PIONIR Jurnal Pendidikan
P-ISSN 2339-2495
E-ISSN 2549-6611

PIONIR: JURNAL PENDIDIKAN
VOLUME 13 No 3 2024
P-ISSN 2339-2495/E-ISSN 2549-661

COMPARING CLASSICAL TEST THEORY AND ITEM RESPONSE THEORY FOR RESPIRATORY SYSTEM QUESTION INSTRUMENTS

Adha Nisfatulsanah*

*Universitas Sebelas Maret, Surakarta, Indonesia

adhanisfatulsanah01@student.uns.ac.id

Bowo Sugiharto

Universitas Sebelas Maret, Surakarta, Indonesia

bowo@fkip.uns.ac.id

Received 05 August 2024, Accepted 14 December 2024, Published 23 December 2024

Abstract

This study aims to compare the results of instrument testing methods between applying classical test theory and item response theory using the Rasch model in question instruments on respiratory system material. This study employed descriptive quantitative methodology, with a sample involving 36 students. The analyzed instrument consisted of 40 multiple-choice questions on respiratory system material. Instrument analysis utilized classical test theory with Microsoft Excel and item response theory with Winstep Rasch ver 4.5.2.0. The data analysis from classical test theory and item response theory offers slightly different interpretations but is mutually complementary. Both classical test theory and item response theory may assess the validity, reliability, distractor effectiveness, difficulty level, and discriminating power of questions. Item response theory provides a comprehensive analysis of test results through the use of the Wright map as a bar which helps determine a student's ability about the difficulty level of the question. Scalogram is used to identify patterns in students' responses, allowing for the detection of cheating and inaccuracies in answering questions. Additionally, DIF items are employed to identify item bias. This study concludes that any developed instrument must possess the characteristics that meet the requirements to measure competency effectively. The requirements for an instrument can be analyzed using item response theory with the Rasch model, which provides in-depth interpretation.

Keywords: Classical test theory, item response theory, Rasch model, instrument

Abstrak

Tujuan dari penelitian ini adalah mengeksplorasi metode pengujian instrument berupa teori tes klasik dan teori respon butir (Rasch) dalam menganalisis instrument soal materi sistem respirasi. Metode dalam penelitian yaitu deskriptif kuantitatif dengan menggunakan sampel sebanyak 36 siswa. Instrument yang dianalisis berupa 40 soal pilihan ganda materi sistem respirasi. Analisis instrument menggunakan teori tes klasik dengan Microsoft Exel dan teori respon butir dengan Winstep Rasch ver 4.5.2.0. Kesimpulan penelitian ini yaitu suatu instrument yang dikembangkan harus memiliki karakteristik memenuhi persyaratan dalam mengukur suatu kompetensi. Data hasil analisis baik teori tes klasik maupun teori respon butir memiliki interpretasi yang sedikit berbeda namun saling melengkapi. Teori tes klasik dan teori respon butir dapat menganalisis validitas, reliabilitas, efektifitas distraktor, tingkat kesukaran, dan daya pembeda soal. Teori respon butir memberikan hasil interpretasi yang lebih mendalam dengan adanya wright map sebagai mistar dalam menentukan kemampuan siswa terhadap tingkat kesulitan butir soal, scalogram untuk melihat pola jawaban siswa sehingga dapat mengetahui kecurangan serta kurang telitinya siswa dalam menjawab soal, dan item DIF dalam mendeteksi bias soal.

Kata kunci: Teori tes klasik, teori respon butir, rasch, instrumen

INTRODUCTION

Educators implement assessments to evaluate the level to which students accomplish desired learning outcomes. (Gronlund & Waugh, 2013). Assessments may include written tests and performance assessments. The main method of assessment used in schools is written tests. Developing high-quality test instruments is crucial for evaluating students' learning outcomes. Test instruments require preliminary testing to get accurate calibration. A high-quality assessment instrument must fulfill several criteria, including good validity in measuring what it is supposed to measure*, the reliability of a good question determined by its consistency when used repeatedly, the level* of difficulty of the items that may vary, the discriminating power of the question to effectively distinguish between students with have high and low abilities, and the distractors that can outwit student responses (Erfan et al., 2020).

There are two approaches to evaluating the quality of instruments, namely classical test theory (CTT) and item response theory (IRT) (Bichi et al., 2015). Classical test theory employs two components of the assessment score: the true score and measurement error. The true score represents the score obtained if there is no error in measurement, whereas the

measurement error indicates the difference between the true score and the observed score (Sumaryanta, 2021). The parameters included in classical test theory encompass reliability, item difficulty, distinguishing power, and distractor effectiveness. The components derived from classical test theory cannot assess individual students' item responses and performance on specific items.

Measurement does not involve initiating a competition among students by categorizing them based on their performance, with some regarded as superior and others inferior. The purpose of measurement in an assessment is to convey the total range of student knowledge, the distribution of the range of expertise, the challenges each student faces, and the areas where students excel in their learning (Boonee· et al., 2014). Rasch analysis enables the assessment of students' abilities by analyzing question items, allowing for a more comprehensive evaluation regarding the quality of question items. The level of student success in answering questions depends on the level of ability and difficulty of the questions, not just on the final score (Septiliana, 2023).

This study aims to explore the methodologies of instrument testing, namely classical test theory and item response theory (Rasch), in analyzing question instruments related to the respiratory system. This research is important to know the difference between the two methods of instrument analysis may be beneficial as a reference for instrument designers to assess the instrument's quality and implement enhancements.

METHOD

Research Methodology

The method of this study is descriptive quantitative. A descriptive quantitative methodology is undertaken by collecting and analyzing empirical data, including statistical data, and later describing or interpreting the results obtained (Mohajan, 2020).

Population and Sample

The population in this study was high school students. The sampling technique used is purposive sampling, which involves selecting participants based on specific criteria, specifically classes with more effective biology class hours than others. Purposive sampling is a technique of selecting samples based on the characteristics of a subject and the willingness to participate in the research (Thomas, 2022). The research sample consisted of 36 students from class XI MIPA 5 who were tested on their knowledge of the respiratory system material.

Data Collection and Analysis

The respiratory system material test instrument consists of 40 multiple-choice questions, with five answer choices. Data were collected using a test technique. The collected data included students' final scores and the selected answers for each question item. Instrument testing and analysis were conducted using classical test theory with Microsoft Excel and item response theory with Winstep Rasch ver 4.5.2.0.

The analyses of each parameter are validity testing in classical test theory is conducted using Pearson's product-moment correlation formula for raw scores (Arikunto, 2012) Meanwhile, item response theory employs item statistics measure order, considering criteria such as outfit MNSQ, outfit ZSTD, and Pt Measure Corr (Boonee· et al., 2014) in the output table 13 item measure menu. Reliability testing in classical test theory employs Cronbach's Alpha reliability formula (r11). In item response theory, reliability is assessed using data on real person reliability, real item reliability, and Cronbach's Alpha (KR-20) from the summary of 36 measured persons and 40 measured items, as shown in the output table 3.1 summary statistics. The distractor effectiveness in classical test theory is determined by calculating the average percentage of students choosing each answer option, divided by the total number of students. The item response theory utilizes data on item category/option/distractor frequencies from the output table 13-item measure. The difficulty level in classical test theory is measured by dividing the number of students who answered an item correctly by the total number of students. In item response theory, the difficulty level is assessed by analyzing data from the Wright map analysis in the output table 1 variable map and the scalogram in the output table 22 scalograms. These data illustrate the difficulty level of the items based on student ability. The discriminating power of items in classical test theory is evaluated using the discriminating power (DP) formula, which measures the performance of the top 27% and bottom 27% of students (Azmi & Salam, 2020). Item response theory uses data on real separation from the summary of 36 measured persons and the summary of 40 measured items in the output table 3.1 summary statistics. The bias index (DIF) is exclusively used in item response theory, assessed by examining probabilities in the DIF class/group specification in the output table 30 item DIF between/within.

RESULTS AND DISCUSSION

1. Validity Test

The validity test in classical test theory is measured by Pearson correlation with a significance level of 0.05. The question is considered to be valid if r count>r table (Arikunto, 2012). The calculation results show 23 valid questions and 17 invalid questions.

Validity in the Rasch model requires the following criteria (Boonee· et al., 2014):

a. The accepted Outfit MNSQ (Mean Square) values are: 0.5 < Outfit – MNSQ < 1.5

b. The accepted Outfit ZSTD (Z – Standard) values are: -2.0 < ZSTD < +2.0

c. The Pt Measure Corr (Point Measure Correlation) value: 0.4 < Point Measure Corr < 0.85

Items are valid if they fulfill a minimum of two criteria, revised if they fulfill only one criterion, and discarded if they do not meet all criteria. Table 1 shows the validity results for each item using both classical and Rasch test theory.

Table 1: Validity Test

Validity Result	Classical Test Theory		Item Response Theory (Rasch)	
	Number of Questions	Question Numbers	Number of Questions	Question Numbers
Valid	23 questions	1, 2, 3, 4, 7, 8, 9, 10, 11, 13, 15, 18, 21, 22, 23, 28, 29, 33, 35, 36, 38, 39, 40	38 questions	1-12, 14, 15, 17-40
Invalid	17 questions	5, 6, 12, 14, 16, 17, 19, 20, 24, 25, 26, 27, 30, 31, 32, 34, 37	2 questions	13 (revised) and 16 (discarded)

Research instruments require validity, meaning they are capable of accurately measuring the intended variables (Azmi & Salam, 2020). Table 1 displays the application of Rasch testing which contains more valid items compared to classical test theory. The Rasch model is more accurate due to the requirement that a question item must meet at least two criteria, namely the MNSQ Outfit value, the ZSTD Outfit value, and the Point Measure Correlation value, then the question is considered valid (Jumini et al., 2023). The Rasch analysis results reveal that 38 questions (95%) meet the criteria, while 2 questions (5%) do not match the criteria. Validity testing employs an item fit test to assess the conformity of the items and to determine whether the items function normally or not. Results that meet the item fit test criteria can be concluded as valid (Darmana et al., 2021).

2. Reliability Test

The reliability test in classical test theory consists only of Cronbach Alpha with a result of 0.832 (very high). Reliability in item response theory includes person reliability, item reliability, and Kuder-Richardson Formula (KR-20). There are two types of reliability, namely real reliability and model reliability, specifically in the field of education using real reliability (Boonee· et al., 2014). KR20 and Cronbach Alpha are used in classical test theory. Meanwhile, person reliability is used in modern theory (Anselmi et al., 2019)

Table 2: Reliability Test

Reliability Results	Classical Test Theory		Reliability Results	Item Response Theory (Rasch)	
	Value	Description		Value	Description
Cronbach Alpha	0.832	Very high	Person reliability	0.81	Good
			Item reliability	0.76	Adequate
			Cronbach Alpha (KR-20)	0.83	Very good

Research instruments must possess reliability, meaning they can be used multiple times and produce consistent data (Sugiyono, 2016). Reliability, also referred to as precision, refers to the consistency of a test in measuring a certain competency and the absence of measurement errors that can cause student scores to deviate (Popham, 2017). Table 2 shows the results of KR20 and Cronbach Alpha derived from student test scores in calculating the observed variance. Errors may arise because student test scores are not a linear representation of the variable, whereas the calculation of variance requires linearity. The Rasch model uses student respondent measurements on a linear scale, making it suitable for calculating observed variance. Person reliability is expected to be a more consistent index than KR20 and Cronbach Alpha (Anselmi et al., 2019) Item reliability reflects the extent to which the item hierarchy can be used across different populations (Wongpakaran et al., 2020). The results of item reliability in the adequate category (0.76).

3. Distractor Effectiveness Test

Classical test theory provides information on the percentage of each distractor in the answer choices, but it does not indicate whether the distractor is effective or not. Table 3 displays the results of testing the distractor effectiveness in classical test theory.

Table 3: Distractor Effectiveness Test of Classical Test Theory

Question Number	Percentage of Classical Test Theory Distractors on Answer Choices				
	A	B	C	D	E
1	6%	**83%**	6%	0%	6%
2	**44%**	42%	3%	3%	8%
3	14%	25%	3%	**58%**	0%
4	22%	11%	6%	0%	**61%**
36	8%	8%	3%	**78%**	3%
37	0%	33%	0%	**64%**	3%
38	3%	19%	3%	**69%**	6%
39	25%	3%	3%	0%	**69%**
40	**47%**	19%	8%	25%	0%

Distractor effectiveness in the Rasch model shows the average abilities in each answer choice. The higher the utility towards the correct option, the more effective the distractor of a question, as indicated in Table 5. Distractors do not function properly if the average ability towards the correct option decreases, as found in Table 6.

Table 4: Distractor Effectiveness Test in Item Response Theory (Rasch)

Item Response Theory (Rasch)		
Distractor Effectiveness	Number of Questions	Question Numbers
Effective	32 questions	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 15, 17, 18, 20, 21, 22, 23, 25, 26, 27, 28, 29, 32, 33, 34, 35, 36, 37, 38, 40
Ineffective	8 questions	12, 14, 16, 19, 24, 30, 31, 39

Table 5: Sample of Effective Distractor in Item Response Theory (Rasch)

Question Number	Answer Choices	Answer Score	Number of Data	Data Percentage	Average Ability
18	C	0	2	6	.08
	D	0	6	17	.37
	E	0	8	22	.96
	A	0	7	19	1.28
	B	1	13	36	2.30

Table 6: Sample of Ineffective Distractor in Item Response Theory (Rasch)

Question Number	Answer Choices	Answer Score	Number of Data	Data Percentage	Average Ability
14	A	0	3	8	0.54
	C	0	4	11	0.76
	E	0	3	8	1.07
	B	0	1	3	2.87
	D	1	25	69	**1.53***

Multiple choice questions consist of five options, one of which is the correct answer, while the remaining four options are distractors. Distractors function as a checker for students when answering questions. A good distractor implies that there are students who choose the distractor option. Rasch demonstrates the analysis of each answer choice by using the average ability. The Rasch analysis results with 32 questions having good distractors, while 8 questions had less effective distractors and did not serve their intended purpose. Tables 5 and 6 display that the higher the ability directed towards the correct option, the better the distractor of a question. Conversely, the distractor does not function properly if the average abilities decrease (Boonee· et al., 2014).

4. Level of Difficulty Test

The level of difficulty test in classical test theory is based on the question difficulty index. The difficulty index is calculated by dividing the number of students who answer correctly on the item by the total number of students. The calculation results are grouped into five categories of difficulty level. The difficulty level in Rasch analysis is determined by grouping items depending on their logit value and standard deviation. The categories of question difficulty are listed in Table 7.

Table 7: Level of Difficulty Test

Level of Difficulty	Classical Test Theory		Item Response Theory (Rasch)	
	Number of Questions	Question Numbers	Number of Questions	Question Numbers
Very easy	0 question	-	0 question	-
Easy	21 questions	1, 6, 8, 10, 12, 13, 15, 16, 20, 21, 24, 25, 26, 27, 29, 30, 31, 33, 34,	6 questions	13, 15, 20, 21, 24, 25

PIONIR: JURNAL PENDIDIKAN VOLUME 13 No 3 2024
P-ISSN 2339-2495/E-ISSN 2549-6611
Journal page is available to: https://jurnal.ar-raniry.ac.id/index.php/Pionir/index

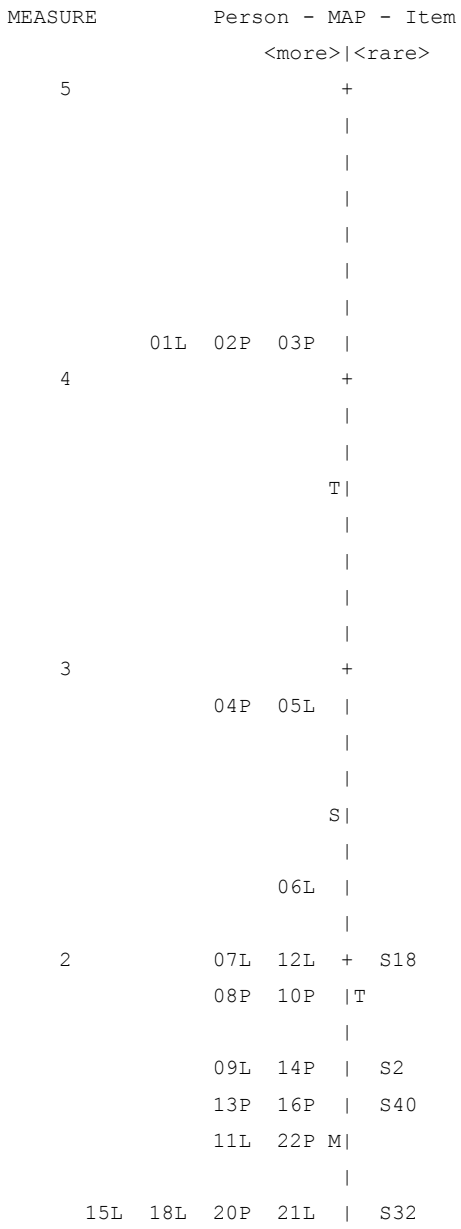
		35, 36		
Moderate	19 questions	2, 3, 4, 5, 7, 9, 11, 14, 17, 18, 19, 22, 23, 28, 32, 37, 38, 39, 40	14 questions	1, 6, 8, 12, 16, 26, 27, 29, 30, 31, 33, 34, 35, 36
Hard	0 question	-	14 questions	3, 4, 5, 9, 10, 14, 17, 19, 22, 23, 28, 37, 38, 39
Very hard	0 question	-	6 questions	2, 7, 11, 18, 32, 40

The level of item difficulty indicates the probability of respondents who can answer a question correctly (Arikunto, 2012). Classical test theory categorizes each question into difficulty criteria ranging from very easy, easy, moderate, hard, and very hard. The data shown in Table 7 regarding the level of item difficulty in classical test theory does not provide information on which items are difficult or easy for certain students, even though each student has different abilities in answering each item. Classical test theory only classifies questions ranging from very easy to very hard. However, it does not provide information on which individuals find the questions difficult or easy. The Rasch model provides a solution to this limitation with the results of the Wright map analysis found in Chart 1.

5. Wright Map

The Wright Map of the Rasch model displays the distribution of student abilities on the left chart and the level of item difficulty on the right chart based on logit values (Boonee· et al., 2014). The result of the Wright map can be seen in Chart 1. The logit mapping indicates that S18 is the most challenging question, however its difficulty may vary among students. Question S18 can be answered correctly and is considered easy for students with the bar positioned above S18, namely 01L, 02P, 03P, 04P, 05L, and 06L. Students whose crossbar position is parallel to S18, namely 07L and 12L, have the potential to answer the question either correctly or incorrectly. The position of students below the S18 bar indicates that question S18 is difficult, as their abilities fall below the logit value of S18. Students with high ability, namely 01L, 02P, 03P, 04P, 05L, and 06L are positioned above the logit of the most difficult problem (S180). This suggests that these six students can easily solve all the 40 questions. Students with the lowest ability, 36L, were able to answer correctly just the easy questions found on the bar below, namely only questions S21, S15, S24, S13, S20, and S25. On the other hand, all questions positioned on the logit bar above are considered difficult by the student.

The Rasch model also produces a Wright map which allows for a more in-depth analysis of item difficulty (Planinic et al., 2019). The Wright map shows whether the instrument tests students well below or above their knowledge level, allowing for a genuine comprehension of students' abilities (Boonee· et al., 2014). Wright maps utilize a logit bar that places students based on ability and item difficulty (Chitaree et al., 2024). The Wright map analysis is employed to gain a more comprehensive knowledge of the instrument's ability to accurately assess students' comprehension (Intasoi et al., 2020). Wright's map ensures that the assessment is valid, reliable, and aligned with educational objectives. This procedure can assist educators in accurately assessing and measuring instruments, making data-driven decisions, and improving the overall quality of education (Chitaree et al., 2024).



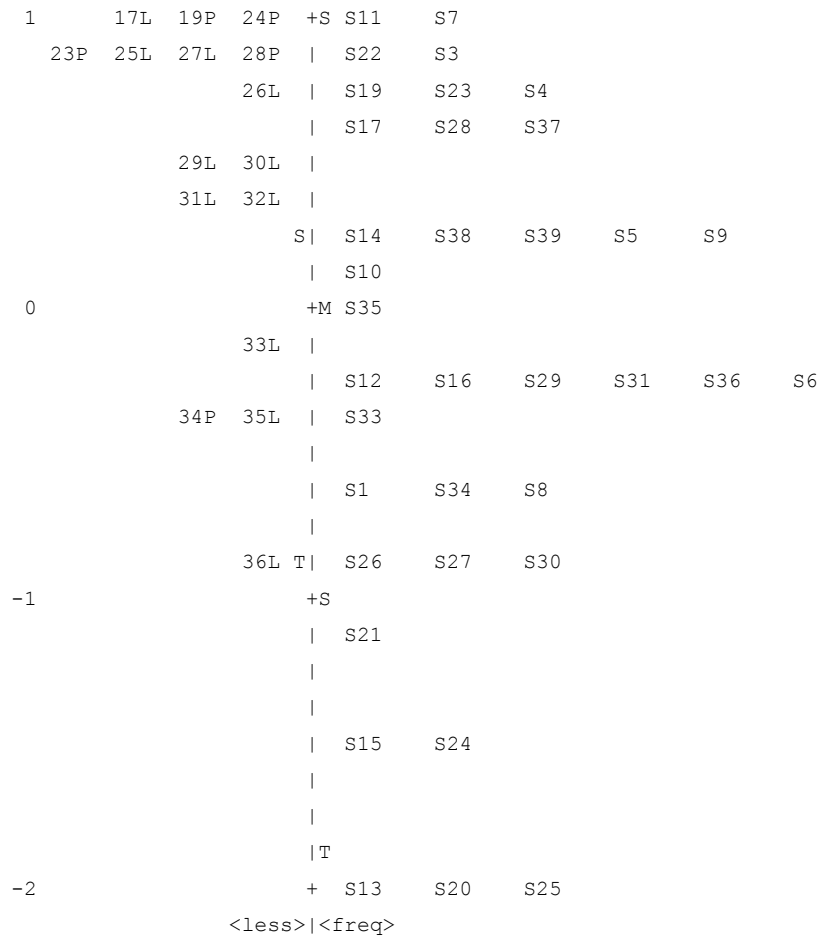


Chart 1. Wright Map

6. Scalogram

The scalogram of Rasch analysis results can be used to identify cases of academic cheating and inaccuracies in students' responses while completing assignments (Nurjanah et al., 2024). The numbers displayed at the top and bottom indicate the order of the items based on their logit values, from lowest to highest, arranged from left to right. The results of the scalogram can be found in Chart 2. Students 01L and 03P exhibit identical scalogram patterns, which is probable that they sit in close distance and collaborate on answering questions. Student 24P who possesses lower position abilities demonstrated capability to answer the most difficult questions S18 and S40. There are two possible explanations for this achievement: either the student answered both questions by making up but unexpectedly getting the correct results or the student engaged in cheating behavior. Students 32L and 35L answered incorrectly on the easiest question, namely S13. This suggests that these students may have lacked comprehension of the concepts in the respiratory system material, indicating a need for further clarification and understanding. Student 18L, who possesses high ability,

provided a wrong response to the easy question S20, perhaps due to a lack of attentiveness in answering the question.

```
GUTTMAN SCALOGRAM OF RESPONSES:
Person |Item
       |122122223  33 1123331  133123 12 2 134 1
       |3055416701843626916505948978749332712028
       |----------------------------------------
     1 +11111111111111101111111111111111111111111  01L
     2 +11111111111111111111111111111111111110111  02P
     3 +11111111111111101111111111111111111111111  03P
     4 +11111111111111111011110111111111111110111  04P
     5 +11111111111111111111101111011111111011111  05L
     6 +11111110111111101111111111111110111110110  06L
     7 +11111101111111011101111111110111101111110  07L
    12 +11111111011111111111111101111011110110101  12L
     8 +11111111111011110111111111011111001101110  08P
    10 +11110111111110111100111110111110101111111  10P
     9 +11111010111111111111101111111110111001001  09L
    14 +11111111100111101011111111111110010111011  14P
    13 +11111111111111011111110011111010110101001  13P
    16 +11111111111111111111010110101111000011110  16P
    11 +11111111111011011011111011001111010011010  11L
    22 +11111111111111011111011110111001011000110  22P
    15 +11111101111101111111011010101101111000010  15L
    18 +10111111101111101110110111111010010101101  18L
    20 +11111111111111011100110011011001011111001  20P
    21 +11011111111111111111111000110001011110000  21L
    17 +11111101111110111110110111100101000111100  17L
    19 +11111111111111101111010111001001100110000  19P
    24 +11110111111111010111011100011001111 00101  24P
    23 +11111111101111101111100110010011010001 10  23P
    25 +11111100111011101111001011101100111001 0  25L
    27 +11111111111011111110111110000110110000000  27L
    28 +11111110101101111101111001001101101011 00  28P
    26 +11111110111111011010100111001011110010 00  26L
    29 +11111111111111001111101111100000100000000  29L
    30 +11111101101100111001101101110010001110 00  30L
    31 +11111011110001010111001101111001001010 00  31L
    32 +01100111101111111100011111100111000000000  32L
    33 +10101010110101111111010000100110100000000  33L
    34 +11111111110010100000010000011000100100 00  34P
    35 +01111101011000010011100100001010011000 00  35L
    36 +11001011000111100100000010101000000000000  36L
       |----------------------------------------
       |122122223  33 1123331  133123 12 2 134 1
       |3055416701843626916505948978749332712028
```

Chart 2. Scalogram

PIONIR: JURNAL PENDIDIKAN VOLUME 13 No 3 2024
P-ISSN 2339-2495/E-ISSN 2549-6611
Journal page is available to: https://jurnal.ar-raniry.ac.id/index.php/Pionir/index

Rasch analysis also generates a scalogram by ranking students and item difficulty based on the Guttmann Scale (Robinson et al., 2019). The scalograms are a useful tool for identifying cases of student inconsistency, cheating on tests, and lack of accuracy in answering questions (Nurjanah et al., 2024). The Guttmann Scale can also assess the accuracy of students in answering questions and identify students who rely on guessing as a strategy for answering questions (Syafrial et al., 2022).

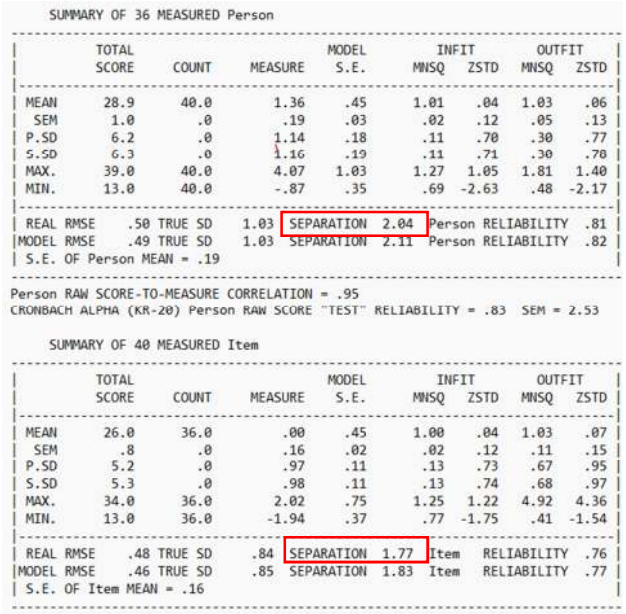
7. Distinguishing Power Test

The distinguishing power of questions in classical test theory is shown in Table 8. The Rasch analysis model has two distinguishing indices, namely person separation and item separation as shown in Figure 1.

Table 8: Distinguishing Power Test in Classical Test Theory

Classical Test Theory		
Distinguishing Power	Number of Questions	Question Numbers
Very low	3 questions	16, 24, 30
Low	7 questions	13, 14, 20, 25, 26, 32, 34
Moderate	18 questions	1, 3, 5, 6, 8, 12, 15, 17, 19, 21, 23, 27, 29, 31, 35, 36, 37, 39
High	10 questions	4, 7, 9, 10, 11, 18, 22, 28, 33, 38
Very high	2 questions	2, 40

Item discriminating power is used to distinguish students with high and low abilities in answering questions. Rasch's analysis assesses two types of discriminating power, namely person separation and item separation (Table 9). Person separation is used to categorize students. The result of high person separation can distinguish students into 3 groups including students with low, moderate, and high abilities (person reliability > 0.8). Item separation is used to check the hierarchy of a question item. A low item separation result indicates that the sample is not large enough to confirm the three difficulty hierarchies: high, medium, and low item difficulties (item reliability < 0.9) (Wongpakaran et al., 2020).



SUMMARY OF 36 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	28.9	40.0	1.36	.45	1.01	.04	1.03	.06
SEM	1.0	.0	.19	.03	.02	.12	.05	.13
P.SD	6.2	.0	1.14	.18	.11	.70	.30	.77
S.SD	6.3	.0	1.16	.19	.11	.71	.30	.78
MAX.	39.0	40.0	4.07	1.03	1.27	1.05	1.81	1.40
MIN.	13.0	40.0	-.87	.35	.69	-2.63	.48	-2.17

REAL RMSE	.50	TRUE SD	1.03	SEPARATION 2.04	Person RELIABILITY .81
MODEL RMSE	.49	TRUE SD	1.03	SEPARATION 2.11	Person RELIABILITY .82
S.E. OF Person MEAN = .19					

Person RAW SCORE-TO-MEASURE CORRELATION = .95
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .83 SEM = 2.53

SUMMARY OF 40 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	26.0	36.0	.00	.45	1.00	.04	1.03	.07
SEM	.8	.0	.16	.02	.02	.12	.11	.15
P.SD	5.2	.0	.97	.11	.13	.73	.67	.95
S.SD	5.3	.0	.98	.11	.13	.74	.68	.97
MAX.	34.0	36.0	2.02	.75	1.25	1.22	4.92	4.36
MIN.	13.0	36.0	-1.94	.37	.77	-1.75	.41	-1.54

REAL RMSE	.48	TRUE SD	.84	SEPARATION 1.77	Item RELIABILITY .76
MODEL RMSE	.46	TRUE SD	.85	SEPARATION 1.83	Item RELIABILITY .77
S.E. OF Item MEAN = .16					

Figure 1: Summary of Statistics Rasch

Strata separation equation (H):

$$H = \frac{[(4 \times Separation) + 1]}{3}$$

(Sumintono & Widhiarso, 2013)

Table 9: Distinguishing Power Test in Item Response Theory (Rasch)

Item Response Theory (Rasch)		
Distinguishing Power	Value	H
Person separation	2.04	3.05 = 3 groups of students
Item separation	1.77	2.69 = 3 groups of questions

8. Item Bias Test

Item bias test was conducted on gender parameters, specifically focusing on men and women. The analysis using the Rasch model enables the identification of gender bias in the questions, as found in Table 10. The results show that 40 questions portray no gender bias.

Table 10: Item DIF (Gender-based item bias)

Item DIF	Number of Questions
Biased (Prob < 5%)	-
Unbiased	40 questions

DIF (Differential Item Functioning) analysis is important for maintaining the validity scale of the instrument. DIF helps in recognizing question items that show bias (Au et al., 2023). Bias in a question regarding gender, ethnicity, religion, education, income, etc., can lead to unfairness for certain groups of respondents (Pellegrino et al., 2001). In this study,

item bias testing was conducted on the gender parameter. The items on the instrument must be invariant across all groups, meaning that the calibration of the questions must be the same for each different group (Wongpakaran et al., 2020). The Rasch Wright map also displays differences in item logits and the average student ability levels, marked by L (male) and P (female). Although there are significant differences between men and women, it does not necessarily mean that there is bias/DIF in the questions (Boonee· et al., 2014).

The Rasch model is used as a measurement approach in investigating the psychometric properties of an instrument due to its advantages over classical test theory (Mitchell-Parker et al., 2018). Rasch analysis can address certain limitations of classical test theory, such as the ability to scale item difficulty and rank student ability categories using an appropriate ordinal scale (Rahayu et al., 2021). The concept of objective assessment is achieved through the use of Rasch analysis (Isnani et al., 2019). This method allows for the production of linear measurements with equal intervals, precise estimation processes, identification of misfit and outlier items, handling of missing data, and generating measurements that are not dependent on personal parameters (absence of bias).

CONCLUSION

The conclusion that can be inferred from the given explanation is that a developed instrument must possess specific characteristics that meet the requirements to measure competency effectively. The data analysis from classical test theory and item response theory (Rasch) have slightly different interpretations but are mutually complementary. Classical test theory and item response theory can analyze the validity, reliability, distractor effectiveness, difficulty level, and discriminating power of questions in the respiratory system instrument. The item response theory (Rasch) offers a deeper interpretation through the use of the Wright map as a ruler which helps determine students' ability levels about the difficulty of the questions. The scalogram allows for observing the patterns in students' answers, enabling the detection of cheating and inaccuracies in answering questions. Additionally, DIF items are used to identify any biases present in the questions.

REFERENCES

Anselmi, P., Colledani, D., & Robusto, E. (2019). A Comparison of Classical and Modern Measures of Internal Consistency. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02714

Arikunto, S. (2012). *Prosedur Penelitian*. Jakarta: Rineka Cipta.

Au, M. L., Li, Y. Y., Tong, L. K., Wang, S. C., & Ng, W. I. (2023). Chinese version of Yoon Critical Thinking Disposition Instrument: validation using classical test theory and Rasch analysis. *BMC Nursing*, *22*(1). https://doi.org/10.1186/s12912-023-01519-y

Azmi, M. P., & Salam, A. (2020). Pengembangan Instrumen Tes Kemampuan Komunikasi Matematis pada Materi Segi Empat. *Journal for Research in Mathematics Learning) p*, *3*(3), 181–192.

Bichi, A. A., Embong, R., Mamat, M., & Maiwada, D. A. (2015). Comparison of Classical Test Theory and Item Response Theory: A Review of Empirical Studies. *Australian Journal of Basic and Applied Sciences*, *9*(7), 549–556. https://doi.org/10.13140/RG.2.1.1561.5522

Boonee·, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. New York : Springer Dordrecht Heidelberg.

Chitaree, D., Junpeng, P., Sonsilphong, S., & Tang, K. N. (2024). Using Machine Learning to Score Multidimensional Assessments of Students' Skill Levels in Mathematics. *Pertanika Journal of Social Sciences and Humanities*, *32*(1), 217–235. https://doi.org/10.47836/pjssh.32.1.10

Darmana, A., Sutiani, A., Nasution, H. A., Ismanisa*, I., & Nurhaswinda, N. (2021). Analysis of Rasch Model for the Validation of Chemistry National Exam Instruments. *Jurnal Pendidikan Sains Indonesia*, *9*(3), 329–345. https://doi.org/10.24815/jpsi.v9i3.19618

Erfan, M., Archi Maulyda, M., Hidayati, V. R., Astria, F. P., Ratu, T., Studi, P., Guru, P., & Dasar, S. (2020). Analisis Kualitas Soal Kemampuan Membedakan Rangkaian Seri dan Paralel melalui Teori Tes Klasik dan Model Rasch. *Indonesian Journal of Educational Research and Review*, *3*(1), 11.

Gronlund, N. E., & Waugh, C. K. (2013). *Assessment of Student Achievement*. New Jersey : Pearson Education.

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

Isnani, I., Utami, W. B., Susongko, P., & Lestiani, H. T. (2019). Estimation of college students' ability on real analysis course using Rasch model. *REID (Research and Evaluation in Education)*, *5*(2), 95–102. https://doi.org/10.21831/reid.v5i2.20924

Jumini, S., Madnasri, S., Cahyono, E., & Parmin, P. (2023). Analisis Kualitas Butir Soal Pengukuran Literasi Sains Melalui Teori Tes Klasik Dan Rasch Model. *Prosiding Seminar Nasional Pascasarjana Universitas Negeri Semarang*, 758–765. http://pps.unnes.ac.id/pps2/prodi/prosiding-pascasarjana-unnes758

Mitchell-Parker, K., Medvedev, O. N., Krägeloh, C. U., & Siegert, R. J. (2018). Rasch analysis of the Frost Multidimensional Perfectionism Scale. *Australian Journal of Psychology*, *70*(3), 258–268. https://doi.org/10.1111/ajpy.12192

Mohajan, H. K. (2020). Quantitative Research: A Successful Investigation in Natural and Social Sciences. *Journal of Economic Development, Environment and People*, *9*(4), 52–79.

Nurjanah, W. L., Sari, I. M., & Saepuzaman, D. (2024). Analisis Pemahaman Konsep Peserta Didik Pada Topik Perambatan Kalor Menggunakan Rasch Model. *Jurnal Kependidikan*, *13*(2). https://jurnaldidaktika.org

Pellegrino, J. W., Chudowsky, Naomi., Glaser, R., & National Research Council (U.S.). Committee on the Foundations of Assessment. (2001). *Knowing what students know : the science and design of educational assessment*. Washington, DC : National Academy Press.

Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research*, *15*(2). https://doi.org/10.1103/PhysRevPhysEducRes.15.020111

Popham, W. James. (2017). *Classroom assessment : what teachers need to know*. University of California, Los Angeles.

Rahayu, W., Putra, M. D. K., Rahmawati, Y., Hayat, B., & Koul, R. B. (2021). Validating an indonesian version of the what is happening in this class? (wihic) questionnaire using a multidimensional rasch model. *International Journal of Instruction*, *14*(2), 919–934. https://doi.org/10.29333/iji.2021.14252a

Robinson, M., Johnson, A. M., Walton, D. M., & MacDermid, J. C. (2019). A comparison of the polytomous Rasch analysis output of RUMM2030 and R (ltm/eRm/TAM/lordif). *BMC Medical Research Methodology*, *19*(1). https://doi.org/10.1186/s12874-019-0680-5

Septiliana, L. (2023). Analisis Item Soal dengan Menggunakan Rasch Model sebagai Ukuran Kualitas Madrasah Ibtidaiyah pada Mata Pelajaran IPA. *Pionir: Jurnal Pendidikan*, *12*(2), 1–12.

Sugiyono. (2016). *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*. Bandung : Alfabeta.

Sumaryanta. (2021). *Teori Tes Klasik & Teori Respon Butir : Konsep & Contoh Penerapannya*. Cirebon : Confident.

Sumintono, B., & Widhiarso, W. (2013). *Aplikasi Model Rasch Untuk Penelitian Ilmu-Ilmu Sosial*. Trim Komunikata Publishing House.

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

Thomas, F. B. (2022). The Role of Purposive Sampling Technique as a Tool for Informal Choices in a Social Sciences in Research Methods. *Just Agriculture*, *2*(5), 1–8. www.justagriculture.in

Wongpakaran, N., Wongpakaran, T., Pinyopornpanish, M., Simcharoen, S., Suradom, C., Varnado, P., & Kuntawong, P. (2020). Development and validation of a 6-item Revised UCLA Loneliness Scale (RULS-6) using Rasch analysis. *British Journal of Health Psychology*, *25*(2), 233–256. https://doi.org/10.1111/bjhp.12404