OPEN ACCESS
Jurnal Ilmiah Ilmu Terapan Universitas Jambi
Vol. 10 No. 1. Ferbruary 2026, pp. 29-39 DOI. 10.22437/jiituj.v10i1.41234
Research Article

DIFFERENTIAL ITEM-PERSON FUNCTIONING (DIPF) ON QUIZIZZ-ASSISTED PHYSICS MEASUREMENT QUESTIONS: A RASCH MODEL ANALYSIS

Mohd Zaidi Bin Amiruddin[1], Achmad Samsudin[1,*], Andi Suhandi[1], Nila Apriliyanti[3], Bayram Costu[3]

[1] Faculty of Mathematics and Science Education, Universitas Pendidikan Indonesia, Jawa Barat, Indonesia
[2] Al-Islam Krian High School, Jawa Timur, Indonesia
[3] Department of Science Education, Yildiz Technical University, Istanbul, Turkey
Corresponding author email: achmadsamsudin@upi.edu

Article Info
Received: Jan 19, 2025
Revised: Aug 17, 2025
Accepted: Feb 02, 2026
OnlineVersion: Feb 19, 2026

Abstract
Assessment and measurement have always been crucial topics, especially in physics education, where accurate evaluation is needed to measure students' understanding and mastery of the subject. This study tested the validity and reliability of physics measurement questions administered through the Quizizz platform and identified Differential Item-Person Functioning (DIPF) using the Rasch model-assisted Winstep software. This research design used Item Response Theory (IRT). The study involved 34 high school students from Sidoarjo, East Java, Indonesia. The instrument consisted of 15 multiple-choice questions on basic physics measurements. The results showed that the instrument had good construct validity, with a raw variance explained by the measurement of 22.8%, indicating an effective measure to gauge students' ability. Reliability analysis showed moderate consistency, with a Cronbach Alpha of 70%, although person and item reliabilities were weaker at 63% and 46%, respectively. DIF analysis showed no significant gender bias. Future research should improve the instrument's reliability and consider a broader range of external factors to understand student performance comprehensively.

Keywords: Gender, Measurement, Physics, Rasch Model, Quizizz.

INTRODUCTION

Assessment and measurement have always been something crucial to discuss. In physics learning, accurate assessment is needed primarily to measure students' abilities and understanding related to mastery of the material and concepts learned (Kauertz & Fischer, 2006; Mi et al., 2023; Pals et al., 2023). In this case, the material tested is essential, namely physics measurement with the help of the Quizizz platform. The rapid development of technology also affects the course of education, which often integrates technology into learning. One of the technologies utilized in the assessment process is Quizizz. According to Moreira and Freire (2024); Zainuddin et al. (2020), digital platforms like Quizizz provide an interactive and efficient way to measure student knowledge. In this study, Quizizz was used to simplify the process of completing questions and conducting evaluations, with results that could be automatically

mapped immediately. Thus, Quizizz not only helped improve efficiency in the assessment process but also provided a quicker overview of the results of the questions given. However, alongside these benefits, new challenges arise regarding the accuracy and fairness of assessment, particularly in ensuring students' honesty in answering questions. These challenges highlight the need for careful consideration and continuous improvement when integrating new technologies like Quizizz into the assessment process.

Assessment and measurement must be distinct from every process, especially in learning. According to Pellegrino (2014); Intasoi et al. (2020), the progress and process of teaching and learning can be evaluated through assessment and measurement. Question instruments always follow the assessment process in essays, multiple choice, multitier, and level of understanding. That way, the instruments must be valid and reliable so as not to bias the question items and the person. The validity and reliability of questions ensure that the feedback given to students is accurate and valid (Harlen, 2005; Kinyua & Okunya, 2014). In addition, it also ensures that the assessment is conducted fairly for both teachers and learners. That way, decision-making has a strong basis related to policies that will be carried out in the future.

One aspect that often becomes a focus in the validity and reliability of the instrument is bias in the items or persons being tested. Item bias can occur when questions or tests systematically favor or disadvantage certain groups of learners, not based on differences in the abilities being measured but because of other characteristics such as cultural background, language, gender, or other factors (Willingham & Cole, 2013; Reynolds et al., 2021). In modern measurement, item-person bias can be measured using the Rasch model with the Differential item functioning (DIF) view. According to Goldhammer (2015), Differential Item-Person Functioning (DIPF) is an analytical method used to detect the presence of bias in judgment. Using the Rasch model, DIPF can evaluate how each item in a test functions differently for different groups of students based on specific characteristics (e.g, gender, cultural backgroun, and demographic factor) (Khalid, 2023). This analysis is critical in using digital platforms such as Quizizz, as it helps to ensure that the assessments made are free from bias and reflect the true abilities of all students.

Previous research has conducted item validity and reliability (Song et al., 2023; Dianovinina et al., 2024; Lechien et al., 2024; Razali et al., 2024), as well as bias detection (Lee & Geisinger, 2014; Lupi et al., 2017; Büyükkıdık, 2023). These studies have provided important insights into the validity, reliability, and detection of bias in assessment instruments, particularly related to gender and domicile. However, they have not fully explored the integrated relationship between validity, reliability, and item-person bias within a single framework using the Rasch Model supported by Winstep software. This study contributes by addressing this gap: not only does it evaluate the validity and reliability of physics exam questions, but it also investigates DIF to detect potential gender-based bias. By using Winstep software for Rasch analysis and administering questions through the Quizizz platform, this research offers a unique perspective that connects psychometric rigor with the practical use of the Quizizz digital assessment tool. That way, this study measures the validity and reliability of the tested question instruments and identifies DIPF in physics questions with the help of the Quizizz platform. The questions that will be answered in this study are as follows:

RQ1: What is the validity and reliability of questions on physics measurement material?

RQ2: What is the item, person, and gender bias of the physics measurement test?

RESEARCH METHOD

This research design uses Item Response Theory (IRT). According to Van der Linden and Hambleton (1997); Maldonado-Murciano et al. (2023), IRT is a theoretical and methodological framework in psychometrics used to design, analyze, and interpret tests and questionnaires. IRT was chosen because it can model the relationship between individuals' latent abilities (traits) and their responses to the items in a test. In addition, IRT makes it possible to understand the characteristics of each item in the test as well as individual abilities in more detail. This is in line with Rasch analysis, which can present the suitability of items with persons (e.g., Van Zile-Tamsen, 2017; Müller, 2020; Nisa et al., 2024). In addition, IRT has been effectively applied to analyze the quality of test items, providing more in-depth information about the level of difficulty, discriminating power, and guessing opportunities for each item (e.g., Quaigrain & Arhin, 2017; Ranyard et al., 2020). Through IRT, it is possible to evaluate model fit and analyze DIF to ensure measurement fairness (Oliveri et al., 2016; Martinková et al., 2017; Bauer, 2023). The results of the analyses are then interpreted to identify items that need revision and

ensure the validity and reliability of the instrument, with the ultimate goal of producing an accurate, valid, and reliable test.

A total of 15 questions on physics measurement material were tested in the form of multiple choice through the help of Quizizz. The questions tested are divided into several parts: an international system of quantities, units, and measurement dimensions. This study involved 34 high school students from Sidoarjo, East Java, Indonesia. A total of 15 questions were selected because they were considered to represent the basic measurement topics contained in the learning outcomes. This number was deemed sufficient to obtain a proportional picture of students' abilities without causing excessive cognitive load. Before being administered to students, the questions underwent content validation by several experts in the field of physics education to ensure the appropriateness of the material, clarity of language, and relevance to learning outcomes. Students were given 45 minutes to work on the questions. Measurement was selected as the instructional material because it is the first topic introduced to upper secondary students at the introductory level and serves as a foundational concept that is continuously applied in subsequent physics learning. One example of the question was used in this study is presented in Figure 1.

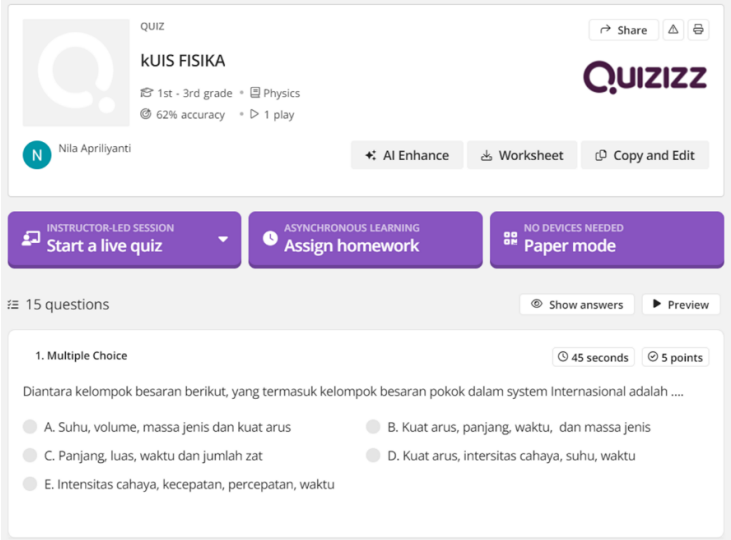

Figure 3. Examples of questions in the study

The data that has been obtained is corrected with a Guttman scale: true is worth 1, and false is worth 0. Then it is mapped in Microsoft Excel. The data used has been adjusted to the “formatted text” format. Then, coding for male (L) and female (P) gender categories. Meanwhile, the question items were coded (Q1-Q15). The validity and reliability of the instrument were also measured using the Rasch Model with principal component analysis. After that, DIPF also used the help of Rasch analysis software Winstep. The criteria for bias in items for DIPF can be known when the probability value of an item is below 5% (e.g., Köhler et al., 2020; Fährmann et al., 2022). The criteria for validity, reliability, and Cronbach alpha values are presented in Table 1, Table 2, and Table 3.

Table 1. Interpretation of instrument validity

Interpretation	Raw variance explained by measures
Fulfilled	>20%
In accordance	>40%
Special	>60%

Table 2. Interpretation of person and item reliability (see. Sumintono & Widhiarso, 2014)

Value Range	Interpretation
value ≥ 0.95	Excellent
0.91 < value ≤ 0.95	Very Good
0.81 < value ≤ 0.91	Good
0.68 < value ≤ 0.81	Moderate
value < 0.68	Weak

Table 3. Interpretation of Cronbach's Alpha values (see. Sumintono & Widhiarso, 2014)

Cronbach Alpha Range (α)	Interpretation
$\alpha \geq 0.8$	Very Good
$0.7 < \alpha \leq 0.8$	Good
$0.6 < \alpha \leq 0.7$	Moderate
$0.5 < \alpha \leq 0.6$	Bad

RESULTS AND DISCUSSION

Validity of questions on physics measurement material

The validity of the tested question instrument is analyzed using the Rasch Model with Winstep software. The output table used to determine the validity of the item-person instrument is dimensionality. According to Clark and Watson (2019), the instrument's unidimensionality is a parameter that can be used to measure the construction of the question instrument related to measuring what should be measured. The results of unidimensionality are presented in Figure 2.

STANDARDIZED RESIDUAL variance in Eigenvalue units = Person information units

	Eigenvalue	Observed		Expected
Total raw variance in observations =	44.0238	100.0%		100.0%
Raw variance explained by measures =	10.0238	22.8%		23.3%
Raw variance explained by persons =	7.5080	17.1%		17.5%
Raw Variance explained by items =	2.5158	5.7%		5.9%
Raw unexplained variance (total) =	34.0000	77.2%	100.0%	76.7%
Unexplned variance in 1st contrast =	5.5332	12.6%	16.3%	
Unexplned variance in 2nd contrast =	5.1282	11.6%	15.1%	
Unexplned variance in 3rd contrast =	4.0516	9.2%	11.9%	
Unexplned variance in 4th contrast =	4.0388	9.2%	11.9%	
Unexplned variance in 5th contrast =	3.4574	7.9%	10.2%	

Figure 2. Instrument validity

Figure 2 summarises the results of the Rasch analysis, which focuses explicitly on the standardized residual variance in eigenvalue units, which is crucial for assessing the dimensionality of the measurement instrument. The total raw variance in the observations is 44.0238 (100%), which serves as the baseline. The raw variance explained by the measurement is 10.0238, representing 22.8% of the total variance, slightly less than the expected 23.3%. This indicates that most of the variance is due to the primary constructs measured, indicating a good model fit. The raw variance explained by the person was 7.5080, accounting for 17.1%, close to the expected 17.5%, highlighting the differences in individual ability levels. The raw variance explained by items was 2.5158 (5.7%), also close to the expected 5.9%, indicating that items make a relatively small but meaningful contribution to the variance.

The total raw unexplained variance was 34.0000 (77.2%), which is typical as it includes all the residual variance not explained by the main dimensions. The unexplained variance in the first contrast was 5.5332 (12.6%), smaller than the expected 16.3%, indicating the presence of potential secondary dimensions, although not very strong. The second contrast showed an eigenvalue of 5.1282 (11.6%), again less than the expected 15.1%, indicating a weaker secondary dimension. The third contrast has an eigenvalue of 4.0516 (9.2%), slightly below the 11.9% expectation, and the fourth contrast is similar to an eigenvalue of 4.0388 (9.2%). The fifth contrast shows the weakest potential of the secondary dimension with an eigenvalue of 3.4574 (7.9%), less than the expected 10.2%.

In conclusion, the primary dimension explains most of the variance (22.8%), close to the expected 23.3%, indicating a good model fit and unidimensionality. The presence of contrasts indicates the presence of potential secondary dimensions, but they are generally weaker than expected, indicating that they do not strongly influence the primary dimension. Based on Figure 2, the raw variance value explained by the measurement data is 22.8%. This means that the instrument used to test students' abilities in physics measurement material is in the fulfilled category (>20%).

Reliability of questions on physics measurement material

Through Rasch analysis with summary statistic output, the reliability of items and persons and interactions between persons can be explored. The overall output is presented in Figure 3.

(a) SUMMARY OF 34 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	10.1	15.0	.94	.68	1.00	.17	.95	.12
SEM	.5	.0	.21	.03	.02	.11	.04	.12
P.SD	3.0	.0	1.18	.18	.13	.65	.22	.70
S.SD	3.0	.0	1.20	.18	.13	.66	.23	.71
MAX.	14.0	15.0	2.78	1.05	1.37	2.45	1.48	2.57
MIN.	1.0	15.0	-2.79	.54	.79	-1.27	.49	-1.22

REAL RMSE .72 TRUE SD .94 SEPARATION 1.31 Person RELIABILITY .63
MODEL RMSE .70 TRUE SD .95 SEPARATION 1.36 Person RELIABILITY .65
S.E. OF Person MEAN = .21

(b) SUMMARY OF 15 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	23.0	34.0	.00	.42	1.00	.06	.95	-.05
SEM	.9	.0	.16	.01	.05	.27	.09	.27
P.SD	3.4	.0	.60	.04	.17	1.02	.33	.99
S.SD	3.5	.0	.62	.04	.18	1.06	.34	1.03
MAX.	29.0	34.0	1.26	.54	1.34	2.12	1.72	1.97
MIN.	15.0	34.0	-1.21	.39	.77	-1.57	.56	-1.32

REAL RMSE .44 TRUE SD .41 SEPARATION .93 Item RELIABILITY .46
MODEL RMSE .43 TRUE SD .42 SEPARATION .99 Item RELIABILITY .50
S.E. OF Item MEAN = .16

Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .70 SEM = 1.65

Figure 3. Instrument reliability

The provided Figure summarizes the Rasch analysis results for 34 measured persons and 15 measured items, offering insights into the reliability and fit of the measurement model. For the person measurement, the descriptive statistics reveal a mean total score of 10.1 with a standard error of the mean (SEM) of 0.5, and a standard deviation (SD) of 3.3, indicating a reasonable variation in person ability. The mean measure is 0.94, suggesting participants scored higher than the average item difficulty, with a spread (SD) of 0.94. Fit statistics, including infit and outfit mean squares (MNSQ), hover around 1.0, indicating an acceptable fit, though the slightly higher outfit MNSQ hints at some unexpected responses.

The reliability and separation indices show a true person reliability of 0.63, indicating moderate consistency in the ability measures of people who fall into the weak category (<67%) and separation of 1.31, indicating that the sample can be divided into approximately 1.3 different ability levels. The reliability of the person model was slightly higher at 0.65. The overall reliability was moderate, reflected by the correlation of raw scores to measures of 0.90 and cronbach alpha (KR-20) of 0.70. Descriptive statistics showed a mean total score of 23.0 for the item measures, with an SEM of 0.9 and an SD of 3.4, indicating variability in item difficulty. A mean measure of 0.00 is typical, as item measures are usually standardized around zero, with an SD of 0.41. The fit statistics indicated a good fit, with MNSQ infit and outfit values close to 1.0. However, the lower item reliability of 0.46 < 0.67 indicates a less consistent measure of item difficulty.

Item, person, and gender bias from the physics measurement test results

The presented person item bias refers to gender, male (L) and female (P). It is said to experience item bias against gender when the probability value is less than 0.05 or 5%. The gender DIF results are presented in Figure 4.

DIF class/group specification is: DIF=$S3W1

Person CLASSES	SUMMARY DIF CHI-SQUARED	D.F.	PROB.	BETWEEN-CLASS/GROUP Item UNWTD MNSQ	ZSTD	Number	Name
2	.0000	1	1.0000	.0000	.0000	1	Q1
2	.2945	1	.5873	.3082	-.22	2	Q2
2	.1717	1	.6786	.1791	-.45	3	Q3
2	.0950	1	.7579	.0993	-.67	4	Q4
2	1.1587	1	.2817	1.2470	.63	5	Q5
2	1.6004	1	.2058	1.7617	.91	6	Q6
2	1.9406	1	.1636	2.1480	1.09	7	Q7
2	.0135	1	.9074	.0140	-1.14	8	Q8
2	.6129	1	.4337	.6484	.19	9	Q9
2	.0819	1	.7748	.0849	-.72	10	Q10
2	.0236	1	.8779	.0247	-1.03	11	Q11
2	1.9514	1	.1624	2.1542	1.09	12	Q12
2	.0819	1	.7748	.0849	-.72	13	Q13
2	.3078	1	.5790	.3230	-.19	14	Q14
2	2.4126	1	.1204	2.7778	1.33	15	Q15

Figure 4. DIPF on Gender

Figure 4 presents summarising the DIF analysis. Variations between the lines indicate the item difficulty for the male and female genders, with notable differences indicating potential DIF, where certain items may be more challenging for one group compared to another. Figure 4 (right side) provide detailed statistical measures for each item, including summary DIF, chi-square values, unweighted mean squares (UNWTD MNSQ), and standardized residuals (ZSTD). Based on the probability values presented, it states that there is no item bias towards gender. This refers to the probability of each item not being below 0.05 or 5%. The combination of DIF plots (left side) and statistical (right side) highlighted

items with significant DIF, both visually and statistically, indicating that these items may require review or adjustment to ensure fairness for further review.

Differential Item-Person Functioning (DIPF) is essential to understand and explore in identifying item-person bias, especially against gender. However, the identification process must undoubtedly be structured in line with the validity and reliability of the instrument that has been made. Instrument validity usually only relies on content validity through expert validators. It is essential to do content and construct validity based on the results of responses from limited trials before being carried out on a larger scale. According to MacKenzie et al. (2011), content and construct validity are essential to ensuring the quality of research instruments, but they have different purposes and are not interchangeable. Content validity evaluates how well a test or instrument covers all relevant aspects of the construct it is intended to measure, whereas construct validity measures how well a test measures the concept or construct it is designed to measure (Cook & Beckman, 2006; Cheung et al., 2023).

Construct validity is more important in certain situations, such as when researching intangible or abstract concepts that cannot be measured directly (Molloy et al., 2011; Mian et al., 2020). Construct validity is essential to ensure the test accurately measures the intended concept (Clark & Watson, 2019; Owan et al., 2023). Conversely, content validity is more relevant when the construct is more tangible and can be measured directly, such as in tests on products. This study focuses on construct validity, which measures students' abilities and knowledge related to physics measurement materials such as international units, area measurement, volume, and dimensions. With the help of Rasch analysis, content and construct validity can be determined so that the validity results are powerful (Boone et al., 2013; Amiruddin et al., 2023). The results of construct validity with what is done refer to unidimensionality, which gets a raw variance measured value of 22.8% in the fulfilled category. According to Yildiz and Kara (2017), the unidimensionality of the instrument can be a parameter used to assess whether the instrument effectively measures the construct that should be measured.

In addition, more than validity is required if it is complemented by instrument reliability. In determining reliability, many ways can be used, namely by test-retest (Polit, 2014), parallels method (Madansky, 1965; Chen et al., 2022), split half method (Chakrabartty, 2013; Pronk et al., 2022) and Rasch analysis (Van Zile-Tamsen, 2017; Cordier et al., 2018). This study uses Rasch analysis as a way to determine reliability because it is able to measure item, person, and item-person. Based on the measurement results, person reliability obtained a value of 63% and item reliability of 46% in the weak category. As for the results of inter-item-person measurements obtained through the Cronbach alpha (KR-20) value of 70% in the sufficient category. Referring to this category states that the consistency of student answers to items is weak. The sufficient category needs to be improved so that the quality of validity and reliability is able to measure what should be measured, and the consistency is at least in the excellent category (Kimberlin & Winterstein, 2008; Bouwer et al., 2023). Rasch analysis can identify items against gender.

Through Rasch analysis, we be able to identify the item on gender. Based on the results of the person DIF plot (DIF = $S3W1), which is the code for male (L) and female (P) gender. The results presented in Figure 4 state that the abilities of men and women differ, but further analysis states that there is no probability value below 5%. This means there is no item bias against both female and male gender. According to Saxena et al. (2001); Lohr (2002), DIF, which refers to gender is crucial to be researched to improve the quality of test instruments and ensure that the results obtained genuinely reflect the abilities or characteristics to be measured so that the test results are fair and equal. Further, DIF analysis also be able to look at the domicile of students (Kusuma et al., 2022; Saputro, 2022). However, in the study, the students came from the same area and domicile and were of the same ethnicity.

Although content and construct validity have been evaluated and shown promising results, and there is no item bias against gender-based on DIF analysis, other aspects are also essential, namely learning environment factors and socioeconomic influences on student performance (Rabgay, 2015; Akukwe & Schroeders, 2016; Rodríguez-Hernández et al., 2020). In the measurement context, it is essential to realize that external factors such as teaching quality, access to learning resources, and family support also be able to affect test results. This has been proven by various studies showing that student performance is not only determined by cognitive abilities alone, but also by the learning environment and socio-economic conditions surrounding them (Malik, 2018; Munir et al., 2023).Therefore, even if the instrument has undergone a rigorous validation process and has sufficient reliability, the results should be analyzed by considering the broader context. This will provide a more comprehensive picture of student performance and ensure that the instrument truly reflects the ability it seeks to measure fairly and

accurately. Thus, education policymakers be able to make more informed decisions in designing and implementing effective learning programs. In addition, the findings of this study highlight the potential of integrating digital platforms such as Quizizz into assessment reform in physics education. Other educators and institutions may benefit from these results by adopting similar tools to enhance formative and summative evaluations, while also remaining aware of the challenges related to validity, reliability, and bias. Furthermore, the insights gained be able to inform the development of other digital assessment models beyond Quizizz, contributing to the broader effort of modernizing assessment practices in line with 21st-century learning needs.

CONCLUSION

This research shows that the physics measurement instrument tested through the Quizizz platform has good construct validity, with a raw variance explained by measures value of 22.8%, indicating that this instrument effectively measures student ability and understanding. Although the item and person reliabilities were weak, with person reliability of 63% and item reliability of 46%, the Cronbach alpha value of 70% showed sufficient consistency in this learning context. DIF analysis indicated no significant item bias toward gender, so the questions were fair for both male and female students. This study emphasizes the importance of construct and content validity, as well as instrument reliability, because accurate and fair results are essential for effective learning evaluation. Furthermore, the study recommends improving the instrument by piloting on a more extensive and diverse sample and considering external factors such as the learning environment and socio-economic influences to get a more comprehensive picture of student performance. In addition, adopting analytical approaches such as IRT in the future research could strengthen the evaluation of item quality and enhance the precision of measurement. Then, to relying on Rasch validity, further research is recommended to conduct construct validity testing using other statistical techniques. Thus, digital platforms Quizizz not only measure students' knowledge effectively but also provide timely feedback, support fair assessment practices, and contribute to the development of innovative physics learning processes. Ultimately, this study highlights the potential of integrating gamified digital assessment tools to improve both the quality and equity of educational evaluation.

ACKNOWLEDGMENTS

The authors would like to thank the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia with Program Magister Menuju Doktor untuk Sarjana Unggul (PMDSU) Batch VII and Peningkatan Kualitas Publikasi Internasional (PKPI) PMDSU, which has provided funding support and opportunities [Contract Number: 101/C3/DT.05.00/PL/2025].

AUTHOR CONTRIBUTIONS

MZBA and NPL: Research design and concept, data acquisition, drafting research manuscript, revision, supervision. MZBA: Drafting manuscript, research data analysis, technical and material support, research data acquisition. ASH: Technical and material support, data acquisition. BC: Translating, proofreading, final approval. ASM: Reviewing the total before submitting.

CONFLICTS OF INTEREST

The author(s) declare no conflict of interest.

USE OF ARTIFICIAL INTELLIGENCE (AI)-ASSISTED TECHNOLOGY

The authors declare that no artificial intelligence (AI) tools were used in the generation, analysis, or writing of this manuscript. All aspects of the research, including data collection, interpretation, and manuscript preparation, were carried out entirely by the authors without the assistance of AI-based technologies.

REFERENCES

Akukwe, B., & Schroeders, U. (2016). Socio-economic, cultural, social, and cognitive aspects of family background and the biology competency of ninth-graders in Germany. *Learning and Individual Differences*, *45*, 185–192. https://doi.org/10.1016/j.lindif.2015.12.009.

Amiruddin, M. Z. Bin, Samsudin, A., Suhandi, A., Kaniawati, I., COŞTU, B., Aminuddin, A. H., & Kuniawan, F. (2023). Validity and reliability of the global warming instrument: A pilot study

using rasch model analysis. *Jurnal Pendidikan MIPA*, *24*(4), 912–922. https://doi.org/10.23960/jpmipa/v24i4.pp912-922.

Bauer, D. J. (2023). Enhancing measurement validity in diverse populations: Modern approaches to evaluating differential item functioning. *British Journal of Mathematical and Statistical Psychology*, *76*(3), 435–461. https://doi.org/10.1111/bmsp.12316.

Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch analysis in the human sciences*. Springer. https://doi.org/10.1007/978-94-007-6857-4.

Bouwer, R., Koster, M., & Van den Bergh, H. (2023). Benchmark rating procedure, best of both worlds? Comparing procedures to rate text quality in a reliable and valid manner. *Assessment in Education: Principles, Policy & Practice*, *30*(3–4), 302–319. https://doi.org/10.1080/0969594X.2023.2241656.

Büyükkıdık, S. (2023). Purification procedures used for the detection of gender DIF: Item bias in a foreign language test. *International Journal of Assessment Tools in Education*, *10*(4), 765–780. https://doi.org/10.21449/ijate.1250358.

Chakrabartty, S. N. (2013). Best split-half and maximum reliability. *IOSR Journal of Research & Method in Education*, *3*(1), 1–8. https://doi.org/10.9790/7388-0310108.

Chen, Z., Li, G., He, J., Yang, Z., & Wang, J. (2022). A new parallel adaptive structural reliability analysis method based on importance sampling and K-medoids clustering. *Reliability Engineering & System Safety*, *218*, 108124. https://doi.org/10.1016/j.ress.2022.108639.

Cheung, G. W., Cooper-Thomas, H. D., Lau, R. S., & Wang, L. C. (2023). Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations. *Asia Pacific Journal of Management*, 1–39. https://doi.org/10.1007/s10490-023-09871-y.

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412. https://doi.org/10.1037/pas0000626.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *The American Journal of Medicine*, *119*(2), 166-e7. https://doi.org/10.1016/j.amjmed.2005.10.036.

Cordier, R., Speyer, R., Schindler, A., Michou, E., Heijnen, B. J., Baijens, L., Karaduman, A., Swan, K., Clave, P., & Joosten, A. V. (2018). Using Rasch analysis to evaluate the reliability and validity of the swallowing quality of life questionnaire: an item response theory approach. *Dysphagia*, *33*, 441–456. https://doi.org/10.1007/s00455-017-9873-4.

Dianovinina, K., Surjaningrum, E. R., & Wulandari, P. Y. (2024). Adaptation and validation of the children's cognitive triad inventory for Indonesian students. *International Journal of Evaluation and Research in Education (IJERE)*, *13*(3), 1356–1362. https://doi.org/10.11591/ijere.v13i3.28038.

Goldhammer, F. (2015). Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: Interdisciplinary Research and Perspectives*, *13*(3–4), 133–164. https://doi.org/10.1080/15366367.2015.1100020.

Fährmann, K., Köhler, C., Hartig, J., & Heine, J.-H. (2022). Practical significance of item misfit and its manifestations in constructs assessed in large-scale studies. *Large-Scale Assessments in Education*, *10*(1), 7.

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, *20*(3), 245–270. https://doi.org/10.1080/02671520500193744.

Intasoi, S., Junpeng, P., Tang, K. N., Ketchatturat, J., Zhang, Y., & Wilson, M. (2020). Developing an assessment framework of multidimensional scientific competencies. *International Journal of Evaluation and Research in Education*, *9*(4), 963–970. https://doi.org/10.11591/ijere.v9i4.20542.

Kauertz, A., & Fischer, H. E. (2006). Assessing students' level of knowledge and analysing the reasons for learning difficulties in physics by Rasch analysis. *Applications of Rasch Measurement in Science Education*, 212–246.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, *65*(23), 2276–2284. https://doi.org/10.2146/ajhp070364.

Kinyua, K., & Okunya, L. O. (2014). Validity and reliability of teacher-made tests: Case study of year 11

physics in Nyahururu district of Kenya. *African Educational Research Journal*, *2*(2), 61–71.
Köhler, C., Robitzsch, A., & Hartig, J. (2020). A bias-corrected RMSD item fit statistic: An evaluation and comparison to alternatives. *Journal of Educational and Behavioral Statistics*, *45*(3), 251–273.
Kusuma, I. Y., Triwibowo, D. N., Pratiwi, A. D. E., & Pitaloka, D. A. E. (2022). Rasch modelling to assess psychometric validation of the knowledge about tuberculosis questionnaire (KATUB-Q) for the general population in Indonesia. *International Journal of Environmental Research and Public Health*, *19*(24), 16753. https://doi.org/10.3390/ijerph192416753.
Lechien, J. R., Maniaci, A., Gengler, I., Hans, S., Chiesa-Estomba, C. M., & Vaira, L. A. (2024). Validity and reliability of an instrument evaluating the performance of intelligent chatbot: The artificial intelligence performance instrument (AIPI). *European Archives of Oto-Rhino-Laryngology*, *281*(4), 2063–2079. https://doi.org/10.1007/s00405-023-08219-y.
Lee, H., & Geisinger, K. F. (2014). The effect of propensity scores on DIF analysis: Inference on the potential cause of DIF. *International Journal of Testing*, *14*(4), 313–338. https://doi.org/10.1080/15305058.2014.922567.
Lohr, K. N. (2002). Assessing health status and quality-of-life instruments: attributes and review criteria. *Quality of Life Research*, *11*, 193–205. https://doi.org/10.1023/A:1015291021312.
Lupi, J. B., Carvalho de Abreu, D. C., Ferreira, M. C., Oliveira, R. D. R. de, & Chaves, T. C. (2017). Brazilian Portuguese version of the revised fibromyalgia impact questionnaire (FIQR-Br): cross-cultural validation, reliability, and construct and structural validation. *Disability and Rehabilitation*, *39*(16), 1650–1663. https://doi.org/10.1080/09638288.2016.1207106.
MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, 293–334. https://doi.org/10.2307/23044045.
Madansky, A. (1965). Approximate confidence limits for the reliability of series and parallel systems. *Technometrics*, *7*(4), 495–503. https://doi.org/10.1080/00401706.1965.10490293.
Maldonado-Murciano, L., Pontes, H. M., Barrios, M., Gómez-Benito, J., & Guilera, G. (2023). Psychometric validation of the Spanish Gaming Disorder Test (GDT): Item response theory and measurement invariance analysis. *International Journal of Mental Health and Addiction*, *21*(3), 1973–1991. https://doi.org/10.1007/s11469-021-00704-x.
Martinková, P., Drabinová, A., Liaw, Y.-L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education*, *16*(2), rm2. https://doi.org/10.1187/cbe.16-10-0307.
Malik, R. S. (2018). Educational challenges in 21st century and sustainable development. *Journal of Sustainable Development Education and Research*, *2*(1), 9–20.
Mi, S., Ye, J., Li, Y., & Bi, H. (2023). Development and validation of a conceptual survey instrument to evaluate senior high school students' understanding of electrostatics. *Physical Review Physics Education Research*, *19*(1), 10114. https://doi.org/10.1103/PhysRevPhysEducRes.19.010114.
Mian, S. H., Salah, B., Ameen, W., Moiduddin, K., & Alkhalefah, H. (2020). Adapting universities for sustainability education in industry 4.0: Channel of challenges and opportunities. *Sustainability*, *12*(15), 6100. https://doi.org/10.3390/su12156100.
Molloy, J. C., Chadwick, C., Ployhart, R. E., & Golden, S. J. (2011). Making intangibles "tangible" in tests of resource-based theory: A multidisciplinary construct validation approach. *Journal of Management*, *37*(5), 1496–1518. https://doi.org/10.1177/0149206310394185.
Moreira, H., & Freire, M. L. L. (2024). Promoting formative assessment with quizizz: A classroom action research study. *Ciencia Latina Revista Científica Multidisciplinar*, *8*(2), 590–604.
Müller, M. (2020). Item fit statistics for Rasch analysis: can we trust them? *Journal of Statistical Distributions and Applications*, *7*(1), 5.
Munir, J., Faiza, M., Jamal, B., Daud, S., & Iqbal, K. (2023). The impact of socio-economic status on academic achievement. *Journal of Social Sciences Review*, *3*(2), 695–705.
Nisa, K., Suprapto, N., Amiruddin, M. Z., Sari, E. P. D. N., & Athiah, B. D. (2024). Ethnoscience-Quizizz test to measure problem-solving skills: a Rasch analysis. *Int J Eval & Res Educ*, *13*(6), 4247–4255.
Oliveri, M. E., Ercikan, K., Lyons-Thomas, J., & Holtzman, S. (2016). Analyzing fairness among linguistic minority populations using a latent class differential item functioning approach. *Applied*

Measurement in Education, *29*(1), 17–29. https://doi.org/10.37811/cl_rcm.v8i2.10511.

Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *EURASIA Journal of Mathematics, Science and Technology Education*, *19*(8), em2307. https://doi.org/10.29333/ejmste/13428.

Pals, F. F. B., Tolboom, J. L. J., & Suhre, C. J. M. (2023). Development of a formative assessment instrument to determine students' need for corrective actions in physics: Identifying students' functional level of understanding. *Thinking Skills and Creativity*, *50*, 101387. https://doi.org/10.1016/j.tsc.2023.101387.

Pellegrino, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa*, *20*(2), 65–77. https://doi.org/10.1016/j.pse.2014.11.002.

Polit, D. F. (2014). Getting serious about test–retest reliability: a critique of retest research and some recommendations. *Quality of Life Research*, *23*, 1713–1720. https://doi.org/10.1007/s11136-014-0632-9.

Pronk, T., Molenaar, D., Wiers, R. W., & Murre, J. (2022). Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. *Psychonomic Bulletin & Review*, *29*(1), 44–54. https://doi.org/10.3758/s13423-021-01948-3.

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1), 1301013.

Rabgay, T. (2015). A study of factors influencing students' academic performance in a Higher secondary school in Bhutan. *Rabsel-the CERD Educational Journal*, *16*(2), 74–96.

Ranyard, R., McNair, S., Nicolini, G., & Duxbury, D. (2020). An item response theory approach to constructing and evaluating brief and in-depth financial literacy scales. *Journal of Consumer Affairs*, *54*(3), 1121–1156.

Razali, M. N. M., Hamid, A. H. A., Alias, B. S., & Mansor, A. N. (2024). The validity and reliability of culturally responsive leadership practice instruments in small schools peninsular Malaysia. *Int J Eval & Res Educ*, *13*(1), 1–8. https://doi.org/10.11591/ijere.v13i1.26274.

Reynolds, C. R., Altmann, R. A., & Allen, D. N. (2021). The problem of bias in psychological assessment. In *Mastering modern psychological testing: Theory and methods* (pp. 573–613). Springer. https://doi.org/10.1007/978-3-030-59455-8_15.

Rodríguez-Hernández, C. F., Cascallar, E., & Kyndt, E. (2020). Socio-economic status and academic performance in higher education: A systematic review. *Educational Research Review*, *29*, 100305. https://doi.org/10.1016/j.edurev.2019.100305.

Saputro, S. (2022). Trend creative thinking perception of students in learning natural science: Gender and domicile perspective. *International Journal of Instruction*, *15*(1), 701–716. https://doi.org/10.29333/iji.2022.15140a.

Saxena, S., Carlson, D., Billington, R., & Orley, J. (2001). The WHO quality of life assessment instrument (WHOQOL-Bref): the importance of its items for cross-cultural research. *Quality of Life Research*, *10*, 711–721. https://doi.org/10.1023/A:1013867826835.

Song, J., Howe, E., Oltmanns, J. R., & Fisher, A. J. (2023). Examining the concurrent and predictive validity of single items in ecological momentary assessments. *Assessment*, *30*(5), 1662–1671. https://doi.org/10.1177/10731911221113563.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan*. Trim komunikata.

Sumintono, B., & Widhiarso, W. (2014). Aplikasi Model Rasch untuk penelitian ilmu-ilmu sosial [Application of the Rasch Model for Social Sciences Research]. *Bandung, Indonesia: Trimkom Publishing House*.

Van der Linden, W. J., & Hambleton, R. K. (1997). Handbook of item response theory. *Taylor & Francis Group. Citado Na Pág*, *1*(7), 8. https://doi.org/10.1007/978-1-4757-2691-6_1.

Van Zile-Tamsen, C. (2017). Using Rasch analysis to inform rating scale development. *Research in Higher Education*, *58*(8), 922–933. https://doi.org/10.1007/s11162-017-9448-0.

Willingham, W. W., & Cole, N. S. (2013). *Gender and fair assessment*. Routledge. https://doi.org/10.4324/9781315045115.

Yildiz, S. M., & Kara, A. (2017). A unidimensional instrument for measuring internal marketing concept in the higher education sector: IM-11 scale. *Quality Assurance in Education*, *25*(3), 343–361.

https://doi.org/10.1108/QAE-02-2016-0009.

Zainuddin, Z., Shujahat, M., Haruna, H., & Chu, S. K. W. (2020). The role of gamified e-quizzes on student learning and engagement: An interactive gamification solution for a formative assessment system. *Computers & Education*, *145*, 103729. https://doi.org/10.1016/j.compedu.2019.103729.