International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-05, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

Validation of Applied Mathematics Exam Test Using Rasch Model Approach: Case Study in Diploma 3 Study Program of Refrigeration and Air Conditioning Engineering Politeknik Negeri Bali

I Ketut Darma[1], I Gde Nyonan Sangka[2]
[1,2] Kampus Politeknik Negeri Bali, Bukit Jimbaran, Kuta Selatan, Tuban, Badung Bali, Indonesia.

ABSTRACT: This study validates the Applied Mathematics exam (AME) test on Diploma 3 (D3) Study Program of Refrigeration and Air Conditioning Engineering (RACE), Politeknik Negeri Bali (PNB) to determine the validity, reliability, unidemensional, level of difficulty, and discriminatory power of the test. Validation uses the modern test theory approach of the Rasch model. Data were collected during the online final exam of the even semester of the 2023/2024 academic year. The instrument uses a multiple-choice test form with 5 answer options. The sample involved 73 even semester students of the D3 RACE study program who took applied mathematics courses. The data collected were analyzed using the Rasch Model assisted by the Winsteps application. The results of the analysis show that the AME test has an adequate level of validity, most of the questions meet the fit criteria for the Rasch Model. The level of test reliability is categorized as very good with a person and item reliability value of 0.90. Several questions still show misfits that require improvement. Item difficulty and person ability show a proportional distribution between the level of difficulty of the questions and the students' abilities. Overall, the test's discriminatory power is categorized as good, although there is one question that needs to be reviewed further for improvement. The implication is that the use of the Rasch Model in validating online test instruments can help teachers in compiling questions that are more valid, reliable, and in accordance with the level of student ability. The implication is that the application of the Rasch Model in validating test instruments can help lecturers in constructing more valid and reliable tests. The results of this study can be used as an empirical example of the application of the Rasch model theory to produce more valid and reliable measurements. It is recommended that the development of future test tests really needs to pay attention to the balance between the level of difficulty of the items and the abilities of the students, and ensure that the measurements are more valid and reliable, especially in the context of polytechnic education.

KEYWORDS: Applied Mathematics, Polytechnic, Reliability, Rasch Model, Test, Validity.

INTRODUCTION

Measuring students' ability to understand lecture materials is a crucial aspect in the learning evaluation process in higher education. One of the evaluation instruments used is the final semester exam test, which aims to measure learning outcomes in the course. However, the effectiveness and validity of these examination instruments are often debated. Examination instruments that do not meet validity and reliability standards can provide biased evaluation results and do not reflect actual abilities [1–4].

An evaluation instrument including an Applied Mathematics test can be said to be of quality, it must meet several main criteria, namely validity, reliability, discrimination power, and appropriate test difficulty level [6–11]. Validity refers to the extent to which an instrument actually measures what it is supposed to measure [12–14]. A valid test ensures that the test items are in accordance with the learning objectives that have been set and cover all aspects of the material that is the focus of the evaluation. Without adequate validity, evaluation results tend to be irrelevant to the abilities to be measured. Meanwhile, Reliability refers to the consistency of measurement results [12–15]. A reliable instrument will provide stable results even if tested at different times or groups of students. This is important to ensure that the test results truly reflect students' abilities, not influenced by random factors such as technical errors or variations in the implementation of the exam. In addition, the test must also have good discriminatory power. Discriminatory power is the ability to distinguish between students who have better understanding and those who have less [12,14]. Tests with low discriminatory power tend to provide uninformative results because they fail to map variations in ability between students.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-05, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

The level of test difficulty also plays an important role in ensuring the effectiveness of the test. The level of test difficulty is the proportion of participants who answer a question correctly [12,14]. A good test should include a distribution of questions with various levels of difficulty, ranging from easy, medium, to difficult. The goal is to ensure that the test is able to measure students' abilities at various levels. If all the questions are too easy, the test will fail to challenge students who have a deep understanding. Conversely, if all the questions are too difficult, students with basic understanding cannot answer well, so that the evaluation results are not representative. A proportional combination of difficulty levels will help produce a comprehensive picture of students' abilities [16–18]. Tests that meet these requirements will contribute to a fair, accurate, and representative evaluation. Therefore, the process of preparing an examination instrument requires not only a deep understanding of the material, but also the application of good evaluation principles [14] . This ensures that tests can function as effective measuring tools in supporting improvements in the quality of learning and achievement of student learning outcomes.

Applied Mathematics is one of the basic courses in the D3 study program of RACE at the PNB which focuses on the application of mathematical concepts in solving real problems. The characteristics of mathematics courses require good conceptual understanding and analytical skills, quality evaluation instruments are needed and are able to measure students' abilities accurately.

Test validity can be determined through various approaches, one of which is the Rasch model approach. The Rasch model is one method in item response theory that is able to evaluate the extent to which test items on the test instrument function consistently and fairly in measuring student abilities [19]. This approach is not only able to evaluate the quality of test items in terms of their level of difficulty, but also provides an in-depth analysis of test reliability, unidimensionality, and potential bias that may be present in the test items [20,21].

The advantage of the Rasch model approach compared to other approaches in educational measurement practice lies in its ability to produce more objective and independent measures of specific test items. This model allows the estimation of test item parameters and student abilities that are not influenced by sample characteristics, so that measurement results are more consistent and generalizable [22]. In addition, the Rasch model can detect misfitting items and provide clear information about the overall quality of the instrument, which is difficult to achieve by other approaches such as the Classical Test Theory [23–26]. Thus, the Rasch model approach offers a higher level of accuracy and validity in evaluating measurement instruments.

Research conducted by Bond & Fox shows that the Rasch model is able to identify invalid test items and provide a more accurate estimate of the test taker's ability [15]. Research conducted by Boone also states that the use of the Rasch model helps instrument developers improve the reliability and validity of measurements through item fit and person fit analysis [27]. In addition, research in various educational contexts shows that the Rasch model is more effective in detecting Differential Item Functioning (DIF) than the classical approach, thus minimizing bias in measuring student ability [24,26].

This study validates the AME test on the D3 RACE study program of PNB using the Rasch model approach. The aim is to determine the validity, reliability, unidemensional, level of difficulty, and discriminatory power of the test. Through this analysis, it is expected to obtain a comprehensive picture of the quality of the test items in the exam as a whole. In addition, the results of this study are expected to be input for the development of better and more reliable examination instruments in the context of polytechnic education.

RESEARCH METHODS

This study was conducted on students of the D3 TPTU study program, Department of Mechanical Engineering of PNB, academic year 2023/2024. Data were collected during the Even Semester Final Exam. The instrument used was a multiple-choice type applied mathematics final semester exam, consisting of 40 items with 5 answer options.

The test was designed to measure the achievement of applied mathematics learning. The instrument includes three types of learning achievement sub-competencies, namely: limit, differential, and integral. The data used were secondary data, namely the results of students' answers on the Even Semester Final Exam in Applied Mathematics collected through the documentation method. The collected data were analyzed using the modern test theory approach of the Rasch model with the help of Winsteps software. The Rasch model can detect misfitting items and provide clear information about the overall quality of the instrument. This model allows estimating item parameters and student abilities that are not influenced by sample characteristics, so that measurement results are more consistent and generalizable [22,28,29].

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-05, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

Test validity and reliability play an important role in validating test items and overall test performance. In the Rasch Model approach, assisted by the Winsteps application program, validity is evaluated through several indicators, namely: Fit Statistics, Unidimensionality, Differential Item Functioning (DIF), and Point Measure Correlation (Pt. Measure Corr.) [15,18,21,30]

Fit Statistics is used to examine the extent to which participants' responses to test items fit the Rasch model. Its two main measures are Infit and Outfit. Infit is sensitive to participants' responses at ability levels that are close to the level of item difficulty. It measures the expected mismatch based on the level of item difficulty faced by the participants. Outfit is more sensitive to extreme responses such as unexpected answers to very easy or very difficult items. It measures the mismatch in unexpected responses, used to detect answers that deviate from the model (anomalies). The criterion for item validity is that the Mean-Square (MNSQ) value is in the range between 0.5 and 1.5 is considered a good fit indicator. Values outside this range indicate misfit and the item needs to be reviewed. The ZSTD (Z-Standardized) value is the standard version of the fit statistic. Acceptable ZSTD values are in the range between -2.0 and +2 [18,30,31]. Values close to 0 indicate that there is no significant discrepancy. Positive values that are too high indicate overfit, the response is too appropriate to the model. While negative values that are too low indicate underfit, the response does not match the model. Another measure is Point Measure Correlation (Pt Measure Corr). Pt Measure Corr is the correlation value between the score on a particular item and the total score of the participant's ability estimated by the Rasch Model. The goal is to ensure that each item contributes positively to the measurement of the construct being measured. Positive values indicate consistency of answers with the Rasch model. while low or negative values indicate inappropriate answer patterns. The accepted normal standard value is in the range between 0.4 and 0.8 [15,18,21,30].

Instrument validity in the Rasch Model assumes that the test is unidimensional, that is, it only measures one construct or attribute [15,31]. Unidimensionality in the context of the Rash model is evaluated using Principal Component Analysis (PCA) of Residuals. The indicator, the variance explained by the Rasch model is at least 40%, indicating strong unidimensionality. While for the variance of the residual (unexplained variance) is a maximum of 15% [15,20,31]. Differential Item Functioning (DIF), is used to evaluate whether a question item provides an advantage or disadvantage to a particular group. DIF occurs when an item in a test functions differently for different groups of participants such as gender, ethnicity, educational background, or other factors, even though they have the same abilities. If there is significant DIF, it will affect validity and reduce fairness in measurement. A test is said to have bias if the p-value is less than 0.05 [20].

Reliability is related to the consistency of instrument measurement, namely the extent to which the instrument produces stable and reliable results. In the Rasch model approach, reliability includes person and item reliability as well as person and item separation index [15,31]. Person reliability, shows the consistency of participants' ability to answer questions. The higher the person reliability value, the more consistent the participants' responses. Item reliability, indicates how consistently the instrument measures the level of difficulty of the questions. High values indicate sufficient variation in the level of difficulty between questions. Person separation index, shows the extent to which the instrument can separate participants based on their ability level. Item separation shows the extent to which the questions are spread based on their level of difficulty. A high separation value indicates a better test ability to distinguish participants and questions. The higher the person separation index value, the better the instrument is at separating test participants based on their abilities [15,20,21].

Further analysis is related to the level of item difficulty, respondent ability level, and discrimination power. Item difficulty is one of the main components that determines how difficult or easy an item in a test is for test takers. In the context of the Rash model, item difficulty is a numerical estimate expressed in a logit scale, describing how difficult a test item is to be answered correctly by participants. The level of item difficulty on the logit scale is mapped variably from negative to positive. A logit value of 0 is considered the average level of difficulty [15,21,32]. A positive value indicates that the item is more difficult than other items. While a negative value indicates that the item is easier, more participants are able to answer it correctly. The level of item difficulty is seen in the Item Measure table output, detailing the logit information of each question item. A high logit value indicates a high level of question difficulty, the higher the logit value, the higher the level of question difficulty, and vice versa. In addition, Item Measure can also inform the standard deviation (SD) value. SD value combined with the average logit value, the level of item difficulty can be grouped into: 0.0 + 1SD is a group of difficult questions, greater than 0.00 + 1 SD is a group of very difficult questions, 0.0 logit - 1SD is a group of questions categorized as easy, and less than 0.00 - 1 SD is a group of questions categorized as very easy [9,15,21,32].

The level of individual ability in completing the test is measured through Person Measure analysis. Person Measure details the logit information of each individual. A high logit value indicates a high level of ability to solve the problem. This corresponds to

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-05, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

the total score column, which states the number of correct answers. Furthermore, the mean and SD values can be used to classify respondents based on their abilities. The mean value ± 1 SD is used as the limit for classifying high, medium, or low ability levels [21,32]. The respondent's ability level is classified into: 1) high ability for respondents who have a person measure value higher than the mean + 1 SD, 2) medium ability for respondents with a person measure value around the mean, namely between mean - 1 SD to mean + 1 SD, and 3) low ability for respondents with a person measure value lower than the mean - 1 SD [9,15,21,32].

Item discrimination power is the ability of a test item to distinguish test takers with high ability from those with low ability. In the rash model, Standard Error (S.E) is one of the tools that can be used to assess whether an item has good or bad discrimination power. S.E indicates the level of accuracy of the item difficulty parameter estimate. A lower S.E value indicates a more accurate item parameter estimate and is better at distinguishing between good and bad test takers [9,15]. An S.E value of about 0.5 to 1 indicates that the test has sufficient discrimination power. While an S.E value of more than 1 indicates that the test has poor discrimination power.

The Wright Map graph displays the distribution of participant ability and item difficulty on the same scale. Its benefits are to see the distribution of participant ability and item difficulty, and whether there are gaps in the test instrument. In the Wright Map, Person and Item are divided into 3 parts, namely person, map, and item. Person shows student ability, map shows student ability mapping and item shows the level of question difficulty. The higher the position, the higher the person's ability and the difficulty of the question, and vice versa. Validation of the AME test using the Rasch model is seen from various aspects and, the summary is presented in Table 1.

Table 1. Summary of Validation of the Mathematics exam test seen from various aspects and criteria of the Rasch model

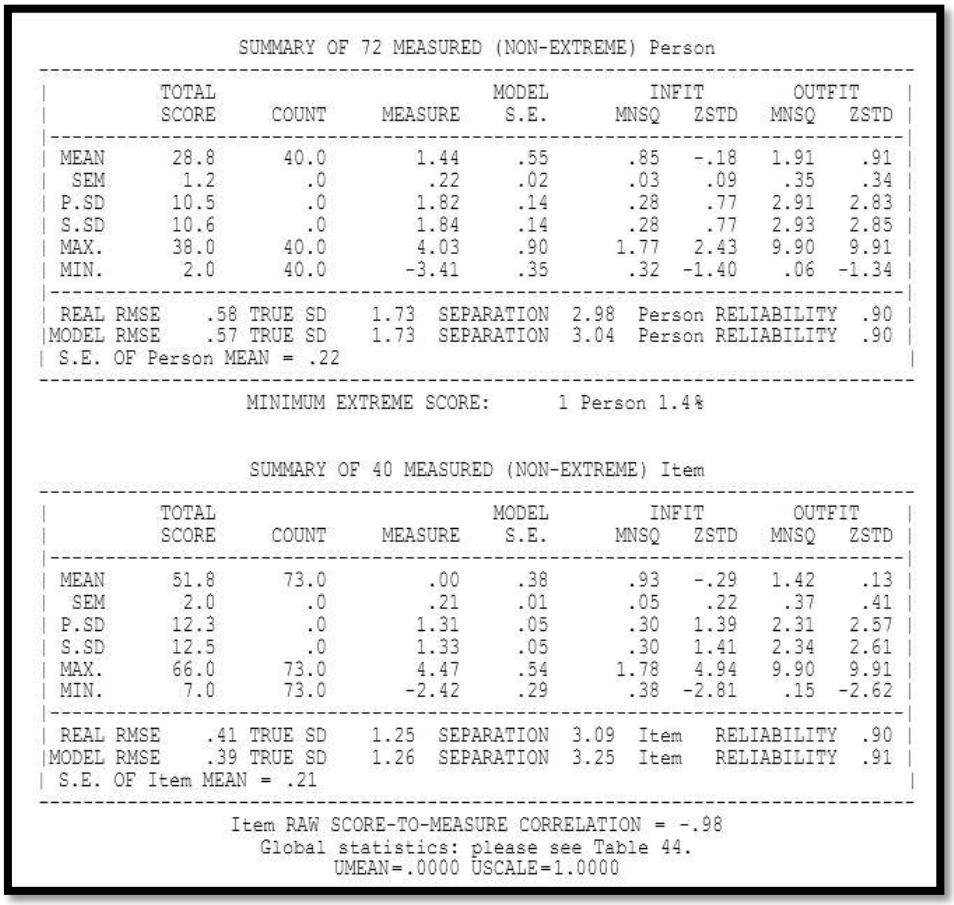
Aspect of the Item	Measurement Parameters	Criteria
Fit item test	Outfit Mean Square Value (MNSQ)	
	Outfit ZStandard Value (ZSTD) Point Measure Correlation Value (Pt Measure Corr)	0.5 < MNSQ < 1.5 -2.0 < ZSTD<2.0 0.4 < Pt Measure Corr < 0.85
Reability		measure < 0.67 :weak
	Person reability	0.67 ≤ measure < 0.80 :sufficient
	Item reability	0.80 ≤ measure < 0.90 :good
		0.90 ≤ meansure ≤ 0.94 :very good
		meansure > 0.94 :excellen
Overall reliability	Alhpa Cronbach	measure < 0.5 : very bad 0.5 ≤ measure ≤ 0.6 : bad 0.6 ≤ measure ≤ 0.7 : adequate 0.7 ≤ meansure ≤0.8 : good meansure ≤ 0.8 : very good
Separation	*Separation Index (Si)*	0 < Si <1.5 : low. 1.5 ≤ Si < 2.0 : sufficient 2.0 ≤ Si ≤ 3.0 : good Si < 3.0 : very Good
Item suitability Item bias	*Dif. Person Functioning (DPF)*	The probability value of the question item is > 5% or 0.05
Unidimensional item test	Raw (crude) variance value	The variance explained by the Rasch model was at least 40%, indicating strong unidimensionality.
	Unknown variance value	Maximum reach 15%
Item difficulty	Item Measure	Measure logit > 1SD :very difficult 0.00 < Measure logit ≤ 1 SD :difficult

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-05, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

Aspect of the Item	Measurement Parameters	Criteria
Level of quality of test participants' abilities	Person Measure	–1 SD ≤ Measure logit ≤ 0.00 :easy Measure logit < – 1SD :very easy measure < 0.5 :very bad 0.5 ≤ measure ≤ 0.6 :bad 0.6 ≤ measure ≤ 0.7 :adequate 0.7 ≤ meansure ≤0.8 :good meansure ≤ 0.8 :very good
Different power	Standard Error (S.E)	S.E < 0.5 good discrimination power 0.5 ≤ S.E ≤ 1 sufficient discrimination power S.E > 1 poor discrimination power

RESULTS AND DISCUSSION

This study validates the AME test of the D3 RACE Study Program, PNB. The aim is to determine the level of validity and reliability reviewed from the modern test theory of the Rasch model. The collected data were analyzed using the Rasch model assisted by the Winsteps application, the results of the statistical summary are presented in Figure 1 and Table 2.



SUMMARY OF 72 MEASURED (NON-EXTREME) Person

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	28.8	40.0	1.44	.55	.85	-.18	1.91	.91
SEM	1.2	.0	.22	.02	.03	.09	.35	.34
P.SD	10.5	.0	1.82	.14	.28	.77	2.91	2.83
S.SD	10.6	.0	1.84	.14	.28	.77	2.93	2.85
MAX.	38.0	40.0	4.03	.90	1.77	2.43	9.90	9.91
MIN.	2.0	40.0	-3.41	.35	.32	-1.40	.06	-1.34

REAL RMSE .58 TRUE SD 1.73 SEPARATION 2.98 Person RELIABILITY .90
MODEL RMSE .57 TRUE SD 1.73 SEPARATION 3.04 Person RELIABILITY .90
S.E. OF Person MEAN = .22

MINIMUM EXTREME SCORE: 1 Person 1.4%

SUMMARY OF 40 MEASURED (NON-EXTREME) Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	51.8	73.0	.00	.38	.93	-.29	1.42	.13
SEM	2.0	.0	.21	.01	.05	.22	.37	.41
P.SD	12.3	.0	1.31	.05	.30	1.39	2.31	2.57
S.SD	12.5	.0	1.33	.05	.30	1.41	2.34	2.61
MAX.	66.0	73.0	4.47	.54	1.78	4.94	9.90	9.91
MIN.	7.0	73.0	-2.42	.29	.38	-2.81	.15	-2.62

REAL RMSE .41 TRUE SD 1.25 SEPARATION 3.09 Item RELIABILITY .90
MODEL RMSE .39 TRUE SD 1.26 SEPARATION 3.25 Item RELIABILITY .91
S.E. OF Item MEAN = .21

Item RAW SCORE-TO-MEASURE CORRELATION = -.98
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000

Figure 1. Summary of Statistics

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-05, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

Table 2. Summary of Statistics of Validity and Reliability Testing Results of AME Tests in the D3 RACE Study Program, PNB.

Measurement Parameters	Results	
	Person (76)	item (40)
Mean logit	1.44	0.00
Maximum	4.03	4.47
Minimum	-3.41	-2.42
Standard Deviation (SD)	1.84	1.31
Validity		
Item-fit and Person-fit Statistik		
Outfit MNSQ	1.91	0.96
Outfit ZSTD	0.91	-0.04
Reliability	0.90	0.90
Separation	2.98	3.09
Cronbach Alpha Reliability	0.89	
Unidimensionality		
Raw Variance Explained by Measures	54.2% %.	
Unexplained Variance	None has a value of more than 15%,	
Item-Person Map (Wright Map)		
Distribution of Participant Ability and Item Difficulty	The scale moves from the highest (+4) to the lowest (-2) logit. Most participants are around the 0 to +2 logit.	
Match between Participants and Items	The distribution of participants is quite even in the logit range of 0 to +3, but there are some areas where items or persons are not evenly distributed. namely: Logit +3 to +4 and Logit -1 to -2 There are no items in logit +2 to +3	
Items with Low or High Match:	Items with a difficulty level that is much higher than the participant's ability are slightly at level +4, while items with a low level of difficulty are spread across levels -1 to -2. The distribution of item difficulty levels shows a gap at logit +1 to +2	
No Differential Item Functioning (DIF) Bias	There are no items that have a probability value of less than 5% ($p < 0.05$)	

Table 2 shows the mean logit value of the person measure of 1.44 and the item measure value of 0.0 logit, which means that the person measure value is higher than the item measure value. The test taker's ability is higher than the level of difficulty of the questions. Test takers have the potential to answer all questions correctly. So test takers who have high ability are able to answer the most difficult and easiest questions correctly.

Person reliability of 0.90 is categorized as very good [15,21]. In general, students provide very stable answers across all items. The ability of students taking the test is very reliable in answering the test. Item reliability of 0.90 is categorized as very good [15,21]. This applied mathematics exam test is very good at differentiating abilities between test participants. The items have very good quality in measuring the targeted competencies. Furthermore, Cronbach's Alpha measures the overall internal consistency of the test [33]. The Cronbach's Alpha value of 0.97 is categorized as excellent [15,21]. This AME test is very reliable. The items in it are very well related to each other in measuring the same concept in applied mathematics, so that the measurement results can be considered very accurate.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-05, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

Person separation of 2.98 is categorized as good and item separation of 3.09 is categorized as very good [15,21]. The person separation index value of 2.98 indicates that the test has good ability in differentiating respondents based on their abilities. The test is able to differentiate students' abilities into about 4 groups. The item separation index value of 3.09 indicates that the items in this applied mathematics exam test can be separated into about 4-5 groups of different levels of difficulty [15,21]. This value is a good value, and proves that this test has a good variation in difficulty levels in measuring the abilities of test participants with various levels of ability. Overall, this Applied Mathematics exam test shows very good quality in measuring students' abilities, with a high level of consistency both from the participant side and from the item side.

The Outfit MNSQ person and item outfit values were 1.91 and 0.96 respectively. Both met the fit criteria because they were in the range of 0.5 < MNSQ < 1.5. This indicates that the test used in the Final exams is in accordance with the model for measuring competencies formulated in the learning outcomes of Applied Mathematics. Furthermore, the Outfit ZSTD person and item values were 0.91 and -0.04 respectively, both also met the fit criteria, because they were in the range of -0.2 < ZSTD < 0.2. This condition indicates that the data has the potential for rational values according to the expectations of the Rasch model. Therefore, all test items are declared suitable for use as test instruments in subsequent testing.

In the context of the Rasch model, test validity refers to the extent to which a test consistently measures a particular construct or competency formulated in learning outcomes [15,21,31]. The suitability of the data to the Rasch Model is one aspect used to evaluate test validity. The MNSQ outfit value in the range between 0.5 and 0.15; the ZSTD outfit value in the range between -2 and 2, and the Point Measure Correlation value in the range between 0.4 and 0.85 are the criteria used to measure the level of item fit [15,19,34]. If these three criteria are not met by an item, then it is certain that the item does not match the expectations of the Rasch model and is not good enough so that it needs to be improved or replaced. The test items are said to fit the Rash model if they cover either one or both criteria are met [30]. The level of suitability of the Applied Mathematics test items can be seen from the results of the item fit order analysis in Figure 2 below (the image display is not full).

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
4	7	73	4.47	.42	1.38	1.26	9.90	9.91	A-.33	.24	90.3	90.3	B4
5	8	73	4.31	.40	1.47	1.65	9.90	9.91	B-.50	.25	88.9	88.9	B5
7	21	73	2.90	.29	1.66	4.94	8.29	5.56	C-.11	.40	70.8	74.4	B7
9	66	73	-2.42	.54	1.78	2.13	2.86	1.47	D .38	.61	87.5	93.3	B9
32	54	73	-.10	.37	1.39	1.52	1.37	.96	E .55	.67	83.3	86.8	B32
24	55	73	-.23	.38	1.02	.15	1.21	.60	F .66	.68	87.5	87.4	B24
19	60	73	-1.05	.43	1.01	.13	1.20	.51	G .66	.68	91.7	90.3	B19
15	55	73	-.23	.38	1.19	.79	1.19	.56	H .62	.68	84.7	87.4	B15
36	50	73	.40	.34	1.15	.78	1.04	.22	I .60	.65	80.6	83.9	B36
25	62	73	-1.44	.46	1.14	.55	.52	-.50	J .66	.66	88.9	91.1	B25
1	53	73	.04	.36	.98	-.02	1.13	.46	K .67	.67	84.7	86.1	B1
31	54	73	-.10	.37	1.12	.55	.94	-.02	L .64	.67	86.1	86.8	B31
40	50	73	.40	.34	1.10	.52	1.12	.46	M .61	.65	83.3	83.9	B40
3	49	73	.51	.33	1.10	.54	1.01	.14	N .62	.64	81.9	83.1	B3
12	49	73	.51	.33	1.05	.33	.98	.05	O .63	.64	81.9	83.1	B12
6	59	73	-.87	.42	.97	-.03	.65	-.51	P .71	.68	87.5	89.8	B6
30	55	73	-.23	.38	.95	-.11	.70	-.64	Q .71	.68	87.5	87.4	B30

Figure 2. Misfit Order

In the Test of the suitability of the items of the Applied Mathematics exam test in Figure 1, it shows that items B4 and B5 both have an outfit value of MNSQ = 9.90, an outfit value of ZSTD = 9.91, while the Pt Mean Corr value = -0.33 and -0.50. The outfit value of MNSQ 9.90 is far outside the range of 0.5 < MNSQ < 1.5, indicating that items B4 and B5 are very less in accordance with the Rasch model [15,21,30,31]. The outfit value of ZSTD = 9.91 is also far from the reasonable range of -0.2 < ZSTD < 0.2, indicating a

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-05, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

very insignificant deviation. Furthermore, Pt Mean Corr = -0.33 and -0.50 are categorized as weak and negative, indicating that the relationship between items B4 and B5 with the total test score is very weak and inappropriate. The low Pt Mean Corr value proves that these two items do not make a significant contribution to the measured construct. Items B4 and B5 are inconsistent with the measured construct, do not match the Rasch model (misfit), both in terms of item suitability and their contribution to measurement [15,21,30,31]. These two items really need major revision, even considered for replacement.

The same condition also occurs for items B7 and B9, but item B9 has an outfit value of MNSQ = 2.86 which is quite far from the normal range accepted, namely 0.5 <MNSQ <1.5 [15,21,30,31], although slightly closer than B4 and B5. This condition indicates that item B9 is slightly inconsistent with the rash model, but not at an extreme level. While the outfit value of ZSTD = 1.47 is in the normal range [15,21], identifying that there is a match in item B9 to the rasch model. Furthermore, Pt. Mean Corr = 0.38, the positive correlation is categorized as quite strong. Item B9 has a positive and significant relationship with the total score, and contributes sufficiently to the measurement. Item B9 has a positive correlation with the overall ability of participants [35–37] although it is still not fully within acceptable limits. Item B9 is quite effective in differentiating participants based on their abilities. Item B9 can be said to have conformity with the Rasch model, has a positive contribution in differentiating participant abilities, but is still within acceptable limits. Item B9 needs minor revision, especially to achieve more accurate and efficient measurement results. Although the Outfit MNSQ is slightly higher than ideal, the ZSTD value is in the normal range, and Pt. Mean Corr shows a fairly strong and positive correlation with the overall ability of participants, item B9 can be said to be in accordance with the expectations of the Rash model. The values of outfit MNSQ, outfir ZSTD, and PT. Measure Corr. for the other items are within the normal range. So there are 37 out of 40 items in accordance with the Rasch model and there are 3 items that are less in accordance with the Rash model, namely items B4, B5, and B7.

Unidimensionality testing aims to ensure that the instrument only measures one dimension that is intended to be measured. A summary of the results of the unidimensional test is presented in Table 3.

Table 3. Summary of the Results of the Unidimensional Test of the Applied Mathematics Examination Test in the D3 RACE Study Program, PNB

Table 3. Summary of Unidimensional Test Results

Raw Variance	Eigenvalue	Observed	Expected	Status
raw variance explained by measure	47.4261	54.2%	51.1%	Accepted in good category
unexplained variance in 1st contrast	3.4475	3.9%	8.6%	Accepted in good category
unexplained variance in 2nd contrast	2.8251	3.2%	7.1%	Accepted in good category
unexplained variance in 3rd contrast	2.4588	2.8%	6.1%	Accepted in good category
unexplained variance in 4th contrast	2.3245	2.7%	5.8%	Accepted in good category
unexplained variance in 5th contrast	2.2301	2.6%	5.6%	Accepted in good category

Unidimensional testing of the results of the AME found a Raw Variance Explained by measures value of 54.2% more than 50%. In the unidimensionality test, the Raw Variance Explained by measures value above 40-50% is categorized as good [38]. This figure shows that 54.2% of the total variance can be explained by the main factor, namely the applied mathematics competency dimension. This value indicates that the applied mathematics exam test that was tested has a good unidimensional structure. This means that most of the variance is explained by one dimension of the ability being measured. The variance not explained by the main dimension (residuals) shows how much other factors may exist besides the main ability dimension being measured. The unexplained variance value for each component is less than 15%, this indicates that there are no significant secondary dimensions or suspicious patterns in the data [38]. There is no strong evidence of additional dimensions to worry about. The items in the final exam in mathematics were proven to measure one construct described in the learning outcomes of applied mathematics, namely knowledge of: limits, differentials, and integrals.

Bias element testing (unidimensionality) also found that no questions had a probability value of less than 5%. All of these questions showed no indication of bias or item function differences (DIF) [38]. All are consistent with the assumption of unidimensionality, the test is not biased towards certain groups in the same measurement. So the AME test that was tested was not only valid in content but also fair to all participants.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-05, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

The analysis of item difficulty level refers to the results of the item measure analysis showing an average item value of 0.00 and SD = 1.31. The average item difficulty level is 0.00 logit, in the context of the Rasch model, this value indicates that overall this AME test has a balanced level of difficulty, most items are neither too difficult nor too easy, and are at the midpoint of the logit scale [15,38]. This indicates that the exam test has a distribution of questions that cover various levels of difficulty well. Furthermore, SD is 1.31, then the level of difficulty of the AME test items can be classified into 4 categories, namely: 1) measure logit more than 1.31 is categorized as very difficult, 2) measure logit in the range between 0.00 and less than or equal to 1.31 is categorized as difficult, 3) measure logit in the range between -1.31 and less than or equal to 0.00 is categorized as easy, and 4) measure logit less than or equal to -1.31 is categorized as very easy. So that the distribution of the level of difficulty of the AME test items, namely: 3 items or 7.5% are in the very difficult category, 12 or 30% of items are in the difficult category, 23 or 57.5% of items are categorized as easy, and 2 or 5% of items are in the very easy category. Overall, the distribution of the level of difficulty of the AME test items tends to be easy, but still contains some very difficult, difficult and very easy items that can help in measuring higher or lower abilities.

In order to improve the test's ability to differentiate participants with higher abilities, test developers need to consider increasing the proportion of items in the difficult category. In addition, it is also necessary to add some items with a very easy level of difficulty to ensure that the range of difficulty levels is more complete and accommodates participants with lower abilities. Increasing the proportion of balanced items needs to be done. Items should represent a spectrum of difficulty that reflects variations in the abilities of test participants so that the results can be more valid, reliable, and informative. The ideal percentage of difficulty categories is: 15-25% easy, 40-50% moderate, 20-30% difficult, and 5-10% very difficult [16–18]. This proportion is to ensure that each participant has the opportunity to work on questions at a level of difficulty that suits their abilities. In turn, the test will be expected to be able to measure the range of test participants' abilities fairly and balanced.

The results of Person Measure show the mean and SD values of 1.35 and 1.98 respectively. The average respondent's ability is at 2.19 logit categorized as moderate, higher than the average question difficulty (0.00) logit. The average respondent's ability is above the average question difficulty, indicating that overall, students have quite good abilities, above the average question difficulty [15,38]. The SD of the respondent's ability is 1.98, so that the test participants' abilities can be classified into three groups: high, moderate, and low. namely: 1) measure logit more than 3.33 is categorized as high, 2) measure logit in the range between -0.63 and less than or equal to 3.33 is categorized as moderate, 3) and measure logit in the range between -0.89 and less than or equal to -0.63 is categorized as low. There are 10.9% of test takers categorized as high ability, 72.6% as medium ability, 15.1% as low ability, and 1.4% as outliers. Information about this distribution can be a positive input for lecturers in charge of the course in an effort to improve the effectiveness of learning, the need for more varied and innovative learning strategies, so that the gap in ability differences between students can be minimized [15,39,40]. Overall, the majority of test takers have medium to high abilities, but there are also a small number that need more attention to improve their abilities.

Wright's map shows several important findings regarding the match between item difficulty and participant ability. Items with high logits (greater than +4 logits), such as B4 and B5, are too difficult for most participants, and only a few participants are able to answer them. In contrast, items with low logits (less than -2 logits), such as B39 and B16, tend to be too easy and therefore do not provide enough diagnostic information to distinguish between low and medium ability participants. The distribution of participants appears fairly even across the 0 to +3 logit range, but there are significant gaps in the distribution of participants and items across certain ranges.

In the logit range +3 to +4, there are no participants with high ability, while there are several items, such as B4 and B5, which are in that range. This indicates that above-average participants are not well measured by this test. Conversely, in the logit range -1 to - 2, there are very few participants with low ability, so items that are too easy are less relevant to the population being tested. In addition, the distribution of item difficulty levels shows a gap in the logit +2 to +3, where no items are available, even though many participants are in this ability range. This indicates that participants at this level are not optimally measured. The absence of items in this range is a critical problem because the test is unable to accurately distinguish participants with moderate to high ability. It is necessary to add items in the logit +2 to +3 to cover the abilities of participants at the moderate to high level. In addition, evaluation is needed for items in the logit +3 to +4 to ensure their relevance and effectiveness, because only a few participants can answer items in this range. With these improvements, it is hoped that the distribution of items will be more even and the test can provide a more accurate measurement of the entire range of participant abilities.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-05, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

This AME test needs to be improved by adding more items that are around the +1 to +3 logit to accommodate medium to high ability participants. Also, more items above the +3 logit can be added to challenge more talented participants. With a more even distribution of items, the test will be more effective in measuring the abilities of all participants, without anyone feeling too difficult or too easy. This AME test is generally suitable for D3 RACE Study Program PNB students

The results of the S.E examination on the fit statistics show that there are 39 items with S.E values below 0.5, meaning that the 39 items have good discrimination power. A low S.E value below 0.5 indicates that the item parameter estimation is quite accurate, and these items are able to differentiate the abilities of test participants well [15,41]. There are two items, namely items B3 and B40, which have S.E values in the range between 0.5 and 1, meaning that these two items are still considered quite good, although their discrimination power is not as good as items with S.E below 0.5. The S.E value in this range indicates that the item parameter estimation is less accurate, but still acceptable, B3 and B40 can still differentiate the abilities of test participants, but with a lower level of accuracy [15]. The test has a relatively good level of accuracy. Overall, the final exam in applied mathematics has good discriminatory power, although there are several items, namely B3 and B40, that need further attention.

Based on the results of the analysis above, this applied mathematics exam test is proven to be able to measure one construct that is in accordance with the Applied Mathematics Learning Achievement. There is one question item that shows a significant DIF problem, while the other items do not show any significant DIF indications. The data obtained have met the expectations of the Rasch model, although some questions need to be reviewed. This test has a fairly good level of validity, reliability for both person and item is in the very good category, a good separation index, a balanced level of item difficulty (0.0 logit), and a good level of discrimination power. This proves that the test can differentiate participants based on their ability levels, and the questions have been formulated well. Overall, the AME test is proven to be valid and reliable as a test to measure the achievement of applied mathematics learning in the D3 RACE Study Program, PNB. Practically, these results show that the test has fairly good and very good validity and reliability, although it is necessary to review several items with extreme MNSQ outfit values to ensure there is no bias. The resulting data provides a strong basis for interpreting the results of ability measurements. Participants with high MNSQ outfit scores need to be analyzed further. Theoretically, these findings strengthen the theory and principles of the Rasch model, that measurement and evaluation in education can be carried out more efficiently and effectively with an adaptive approach [42]. So these findings have a strong theoretical basis in the context of educational measurement, and can be used as an empirical example of the application of Rasch theory to produce more valid, reliable, and informative measurements.

CONCLUSION

The AME test has an adequate level of validity, most of the questions meet the fit criteria for the Rasch Model. The level of test reliability is categorized as very good with a person and item reliability value of 0.90 categorized as good. Several questions show misfits that require improvement. Item difficulty and person ability show a proportional distribution between the level of question difficulty and student ability. Overall, the test's discrimination power is categorized as good, although there is one question that needs to be reviewed further for improvement. The implication is that the application of the Rasch Model in validating exam tests can help lecturers construct tests that are more valid, reliable, and in accordance with the level of student ability. The results of this study have a strong theoretical basis in the context of educational measurement, and can be used as an empirical example of the application of the Rasch model theory to produce more valid, reliable, and informative measurements. It can be recommended to test developers, for the development of exam tests in the future, it is very necessary to pay attention to the balance between the level of question difficulty and student ability, and to ensure that measurements are more valid and reliable in the context of higher education, especially polytechnics.

ACKNOWLEDGEMENT

The author would like to thank all parties who have contributed to the completion of this research. In particular, deep appreciation is conveyed to the D3 RACE PNB Study Program and all teaching staff, for the support, facilities, and input provided during the research process. The author would also like to thank the respondents and parties who have taken the time to participate in this research. Finally, the author expresses gratitude for all the constructive input from the IJCSRR editor and reviewer team, so that this article is worthy of publication.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-05, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

REFERENCES

1. Bernard, R. M., Borokhovski, E., Schmid, R. F., Tamim, R. M. & Abrami, P. C. A meta-analysis of blended learning and technology use in higher education: From the general to the applied. *J. Comput. High. Educ.* 26, 87–122 (2014).
2. Li, T., Yeung, M., Li, E. & Leung, B. How formative are assessments for learning activities towards summative assessment? *Int. J. Teach. Educ.* 9, 42–57 (2021).
3. Rawlusyk, P. . Assessment in Higher Education and Student Learning. *J. Instr. Pedagog.* 21, 1–34 (2018).
4. Sekyi, E. T. A. Assessment, Student Learning and Classroom Practice: A Review. *J. Educ. Pract.* 7, 1–6 (2016).
5. Brookhart, S. M. & Nitko, A. J. *Educational Assessment of Students*. (Pearson Education, 2011).
6. Crocker, L. M. & Algina, J. *Introduction to Classical and Modern Test Theory*. (Wadsworth Publishing, 1986).
7. Nunnally, J. C. & Bernstein, I. H. *Psychometric Theory*. (McGraw-Hill Book Company, 1994).
8. Allen, M. J. & Yen, W. M. *Introduction to Measurement Theory*. (Waveland Press, 2001).
9. Wright, B. D. & Stone, M. H. *Best Test Design*. *University of Chicago* (Mesa Press, 1979). doi:10.2307/jj.3685368.13.
10. Bond, T. G. & Fox, C. M. *Applying the Rasch Model : Fundamental Measurement in the Human Sciences Second Edition*. (Lawrence Erlbaum Associates, Inc, 2007).
11. Baghaei, P. & Amrahi, N. Validation of a Multiple Choice English Vocabulary Test with the Rasch Model. *J. Lang. Teach. Res.* 2, (2011).
12. Anastasi, A. & Urbina, S. *Psychological Testing*. (PT. Indeks, Gramedia Grup, 2007).
13. Gronlud, N. E. & Linn, R. L. *Meansurment and Evaluation in Teaching*. (Macmillan Publishing Co., Inc, 1990).
14. Mardapi, D. *Pengukuran, penilaian dan evaluasi pendidikan*. (: Nuha Medika, 2016).
15. Bond, T. G. & Fox, C. M. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences,*. (Routledge, 2015).
16. Thompson, B. *Exploratory and Confirmatory Factor Analysis*. *American Psychological Association* (American Psychological Association, 2004).
17. Anderson, L. W. & Krathwohl, D. R. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives: A Revision of Bloom's Taxonomy of Educational Objectives*. (Pearson, 2001).
18. Boone, W. J., Staver, J. R. & Yale, M. S. *Rasch Analysis in the Human SciencesNo Title*. (Springer., 2014).
19. Boone, W. J. Rasch analysis for instrument development: Why,when,and how? *CBE Life Sci. Educ.* 15, 1–7 (2016).
20. Linacre, J. M. Data Variance Explained by Rasch Measures. *Rasch Meas. Trans.* 20, 1045 (2006).
21. Linacre, J. M. A User's Guide to Winsteps, Ministep Rasch-Model Computer Programs. *Winsteps* https://www.winsteps.com/winman/copyright.htm (2019).
22. Kubinger, K. D., Rasch, D. & Yanagida, T. A new approach for testing the Rasch model. *Educ. Res. Eval. An Int. J. Theory Pract.* 17, 321–333 (2011).
23. Zanon, C., Hutz, C. S., Yoo, H. & Hambleton, R. K. An application of item response theory to psychological test development. *Psicol. Reflex. e Crit.* 29, (2016).
24. Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P. & Ratu, T. Tes Klasik Dan Model Rasch. *Indones. J. Educ. Res. Rev.* 3, 11–19 (2020).
25. Susdelina, Perdana, S. A. & Febrian. Analisis Kualitas Instrumen Pengukuran Pemahaman Konsep Persamaan Kuadrat Melalui Teori Tes Klasik Dan Rasch Model. *J. Kiprah* 6, 41–48 (2018).
26. Bichi, A. A., Talib, R., Atan, N. A., Ibrahim, H. & Yusof, S. M. Validation of a developed university placement test using classical test theory and Rasch measurement approach. *Int. J. Adv. Appl. Sci.* 6, 22–29 (2019).
27. Boone, W. J. & Noltemeyer, A. Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Educ.* 4, (2017).
28. Mahmud, Z. & Porter, A. Using rasch analysis to explore what students learn about probability concepts. *J. Math. Educ.* 6, 1–10 (2015).
29. Runnels, J. & Bunkyo, H. Using the Rasch model to validate a multiple choice English achievement test. *Int. J. Lang. Stud.* 6, 141–153 (2015).
30. Sumintono, B. & Widhiarso, W. *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. (Trim Komunikata, 2015).

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 08 Issue 01 January 2025
DOI: 10.47191/ijcsrr/V8-i1-05, Impact Factor: 7.943
IJCSRR @ 2025
www.ijcsrr.org

31. Linacre, J. M. What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Meas. Trans.* 6, 878 (2012).
32. Linacre, J. M. Optimizing rating scale category effectiveness Optimizing Rating Scale Category Effectiveness University of Chicago. *J. Appl. Meas.* 3, 85–106 (2002).
33. Gliem, J. A. & Gliem, R. R. Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales. in *Midwest Research to Practice Conference in Adult, Continuing, and Community Education* 83–88 (The Ohio State University, Columbus, OH, 2003).
34. Planinic, M., Boone, W. J., Susac, A. & Ivanjek, L. Rasch analysis in physics education research: Why measurement matters. *Phys. Rev. Phys. Educ. Res.* 15, 20111 (2019).
35. Taylor, R. Interpretation of the Correlation Coefficient: A Basic Review. *J. Dianostic Med. Sonogr.* 6, 35–39 (1990).
36. Schober, P. & Schwarte, L. A. Correlation coefficients: Appropriate use and interpretation. *Anesth. Analg.* 126, 1763–1768 (2018).
37. Sugiyono. *Metode Penelitian Kuantitatif, Kualitatif dan R&D*. (Alfabeta, 2017).
38. Linacre, J. M. Predicting responses from rasch measures. *J. Appl. Meas.* 11, 1–10 (2010).
39. Wright, B. D. & Masters, G. N. *RATING SCALE ANALYSIS-n Item Response Modeling Approach*. (1982).
40. Coladarci, T. & Cobb, C. D. *Fundamentals of Statistical Reasoning in Education*. (John Wiley & Sons, Inc., 2013).
41. Wright, B. D. & Stone, M. H. *Measurement Essentials*. *Measurement* (WIDE RANGE, INC, 1999).
42. Glas, C. A. W. & Vos, H. J. *Adaptive Mastery TestingUsing the Rasch Model and Bayesian Sequential Decision Theory*. *ERIC* https://eric.ed.gov/?id=ED417056 (1998).

Cite this Article: I Ketut Darma, I Gde Nyonan Sangka (2025). Validation of Applied Mathematics Exam Test Using Rasch Model Approach: Case Study in Diploma 3 Study Program of Refrigeration and Air Conditioning Engineering Politeknik Negeri Bali. International Journal of Current Science Research and Review, 8(1), 33-44, DOI: https://doi.org/10.47191/ijcsrr/V8-i1-05