Kasuari: Physics Education Journal (KPEJ) 9(1) (2026) 1-10
P-ISSN: 2615-2681
E-ISSN: 2615-2673

Kasuari: Physics Education Journal (KPEJ)
Universitas Papua
website: https://journalfkipunipa.org/index.php/kpej

Analysis of Multiple Choice Questions on Impulse Momentum Material to See the Level of Difficulty of the Questions

Nur Ainayah* & Rida Siti Nur'aini Mahmudah
Department of Physics Education, Faculty of Mathematics and Natural Sciences, Yogyakarta State University, Yogyakarta, Indonesia
*Corresponding author: ainayah306@gmail.com

Abstract: *This research aims to analyze the suitability, difficulty level, and reliability of impulse-momentum test instruments in physics subjects using the Rasch model. The background to this research is based on the importance of ensuring that the assessment instruments used in education have adequate validity and reliability. Using descriptive research methods with a quantitative approach, data were collected from 275 students from three schools in Bengkulu Province. The impulse-momentum test instrument was created and validated with the help of physics teachers from each school, and the research was conducted over 5 days. Data analysis was carried out using the QUEST application to assess the suitability of the questions under the Rasch model, the level of difficulty of the questions, and the instrument's reliability. The research results show that the majority of questions are of moderate difficulty, with instrument reliability at 0.97, indicating very good reliability. However, two items were found to have outfit values exceeding 2, indicating a mismatch with the Rasch model. In conclusion, the momentum impulse test instrument has good reliability, but several items need improvement to better align with the analytical model used. This emphasizes the importance of understanding item parameters and ensuring consistency with analytical models to ensure the validity and reliability of assessment instruments in education.*

Keywords: assessment, level of difficulty, Rasch model, reliability, validity

Analisis Soal Pilihan Ganda Momentum dan Impuls untuk Melihat Tingkat Kesukaran Soal

Abstrak: Penelitian ini bertujuan untuk menganalisis kesesuaian, tingkat kesukaran, dan reliabilitas instrumen tes momentum impuls pada mata pelajaran fisika menggunakan metode model Rasch. Latar belakang penelitian ini didasarkan pada pentingnya memastikan bahwa instrumen penilaian yang digunakan dalam pendidikan memiliki validitas dan reliabilitas yang memadai. Menggunakan metode penelitian deskriptif dengan pendekatan kuantitatif, data dikumpulkan dari 275 siswa di tiga sekolah di Provinsi Bengkulu. Instrumen tes momentum impuls dibuat dan divalidasi dengan bantuan guru fisika dari masing-masing sekolah, kemudian penelitian dilakukan selama 5 hari. Analisis data dilakukan menggunakan aplikasi QUEST untuk melihat kesesuaian soal dengan model Rasch, tingkat kesukaran soal, serta reliabilitas instrumen. Hasil penelitian menunjukkan bahwa sebagian besar soal memiliki tingkat kesukaran sedang, dengan reliabilitas instrumen mencapai 0.97, sehingga menunjukkan reliabilitas yang sangat baik. Namun demikian, terdapat dua butir soal yang ditemukan memiliki nilai kesukaran melebihi 2, sehingga menunjukkan ketidaksesuaian dengan model Rasch. Kesimpulannya, instrumen tes momentum impuls memiliki reliabilitas yang baik, tetapi beberapa butir perlu ditingkatkan agar sesuai dengan model analisis yang digunakan. Hal ini menekankan pentingnya memahami parameter butir dan memastikan konsistensinya dengan model analisis untuk memastikan validitas dan reliabilitas instrumen asesmen dalam pendidikan.

Kata kunci: model Rasch, penilaian, reliabilitas, tingkat kesulitan, validitas

INTRODUCTION

Assessment is an important part of a learning process. Assessment is important to carry out because this assessment will be able to provide a clear and accurate picture of learning outcomes and facilitate effective communication between educational researchers (Palm, 2008). Assessment itself consists of two types, namely formative and summative assessment. Formative assessment is used to monitor the learning process and assist in continuous improvement and enhancement of the learning process. Formative assessment plays an important role in providing feedback to teachers, students and educational stakeholders and in improving the learning process (Dunn & Mulvenon, 2009). Meanwhile, summative assessment is important for assessing the success of learning programs, evaluating student achievement, and ensuring consistency with curriculum standards (Vero & Chukwuemeka, 2019). So, formative and summative assessments are equally important in viewing and improving the learning process.

To be able to assess well, a good assessment instrument is also needed. A good instrument can be seen from its characteristics (Elbes & Oktaviani, 2022). The characteristics of a good instrument must have high validity, meaning it measures exactly what is intended. Apart from that, it must also have high reliability, so that it provides consistent results every time it is used. Other characteristics are non-invasive, portable, cost effective, and easily accessible (Leigheb et al., 2021). Non-invasive means more accurate, repeatable with consistent results, and preferred by students and samples (Loomba & Adams, 2020). In addition, the instrument should be portable because it allows rapid evaluation without complicated protocols (Kademi et al., 2019). Good instruments must also be cost effective because this allows for the development of affordable equipment and allows innovation in design and production (Culmone et al., 2019). Finally, easy access to assessment instruments allows researchers to quickly analyze and repeat evaluations of their psychometric properties (Osborne & Fitzpatrick, 2012). Understanding the characteristics of assessment instruments well can be the basis for making assessment instruments correctly.

The instrument is made as well as possible, of course with the aim of achieving the objectives of an assessment. The purpose of assessment is to provide feedback to individuals in order to improve their learning and performance (Andrade, 2019). The general aim of the assessment is to discuss the development and application of the sampling framework in relation to qualitative evidence synthesis, as well as to evaluate the approach used in conducting qualitative evidence synthesis. In addition, the assessment also aims to address methodological issues related to the large number of studies in this synthesis and to draw lessons from the process (Ames et al., 2019). So, the purpose of assessment is to be able to get feedback from something being researched.

Educational assessment generally encompasses three domains, namely cognitive, affective, and psychomotor, and each requires different instruments to measure learning outcomes accurately (Naughton & Shumaker, 2003). Likewise with the cognitive domain. The instruments that are often used to determine students' cognitive abilities are tests. Tests are used to collect data on the development of student learning outcomes (Yuniasih et al., 2021). A good instrument must of course have certain characteristics. Likewise with cognitive test instruments. A good cognitive test instrument must be multidimensional, include subjective complaints, have been validated, be holistic, be suitable for elderly populations, be easy to use, and be able to detect changes in cognitive status over time (De-Roeck et al., 2018). A good test must be multidimensional because multidimensional experiences involve qualitative variations and different intensities as well as various emotional responses (Banzett et al., 2015). Test instruments should be able to incorporate

subjective complaints as this is important to understand the individual's overall experience, including quality of life, mental well-being and functional impact of the condition being assessed, thus providing a more complete picture and assisting in better diagnosis and intervention planning which is more appropriate (Bankstahl & Görtelmeyer, 2013). In addition, the test instrument must be holistic because it measures students' understanding of holistic competencies, which include aspects such as motivation, insurance quality, and transferable abilities that are not limited to a particular discipline. Thus, providing a more complete picture of student progress as a whole. Holistic in this context refers to a comprehensive and thorough understanding of student competence in various cognitive, emotional and social aspects (Chan & Luk, 2021). Other characteristics are the same as other instruments such as validity and reliability.

To be able to determine whether the instrument is valid and reliable, of course it is necessary to carry out a test on the test instrument that has been created. To see whether a test instrument is valid or not, it is necessary to carry out a validity test (Surucu & Maslacki, 2020). Meanwhile, to see whether an instrument is reliable or not, it needs to be tested using certain methods, for example the Rasch model. The Rasch model is a measurement model used to evaluate the quality of items in a test and obtain empirical evidence about the validity and reliability of the test (Mokshein et al., 2019). Interpretation of the results from the application of the Rasch model in the WIHIC Indonesia research will involve analysis of item fit, item size, reliability, category function, person-item map, and person size. By using the Rasch model, the research can evaluate the extent to which the WIHIC items fit the model, the level of difficulty of each item, the reliability of the scale, the function of response categories, as well as mapping the relationship between respondent characteristics and test items (Rahayu et al., 2021). From the explanation above, it can be seen that the Rasch model uses 1 logistic parameter to measure the test, namely the level of difficulty while still providing validity and reliability results (Darmana et al., 2021). So, it is necessary to understand the purpose of the research so as not to make mistakes in determining the many parameters to be measured and the type of method to measure them.

In this research on the analysis of multiple choice questions on impulse momentum material, the Rasch model method was used. This method was chosen because it is in accordance with the research objective of seeing the suitability of the question items with the chosen measurement method. Apart from that, the research also wanted to see the level of difficulty of the questions and the reliability of the questions. From this objective, it can be seen that there is only one logistics parameter that we want to measure, namely the level of difficulty. Therefore, the Rasch model method is suitable for this research.

METHOD

In this research, descriptive research methods (descriptive statistics) were used. The approach used is a quantitative approach. Descriptive research with a quantitative approach is research that aims to explain phenomena by collecting and analyzing numerical data systematically. The steps include formulating a specific hypothesis, structured research design, collecting data through instruments or surveys, statistical data analysis, and interpreting the results at the end of the study (Ahmad et al., 2019). The research was carried out using these steps. In short, starting from pre-research, research, and post-research.

Research was conducted to see the suitability and level of difficulty of the momentum and impulse test instruments. The subject of this research is physics. Especially impulse momentum material. The research used 275 samples from three schools in Bengkulu Province. Namely 106 students from SMA Negeri 8 North Bengkulu, 94 students from

SMA Negeri 2 Kaur, and 75 students from SMA Negeri 7 Bengkulu Selatan. Pre-research was carried out during May 2024. What was done was to create a test instrument and validate the test instrument. The validation carried out was content validation by 3 experts. Next, enter the research phase. The research was conducted for 5 days starting from June 3 2024 to June 7 2024. In this research, researchers were assisted by physics teachers from each school to distribute the test instruments. The last is post-research. In this phase, data analysis and data interpretation are carried out. As an output, an article related to the research was produced. The time to carry out a series of post-research activities starts from June 8 2024 to June 17 2024.

In this research, there are two variables used, namely the dependent variable and the independent variable. Independent variables refer to factors that are studied or manipulated in an experiment. Meanwhile, the dependent variable is the response or result observed as a result of changes in the independent variable (Mishra et al., 2019). In this research, the independent variable is the impulse momentum question item and the dependent variable is the level of difficulty of the question item. Because the method used is the Rasch model, the suitability of the question items with the model used and the level of reliability of the question items will also be seen. The range of item difficulty values in the Rasch Model can be adjusted to range from -2 to +2, where positive values indicate a higher level of difficulty and negative values indicate a lower level of difficulty (Zuo, 2020). The Rasch model measures the probability of answering correctly to a dichotomous question by comparing the student's ability with the difficulty level of the question, where the student has a 50% chance of answering the question correctly if the student's ability is equal to the difficulty level of the question, and the Rasch model has criteria that allow identification of incorrect responses , predicting the value of missing data, differentiating the ability of respondents with the same raw score, and identifying indications of guessing and cheating, with a range of item difficulty levels starting from -2 to +2 (Darmana et al., 2021).

RESULTS AND DISCUSSION

The application used to carry out suitability analysis between the test items and the analysis model used is the Quest application. To see the suitability between the items and the model used, we can look at the INFIT MNSQ and OUTFIT values (Arasinah et al., 2015). OUTFt t is stricter than INFIT MNSQ. So, items that have passed INFIT MNSQ do not necessarily pass OUTFIT. According to INFIT MNSQ an item is considered inappropriate if simultaneously the item's Mean Square (MNSQ) value is outside the accepted range ($0.5 < MNSQ < 1.5$), z-standard value ($|Z\text{-}STD| < 2.0$), and point correlation -measure or PMC ($0.4 < PMC < 0.8$) (Aysan, 2021). The results of the INFIT MNSQ for each item can be seen in Figure 1.

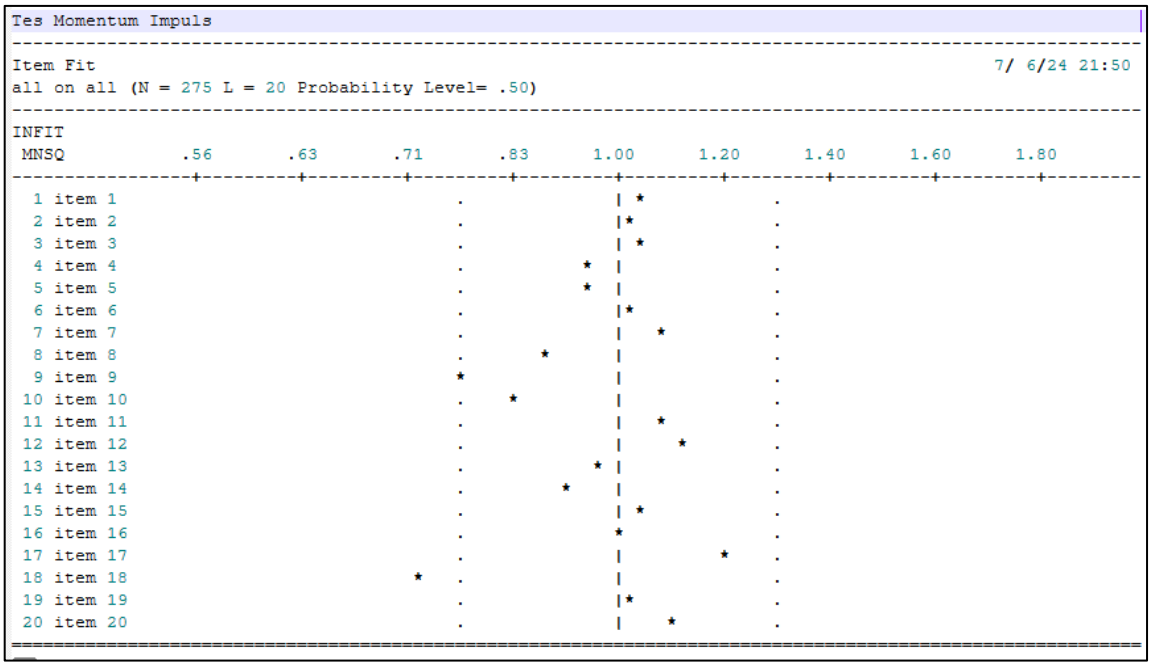

Figure 1. INFIT MNSQ Results

From Figure 1, it can be seen that there is a distribution of stars which describes the position of the item's suitability to the Rasch model used from item 1 to item 20. Then, there is a dividing line between 0.71 and 0.83 with between 1.20 and 1.40. The stars between the dividing lines indicate that the item conforms to the Rasch model. If you use a dividing line, you can see that item 18 is outside the line. However, judging from the applicable rules, item 18 is still above 0.5, meaning the item can still be said to be in accordance with the Rasch model used. So, based on the INFIT MNSQ rules, all items can still be used and are in accordance with the model used to test the items (can make a meaningful contribution to research).

To see whether an item has fallen or not, you can look at the resulting outfit. The criteria for a question to be disqualified is when it has an OUTFIT t value greater than 2. For the test results of 20 impulse momentum items, the results are as shown in Table 1.

Table 1. Result of Outfit T

Items	Outfits t	Items	Outfits t
Item 1	1.5	Item 11	1.1
Item 2	1	Item 12	1.3
Item 3	-0.5	Item 13	-0.5
Item 4	1.9	Item 14	-0.9
Item 5	-0.8	Item 15	0
Item 6	1.5	Item 16	0
Item 7	0.5	Item 17	2.8
Item 8	-1.5	Item 18	-1
Item 9	-2.4	Item 19	0.6
Item 10	-1.6	Item 20	2.8

Based on this Table 1, it can be seen that there are two questions that have an outfit greater than 2. Namely question number 18 with an outfit of 2.8 and question number 20 with an outfit of 2.8 too. This shows that questions 18 and 20 were dropped in the outfit results even though they had passed based on the INFIT MNSQ results. The failure of this item indicates that these two items cannot behave as desired by the Rasch model. Items are

also said to be disqualified because the respondent's actual answers do not match the predictions made by the Rasch model and cannot provide a meaningful contribution to the research (Daher et al., 2015). Meanwhile, other items can be used or not lost. Like items 1 and 6 have outfit t 1.5. Items 15 and 16 have outfit 0. Then there are those who get more than 1 outfit, such as item 4 with outfit 1.9, item 2 with outfit 1, item 4 with outfit 1.9, item 11 with outfit 1.1, and item 12 with outfit 1.3. For outfits under 1 there is item 3 with outfit – 0.5. Item 5 with outfit -0.8, item 7 with outfit 0.5, item 8 with outfit -1.5, item 9 with outfit -2.4, item 10 with outfit -1.6, item 13 with outfit -0.5, item 14 with outfit -0.9, item 18 with outfit-1, and item 19 with outfit 0.6. All items mentioned can make a meaningful contribution to the research and are in accordance with the Rasch model.

The level of difficulty of the questions is also seen from the results obtained from the data entered into the QUEST application. The criteria for the level of difficulty of the questions are based on the THRSHL value, with a value above 2.00 indicating a very high level of difficulty, between 1.00 and 2.00 indicating a high level of difficulty, between -1.00 and 1.00 indicating a medium level of difficulty, between -1.00 and -2.00 indicating low level of difficulty, and less than -2.00 indicates a very low level of difficulty (Yuniasih et al., 2021). For data on the level of difficulty obtained, it can be seen in Table 2.

Table 2. Difficulty Level of Question Items

Items	Difficulty Level	Items	Difficulty Level
Item 1	-2.25	Item 11	0.52
Item 2	-0,55	Item 12	-0.1
Item 3	-0.55	Item 13	-0.46
Item 4	-0.57	Item 14	-0.16
Item 5	-0.41	Item 15	0.89
Item 6	0.82	Item 16	0.94
Item 7	0.55	Item 17	0.17
Item 8	-0.26	Item 18	1.61
Item 9	0.29	Item 19	0.77
Item 10	-1.27	Item 20	-0.21

Based on the Table 2, can classify the difficulty results for each item. Firstly, a value greater than 2 means the question is very difficult and judging from the data there are no questions that are very difficult. Then, a score of 1.00 to 2.00 indicates a difficult question and in the table above there is only one question, namely question number 18. Then, a score between -1.00 to 1.00 indicates a medium level of difficulty. In the case above the question is consisting of item numbers 2, 3,4,5, 6,7,8,9,11,12,13,14,15,16,17,19, and 20. Then the value is between -1.00 up to - 2.00 indicates an easy question such as item number 10. Finally, less than -2.00 indicates a very easy question as shown in question number 1. From the data above, it can be seen that the majority of the questions are at a medium level of difficulty. The most difficult question is number 18 and the easiest question is number 1.

Reliability is the ability of a measurement to consistently rank individuals or produce consistent effects in experiments or correlational research, which is vital to ensure the accuracy and consistency of measurement results (Hedge et al., 2018). The reliability classifications are: > 0.94 (Very Good), 0.91 - 0.94 (Very Good), 0.81 - 0.90 (Good), 0.67 - 0.80 (Fair), and < 0.67 (Poorly Good) (Widyaningsih et al., 2021). The summary of item estimate statistics, including the mean measure, adjusted standard deviation, and reliability coefficient, is presented in Table 3.

Table 3. Summary of Item Estimates

Item Estimate Statistics	Value
Mean	0
elementary school	0.0
SD (adjusted)	0.88
Reliability of estimates	0.97

From the table 3, it can be seen that the Reliability of estimate is at 0.97. If we look at the previous criteria, reliability is in a very good classification because it is at a value of >0.94. This means that this question has a very good ability to measure students' abilities consistently over different periods of time.

When analyzing test items, the logistical parameters you want to measure must first be determined. This is because the number of logistic parameters to be tested determines the type of analysis application used and the analysis model used. The logistic parameters of the test items (item response theory) is a theory that can predict test participants' responses to each question item. Even if these items have never been seen by test participants before. This theory is used in test development, test adjustment, and detection of differences in item function in various educational, military, and industrial institutions (Hedge et al., 2018). In the case of this research, the research was carried out using 1 logistic parameter, namely the level of difficulty of the questions. From the results obtained, the average question item is at a moderate level of difficulty, namely in the range -1.00 to 1.00. and the most difficult question is question number 18, while the easiest question is number 1. Apart from the level of difficulty, other logistical parameters that are often considered in item analysis include discrimination and guessing parameters. An in-depth understanding of these parameters allows researchers to better understand the characteristics of the test items and increase the validity and reliability of the measurements (Embretson & Reise, 2013).

Apart from difficulty analysis, the Rasch model also explains the questions that are appropriate to the analysis model used. Analysis using the Rasch model is an important approach in checking the consistency of test items with the model used (Sumintono & Widhiarso, 2015). Infit Mean Square (MNSQ) and outfit Mean Square (MSQ) are metrics commonly used to evaluate the suitability of test items to the Rasch model. According to general rules, MNSQ values that are outside the ideal range (usually between 0.7 to 1.3) can indicate a mismatch between the items and the model used (Şahin & Anıl, 2017). The results of the MNSQ analysis on item number 18 which shows a value below 0.73 suggests that there is a potential problem in this item. Apart from that, outfit statistics also provide information about the suitability of the question items to the Rasch model. Outfit statistical testing that shows extreme values or outside the expected range, such as items number 18 and 20 which were eliminated, can indicate the existence of items that are inconsistent with the Rasch model (Wright & Masters, 1982). The existence of question items that do not conform to the Rasch model can reduce the validity and reliability of the test results. Lastly, the thing tested is reliability. Reliability is crucial in determining whether an evaluation instrument, such as a test, can provide consistent and reliable results in measuring the same construct at different times. The higher the reliability value of a test, the greater the confidence that the test will provide stable and consistent results for test participants (Hidayati & Listyani, 2010). In the case studied, the question reliability value reached 0.97, indicating a very good or high level of reliability, in accordance with the standards expected for an evaluation instrument.

CONCLUSION AND SUGGESTIONS

Analysis of test items using the Rasch model provides an in-depth understanding of the consistency of test items with the model used, especially through metrics such as INFIT MNSQ and OUTFIT. Even though several items passed the INFIT MNSQ, it was found that several items still failed based on the OUTFIT results, indicating a discrepancy with the Rasch model. Two questions were found, namely items number 18 and 20, which were dropped based on OUTFIT with a value of 2.8. However, the majority of other questions can be used well. In addition, the level of difficulty of the questions varies, with the majority of questions being at a medium level of difficulty. However, there is one question item that is very difficult, namely item number 18, and one item that is very easy, namely item number 1. However, the reliability value of the evaluation instrument which reaches 0.97 shows a very good level, giving confidence that the test provides consistent and reliable results in measuring the same construct at different times. Thus, this analysis makes a meaningful contribution in improving the quality of evaluation instruments and the validity of measurement results.

This study has limitations in the relatively small number of multiple-choice questions on momentum and impulse used, as well as the respondent sample being limited to one or a few schools, so the results cannot be generalized widely. Additionally, the analysis employed the standard Rasch model assuming unidimensionality, whereas the understanding of momentum and impulse concepts may be multidimensional. Therefore, future research is recommended to expand the number and variety of questions, involve a more diverse sample, and consider using a multidimensional Rasch model for deeper analysis. Integrating qualitative methods can also help to gain a more comprehensive understanding of students' difficulties.

ACKNOWLEDGMENTS

The researchers would like to thank the three physics teachers who helped me in distributing the questions, namely Diyah Uswatun Hasanah, Lola Dwi Syafitri, and Wahyu Apridonata. The researcher also would like to thank the UNY Physics Education Masters Lecturers who have supported this activity.

BIBLIOGRAPHY

Ahmad, S., Wasim, S., Irfan, S., Gogoi, S., Srivastava, A., & Farheen, Z. (2019). Qualitative v/s. Quantitative Research-A Summarized Review. *Journal of Evidence Based Medicine and Healthcare*, *6*(43), 2828–2832. https://doi.org/10.18410/jebmh/2019/587

Ames, H., Glenton, C., & Lewin, S. (2019). Purposive Sampling In A Qualitative Evidence Synthesis: A Worked Example From A Synthesis on Parental Perceptions of Vaccination Communication. *BMC Medical Research Methodology*, *19*(1), 1–9. https://doi.org/10.1186/s12874-019-0665-4

Andrade, H. L. (2019). A Critical Review of Research on Student Self-Assessment. *Frontiers in Education*, *4*(August), 1–13. https://doi.org/10.3389/feduc.2019.00087

Arasinah, K., Bakar, A. R., Ramlah, H., Soaib, A., & Zaliza, H. (2015). Using Rasch Model and Confirmatory Factor Analysis to Assess Instrument for Clothing Fashion Design Competency. *International Journal of Social Science and Humanity*, *5*(5), 418–421. https://doi.org/10.7763/ijssh.2015.v5.492

Bankstahl, U. S., & Görtelmeyer, R. (2013). Measuring Subjective Complaints of Attention and Performance Failures-Development and Psychometric Validation in Tinnitus of The Self-Assessment Scale APSA. *Health and Quality of Life Outcomes*,

11(1), 1–12. https://doi.org/10.1186/1477-7525-11-86

Banzett, R. B., O'Donnell, C. R., Guilfoyle, T. E., Parshall, M. B., Schwartzstein, R. M., Meek, P. M., Gracely, R. H., & Lansing, R. W. (2015). Multidimensional dyspnea profile: An instrument for clinical and laboratory research. *European Respiratory Journal*, *45*(6), 1681–1691. https://doi.org/10.1183/09031936.00038914

Chan, C. K. Y., & Luk, L. Y. Y. (2021). Development and Validation of An Instrument Measuring Undergraduate Students' Perceived Holistic Competencies. *Assessment and Evaluation in Higher Education*, *46*(3), 467–482. https://doi.org/10.1080/02602938.2020.1784392

Culmone, C., Smit, G., & Breedveld, P. (2019). Additive Manufacturing of Medical Instruments: A State-of-The-Art Review. *Additive Manufacturing*, *27*(October 2018), 461–473. https://doi.org/10.1016/j.addma.2019.03.015

Darmana, A., Sutiani, A., Nasution, H. A., 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

De-Roeck, E. E., Dury, S., De Witte, N., De Donder, L., Bjerke, M., De Deyn, P. P., Engelborghs, S., & Dierckx, E. (2018). CFAI-Plus: Adding Cognitive Frailty As A New Domain to The Comprehensive Frailty Assessment Instrument. *International Journal of Geriatric Psychiatry*, *33*(7), 941–947. https://doi.org/10.1002/gps.4875

Elbes, E. K., & Oktaviani, L. (2022). Character Building in English for Daily Conversation Class Materials for English Education Freshmen Students. *Journal of English Language Teaching and Learning*, *3*(1), 36–45. https://doi.org/10.33365/jeltl.v3i1.1714

Embretson, S. E., & Reise, S. P. (2013). Item Response Theory for Psychologists. *Item Response Theory for Psychologists*, 1–371. https://doi.org/10.4324/9781410605269

Hedge, C., Powell, G., & Sumner, P. (2018). The Reliability Paradox: Why Robust Cognitive Tasks Do Not Produce Reliable Individual Differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Kademi, H. I., Ulusoy, B. H., & Hecer, C. (2019). Applications of Miniaturized and Portable Near Infrared Spectroscopy (NIRS) For Inspection and Control of Meat and Meat Products. *Food Reviews International*, *35*(3), 201–220. https://doi.org/10.1080/87559129.2018.1514624

Leigheb, M., de Sire, A., Colangelo, M., Zagaria, D., Grassi, F. A., Rena, O., Conte, P., Neri, P., Carriero, A., Sacchetti, G. M., Penna, F., Caretti, G., & Ferraro, E. (2021). Sarcopenia Diagnosis: Reliability of The Ultrasound Assessment of The Tibialis Anterior Muscle As An Alternative Evaluation Tool. *Diagnostics*, *11*(11), 1–10. https://doi.org/10.3390/diagnostics11112158

Loomba, R., & Adams, L. A. (2020). Advances in Non-Invasive Assessment of Hepatic Fibrosis. *Gut*, *69*(7), 1343–1352. https://doi.org/10.1136/gutjnl-2018-317593

Mishra, P., Singh, U., Pandey, C. M., Mishra, P., & Pandey, G. (2019). Application of Student's t-test, Analysis of Variance, and Covariance. Annals of Cardiac Anaesthesia, 22(4), 407–411. https://doi.org/10.4103/aca.ACA_94_19

Mokshein, S. E., Ishak, H., & Ahmad, H. (2019). The Use of Rasch Measurement Model In English Testing. *Cakrawala Pendidikan*, *38*(1), 16–32. https://doi.org/10.21831/cp.v38i1.22750

Naughton, M. J., & Shumaker, S. A. (2003). The Case for Domains of Function in Quality of Life Assessment. Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 12 Suppl 1, 73–80.

https://doi.org/10.1023/a:1023585707046

Rahayu, W., Putra, M. D. K., Rahmawati, Y., Hayat, B., & Koul, R. B. (2021). Validating an Indonesian Version of The What Is Happening in This Class? (Wihic) Questionnaire Using A Multidimensional Rasch Model. *International Journal of Instruction*, *14*(2), 919–934. https://doi.org/10.29333/iji.2021.14252a

Şahin, A., & Anıl, D. (2017). The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory. *Kuram ve Uygulamada Egitim Bilimleri*, *17*(1), 321–335. https://doi.org/10.12738/estp.2017.1.0270

Surucu, L., & Maslacki, A. (2020). Validity and Reliability in Quantitative Research. *Business & Management Studies: An International Journal*, *8*(3), 2694–2726. https://doi.org/10.15295/bmij.v8i3.1540

Vero, M., & Chukwuemeka, O. A. (2019). Formative and Summative Assessment: Trends and Practices in Basic Education. *Journal of Education and Practice*, 1–19. https://doi.org/10.7176/jep/10-27-06

Widyaningsih, S. W., Yusuf, I., Prasetyo, Z. K., & Istiyono, E. (2021). The Development of the HOTS Test of Physics Based on Modern Test Theory: Question Modeling Through E-Learning of Moodle LMS. *International Journal of Instruction*, *14*(4), 51–68. https://doi.org/10.29333/iji.2021.1444a

Yuniasih, N. K., Yudiana, K., & Japa, I. G. N. (2021). The Concept of Heat Transfer measured by Cognitive Domain Assessment Instruments. *Jurnal Ilmiah Sekolah Dasar*, *5*(1), 140. https://doi.org/10.23887/jisd.v5i1.34328

Zuo, Y. (2020). A Comprehensive Simulation Study of Estimation Methods for the Rasch Model. *Stats*, *3*(2), 94–106. https://doi.org/10.3390/stats3020009