Pak. j. life soc. Sci. (2024), 22(2):14571-14579 E-ISSN: 2221-7630;P-ISSN: 1727-4915

Pakistan Journal of Life and Social Sciences
www.pjlss.edu.pk

https://doi.org/10.57239/PJLSS-2024-22.2.001049

RESEARCH ARTICLE

Rasch Analysis in Diagnosing Learning Gaps and Gauge Item Difficulty in Eighth-Grade Mathematics Assessments

Abzonie B. Reño[1], Alvic A. Arnado[2]

[1]Caraga State University, Philippines
[2]PhD, Caraga State University, Philippines

ARTICLE INFO

Received: Sep 29, 2024

Accepted: Nov 11, 2024

Keywords

Learning gaps
Item difficulty
Rasch Model
Mathematics assessment

***Corresponding Author**

abreno@carsu.edu.ph

ABSTRACT

This study used the Rasch model, the item response theory (IRT), to analyze learning gaps, problem difficulty, and test reliability in assessing students' ability to eighth grade mathematics. The analysis was carried out through the Jamovi program, using data from 50 tests - of multiple tests - given to 31 students. The results showed a reliability score of 74.7%, which shows the disparity in student achievement. Of the problems, 96% performed well, although two inconsistent problems were marked for improvement. Difficulty analysis revealed a wide range of problems, although a learning gap was evident, with students performing well on easy problems but struggling on difficult problems. Wright's map showed the relationship between student performance and academic difficulty, indicating areas of study that may require attention. This study demonstrates the utility of the Rasch model in developing measurement tools, identifying knowledge gaps, and making recommendations for curriculum improvement. Despite the limitations of the sample size, the results provide a basis for future research to improve the value of mathematics teaching.

1. INTRODUCTION

Education is the basis of personal and social development. Teachers play an important role in facilitating the learning process, and assessment and evaluation are essential to improve learning outcomes. However, in addition to measuring performance, assessment is also essential to identifying learning gaps. Understanding these gaps allows teachers to effectively design instruction and remediation so that no student is left behind.

The main purpose of teacher assessment is to monitor learning progress and collect information about student progress (Murphy, 2019). By regularly assessing students' progress, teachers can not only evaluate the success of their instructional methods but also detect learning gaps that may hinder students from reaching their full potential. Mathematics assessments evaluate students' comprehension and application of mathematical concepts, problem-solving skills, and logical reasoning. These assessments use diverse formats like multiple-choice questions, problem-solving tasks, and performance-based evaluations to measure proficiency across various domains. Effective assessments in mathematics not only gauge knowledge but also highlight gaps in understanding, guiding instruction and identifying areas where students require further support.

The quality of assessments is closely tied to the effectiveness of their test items, which must accurately evaluate the intended traits (Duckworth & Yeager, 2015). Yet, many assessments rely on raw data rather than standardized measures, often overlooking learning gaps. Classical Test Theory (CTT) posits that test items influence individual student performance (person statistics), while Item Response Theory (IRT), and specifically the Rasch model, offers a more nuanced understanding by linking item characteristics—such as difficulty and discrimination—with student abilities (LeBeau et al., 2020).

Teacher evaluations are important for monitoring students' progress and gathering knowledge about their learning gaps, allowing learning to be adapted over time (van der Steen et al., 2023). Well-developed testing tools based on modern analytical methods, such as the Rasch model, improve the ability to determine where students are struggling and what needs to be addressed to achieve better learning outcomes (Fowler & John's College, 2023). The Rasch model, introduced by (Zeileis, 2024) in his work " Examining Exams Using Rasch Models and Assessment of Measurement Invariance," focuses on the relationship between student ability and product difficulty. This model provides a framework for analyzing test item accuracy and gaps in student learning. It also allows educators to group students by ability while assessing the difficulty level of each test item, which helps reveal areas where students struggle and the extent of those problems. Due to the advantages of the Rasch model, including its consistency in identifying item failures and measuring student proficiency over time, this model is an appropriate tool for analyzing the first quarter mathematics test at Pigdaulan National High School. The aim of this study is to use the Rasch model in this test and provide important information about students' learning disabilities, question difficulty, and test reliability. The results will inform curriculum development, instructional strategies, and assessment practices, ultimately improving student achievement in eighth grade mathematics.

METHODS

This study uses a descriptive approach to critically evaluate the quality of researcher-designed mathematics tests and the mathematical skills demonstrated by Grade 8 students enrolled at Pigdaulan National High School. This research was based on data from previous quarterly assessments provided by 8th graders. Participants in the research included 31 students whose responses were examined on a 50-item multiple-choice test, each with four response options. Research covers various topics such as systems of linear equations in two variables, rational algebraic expressions, deterministic polynomials and linear equations in two variables. Specifically, test item construction was carefully guided by a researcher-designed measurement structure chart to ensure alignment with learning objectives.

Participants were selected through purposive sampling to capture different representations of the mathematical abilities of Year 8 students at Pigdaulan National High School. The assessment uses a variable scoring system, where a correct answer receives a score of 1. While an incorrect answer receives a score of 0, the data are therefore determined by their binary nature. The analysis of test results involves the application of Rasch's item response theory (IRT), facilitated using Jamovi software. This method enables a detailed evaluation of the test results. The analysis yielded several criterion items that demonstrated adherence to the Rasch model and provided valuable insight into the mediation and effectiveness of each item in the experiment. The use of these models and software greatly contributed to the depth and accuracy of the assessments performed in this research.

RESULTS AND DISCUSSION

Table 1 shows a person reliability model. The table comprises various parameters about the person reliability and the data utilized in the person reliability model. Specifically, it shows a 0.747- or 74.7%-person reliability, indicating that the individual's performance on the quarterly examination is moderately consistent.

Table 1: Model fit of the Rasch model analysis

	Person Reliability	**MADaQ3**	**p**
scale	0.747	0.153	0.015

Note. MADaQ3= Mean of absolute values of centered Q_3 statistic with p-value obtained by Holm adjustment; Ho= the data fit the Rasch model.

Their scores on the assessment are likely to be a reasonably accurate reflection of their true abilities. Furthermore, a score of 0.747 is considered a sufficient level of personal reliability, suggesting that the individual's performance is generally consistent but not always perfectly predictable (Demetriou et al., 2023). This "enough" level of reliability suggests room for improvement, prompting us to consider external factors or refine the model for better consistency. Ultimately, while this individual

is generally strong in the assessed area, there is space to make their performance even more predictable.

Table 2 outlines the performance of various items (Q1 to Q50) in terms of their difficulty levels (measured by the "Measure" values), response accuracy (proportions), and model fit indices (Infit and Outfit). The "Proportion" column shows how frequently respondents answered each item correctly, ranging from high (0.9355 for Q26) to low (0.0968 for Q33). Lower "Measure" values indicate easier items, while higher values signify more difficult ones. The "Infit" and "Outfit" columns measure how well the data fit the expected model, with values close to 1 representing a good fit. The items with high correct response rates, such as Q1, Q2, Q26, and Q48, exhibit low "Measure" values (ranging from -2.0179 to -2.8077), indicating that these items are easier for respondents. On the other hand, items like Q33 (proportion of 0.0968) and Q45 (proportion of 0.0968) have high "Measure" values (2.3701 and 2.3701, respectively), showing they are more challenging. The infit and outfit values for most items fall within acceptable ranges (0.7–1.3), suggesting a good fit for the majority of the data. However, some items like Q45 (Outfit = 1.931) and Q33 (Outfit = 1.589) slightly exceed the acceptable fit thresholds, implying potential inconsistencies in how respondents approached these questions, which may indicate issues in question clarity or difficulty.

Table 2. Item statistics

Item	Proportion	Measure	S.E.Measure	Infit	Outfit	Item	Proportion	Measure	S.E.Measure	Infit	Outfit
Q1	0.871	-2.0179	0.547	0.91	0.711	**Q26**	0.9355	-2.8077	0.740	1.06	1.20
Q2	0.871	-2.0179	0.547	1.01	0.89	**Q27**	0.7419	-1.1179	0.424	1.08	1.111
Q3	0.5161	-0.0529	0.373	1.1	1.099	**Q28**	0.2258	1.3306	0.440	1.02	0.960
Q4	0.6452	-0.6272	0.389	0.91	0.888	**Q29**	0.3871	0.5104	0.382	0.87	0.834
Q5	0.4839	0.0861	0.373	1.21	1.245	**Q30**	0.6774	-0.7819	0.398	1.2	1.257
Q6	0.3871	0.5104	0.382	1.15	1.141	**Q31**	0.4194	0.3666	0.377	0.8	0.768
Q7	0.7097	-0.9446	0.409	0.98	0.989	**Q32**	0.129	2.0365	0.545	1.0	1.371
Q8	0.6774	-0.7819	0.398	0.9	0.881	**Q33**	0.0968	2.3701	0.615	1.05	1.589
Q9	0.7742	-1.3052	0.443	1.01	0.967	**Q34**	0.6129	-0.4784	0.383	0.95	0.939
Q10	0.1935	1.5351	0.465	1.05	1.23	**Q35**	0.2903	0.9733	0.408	0.96	0.959
Q11	0.4194	0.3666	0.377	0.93	0.924	**Q36**	0.3548	0.6583	0.388	1.14	1.214
Q12	0.871	-2.0179	0.547	0.85	0.667	**Q37**	0.1935	1.5351	0.465	1.12	1.203

Q13	0.2258	1.3306	0.44	0.97	0.96	**Q38**	0.5484	-0.1926	0.375	0.95	0.944
Q14	0.6774	-0.7819	0.398	0.91	0.898	**Q39**	0.3226	0.8119	0.396	1.08	1.092
Q15	0.7742	-1.3052	0.443	0.84	0.725	**Q40**	0.2581	1.1451	0.422	0.99	0.95
Q16	0.2258	1.3306	0.44	0.93	0.851	**Q41**	0.7419	-1.1179	0.424	0.92	0.849
Q17	0.3871	0.5104	0.382	1.09	1.091	**Q42**	0.5484	-0.1926	0.375	1.15	1.179
Q18	0.4516	0.2256	0.374	0.95	0.924	**Q43**	0.7419	-1.1179	0.424	0.8	0.729
Q19	0.3871	0.5104	0.382	1.11	1.175	**Q44**	0.7419	-1.1179	0.424	1.02	1.055
Q20	0.5806	-0.334	0.378	0.92	0.901	**Q45**	0.0968	2.3701	0.615	1.12	1.931
Q21	0.6452	-0.6272	0.389	1.11	1.133	**Q46**	0.1613	1.7661	0.498	1.01	0.962
Q22	0.8065	-1.5117	0.467	0.89	0.861	**Q47**	0.3548	0.6583	0.388	1.02	1.061
Q23	0.2258	1.3306	0.44	1.07	1.164	**Q48**	0.871	-2.0179	0.547	0.89	0.682
Q24	0.6774	-0.7819	0.398	0.92	0.871	**Q49**	0.129	2.0365	0.545	1.09	1.309
Q25	0.5161	-0.0529	0.373	0.92	0.925	**Q50**	0.3226	0.8119	0.396	1.01	0.967

Note. Infit= Information-weighted mean square statistic; Outfit= Outlier-sensitive means square statistic.

The data reveals a clear learning gap between the easiest and most difficult items, suggesting that while many respondents may grasp simpler concepts (represented by items like Q1, Q2, and Q26), they struggle with more complex material (items such as Q33, Q45). This gap may point to instructional deficiencies, where certain critical areas are not being adequately addressed or practiced. Challenging topics may include advanced topics that require more explanation and greater focus in the course. Techniques for instructors are the need to analyze the content of more difficult questions, and to address specific areas where students have difficulty. This may include targeted interventions such as remedial courses or other resources for these difficult subjects. In addition, items with high variance scores (for example, Q45) indicate the need to change, to adjust the questions mislead or mislead students.

Current research highlights the importance of aligning assessment items with learning outcomes to accurately measure student strengths and learning areas. For instance, (Fährmann et al., 2022) shows that irrelevant items can distort students' understanding of their abilities, thus leading to inaccurate judgments about their knowledge. Similarly, (Jones et al., 2023) emphasizes the

importance of monitoring good statistics such as Infit and Outfit to increase the reliability and validity of tests.

Figure 1 shows a Wright plot showing the relationship between the respondent's ability and the difficulty of the test item. The left side plots the basic characteristics or abilities of the respondent, with higher values indicating stronger individuals and lower values indicating lower responders. Most respondents are in the middle of the scale, just above zero, indicating that most people are in the average ability range. The distribution narrows in both categories, with fewer respondents indicating more or less power.

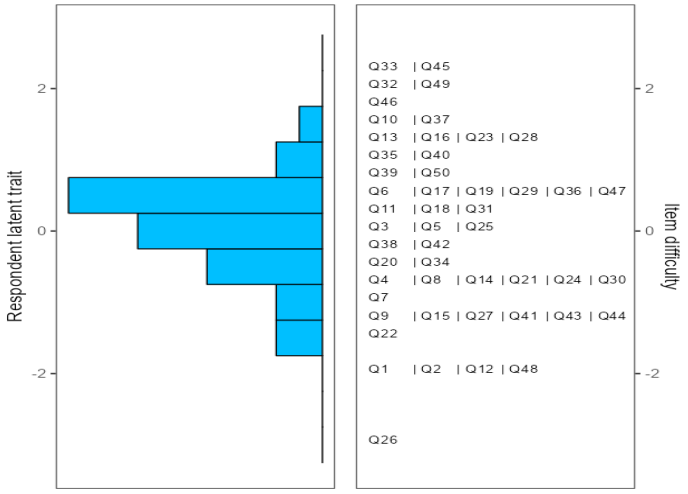

Figure 1. Wright Map using the Rasch Model analysis

This distribution has important implications for assessment design. A well-calibrated test should offer a range of item difficulties that correspond to the varying abilities of test-takers, as emphasized by (Guo et al., 2024). In this case, many items are appropriately centered around the average ability level of respondents, making the test sufficiently challenging for most students. However, the presence of a few very difficult items (e.g., Q33 and Q45) may point to gaps in the instructional approach or areas of the curriculum that are not being adequately covered. These challenging items could represent higher-order thinking skills that students are struggling to master, which aligns with findings by (Wei et al., 2021), who argue that assessments should identify gaps in complex cognitive skills to inform instruction.

Conversely, items like Q1 and Q2 are so easy that they may not differentiate between higher-ability students, suggesting that these items offer little challenge to more proficient test-takers. The need to balance item difficulty is further highlighted by (Karst et al., 2022), who notes that tests should be able to distinguish between different levels of student proficiency, particularly in large, heterogeneous groups. Without sufficient variation in item difficulty, higher-ability students may not be adequately tested, which can limit the assessment's effectiveness in identifying top performers.

From an instructional perspective, this map highlights potential learning gaps that can be addressed. Items at the higher end of the difficulty scale may correspond to concepts or skills that need more attention in the classroom, as suggested by (Kong & Wang, 2021). For example, if these difficult items assess higher-order skills like analysis or evaluation, teachers might need to devote more time to these areas. Likewise, very easy items may need to be revised or replaced to maintain the test's ability to challenge all learners adequately.

The Wright map provides a clear indication of how well the test matches the abilities of the respondents and where improvements in instruction or test design may be needed. By addressing both challenging and overly simple items, educators can create a more balanced and effective assessment that better reflects student understanding across the full range of abilities.

Table 3 presents a recapitulation of the difficulty levels of questions assessed using the Rasch Model, detailing threshold values (b) that categorize each item into varying levels of difficulty. Items range from "very easy" to "very difficult," with thresholds indicating estimated difficulty based on general

test skills. For example, questions 1, 2, and 12 are classified as "very easy" and have a value of -2.0179, indicating that these questions are within the ability of most students. In contrast, items such as 32 and 34, with cutoffs of 2.0365 and 1.5351, were labeled "very difficult" and "difficult" respectively, indicating that these questions may be more challenging for proficient students.

Table 3: Recapitulation of the difficulty level of the Rasch Model questions

Item	Threshold Value (b)	Interpretation	Item	Threshold Value (b)	Interpretation
1	-2.0179	Very Easy	26	-2.8077	Very Easy
2	-2.0179	Very Easy	27	-1.1179	Easy
3	-0.0529	Medium	28	1.3306	Difficult
4	-0.6272	Medium	29	0.5104	Medium
5	0.0861	Medium	30	-0.7819	Medium
6	0.5104	Medium	31	0.3666	Medium
7	-0.9446	Medium	32	2.0365	Very Difficult
8	-0.7819	Medium	33	2.3701	Very Difficult
9	-1.3052	Easy	34	-0.4784	Medium
10	1.5351	Difficult	35	0.9733	Medium
11	0.3666	Medium	36	0.6583	Medium
12	-2.0179	Very Easy	37	1.5351	Difficult
13	1.3306	Difficult	38	-0.1926	Medium
14	-0.7819	Medium	39	0.8119	Medium
15	-1.3052	Easy	40	1.1451	Difficult
16	1.3306	Difficult	41	-1.1179	Easy
17	0.5104	Medium	42	-0.1926	Medium
18	0.2256	Medium	43	-1.1179	Easy
19	0.5104	Medium	44	-1.1179	Easy
20	-0.334	Medium	45	2.3701	Very Difficult
21	-0.6272	Medium	46	1.7661	Difficult
22	-1.5117	Easy	47	0.6583	Medium
23	1.3306	Difficult	48	-2.0179	Very Easy
24	-0.7819	Medium	49	2.0365	Very Difficult
25	-0.0529	Medium	50	0.8119	Medium

$b > 2$: Very Difficult; $1 < b \leq 2$: Difficult; $-1 \leq b \leq 1$: Medium; $-2 \leq b < -1$: Easy; $b < -2$: Very Easy

The distribution of item difficulties reveals several implications for both instructional design and assessment practices. For example, the presence of a few "very easy" things at the beginning shows that the teacher can build students' confidence so that they can learn without getting frustrated. However, the transition to "difficult" and "very difficult" materials demonstrates the need for careful planning in instruction so that students have a solid foundation before facing more challenging challenges. So-called "intermediate" and "easy" tasks can act as a transition point where students can apply their understanding before tackling more complex concepts.

Recent research emphasizes the importance of understanding the problem of objectivity in educational assessment. For instance, (Wang, 2014) emphasized that solutions related to the problems of evaluation materials can significantly change educational outcomes by leading educational change. Similarly, (Rothnie, 2020) advocates the use of diagnostic analysis to identify learning opportunities, emphasizing the importance of the Rasch model in teaching educators about student abilities. Therefore, information from this table can help improve educational strategies, ultimately improving student performance and learning ability.

Table 5 summarizes the distribution of items in the different difficulty groups for this assessment and provides information on all difficulties encountered by students. The table shows that the majority of tasks fall into the category of "medium" difficulty, accounting for 52% of all questions,

indicating that more than half of the questions are designed to match the student's ability. This shows that the assessment is well designed to measure a wide range of skills while still being challenging enough for most students.

Table 5. Category level of difficulty of item

Category	**Frequency**	**Percentage (%)**
Very Easy	5	10
Easy	7	14
Medium	26	52
Difficult	8	16
Very Difficult	4	8
Total	**50**	**100**

In contrast, "very simple" and "other" items were less frequent, accounting for only 10% and 14% of the total, respectively. This classification represents a limited number of introductory questions that may be effective for measuring basic knowledge but are not controllable on the test. The presence of "difficult" (16%) and "very difficult" (8%) schemes suggest that the exam is also designed to challenge successful students, which can encourage broader thinking and deeper engagement with the material.

The results of this publication are of great importance to teachers and curriculum designers. Reviewing items of moderate difficulty can add diagnostic testing, provide students with intermediate levels of understanding, and reveal areas where additional support may be needed. For example, the very high proportion of projects at the intermediate level suggests that teachers can focus on inculcating basic knowledge and skills before introducing more complex concepts. Challenging material allows for variation in instruction for students who may have special developmental needs.

According to (Chang, 2024), different challenges can help determine not only students' understanding of the content, but also their readiness for subsequent learning experiences. This adaptation and development of learning is important for students to successfully meet future educational challenges. Therefore, the ideas in Table 5 can guide teachers in designing tests that not only measure current understanding but also encourage continuous improvement.

Limitations

The descriptive approach of this study is to provide a comprehensive view of the quality of teacher-designed mathematics tests and the abilities of Year 8 students. However, there are several limitations to consider. First, the small sample size of 31 students raises concerns about generalizability and representativeness and may limit the ability to draw general conclusions about study areas in the larger population. The purposeful sampling method also creates a strong scoring system, influencing findings about student understanding and areas for improvement. In addition to that, the validity and reliability of the tests carried out by the researchers cannot be guaranteed because the tests have not been carried out. Lack of testing leads to misidentification of learning gaps because assessments may not accurately capture the full range of strengths and weaknesses of students. In addition, the use of dual assessments ignores diversity of understanding and fails to uncover the specific areas in which students strive, thereby making important learning opportunities.

Finally, although the analysis focuses on the quality of the items in the sample, this does not guarantee the overall quality of the test and the correct identification of the study areas. Although this study provides significant information on student performance and test questions, these limitations should be interpreted with caution and highlight areas for improvement for future research, especially in developing assessments. comprehensive to accurately identify learning opportunities and inform targeted teaching strategies.

CONCLUSION

The results of this study show the effectiveness of using the Rasch model to examine learning areas and assess the difficulty of items in the 8th grade mathematics assessment. Through detailed analysis of student responses, the study showed good reliability of the test one score of 74.7% reliability for each individual, indicating consistency in the assessment of students' abilities The results of the study showed that most of the test items were well connected to the Rasch model, with 96% of the items performing as expected . However, two elements were insufficient and should be reevaluated to improve the quality of the evaluation.

The analysis of the difficulty of the items showed a balanced range, and each item was correctly distributed between the different levels of difficulty, allowing a comprehensive assessment of the students' ability. Identifying what is easy and what is difficult will reveal potential learning opportunities, especially in the more difficult areas of the curriculum. These areas show the need for instructional strategies that address students' difficulties with complex mathematical concepts.

The Rasch model provides important information about student performance and the quality of assessment items, a reliable way to identify learning gaps and improve testing tools. Future research should address the limitations of the study, namely the small sample size, to improve the validity and generalizability of the findings. Through assessment and item analysis, educators can better tailor instruction to the diverse needs of students and promote more effective learning outcomes in mathematics.

REFERENCES

Chang, C. C. (2024). Enhancing EFL University Students' Readiness for Learning Academic Content in English: the Effectiveness of Combining MOOCs with Data-Driven Learning Activities in an English Language Classroom. English Teaching and Learning, 48(3), 439–464. https://doi.org/10.1007/s42321-023-00139-8

Demetriou, A., Spanoudis, G., Christou, C., Greiff, S., Makris, N., Vainikainen, M.-P., Golino, H., & Gonida, E. (2023). Cognitive and personality predictors of school performance from preschool to secondary school: An overarching model. Psychological Review, 130(2), 480–512. https://doi.org/10.1037/rev0000399

Duckworth, A. L., & Yeager, D. S. (2015). Measurement Matters. Educational Researcher, 44(4), 237–251. https://doi.org/10.3102/0013189X15584327

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). https://doi.org/10.1186/s40536-022-00124-w

Fowler, A. T., & John's College, S. (2023). Views from the students' desks: How students experience and comprehend demand and difficulty in GCSE mathematics examination questions.

Guo, H., Johnson, M. S., McCaffrey, D. F., & Gu, L. (2024). Practical Considerations in Item Calibration With Small Samples Under Multistage Test Design: A Case Study. ETS Research Report Series. https://doi.org/10.1002/ets2.12376

Jones, E. A., Wind, S. A., Tsai, C.-L., & Ge, Y. (2023). Comparing Person-Fit and Traditional Indices Across Careless Response Patterns in Surveys. Applied Psychological Measurement, 47(5–6), 365–385. https://doi.org/10.1177/01466216231194358

Karst, K., Bonefeld, M., Dotzel, S., Fehringer, B. C. O. F., & Steinwascher, M. (2022). Data-based differentiated instruction: The impact of standardized assessment and aligned teaching material on students' reading comprehension. Learning and Instruction, 79, 101597. https://doi.org/10.1016/j.learninstruc.2022.101597

Kong, S. C., & Wang, Y. Q. (2021). Item response analysis of computational thinking practices: Test characteristics and students' learning abilities in visual programming contexts. Computers in Human Behavior, 122, 106836. https://doi.org/10.1016/j.chb.2021.106836

LeBeau, B., Assouline, S. G., Mahatmya, D., & Lupkowski-Shoplik, A. (2020). Differentiating Among High-Achieving Learners: A Comparison of Classical Test Theory and Item Response Theory on Above-Level Testing. Gifted Child Quarterly, 64(3), 219–237. https://doi.org/10.1177/0016986220924050

Murphy, R. F. (2019). RAND Corporation Artificial Intelligence Applications to Support K-12 Teachers and Teaching: A Review of Promising Applications, Opportunities, and Challenges.

Rothnie, I. P. (2020). A validity study of a medical student objective structured clinical examination using the many-facet Rasch model.

van der Steen, J., van Schilt-Mol, T., van der Vleuten, C., & Joosten-ten Brinke, D. (2023). Designing Formative Assessment That Improves Teaching and Learning: What Can Be Learned from the Design Stories of Experienced Teachers? Journal of Formative Design in Learning, 7(2), 182–194. https://doi.org/10.1007/s41686-023-00080-w

Wang, T.-H. (2014). Developing an assessment-centered e-Learning system for improving student learning effectiveness. Computers & Education, 73, 189–203. https://doi.org/10.1016/j.compedu.2013.12.002

Wei, X., Saab, N., & Admiraal, W. (2021). Assessment of cognitive, behavioral, and affective learning outcomes in massive open online courses: A systematic literature review. Computers & Education, 163, 104097. https://doi.org/10.1016/j.compedu.2020.104097

Zeileis, A. (2024). Examining Exams Using Rasch Models and Assessment of Measurement Invariance. http://arxiv.org/abs/2409.19522