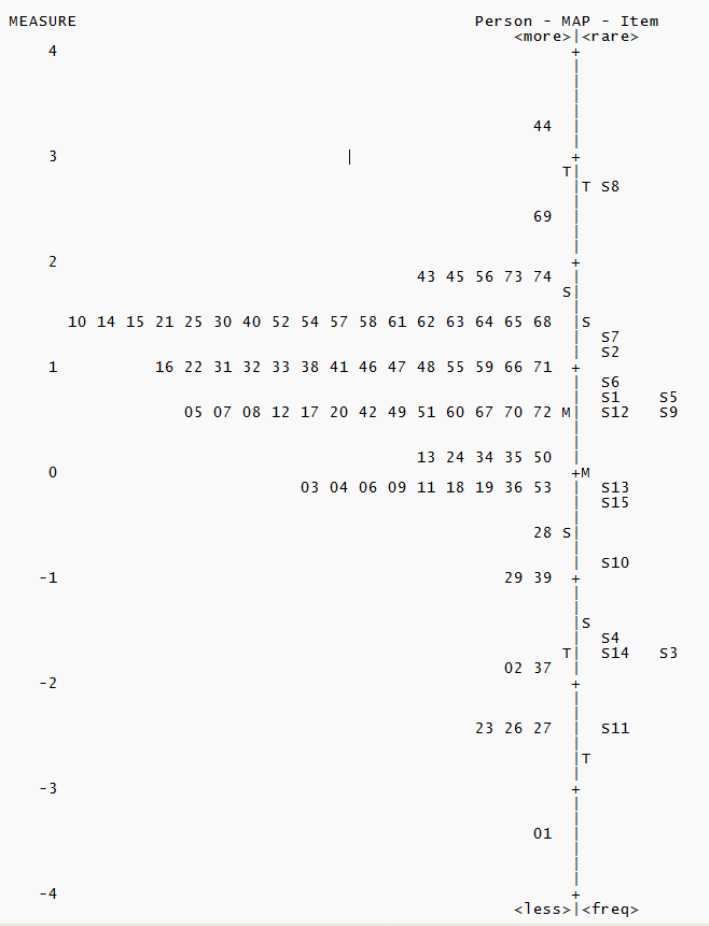

Figure 2. The Wright Map

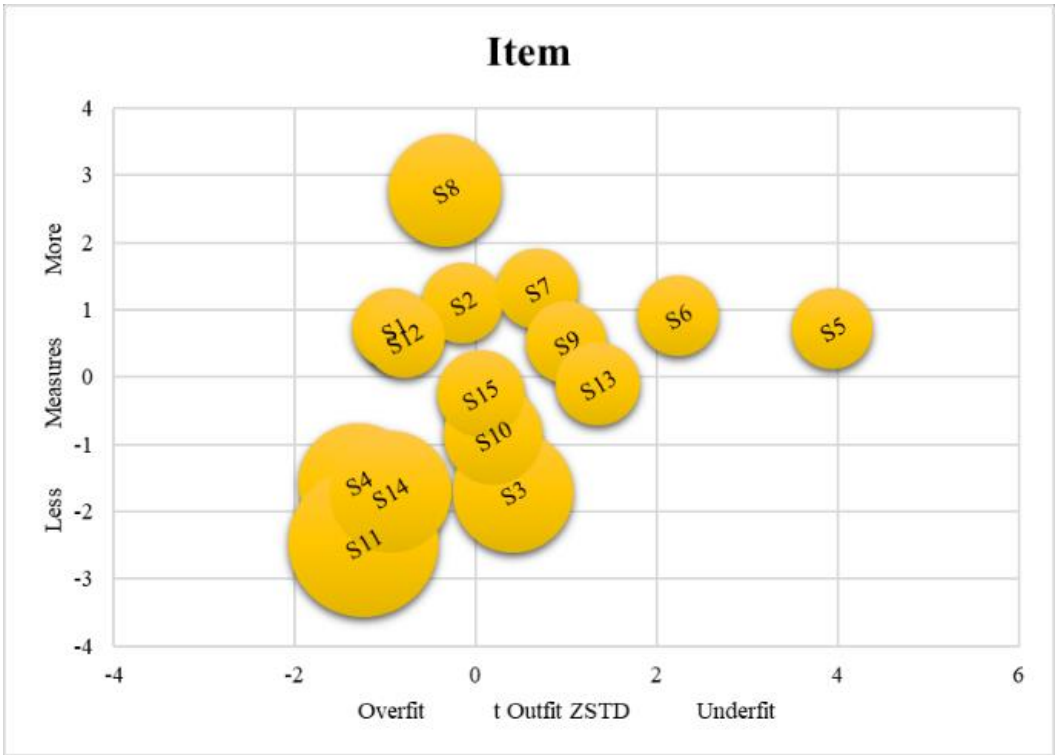

Figure 3. Level of items difficulty

The Wright Map analysis is also supported by the item difficulty plot in Figure 2. The Wright Map (named after its developer, Benjamin Wright), is a comprehensive person-item (Dahl et al., 2023; Kabic & Alexandrowicz, 2023). The largest circle is indicated by a question on number 11 (S11) which has a smaller logit value of -2. This question is in the category of the easiest questions to work on because only one student is unable to do it. Then followed by questions number 14 and 3 which have the same level of difficulty. The most difficult questions are on the top axis (S8) with logic scores greater than +2. This result is consistent with the findings of Hasanah and Purwanto (2023) and Diantoro et al. (2020) who discovered that students were unable to complete spring arrangement analysis problems. Students struggle to grasp the concept of springs, particularly springs coupled in series and parallel. Tumanggor et al. (2020) discovered that 58.3% of people had misunderstandings regarding assessing spring constants with various spring loads.

Two questions are answered as long as you guess, indicated by a circle in the underfit area, namely S6. Several questions have the same level of difficulty; S1 with S5, S9 with S9, and S3 with S14. Probability 23, 26, and 27 answered question S11 correctly 50% but certainly, S1 cannot answer the question correctly. There are 50% of students whose abilities are above average so that they can answer S1, S2, S5, 6, S7, S9, and twelve questions but are not necessarily correct about S8 questions. It is possible that student 44 can answer all the questions correctly because his ability is already above the second standard deviation.

RQ2: Is there a gender bias in the questions that have been developed?

Gender bias in test item responses was investigated using Differential Item Functioning (DIF) analysis (Figure 2). The t value of a DIF item was less than -2.0 or higher than 2.0, the DIF contrast value was less than -0.5 or greater than 0.5, and the p (probability) value was lower than 0.05 or higher than -0.05.

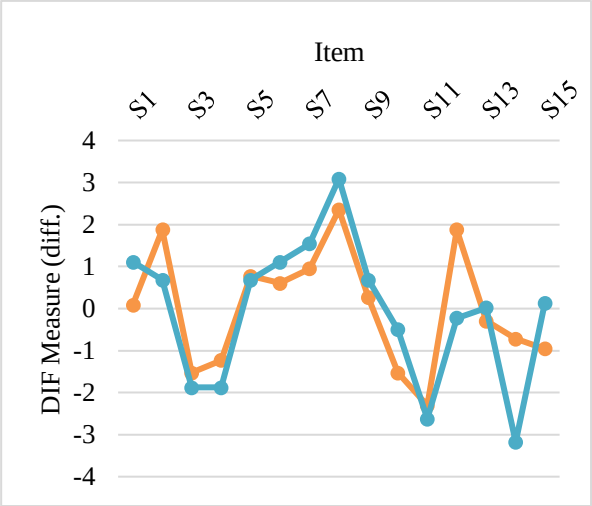

Figure 4. Person DIF plot (Blue: Female and Orange: Male)

The graph shows a curve that is close to the upper limit (as in the S8 question), indicating that S8 has a high level of difficulty, while the curve below, namely S14, shows easy questions. Item S18 asks students to determine the load ratio on a spring circuit so that it produces the same increase in length if it is arranged in series and parallel. Question S14 is the easiest question that can be answered by gender, such as calculating rubber tension. Graph 4 shows that male students excel in solving questions S1, S6, S7, S8, S10, and S15 (40%). Excellent female students solve questions S2, S4, S12, and S14 (27%). Five questions can be answered by both male and female students, namely S3, S5, S9, S11, and S13 (33%)

S12 and S14 are the two questions that include prejudice (DIF). Female pupils find it simpler to answer these two questions. This study's findings are consistent with Gladys et al. (2023) findings on gender bias in first-year multiple-choice physics tests. In their research, of the two multiple-choice questions tested, there were two that were biased towards female students. These questions are categorized as multiple-choice questions with a number (N), equations (E), concept (C), image (I), and visualization (V) scheme. The findings of this research also show that question S12 contains the characteristics of the number (N), equation (E), concept (C), and image (I) schemes, question S14 contains the characteristics of the number (N), equation (E), and concept (C) schemes. These findings can be a reference for future teachers and researchers to develop effective multiple-choice questions without gender bias. Based on Hedgeland et al. (2018), there is a modest male bias for multiple-choice questions, although the difference is not statistically significant. In other words, Hedgeland stated that there was just a hazy indication that men's scores improved as compared to women's when doing multiple-choice questions for the final exam. According to Chen et al, (2020), multiple choice provides an effective alternative method for assessors. Because multiple choice significantly increases the efficiency of assessment.

CONCLUSION AND SUGGESTION

This study was conducted on 74 pupils from a senior high school in Padang City. The findings of question one of the research revealed that the data from this study, which employed the Rasch model of measuring as the EMCI, were appropriate for outstanding measurement and productive for measurement. They were also quite reliable, with no signs of under- or over-predictability. In terms of research question number two, female students answered 53% of the questions better than male students. Furthermore, the DIF plot shows that S8 is the most difficult problem, while S14 is the easiest. There are two gender bias questions, namely S12 and S14. Both questions were easily solved by female students. However, all questions can be used to measure students' abilities in elasticity and Hooke's law. This work contributes to the field of physics assessments by confirming the EMCI using psychometric analysis and Rasch model measurement. This paper addresses a gap in the EMCI literature by providing further validation of EMCI. It also helps researchers and educators confidently employ EMCI in schools. This study will help academics and

scholars examine the capacities of EMCI in students. It provides useful information on gender differences in learning about EMCI. The data gathered will aid researchers and teachers in revising EMCI and establishing additional evaluations. The effects of this study have implications for actual data on the use of EMCI assessments in secondary schools.

AUTHORS' CONTRIBUTIONS

All contributors participated in the article's development. Vivi Mardian and Achmad Samsudin conducted research and completed the final draft of the paper. Vivi Mardian, Judhistira Aria Utama, and Irma Rahma Suwarma are responsible for carrying out revisions according to the editor's input until the publication stage.

ACKNOWLEDGMENT

We would like to express our gratitude to the Indonesian Education Fund Management Agency (LPDP) for funding our education and research.

REFERENCES

Akhtar, H., & Sumintono, B. (2023). Rasch analysis of the International Personality Item Pool Big Five Markers Questionnaire: Is longer better? *Primenjena Psihologija*, *16*(1), 3–28. https://doi.org/10.19090/pp.v16i1.2401

Al-Owidha, A. A. (2018). Investigating the psychometric properties of the Qiyas for L1 Arabic language test using a Rasch measurement framework. *Language Testing in Asia*, *8*(1), 12. https://doi.org/10.1186/s40468-018-0064-5

Borda, E., Schumacher, E., Hanley, D., Geary, E., Warren, S., Ipsen, C., & Stredicke, L. (2020). Initial implementation of active learning strategies in large, lecture STEM courses: Lessons learned from a multi-institutional, interdisciplinary STEM faculty development program. *International Journal of STEM Education*, *7*(1), 4. https://doi.org/10.1186/s40594-020-0203-2

Brassil, C. E., & Couch, B. A. (2019). Multiple-true-false questions reveal more thoroughly the complexity of student thinking than multiple-choice questions: A Bayesian item response model comparison. *International Journal of STEM Education*, *6*(1), 16. https://doi.org/10.1186/s40594-019-0169-0

Carotenuto, G., Di Martino, P., & Lemmi, M. (2021). Students' suspension of sense making in problem solving. *ZDM – Mathematics Education*, *53*(4), 817–830. https://doi.org/10.1007/s11858-020-01215-0

Chan, S.-W., Looi, C.-K., & Sumintono, B. (2021). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. *Journal of Computers in Education*, *8*(2), 213–236. https://doi.org/10.1007/s40692-020-00177-2

Chen, Q., Zhu, G., Liu, Q., Han, J., Fu, Z., & Bao, L. (2020). Development of a multiple-choice problem-solving categorization test for assessment of student knowledge structure. *Physical Review Physics Education Research*, *16*(2), 020120. https://doi.org/10.1103/PhysRevPhysEducRes.16.020120

Dahl, L. S., Staples, B. A., Mayhew, M. J., & Rockenbach, A. N. (2023). Meeting Students Where They Are: Using Rasch Modeling for Improving the Measurement of Active Research in Higher Education. *Innovative Higher Education*, *48*(3), 557–577. https://doi.org/10.1007/s10755-022-09643-4

Diantoro, M., Wartono, W., Leasa, M., & Batlolona, J. R. (2020). Students Mental Models of Solid Elasticity: A

Mixed Method Study. *Turkish Journal of Science Education*, *17*(2), 199–209. https://doi.org/10.36681/tused.2020.21

Dziob, D. (2020). Board Game in Physics Classes—A Proposal for a New Method of Student Assessment. *Research in Science Education*, *50*(3), 845–862. https://doi.org/10.1007/s11165-018-9714-y

Gamage, S. H. P. W., Ayres, J. R., & Behrend, M. B. (2022). A systematic review on trends in using Moodle for teaching and learning. *International Journal of STEM Education*, *9*(1), 9. https://doi.org/10.1186/s40594-021-00323-x

Gladys, M. J., Furst, J. E., Holdsworth, J. L., & Dastoor, P. C. (2023). Gender bias in first-year multiple-choice physics examinations. *Physical Review Physics Education Research*, *19*(2), 020109. https://doi.org/10.1103/PhysRevPhysEducRes.19.020109

Grover, R., & Wright, A. (2023). Shutting the studio: The impact of the Covid-19 pandemic on architectural education in the United Kingdom. *International Journal of Technology and Design Education*, *33*(3), 1173–1197. https://doi.org/10.1007/s10798-022-09765-y

Ha, H. T. (2021). A Rasch-based validation of the Vietnamese version of the Listening Vocabulary Levels Test. *Language Testing in Asia*, *11*(1), 16. https://doi.org/10.1186/s40468-021-00132-7

Hasanah, D., & Purwanto, J. (2023). Rasch model analysis of physics test of HOTS on the topic of elasticity and hooke's law. *THABIEA : JOURNAL OF NATURAL SCIENCE TEACHING*, *6*(1), 25. https://doi.org/10.21043/thabiea.v6i1.20422

Hedgeland, H., Dawkins, H., & Jordan, S. (2018). Investigating male bias in multiple choice questions: Contrasting formative and summative settings. *European Journal of Physics*, *39*(5), 055704. https://doi.org/10.1088/1361-6404/aad169

Henry, M. A., Shorter, S., Charkoudian, L. K., Heemstra, J. M., Le, B., & Corwin, L. A. (2021). Quantifying fear of failure in STEM: Modifying and evaluating the Performance Failure Appraisal Inventory (PFAI) for use with STEM undergraduates. *International Journal of STEM Education*, *8*(1), 43. https://doi.org/10.1186/s40594-021-00300-4

Heritage, B., Ladeira, C., & Steele, A. R. (2023). The development and pilot of the university student embeddedness (USE) scale for student retention within universities: Validation with an Australian student sample. *Higher Education*, *85*(1), 27–54. https://doi.org/10.1007/s10734-022-00813-z

Kabic, M., & Alexandrowicz, R. W. (2023). RMX/PIccc: An Extended Person–Item Map and a Unified IRT Output for eRm, psychotools, ltm, mirt, and TAM. *Psych*, *5*(3), 948–965. https://doi.org/10.3390/psych5030062

Kashihara, S., & Fukaya, T. (2022). Does a self-report questionnaire predict strategy use in mathematical problem solving among elementary school children? Importance of question format depending on the grade. *European Journal of Psychology of Education*. https://doi.org/10.1007/s10212-022-00668-z

Khan, S., & Krell, M. (2019). Scientific Reasoning Competencies: A Case of Preservice Teacher Education. *Canadian Journal of Science*,

Mathematics and Technology Education, *19*(4), 446–464. https://doi.org/10.1007/s42330-019-00063-9

Khine, M. S. (2020). Objective Measurement in Psychometric Analysis. In M. S. Khine (Ed.), *Rasch Measurement* (pp. 3–7). Springer Singapore. https://doi.org/10.1007/978-981-15-1800-3_1

Larrain, M., & Kaiser, G. (2022). Interpretation of Students' Errors as Part of the Diagnostic Competence of Pre-Service Primary School Teachers. *Journal Für Mathematik-Didaktik*, *43*(1), 39–66. https://doi.org/10.1007/s13138-022-00198-7

McKenna, P. (2019). Multiple choice questions: Answering correctly and knowing the answer. *Interactive Technology and Smart Education*, *16*(1), 59–73. https://doi.org/10.1108/ITSE-09-2018-0071

Mierke, C. T. (2020). *Cellular Mechanics and Biophysics: Structure and Function of Basic Cellular Components Regulating Cell Mechanics*. Springer International Publishing. https://doi.org/10.1007/978-3-030-58532-7

Park, M., & Liu, X. (2021). An Investigation of Item Difficulties in Energy Aspects Across Biology, Chemistry, Environmental Science, and Physics. *Research in Science Education*, *51*(S1), 43–60. https://doi.org/10.1007/s11165-019-9819-y

Pedaste, M., Baucal, A., & Reisenbuk, E. (2021). Towards a science inquiry test in primary education: Development of items and scales. *International Journal of STEM Education*, *8*(1), 19. https://doi.org/10.1186/s40594-021-00278-z

Podschuweit, S., & Bernholt, S. (2018). Composition-Effects of Context-based Learning Opportunities on Students' Understanding of Energy. *Research in Science Education*, *48*(4), 717–752. https://doi.org/10.1007/s11165-016-9585-z

Rodríguez-Mora, F., Cebrián-Robles, D., & Blanco-López, Á. (2022). An Assessment Using Rubrics and the Rasch Model of 14/15-Year-Old Students' Difficulties in Arguing About Bottled Water Consumption. *Research in Science Education*, *52*(4), 1075–1091. https://doi.org/10.1007/s11165-020-09985-z

Schwarz, G. (2023). Multiple-Choice Questions for Teaching Quantitative Instrumental Element Analysis: A Follow-Up. *Journal of Chemical Education*, acs.jchemed.3c00061. https://doi.org/10.1021/acs.jchemed.3c00061

Silva, O. H. M., Laburú, C. E., Camargo, S., & Christófalo, A. A. C. (2019). Epistemological Contributions Derived from an Investigative Method in an Experimental Class in the Study of Hooke's Law. *Acta Scientiae*, *21*(2). https://doi.org/10.17648/acta.scientiae.v21iss2id4695

Susanti, Y., Tokunaga, T., Nishikawa, H., & Obari, H. (2018). Automatic distractor generation for multiple-choice English vocabulary questions. *Research and Practice in Technology Enhanced Learning*, *13*(1), 15. https://doi.org/10.1186/s41039-018-0082-z

Suwono, H., Prasetyo, T. I., Lestari, U., Lukiati, B., Fachrunnisa, R., Kusairi, S., Saefi, M., Fauzi, A., & Atho'Illah, M. F. (2021). Cell Biology Diagnostic Test (CBD-Test) portrays pre-service teacher misconceptions about biology cell. *Journal of*

Biological Education, *55*(1), 82–105. https://doi.org/10.1080/00219266.2019.1643765

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education*, *10*(2), 159–169. https://doi.org/10.1080/0950069880100204

Tumanggor, A. M. R., Supahar, S., Ringo, E. S., & Harliadi, M. D. (2020). Detecting Students' Misconception in Simple Harmonic Motion Concepts Using Four-Tier Diagnostic Test Instruments. *Jurnal Ilmiah Pendidikan Fisika Al-Biruni*, *9*(1), 21–31. https://doi.org/10.24042/jipfalbiruni.v9i1.4571

Van Wijk, E. V., Janse, R. J., Ruijter, B. N., Rohling, J. H. T., Van Der Kraan, J., Crobach, S., Jonge, M. D., Beaufort, A. J. D., Dekker, F. W., & Langers, A. M. J. (2023). Use of very short answer questions compared to multiple choice questions in undergraduate medical students: An external validation study. *PLOS ONE*, *18*(7), e0288558. https://doi.org/10.1371/journal.pone.0288558

Vázquez-Bernal, B., & Jiménez-Pérez, R. (2023). Modeling a Theoretical Construct on Pupils' Difficulties in Problem Solving. *Science & Education*, *32*(1), 199–229. https://doi.org/10.1007/s11191-021-00289-w

Wammes, D., Slof, B., Schot, W., & Kester, L. (2022). Pupils' prior knowledge about technological systems: Design and validation of a diagnostic tool for primary school teachers. *International Journal of Technology and Design Education*, *32*(5), 2577–2609. https://doi.org/10.1007/s10798-021-09697-z

Wind, S. A., Alemdar, M., Lingle, J. A., Moore, R., & Asilkalkan, A. (2019). Exploring student understanding of the engineering design process using distractor analysis. *International Journal of STEM Education*, *6*(1), 4. https://doi.org/10.1186/s40594-018-0156-x