International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 08 August 2026
DOI: 10.47191/ijcsrr/V9-i8-07, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Psychometric Evaluation of an Applied Mathematics Examination Instrument Based on The Rasch Model

I Ketut Darma[1*], I Gde Nyoman Sangka[2]
[1] Department of Mechanical Engineering, Politeknik Negeri Bali, Indonesia
[2] Department of Electrical Engineering, Politeknik Negeri Bali, Indonesia

ABSTRACT: Administering the same examination to students from different programs can result in uneven measurement of competence and potentially flawed academic decisions. In many vocational settings, psychometric reviews of examination instruments are uncommon, increasing the risk of inaccurate evaluation. This study investigates the Applied Mathematics Final Semester Examination (FSE) instruments in the Mechanical Engineering Department, Bali State Polytechnic, through the Rasch model. The data came from 206 students who answered 40 multiple-choice questions. The test was delivered online via the institutional e-learning system but taken in the classroom. Winsteps 5.9 was used to assess construct validity, reliability, separation indices, item difficulty in relation to ability, and Differential Item Functioning (DIF) across programs. Demonstrates that the instrument meets all psychometric standards. Nine components that didn't fit nonetheless showed the planned construct. The separation indices were 3.46 and 6.18, and the person and item reliability were both over 0.90. The major dimension explained more than 63% of the variation, which indicated that it was one-dimensional. Usually, the difficulty levels were right for the students, but not at the extremes. Eight things suggested that there might be bias. The result is a one-of-a-kind Rasch evaluation of a vocational mathematics FSE that combines reliability, unidimensionality, and DIF analysis. It can be used as a model for other situations and to make vocational math tests fair, valid, and reliable.

KEYWORDS: applied mathematics, psychometric evaluation, Rasch model, item fit, vocational education

INTRODUCTION

Assessment remains a component of the educational process. Although primarily a tool for measuring learning outcomes, it can also be important in improving learning effectiveness and bringing accountability to education. Thus, psychometric quality places significant requirements on technical vocational education and training assessment instruments to accurately and consistently assess actual competencies across a broad range of student characteristics. Here, quality assessment is the primary check on whether the stated learning outcomes have been achieved, particularly in a competency-based learning environment (Yusop et al., 2023) .

Applied Mathematics is a core course in the Department of Mechanical Engineering, Bali State Polytechnic (BSP). This course is evaluated through a final semester exam (FSE) that generally applies to three study programs: Mechanical Engineering (M), Refrigeration and Air Conditioning Engineering (P), and Utility Engineering Technology (U). The weakness of this approach is that the use of the same instrument in different study programs cannot measure competence proportionally and fairly without the potential for elements of injustice (Holden & Tanenbaum, 2023). Despite this, most vocational institutions do not conduct in-depth psychometric evaluations of their examination instruments. As a result, some inaccurate assessments and misclassification of students' academic characteristics can result in incorrect academic decision (Wiley et al., 2015).

Classical Test Theory (CTT) was long assumed the most effective approach in test development and validation. However, as shortcomings emerged, the method became increasingly unviable. These shortcomings include sample dependence characteristics and inadequate measures to detect item performance. Researchers have therefore shifted toward using an Item Response Theory (IRT) framework with Rasch modelling in the calibration of their research instruments. Rasch modelling offers sample free calibration together with elaborate analysis possible at the item level (Boone, 2016; Ranyard et al., 2020; Sijtsma & van der Ark, 2022). This one-parameter logistic model has strong methodology in terms of evaluating item validity, respondent-item separation, and respondent-item reliability, and is able to triangulate data in the case of potential bias (Huang et al., 2022; Vincent et al., 2015; Woudstra et al., 2019).

Here, the Rasch model by Winsteps 5.9 is considered useful in determining inappropriate items, checking the consistency of responses, checking for biases, and visualizing the relationship between participants' ability and the difficulty of items on the Wright map. The model's advantages in checking validity, reliability, unidimensionality, and scale refinement have been tested in several fields of application (Curtin et al., 2016; Papini et al., 2023; Vincent et al., 2015; Woudstra et al., 2019; Yang et al., 2020). Thus, knowledge about this type of model in the evaluation of Applied Mathematics FSE tools in vocational education, primarily in the Mechanical Engineering Study Program in our case, is limited. This is because most studies have concentrated on general education, while validation in vocational contexts remains scarce and inconclusive. Furthermore, although the Rasch model provides more precise and objective measurement than classical approaches, many vocational institutions continue to rely on Classical Test Theory (CTT) analysis. This highlights the importance of studies on the use of the Rasch model in local settings to ensure that assessments are valid and reliable.

This study attempts to fill the gap outlined above by using the Rasch model to check the Applied Mathematics FSE instrument used in the BSP Mechanical Engineering Department. It examines the following: 1) the construct validity of the Applied Mathematics FSE instrument, 2) its reliability and measurement separation index, 3) distribution of item difficulty levels and their suitability to students' ability levels, and 4) potential bias in the instrument between study programs. This study provides theoretical and practical contributions by providing empirical evidence to improve test development practices in vocational education. It also offers methodological insights for educators and policy makers in implementing fair and valid assessment instruments aligned with the principles of competency-based learning.

LITERATURE REVIEW

Psychometric Evaluation of Instruments

The psychometrics of educational measurement deals with the principles and criteria included in the concept of scientific quality for educational tests and guided by those results, allows for accurate and fair interpretation (Boone, 2016; Boone et al., 2014; Gunawan et al., 2024; Hamzah et al., 2022; Rezai, 2022). The dimensions of psychometric standards include validity and reliability as well as item difficulty, unidimensionality, and fairness in different formulations, all of which ensure the integrity of the assessment (AERA et al., 2014; Ajeigbe & Afolabi, 2014; Joo et al., 2022). Validity is the degree to which an instrument accurately measures what it is intended to measure (Brown et al., 2019; Maric et al., 2023). Reliability refers to the consistency of results over time and across contexts of use (Agala et al., 2020; Bannigan & Watson, 2009; Heale & Twycross, 2015; Huang et al., 2022). The difficulty level of the test items is assessed based on how appropriately the difficulty level of the test items matches the respondent's ability level, as measured by the proportion of respondents who answered each correctly (DeVellis, 2017; Heale & Twycross, 2015; Hoeve, 2022). Unidimensionality requires this test to measure a single construct (Humphris et al., 2018). Finally, fairness requires that the assessment be fair and not disadvantageous or advantageous to any party regarding social values in education (Drake et al., 2017; Maric et al., 2023).

Rasch Model for Instrument Evaluation

Rasch modeling is an IRT model developed by Georg Rasch in the early 1960s as a tool for measuring respondents' latent ability and item difficulty on the same logit scale. The model has been described as one that will yield objective, meaningful measurements by clearly separating the individual (difficulty level) and item parameters. The probability that a person with ability θ will answer a question whose difficulty level is β is adjusted according to the Rasch modeling in the IRT framework (Avinç & Doğan, 2024; Ekstrand et al., 2022; Tan et al., 2024). Rasch modeling is a 1-PL formulation that models the probability of getting a correct response as a function of the difference between the respondent's ability and the difficulty parameter of an item (Bond & Fox, 2015; Sijtsma & van der Ark, 2022). In this study, the mathematically uncomplicated model employed uses only one item parameter to focus the analysis and enable easy implementation (Schumacker & Smith, 2007; Tutz, 2024). Furthermore, Rasch modeling is based on the principle of specific objectivity, which means that comparisons of the ability of two respondents should be independent of the item, and similarly, comparisons of difficulty between two items should be independent of the sample (Andrich et al., 2012; Lamoureux et al., 2009; Tutz, 2024). This allows the model to be used as a fair and invariant measurement across populations.

The methodological advantages of the Rasch model are represented by the dimensions of unidimensionality, invariance, linearity, and the possibility of item bias detectability through analysis (Buntragulpoontawee et al., 2021; Frick et al., 2015; Heaney et al., 2020; Lamoureux et al., 2009). In this regard, unidimensionality indicates that all items respond to one latent construct, thereby

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 08 August 2026
DOI: 10.47191/ijcsrr/V9-i8-07, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

increasing construct validity. Invariance speaks to the fact that measurement properties hold true across different samples and supports generalizability of findings. Linearity permits the transformation of raw scores into interval-level logit scales for use in parametric statistical analyses. In this case, the differential item functioning analysis provides information about the presence of potential item bias between groups and thus ensures the fairness of an assessment. These dimensions have rendered the model favorable among researchers in interpreting measurement results in modern psychometric-based measurement instrument development and evaluation.

Accordingly, the Rasch model is frequently noted in reports on the instrument assessment of educational tools in diverse fields, particularly mathematics (Boone et al., 2014; Linacre, 2012). For instance, results for one test in cultural mathematics were highly reliable and valid (Fitrah et al., 2024). Another numerical aptitude test provided evidence of uniform item functioning by gender and school type (Ridwan et al., 2023). Furthermore, a problem-solving test based on STEM was found to be valid and reliable, although it required clarification of some item language (Amalina & Vidákovich, 2022). In the end, the Indonesian version of the Junior Metacognitive Awareness Inventory showed strong psychometric qualities with no signs of gender bias (Pilotti, 2021). This outcome again points to what has already been seen about how the Rasch model helps make measurements in educational tests more valid, accurate, and just.

RESEARCH METHODS

Research Design

This study adopted a quantitative approach using a cross-sectional design. One of the most common methodologies in educational assessment research, the latter permits simultaneous data collection and analysis at one point in time. It is also considered appropriate in judging the characteristics of assessment instruments and participant performance without having to institute intervention (Creswell & Creswell, 2017; Kesmodel, 2018).

Population and Sample

The population in this study were even-semester students of the 2023/2024 academic year in the Department of Mechanical Engineering, BSP, who took the Applied Mathematics FSE. A total sampling technique was employed, wherein all students who took the exam and provided complete and valid answers were included as the sample of the study. Thus, the total sample was 206 people: 117 students in the Mechanical Engineering Study Program, 72 in Refrigeration and Air Conditioning Engineering, and 17 in Utility Engineering Technology.

Research Instrument

In this study, 40 multiple-choice Applied Mathematics FSE questions were analyzed. The questions were arranged based on the Semester Learning Plan, which included limits, derivatives, and integrals. The questions were arranged by a team of lecturers who teach the courses and were officially used in the implementation of the FSE in the Department of Mechanical Engineering.

Data Collection

Data were collected during the implementation of the even-semester FSE of the 2023/2024 academic year through BSP e-learning, and the questions were presented in Google Forms. Although online, the exam was conducted in class under the direct supervision of two supervisors to maintain the integrity of the exam. Subsequently, the results of the students' answers were exported in the form of a digital spreadsheet and coded with a value of 1 for correct answers and 0 for incorrect answers.

Data Analysis

Data analysis was conducted using the Winsteps application, version 5.9, through a Rasch model approach. The aspects of the analysis are described below.

Item and Respondent Suitability

An analysis was conducted on the values of the outfit mean square (MNSQ), Z-standard (ZSTD), and point-measure correlation (Pt. Measure Corr) to assess whether the test items and student responses conform to the Rasch model (Bond & Fox, 2007; Boone et al., 2014; Linacre, 2012, 2017, 2024; Wright & Linacre, 1994). To evaluate item fit, the following criteria were used: outfit MNSQ values were expected to fall between 0.5 and 1.5, outfit ZSTD values between -2 and 2, and Pt. Measure Corr values within the range of 0.40 to 0 (Bond & Fox, 2015; Boone et al., 2014; Linacre, 2012, 2024)

Reliability and Separation Index

The Rasch model uses three components to measure the reliability of an instrument: Cronbach's alpha (KR-20), person and item reliability, and person and item separation. This study used item and respondent level reliabilities, including the calculation of the separation index, to show how well an instrument can discriminate levels of competence or item difficulty. The "person" and "item" reliability values are interpreted as follows: very high for above 0.90, high from 0.80 to 0.89, adequate from 0.70 to 0.79, low from 0.60 to 0.69, very low (poor) from 0.50 to 0.59, and unacceptable below 0.50. The classification of person and item separation values for the value range is as follows: ≥3.00 is classified as very high, 2.00–2.99 as high, 1.00–1.99 as sufficient, <1.00 as low (questionable) (Linacre et al., 2023; Wright & Masters, 1982).

Unidimensionality Analysis

This measuring tool was checked for a single dimension to ensure that it considered only one main hidden idea, matching the key rules of the Rasch model (Linacre, 2024; Linacre et al., 2023). Principal component analysis (PCA) was used on the remaining items, and changes not caused by the main dimension were used to identify possible extra dimensions (Bond & Fox, 2015; Linacre, 2024; Linacre et al., 2023; Wright & Stone, 1999). There were two indicators. First, the variance explained by the Rasch measure in its raw form was expected to be ≥50%, with the primary dimension dominating the variance. Second, the eigenvalue of 3.0 in the first contrast was less than three, meaning that there was not sufficient strength in a secondary dimension to significantly distort measurement (Supriatna et al., 2025; Tennant et al., 2004; Tennant & Küçükdeveci, 2023).This test is described in terms of deciding whether this tool indeed meets the unidimensionality assumption and whether it can provide valid and consistent measurements on a particular construct.

Distribution of Item Difficulty Levels

The item difficulty distribution analysis indicated how well the items covered and matched participants' ability range on a single logit scale. It is important to ascertain whether information is available from the test instrument for all ranges of participants' ability levels and to locate any possible gaps or redundancies in item difficulties. Specifically, item difficulty distribution was analyzed through a Wright map, a psychometric mapping approach that graphically depicts two major components on the same logit scale, namely person ability distribution and item difficulty distribution (Bond & Fox, 2015; Boone, 2016). Merging these into one, a test researcher can describe test targeting, i.e., how well the test items fit the entire range of abilities within a group taking a test. It seeks to determine whether the test is not too easy or too difficult for the majority of participants and whether it can distinguish between people with different skills. Hence, the results for the item difficulty distribution Wright map analysis are best placed to judge how well the tool matches the target population's ability characteristics. Considering these results can help enhance such a tool in its current form by balancing its difficulty, which can be improved by finding an optimal level.

Test Information Function

A test information function (TIF) analysis was conducted to evaluate the information that could be elicited by the instrument at different levels of participants' ability (θ). The instrument's sensitivity and accuracy in measuring participant ability across an entire range of abilities have been described.

The TIF graph was then reviewed to determine the distribution of information across the ability range by looking at the peaks in the graph. Peaks in the TIF curve will show at what ability level the test information gets to its highest value, i.e., where measurement is best. If the TIF curve $\leq \theta = 0$, then the item level distribution does not correspond with participants' ability distribution; thus, the test only relates to the mean population that will be assessed (Bond & Fox, 2015; Linacre et al., 2023).

Differential Item Functioning

Differential item function (DIF) was applied to identify item bias among the study program groups. An item is considered to have significant potential bias if it has a DIF contrast value ≥0.64 and a significant p-value <0.05 (Linacre, 2024). This ensures that the Applied Mathematics final exam instrument is evaluated with regard to construct validity, consistent measurement, and fairness for all test takers from any study program.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 08 August 2026
DOI: 10.47191/ijcsrr/V9-i8-07, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

RESULTS

Table 1 shows the results of the analysis using the Rasch model for the Applied Mathematics FSE data conducted on 206 students with 40 questions using the Rasch model methodology on Winsteps 5.9 software.

Table 1. Summary of the Statistics of the Applied Mathematics FSE Data

Statistics	Person (Non-extreme)	Person (all)	Item
Number of respondents/item	205	206	40
Mean total score	26.0	25.9	133.2
Mean ability/item (logit)	1.32	1.29	0.00
Standard deviation	2.72	2.76	1.60
MNSQ Infit (mean)	0.95	-	0.96
MNSQ Outfit (mean)	1.25	-	1.26
ZSTD Infit (mean)	−0.05		−0.34
ZSTD Outfit (mean)	0.19	-	0.20
Logit range	−4.46–5.38	−5.70–5.38	−2.08–6.22
Reliability	0.92	0.92	0.97
Separation index	3.46	3.47	6.18
Score metric correlation	0.98	0.98	1.00
Cronbach's alpha (KR-20)	0.98		
Number of respondents (extreme)	1 (0.5%)	1 (0.5%)	

The summary of the psychometric analysis results in Table 1 indicates that the FSE instrument has very high person and item reliability (Boone et al., 2014). High reliability indicates dependable consistency. Additional validation of the instrument's ability to distinguish between students' ability levels was evident in the separation index (Linacre et al., 2023). Extremes were not noted either in the range of student abilities or in the range of item difficulties, thus suggesting that this instrument can measure competencies across a wide spectrum (Engelhard & Wang, 2022; Wilson, 2023). Most of the items were within the productive range. The few minor deviations still fit the Rasch model (Linacre, 2002). The strong correlation between observed and expected scores further reinforced the validity of the construct (Borsboom et al., 2004). In conclusion, this instrument appears poised to serve as a dependable means for assessing applied mathematics competence and meets rigorous psychometric standards (AERA et al., 2014; DeMars, 2010)

Table 2 Provides the results of the PCA.

Table 2. Summary of the Unidimensionality Tests

Variance Components	Eigenvalue	Percentage Observed	Percentage Expected
Total raw variance in observations	110.7713	100.0%	100.0%
Variance is explained by measures	70.7713	63.9%	62.5%
• by persons	50.7008	45.8%	44.8%
• by items	20.0705	18.1%	17.7%
Unexplained variance	40.0000	36.1%	37.5%
• Unexplained variance in the 1st contrast	2.5280	2.3%	6.3%
• Unexplained variance in the 2nd contrast	2.0527	1.9%	5.1%

 *Corresponding Author: I Ketut Darma

• Unexplained variance in the 3rd contrast	1.7257	1.6%	4.3%
• Unexplained variance in the 4th contrast	1.7090	1.5%	4.3%
• Unexplained variance in the 5th contrast	1.6677	1.5%	4.2%
Essential unidimensionality (Rasch)		88.0%	

Results of residual analysis, indicating that this FSE instrument focuses on a single primary construct, with explained variance significantly greater than unexplained variance. In fact, the first contrast value is still well below the commonly used 15% threshold (Bond et al., 2020; Linacre, 2010; Linacre et al., 2023). Essential unidimensionality is also high, indicating that no other construct is strong enough to collaborate with the primary dimension. Near-perfect inter-cluster correlations and balanced distribution of items ensured consistency, though a few items showed relatively high positive or negative loadings (Boone et al., 2014). This provides a solid foundation for further analysis because the tool satisfies the unidimensionality assumption, which is a crucial principle in contemporary measurement theory (Boone et al., 2014).

The item suitability analysis with the Rasch model was conducted using the misfit order output from Winsteps 5.9. The results are summarized in Table 3.

Table 3. Summary of the Misfit Order Analysis

No.	Item Code	Measure (Logit)	MNSQ Infit	Outfit		PT. Meas. Corr	Status
				MNSQ	ZSTD		
1	B31	−2.02	1.47	5.65	3.79	0.48	Severe misfit
2	B14	−2.08	1.09	3.97	2.87	0.58	Severe misfit
3	B34	1.81	1.61	2.74	3.27	0.60	Moderate misfit
4	B36	1.90	1.58	2.07	2.22	0.59	Moderate misfit
5	B6	0.49	0.38	0.19	−3.73	0.89	Moderate misfit
6	B5	1.63	0.96	1.96	2.22	0.74	Almost misfit
7	B32	−0.68	1.54	1.91	1.67	0.63	Needs review
8	B33	−1.95	1.17	1.89	1.30	0.59	Needs review
9	B2	−0.09	1.06	1.82	1.74	0.74	Needs review

Analysis of How Items Fit Most items fit the Rasch model and differentiate students' abilities well, thus supporting the construct validity of the instrument (Bond & Fox, 2015; Boone et al., 2014; Linacre, 2012, 2024). However, several items exhibited varying degrees of misfit. Those with moderate misfits are most probably due to non-functioning distractors or some ambiguity in the wording, and those with severe misfits seem to be off the construct being measured. These need more qualitative validation through an expert review or student input. They have not gone beyond reasonable bounds, which speaks to the fact that even though the tool is already powerful, it can be made much better for future administrations.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 08 August 2026
DOI: 10.47191/ijcsrr/V9-i8-07, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

The distribution of the difficulty level of the test items is shown in the Wright map in Figure 1.

```
TABLE 1.2 Data_TM276                ZOU015WS.TXT  Oct 24 2024 15:56
INPUT: 206 Person  40 Item  REPORTED: 206 Person  40 Item  2 CATS
WINSTEPS 5.9.0.0
---------------------------------------------------------------------

MEASURE    Person - MAP - Item
                <more>|<rare>
    7              .  +
                     T|
                      |
                      | B40
                      |
    6                 +
                      |
                      |
               .####  |
                      |
    5                 +
                      |
                      |
                      |
                     S|
    4    .##########  +
                      |
                      |
            ########  |
                      |T
    3                 +
              .#####  |
               #####  |
                      | B39
                 ###  |
    2             ##  + B36
                   .  | B34
                      |S B5
                     M| B21
                   .  |
    1              #  + B22    B23    B7
                   #  | B38    B6     B9
                   #  | B4     B8
                   .  | B1
                   #  | B19    B25    B3
    0             ##  +M B15    B17    B20    B30
                   #  | B2     B24    B27
                   #  | B28
                   .  | B29    B32
                   .  |
   -1             .#  + B37
                  .#  | B16    B35
                  .# S|
                   #  |S
                   .  | B10    B13    B26
   -2             .#  + B11    B12    B14    B18    B31    B33
                 .##  |
                 .##  |
                  .#  |
                  .#  |
   -3                 +
                   .  |T
                      |
                      |
                .###  |
   -4                 +
                     T|
                   .  |
                      |
                      |
   -5              .  +
                <less>|<freq>
 EACH "#" IS 3: EACH "." IS 1 TO 2
```

Figure 1. The Wright Map

The Wright map shows that the abilities of test takers ranged from −5 to +6 logits, with most individuals falling between −2 and +2. The item distribution also shows a concentration within the −2 to +2 logit difficulty range; few items fall outside this range. No items fall within the range from +3 to +5 logits, meaning the gap in difficulty is large. Furthermore, one item, B40, was found to be very difficult at the logit scale's upper end. In contrast, six items were found to be of low difficulty. Most of these can even be considered very easy since they fall well below the average item difficulty level. This finding strongly suggests that, despite a general balance in the distribution of ability and difficulty, the test items do not accurately represent areas at the lower end of the scale.

The TIF analysis showed a peak in information at the 0-logit level with a maximum information value of 7.5. The minimum measurement error was 0.365, with the precision maximized at the medium level. The DIF analysis shows that several items indicate DIF. Table 4 summarizes the results.

Table 4. Summary of the DIF Analysis Results

Items Code	Study Program Pair	DIF Contrast	p-value	Interpretation of Fairness
B5	M vs. P	−1.17	0.0139	P is more difficult than M
B15	M vs. P	+1.49	0.0035	M is more difficult than P
B16	M vs. U	−2.47	0.0077	U is more difficult than M
B17	M vs. P	−1.40	0.0073	P is more difficult than M
B21	M vs. P	−1.32	0.0065	P is more difficult than M
B34	M vs. P	−1.46	0.0019	P is more difficult than M
B36	M vs. P	−1.96	0.0001	P is more difficult than M
B39	M vs. P	+1.57	0.0009	M is more difficult than P

The summary of the DIF test in Table 4 shows that eight items (B15, B5, B16, B17, B21, B34, B36, and B39) have significant differences across academic programs and are more difficult for certain populations. In particular, Refrigeration and Air Conditioning Engineering students found items B5, B34, B36, and B39 more difficult. Utilities Engineering students found items B16 and B36 more difficult. Results of this study proved that, aside from some items needing further validation, the Applied Mathematics final exam instrument is both valid and reliable in terms of its conformity to the Rasch model. This, therefore, implies that the instrument has beneficial psychometric properties.

DISCUSSION

This study will present convincing evidence of the psychometric quality of the Applied Mathematics FSE instrument under Rasch analysis. The findings also showed a mixed spread of abilities among students, ranging from low to high. This ability distribution confirms that the instrument has attained objectivity in measuring real ability variation among participants. The participant fit scores were generally good, indicating that students' response patterns were consistent with the assumptions of the Rasch model, except for a few unexpected responses. This means that most students completed all items according to the procedure. The consistency of this pattern is supported by previous findings that a good participant fit score reflects integrity in assessment implementation and seriousness by participants in working on the exam questions (Alhadabi & Aldhafri, 2021). From participants' perspective, this tool attained a very high level of reliability (Table 1). This proves that the test has good discriminatory ability in separating students into several different ability strata. The findings are consistent with research showing that instruments with high reliability provide a basis for interpreting assessment results more accurately (Azizah et al., 2022; Darman et al., 2024). Thus far, the internal consistency of our instrument is confirmed, and thus, it meets one of the major criteria in Rasch modeling-based assessments.

The analysis further showed that variation in item difficulties falls under a good category, indicating that the items can differentiate respondents in terms of different ability levels from the lowest to the highest. This distribution means that the instrument items have been developed fairly proportionately at different cognitive levels, ranging from simple to complex. Very high item reliability and the separation index (Table 1) indicate stability and strength in the quality for measuring participants' ability. This supports an earlier finding that tools with a sufficient difference in difficulty are better at Rasch-based measurement (Aarts et al., 2023). However, several items indicated a misfit. These are crucial in the improvement of future instruments. The emergence of items with misfits was previously reported by Linacre, who emphasized that if one wants to maintain the overall quality of the instrument, such findings cannot be ignored (Linacre, 2024). He stated that items demonstrating signs of misfit must be analyzed more thoroughly to identify the root cause. To this end, he recommended that researchers evaluate each problematic item to determine whether the misfits are substantive or just statistical artifacts owing to technical factors. These items can be revised or deleted or, if relevant and important, retained through careful review. In general, the Applied Mathematics FSE instrument demonstrates satisfactory quality in terms of its development and implementation. However, items identified as misfits still require ongoing item analysis. The results confirm the high construct validity and internal consistency of this tool and that it fulfills the principle of unidimensionality, a major attribute in ensuring fair and accurate assessment results in the vocational education environment.

The Rasch model described over 60% of the total variance, and the first contrast described only 2.3%, which means there is no strong secondary dimension; hence, this tool meets one basic requirement of unidimensionality as stressed by Linacre (Linacre, 2002, 2010; Linacre et al., 2023), and Bond and Fox (Bond et al., 2020). The essential unidimensionality value of 88% increases confidence that residual variance is still in line with the primary dimension so that measurement results can be interpreted as representing a single underlying construct (Bond et al., 2020).

The identification of items with relatively high positive and negative residual loadings shows variability in item contributions to the construct. But since the correlation between the clusters is very high, such variability does not constitute a separate dimension that could interfere with the consistency of measurement. This indicates that the items still function coherently within a single construct framework despite differences in residual contributions. Therefore, the tool does not just meet the technical demand for unidimensionality but also offers proof of measurement stability as a crucial element for construct validity in educational assessment (Boone et al., 2014; Tennant et al., 2004; Tennant & Küçükdeveci, 2023; Wright & Stone, 1999). Items demonstrated significant misfit, especially items B14 and B31 (Bond & Fox, 2015; Boone et al., 2014; Linacre, 2002, 2024; Wright & Stone, 1999), implying that some distractors are not working well or item wording is ambiguous. To improve item quality, earlier work suggested using qualitative validation (Boone, 2016), for example, expert review and cognitive interviewing, in addition to quantitative analysis. All these results prove that the instruments works based on the main rules of new measurement ideas by giving scores that always show one hidden concept. This means that the tool can be trusted to assess students' skills without being affected by other unrelated factors that might harm the validity of the main idea.

The analysis of the Wright map indicated gaps at the extremes of the ability distribution; thus, adding items of very high and very low difficulty levels is recommended to minimize ceiling and floor effects. This will also better fine-tune the measurement targeting accuracy. This is crucial in vocational education assessments, which generally face a high heterogeneity of abilities.

The results of the DIF analysis revealed significant differential functioning for several items across study programs, most likely because of differences in the technical context of item content that are not completely neutral to the curricula of each program (Hambleton et al., 1992; Zumbo, 2006; Zumbo & Gelin, 2005) This highlights a bias that undermines the fairness and equivalency of the construct. Thereby, cross-disciplinary stakeholders in item design should ensure neutrality and relevance for all groups of participants (Biggs & Tang, 2011; Hoeve, 2022; Holden & Tanenbaum, 2023; Hope et al., 2018).

In addition, the results are consistent with those of previous studies (Curtin et al., 2016; Papini et al., 2023). This study further confirms Rasch-based analysis as a powerful method for scrutinizing and upgrading assessment tools. The approach supports quality assurance in higher education on an evidential basis by ensuring that appraisal tools are psychometrically sound, just, and related to the learning outcome.

Future studies on item and DIF bias challenges alongside a wider spread of item difficulties are strongly recommended. To this end, this study advocates the systematic infusion of Rasch modeling into the periodic validation cycle of assessment tools. Such an update will take the lead in setting fairer and more precise tools that can bring instructional insight to vocational education setups.

A Rasch-based psychometric evaluation will unequivocally ascertain the validity and reliability of the Applied Mathematics test as a high-caliber measuring tool. This method could assist you in pinpointing specific areas that require improvement. These results not only provide empirical proof for the test's validity but also pave the way for new research avenues, especially in the realm of mixed methods for examining the qualitative elements that affect response pattern deviations and item difficulty characteristics. This study is not merely an additional measure for quality assurance; it will serve as a catalyst for the development of more equitable, transparent, and sophisticated assessments within vocational education.

CONCLUSION

The evaluation results show that the Applied Mathematics FSE instrument generally meets the psychometric requirements based on the Rasch model, with high construct validity, very good internal reliability, and confirmed unidimensionality. Most test items are in the logit range of −2 to +2, in accordance with the abilities of most students. Most items functioned well, although some demonstrated significant misfit and DIF bias between the study program groups.

Accordingly, this instrument is suitable for objectively measuring students' applied mathematics abilities; however, to improve measurement accuracy, improvements to problematic items and the addition of items with extreme difficulty levels are required. The systematic application of Rasch analysis in instrument evaluation is expected to strengthen the quality of fair, valid, and relevant vocational education assessments.

RECOMMENDATIONS

These results helped to beef up the theoretical basis of measurement in mathematics education and framed practically how valid and reliable assessment tools can be developed. It is very handy for teachers, lecturers, and researchers at any level who want to assess applied mathematical competence. Instrument development is a continuous process that requires further refinement and validation (Engelhard & Wang, 2022).

Future research directions are thus guided by the findings of this study and its limitations. The study should be replicated on a broader and more diversified sample drawn from several polytechnics to check the stability of item parameters and generalizability of findings. This was followed by undertaking a longitudinal study to track measurement stability and test invariance over different cohorts also qualitative research that should entail think-aloud or cognitive interviewing whereby respondents would be able to identify the sources of misfit items. DIF analysis shall be the first step toward determining any potential bias of the items by demographic variables in the present study. It is imperative to note that priorities for further instrument development must include creating new items at a high level of difficulty and using an anchor test design to widen the targeting scope while raising measurement precision at higher ability levels.

ACKNOWLEDGEMENTS

Thanks are due to the Mechanical Engineering Study Program, Bali State Polytechnic, and all teaching staff who provided kind support, facilities, and guidance in carrying out this research. We extend our gratitude to the study participants and other individuals who contributed to the completion of this work. The authors wish to thank the editors and anonymous reviewers of IJCSRR for their valuable comments and suggestions, which led to improvements in the quality of this manuscript. This study belongs here under the Knowledge Development umbrella, with a specific psychometric evaluation. Public, commercial, or not-for-profit funding agencies have not specifically funded this study

REFERENCES

1. Aarts, P., Van Huijstee, J. C., Ragamin, A., Reeves, J. L., Van Montfrans, C., Van Der Zee, H. H., & Prens, E. P. (2023). Validity and reliability of two digital wound measurement tools after surgery in patients with Hidradenitis Suppurativa. *Dermatology*, *239*(1), 99–108. https://doi.org/10.1159/000525844
2. AERA, APA, & NCME. (2014). *Standards for educational and psychological testing* (14 ed). American Educational Research Association. https://www.testingstandards.net/
3. Agala, C. B., Fried, B. J., Thomas, J. C., Reynolds, H. W., Lich, K. H., Whetten, K., Zimmer, C., & Morrissey, J. P. (2020). Reliability, validity and measurement invariance of the Simplified Medication Adherence Questionnaire (SMAQ) among HIV-positive women in Ethiopia: A quasi-experimental study. *BMC Public Health*, *20*(1), 1–16. https://doi.org/10.1186/s12889-020-08585-w
4. Ajeigbe, T. O., & Afolabi, E. R. I. (2014). Assessing unidimensionality and differential item functioning in qualifying examination for senior secondary school students, Osun State, Nigeria. *World Journal of Education*, *4*(4), 30–37. https://doi.org/10.5430/wje.v4n4p30
5. Alhadabi, A., & Aldhafri, S. (2021). A rasch model analysis of the psychometric properties of the student-teacher relationship scale among middle school students. *European Journal of Educational Research*, *10*(2), 957–973.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 08 August 2026
DOI: 10.47191/ijcsrr/V9-i8-07, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

https://doi.org/10.12973/EU-JER.10.2.957
6. Amalina, I. K., & Vidákovich, T. (2022). An integrated STEM-based mathematical problem-solving test: Developing and reporting psychometric evidence. *Journal on Mathematics Education*, *13*(4), 587–604. https://doi.org/10.22342/jme.v13i4.pp587-604
7. Andrich, D., Marais, I., & Humphry, S. (2012). Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. *Journal of Educational and Behavioral Statistics*, *37*(3), 417–442. https://doi.org/10.3102/1076998611411914
8. Avinç, E., & Doğan, F. (2024). Digital literacy scale: Validity and reliability study with the rasch model. *Education and Information Technologies*, *29*(nd), 22895–22941. https://doi.org/10.1007/s10639-024-12662-7
9. Azizah, N., Suseno, M., & Hayat, B. (2022). Item analysis of the Rasch model items in the final semester exam Indonesian language lesson. *World Journal of English Language*, *12*(1), 15–26. https://doi.org/10.5430/wjel.v12n1p15
10. Bannigan, K., & Watson, R. (2009). Reliability and validity in a nutshell. *Journal of Clinical Nursing*, *18*(23), 3237–3243. https://doi.org/10.1111/j.1365-2702.2009.02939.x
11. Biggs, J., & Tang, C. (2011). Teaching for quality learning at university. What the student does (4th Edn.). In *Innovations in Education and Teaching International* (4nd editio, Vol. 50, Issue 4). SRHE and Open University Press. https://cetl.ppu.edu/sites/default/files/publications/-John_Biggs_and_Catherine_Tang-_Teaching_for_Quali-BookFiorg-.pdf
12. Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model : Fundamental measurement in the human sciences* (2ed ed.). Lawrence Erlbaum Associates, Inc. https://researchonline.jcu.edu.au/9907/2/9907
13. Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences,* (3rd ed.). Routledge. https://doi.org/https://doi.org/10.1111/j.1745-3984.2003.tb01103.x
14. Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model fundamental measurement in the human sciences* (4th Editio). Routledge. https://doi.org/https://doi.org/10.4324/9780429030499
15. Boone, W. J. (2016). Rasch analysis for instrument development: Why,when,and how? *CBE Life Sciences Education*, *15*(4), 1–7. https://doi.org/10.1187/cbe.16-04-0148
16. Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the human sciences*. Springer. https://doi.org/10.1007/978-94-007-6857-4
17. Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061
18. Brown, A., Nidumolu, A., McConnell, M., Hecker, K., & Grierson, L. (2019). Development and psychometric evaluation of an instrument to measure knowledge, skills, and attitudes towards quality improvement in health professions education: The Beliefs, Attitudes, Skills, and Confidence in Quality Improvement (BASiC-QI) scale. *Perspectives on Medical Education*, *8*(3), 167–176. https://doi.org/10.1007/s40037-019-0511-8
19. Buntragulpoontawee, M., Khunachiva, J., Euawongyarti, P., Wongpakaran, N., Wongpakaran, T., Kaewma, A., & Ashford, S. (2021). Investigating psychometric properties of the arm activity measure – Thai version (ArmA-TH) sub-scales using the Rasch model. *BMC Medical Research Methodology*, *21*(1), 1–9. https://doi.org/10.1186/s12874-021-01238-5
20. Creswell, J. W., & Creswell, J. D. (2017). *Research design qualitative, quantitative, and mixed methods approaches* (4 th ed). SAGE Publications, Inc.
21. Curtin, M., Browne, J., Staines, A., & Perry, I. J. (2016). The Early development instrument: An evaluation of its five domains using Rasch analysis. *BMC Pediatrics*, *16*(1), 1–14. https://doi.org/10.1186/s12887-016-0543-8
22. Darman, D. R., Suhandi, A., Kaniawati, I., Samsudin, A., & Wibowo, F. C. (2024). Development and validation of Scientific Inquiry Literacy Instrument (SILI) using Rasch measurement model. *Education Sciences*, *14*(3), 1–28. https://doi.org/10.3390/educsci14030322
23. DeMars, C. (2010). *Item Response Theory*. Oxford University Press, Inc. https://api.pageplace.de/preview/DT0400.9780199703845_A23607784
24. DeVellis, R. F. (2017). Scale development theory and applications. In *SAGE Publication* (4th ed). SAGE Publications, Inc. All.

25. Drake, D., Kennedy, R., & Wallace, E. (2017). The validity and responsiveness of isometric lower body multi-joint tests of muscular strength: A systematic review. *Sports Medicine - Open*, *3*(1), 1–11. https://doi.org/10.1186/s40798-017-0091-2
26. Ekstrand, J., Westergren, A., Årestedt, K., Hellström, A., & Hagell, P. (2022). Transformation of Rasch model logits for enhanced interpretability. *BMC Medical Research Methodology*, *22*(1), 1–10. https://doi.org/10.1186/s12874-022-01816-1
27. Engelhard, G., & Wang, J. (2022). *Rasch Models for Solving Measurement Problems: Invariant Measurement in the Social Sciences*. SAGE Publications, Inc. https://doi.org/10.4135/9781071878675
28. Fitrah, M., Sofroniou, A., Ofianto, Judijanto, L., & Widihastuti. (2024). Reliability and separation index analysis of mathematics questions integrated with the cultural architecture framework using the Rasch model. *Journal of Education and E-Learning Research*, *11*(3), 499–509. https://doi.org/10.20448/jeelr.v11i3.5861
29. Frick, H., Strobl, C., & Zeileis, A. (2015). Rasch mixture models for DIF detection: A comparison of old and new score specifications. *Educational and Psychological Measurement*, *75*(2), 208–234. https://doi.org/10.1177/0013164414536183
30. Gunawan, B., Rahayu, W., & Aziz, T. A. (2024). Rasch model analysis: Validity and reliability of context-based geometry performance assessment instruments Jakarta Cultural Heritage. *EduMa: Mathematics Education Learning and Teaching*, *13*(2), 131. https://doi.org/10.24235/eduma.v13i2.18136
31. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1992). Fundamentals of item response theory. *American Sosiological Association*, *21*(2), 289–290. http://www.jstor.org/stable/2075521
32. Hamzah, F. M., Rashid, M. N. A., Rahman, M. N. A., & Rasul, M. S. (2022). Evaluating the validity and reliability of authentic learning instruments using Rasch Model. *International Journal of Global Optimization and Its Application*, *1*(3), 182–189. https://doi.org/10.56225/ijgoia.v1i3.69
33. Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence-Based Nursing*, *18*(3), 66–67. https://doi.org/10.1136/eb-2015-102129
34. Heaney, A., McKenna, S. P., Hagell, P., & Sengupta, R. (2020). Improving scoring precision and internal construct validity of the bath ankylosing spondylitis disease activity index using Rasch measurement theory. *Journal of Rheumatology*, *47*(3), 354–361. https://doi.org/10.3899/jrheum.180943
35. Hoeve, K. B. (2022). A validity framework for accountability: Educational measurement and language testing. *Language Testing in Asia*, *12*(1), 1–14. https://doi.org/10.1186/s40468-021-00153-2
36. Holden, L. T. R., & Tanenbaum, G. J. (2023). Modern assessments of intelligence must be fair and equitable. *Journal of Intelligence*, *11*(6), 1–20. https://doi.org/10.3390/jintelligence11060126
37. Hope, D., Adamson, K., McManus, I. C., Chis, L., & Elder, A. (2018). Using differential item functioning to evaluate potential bias in a high stakes postgraduate knowledge based assessment. *BMC Medical Education*, *18*(1), 1–7. https://doi.org/10.1186/s12909-018-1143-0
38. Huang, S. M., Fang, S. C., Hung, C. T., & Chen, Y. H. (2022). Psychometric evaluation of a nursing competence assessment tool among nursing students: A development and validation study. *BMC Medical Education*, *22*(1), 1–9. https://doi.org/10.1186/s12909-022-03439-y
39. Humphris, G. M., Watson, E., Sharpe, M., & Ozakinci, G. (2018). Unidimensional scales for fears of cancer recurrence and their psychometric properties: The FCR4 and FCR7. *Health and Quality of Life Outcomes*, *16*(1), 1–12. https://doi.org/10.1186/s12955-018-0850-x
40. Joo, S., Ali, U., Robin, F., & Shin, H. J. (2022). Impact of differential item functioning on group score reporting in the context of large-scale assessments. *Large-Scale Assessments in Education*, *10*(1), 1–22. https://doi.org/10.1186/s40536-022-00135-7
41. Kesmodel, U. S. (2018). Cross-sectional studies – what are they good for? *Acta Obstetricia et Gynecologica Scandinavica*, *97*(4), 388–393. https://doi.org/10.1111/aogs.13331
42. Lamoureux, E. L., Pesudovs, K., Thumboo, J., Saw, S. M., & Wong, T. Y. (2009). An evaluation of the reliability and validity of the visual functioning questionnaire (VF-11) using rasch analysis in an Asian population. *Investigative Ophthalmology and Visual Science*, *50*(6), 2607–2613. https://doi.org/10.1167/iovs.08-2359

43. Linacre, J. M. (2002). Optimizing rating scale category effectiveness optimizing rating scale category effectiveness university of Chicago. *Journal of Applied Measurement*, *3*(1), 85–106. https://www.researchgate.net/publication/292264351
44. Linacre, J. M. (2010). Predicting responses from rasch measures. *Journal of Applied Measurement*, *11*(1), 1–10. https://www.winsteps.com/a/Linacre-Predicting.pdf
45. Linacre, J. M. (2012). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *6*(2), 878. https://www.rasch.org/rmt/rmt162.htm
46. Linacre, J. M. (2017). Rasch measurement transactions: Teaching Rasch measurement. *Autumn.*, *31*(2), 1630–1642. www.rasch.org/rmt/rmt111n.htm
47. Linacre, J. M. (2024). Reasonable mean-square fit values revisited. *Rasch Measurement Transactions*, *36*(3), 1–6. https://www.rasch.org/rmt/rmt83b.htm?utm_source=chatgpt.com
48. Linacre, J. M., Boone, W., Green, R., Trevor, G., & Christine, M. (2023). *A User's Guide to WINSTEPS® / MINISTEP Rasch-model computer programs* (5.10; pp. 2–4). Winsteps.com. www.winsteps.com/manuals.htm
49. Maric, D., Fore, G. A., Nyarko, S. C., & Varma-Nelson, P. (2023). Measurement in STEM education research: A systematic literature review of trends in the psychometric evidence of scales. *International Journal of STEM Education*, *10*(1), 1–22. https://doi.org/10.1186/s40594-023-00430-x
50. Papini, N. M., Jung, M., Kang, M., Lopez, N. V., & Herrmann, S. D. (2023). Development and Rasch analysis of the 18-Item Health Resilience Profile (HRP). *Nutrients*, *15*(4), 1–17. https://doi.org/10.3390/nu15040807
51. Pilotti, M. A. E. (2021). What lies beneath sustainable education? Predicting and tackling gender differences in STEM academic success. *Sustainability*, *13*(4), 1671. https://doi.org/10.3390/su13041671
52. 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. https://doi.org/10.1111/joca.12322
53. Rezai, A. (2022). Fairness in classroom assessment: development and validation of a questionnaire. *Language Testing in Asia*, *12*(1), 1–27. https://doi.org/10.1186/s40468-022-00162-9
54. Ridwan, M. R., Hadi, S., & Jailani, J. (2023). Measurement of Psychometric Pproperties Numerical Aptitude Assessment Scale for prospective high school students: A Rasch model analysis. *TEM Journal*, *12*(4), 2416–2429. https://doi.org/10.18421/TEM124-54
55. Schumacker, R. E., & Smith, E. V. (2007). Reliability a rasch perspective. *Educational and Psychological Measurement*, *67*(3), 394–409. https://doi.org/10.1177/0013164406294776
56. Sijtsma, K., & van der Ark, L. A. (2022). Advances in nonparametric item response theory for scale construction in quality-of-life research. *Quality of Life Research*, *31*(1), 1–9. https://doi.org/10.1007/s11136-021-03022-w
57. Supriatna, M., Suryana, D., Anzhali, M. N., & Noor, A. M. (2025). Evaluating career self-awareness instruments for victims of violence using the Rasch model. *BMC Psychology*, *13*(1), 1–13. https://doi.org/10.1186/s40359-025-02762-5
58. Tan, L. Y., McLean, S., Kim, Y. A., & Vitta, J. P. (2024). Rasch modelling vs. item facility: Implications on the validity of assessments of Asian EFL/ESL vocabulary knowledge and lexical sophistication modelling. *Language Testing in Asia*, *14*(1), 1–22. https://doi.org/10.1186/s40468-024-00327-8
59. Tennant, A., & Küçükdeveci, A. A. (2023). Application of the Rasch measurement model in rehabilitation research and practice: Early developments, current practice, and future challenges. *Frontiers in Rehabilitation Sciences*, *4*(July), 01–17. https://doi.org/10.3389/fresc.2023.1208670
60. Tennant, A., McKenna, S. P., & Hagell, P. (2004). Application of Rasch analysis in the development and application of quality of life instruments. *Value in Health*, *7*(Suplement. 1), 22–26. https://doi.org/10.1111/j.1524-4733.2004.7s106.x
61. Tutz, G. (2024). Invariance of comparisons: Separation of item and person parameters beyond Rasch models. *Journal of Mathematical Psychology*, *122*(August), 102876. https://doi.org/10.1016/j.jmp.2024.102876
62. Vincent, J. I., MacDermid, J. C., King, G. J. W., & Grewal, R. (2015). Rasch analysis of the Patient Rated Elbow Evaluation questionnaire. *Health and Quality of Life Outcomes*, *13*(1), 1–13. https://doi.org/10.1186/s12955-015-0275-8
63. Wiley, J. A., Martin, J. L., Herschkorn, S. J., & Bond, J. (2015). A new extension of the binomial error model for responses

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 08 August 2026
DOI: 10.47191/ijcsrr/V9-i8-07, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

to items of varying difficulty in educational testing and attitude surveys. *PLoS ONE*, *10*(11), 1–16. https://doi.org/10.1371/journal.pone.0141981
64. Wilson, M. (2023). *Constructing Measures: An Item Response Modeling Approach* (Second Edi). Routledge Falmer Taylor & Francis Group. https://doi.org/10.4324/9781003286929
65. Woudstra, A. J., Meppelink, C. S., Pander Maat, H., Oosterhaven, J., Fransen, M. P., & Dima, A. L. (2019). Validation of the short assessment of health literacy (SAHL-D) and short-form development: Rasch analysis. *BMC Medical Research Methodology*, *19*(1), 1–10. https://doi.org/10.1186/s12874-019-0762-4
66. Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370. https://www.rasch.org/rmt/rmt83b.htm
67. Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis item response modeling approach*. Mesa Press.
68. Wright, B. D., & Stone, M. H. (1999). Measurement essentials. In *Measurement* (2 Ed). WIDE RANGE, INC. https://www.rasch.org/measess/me-all.pdf
69. Yang, H., Su, J., & Bradley, K. D. (2020). Applying the Rasch model to evaluate the Self-Directed Online Learning Scale (SDOLS) for graduate students. *International Review of Research in Open and Distributed Learning*, *21*(3), 99–120. https://doi.org/https://doi.org/10.19173/irrodl.v21i3.4654
70. Yusop, S. R. M., Rasul, M. S., Yasin, R. M., & Hashim, H. U. (2023). Identifying and validating vocational skills domains and indicators in classroom assessment practices in TVET. *Sustainability (Switzerland)*, *15*(6), 1–27. https://doi.org/10.3390/su15065195
71. Zumbo, B. D. (2006). *Psychometric methods for investigating DIF and test bias during test adaptation across languages and cultures* (pp. 1–85). Lawrence Erlbaum Associates. https://faculty.educ.ubc.ca/zumbo/talks/Zumbo_ITC.PDF
72. Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bBias in educational policy research: Bringing the context into picture by iInvestigating sociological / community moderated (or mediated) test and item bias. *Journal of Educational Research & Policy Studies*, *5*(1), 1–23. https://faculty.educ.ubc.ca