

## Comparison of the Classical Test Theory and the Rasch Model in the Analysis of Mastery Test of Concept Systems of Linear Equations of Two Variable (SPLDV)

Putu Aulia Ciptari<sup>1</sup>, Ni Kadek Rini Purwati<sup>2</sup>, Ni Ketut Erawati<sup>3</sup>

<sup>123</sup>Universitas PGRI Mahadewa Indonesia, Jl. Seroja 57, Indonesia  
e-mail: [ciptariaulia@gmail.com](mailto:ciptariaulia@gmail.com)

### Abstract

This study aims to compare the analysis of the quality of a measurement instrument for concept understanding using the classical test theory approach and the Rasch model. The analysis methods for instrument quality, including validity, reliability, item difficulty, and item discrimination, were applied to the SPLDV concept mastery instrument. The research findings reveal differences in the instrument quality between the two approaches. The analysis through the classical test theory approach shows good instrument quality regarding validity and reliability, although item difficulty still needs improvement. On the other hand, the analysis through the Rasch model shows variation in item difficulty and sufficient reliability. While the item discrimination through the classical test theory approach is categorized as good, the analysis through the Rasch model yields unsatisfactory results. In conclusion, the quality of the instrument can differ depending on the approach used. This study provides a better understanding of the differences between the classical test theory approach and the Rasch model in analyzing the quality of a measurement instrument for conceptual understanding.

**Keywords:** instrument quality analysis, concept understanding, classical test theory, rasch model

### Abstrak

Penelitian ini bertujuan untuk membandingkan analisis kualitas instrumen pengukuran pemahaman konsep melalui pendekatan teori tes klasik dan model Rasch. Metode analisis kualitas instrumen, termasuk validitas, reliabilitas, tingkat kesukaran, dan daya pembeda, diaplikasikan pada instrumen penguasaan konsep SPLDV. Hasil penelitian menunjukkan perbedaan dalam kualitas instrumen antara kedua pendekatan tersebut. Analisis melalui pendekatan teori tes klasik menunjukkan kualitas instrumen yang baik dari segi validitas dan reliabilitas, tetapi tingkat kesukaran butir soal masih perlu perbaikan. Sementara itu, analisis melalui model Rasch menunjukkan variasi tingkat kesukaran dan reliabilitas yang cukup. Meskipun daya pembeda instrumen melalui pendekatan teori tes klasik tergolong baik, analisis melalui model Rasch menunjukkan hasil yang belum memuaskan. Dalam kesimpulannya, kualitas instrumen dapat berbeda tergantung pada pendekatan yang digunakan. Penelitian ini memberikan pemahaman yang lebih baik tentang perbedaan antara pendekatan teori tes klasik dan model Rasch dalam analisis kualitas instrumen pengukuran pemahaman konsep.

**Kata kunci:** analisis kualitas instrumen, pemahaman konsep, teori tes klasik, model rasch

### INTRODUCTION

Evaluation in learning can be assessed to measure the understanding and mastery of students regarding the delivered material. In other words, evaluation can provide an explanation of each student's information about their achievements and competency attained during the teaching and learning process (Marjiastuti and Wahyuni, 2014). Evaluation plays a crucial role in accommodating students according to the educational objectives and can serve as a means to motivate students to develop positive learning habits, identify strengths and weaknesses, and provide feedback (Pisca, 2014). These statements

indicate that effective evaluation contributes to effective learning. Therefore, educators have an important responsibility to conduct appropriate evaluation to measure student's academic performance. The common method used to assess student's skills is the use of tests or non-test instruments. Test instruments are commonly employed in schools to measure student's abilities. Test instruments are considered acceptable for all subjects taught by teachers to students through the distribution of questions that explain the concepts that have been taught. It is important that the questions used in evaluation have good quality. This is because high-quality questions can provide accurate information about student's ability to understand the material taught by the teacher.

Hayati and Lailatussaadah (2016) stated in their research that a test instrument is considered to have good quality if its validity and reliability levels are high. The research data will be more accurate if the validity and reliability levels are high. Tri Wahyuningsih (2015) also supports this statement, emphasizing that validity and reliability are fundamental aspects in determining the quality of a test instrument. In addition to these factors, difficulty level and item discrimination are other aspects that go hand in hand in determining the quality of a test instrument (Tri Wahyuningsih, 2015). Therefore, it is crucial to conduct validity and reliability tests, as well as assess the difficulty level and item discrimination of the test instrument to ensure its quality. This applies not only to researchers in the field of education but also to anyone conducting research. The test instrument used in research must have high validity and reliability to accurately and precisely measure the variables under study. Additionally, test developers should have knowledge about the difficulty level and item discrimination of the test instrument because it provides information on the relationship between student's ability levels and the difficulty level of the test instrument. However, in practice, the testing of test instrument quality is rarely performed, resulting in many test instruments whose quality is unknown. As a result, the assessment of students' abilities becomes inaccurate and unmeasurable.

There are two approaches that can be used to analyze test instruments in the field of education. One commonly used approach is the classical test theory (CTT). In this approach, the significant factors in assessing the quality of test items are the difficulty level and the ability of the items to discriminate among test takers. However, the characteristics of test items produced by classical test theory can vary according to the abilities of the test takers.

According to Marjiastuti and Wahyuni (2014), in classical test theory, measurement errors can only be traced back to groups of test takers and not individuals.

Another approach is to use a modern approach known as the Rasch model. This approach was introduced as a solution to address the limitations of the classical theory. The Rasch model provides a different method for utilizing raw scores or data in the context of educational assessment. The purpose of using the Rasch model on raw test data is to produce a measurement scale that has consistent intervals, thus providing accurate information about the abilities of test takers and the quality of items answered by students. Through the analysis of test items using the Rasch model, we can obtain information about the characteristics of test items and students, which are transformed into a uniform measurement scale (Sumintono & Widhiarso, 2015).

The purpose of this study is to compare the results of the analysis between classical test theory and the Rasch model in evaluating the characteristics of test instruments, including validity, reliability, item discrimination, and difficulty level, in order to ensure the quality of the test instrument. The object of analysis in this study is the test instrument for assessing the mastery of concepts in Systems of Linear Equations with Two Variables (SPLDV).

## **METHODS**

In this study, secondary data from the test instrument measuring the mastery of concepts in Systems of Linear Equations with Two Variables (SPLDV), conducted in Class VIII G and VIII H at SMP Negeri 3 Abiansema in May 2023, were used. The secondary data were obtained through documentation, where written documents consisting of a multiple-choice test instrument with 15 items and 64 written answer sheets were collected. This qualitative descriptive study aims to obtain information and data that can be used to describe the quality of the test instrument empirically. The analysis is conducted using the classical test theory approach using the elements of validity, reliability, difficulty level, and item discrimination. The data were analyzed using SPSS version 23 and the Rasch Model with the assistance of Minsteps software.

## **RESULTS AND DISCUSSION**

This study aims to compare the analysis results between classical test theory and the Rasch model in evaluating the characteristics of the test instrument, including validity,

reliability, item discrimination, and difficulty level. The test instrument analyzed is the SPLDV concept mastery test instrument developed by the researcher through the test instrument development procedure.

### Validity Test

A validity test is a method used to determine the extent to which a measuring instrument or tool can accurately measure what it intends to measure (Sugiyono, 2018). In classical test theory, validity testing is conducted using the product-moment correlation method with the assistance of SPSS version 23. The testing is performed at a significance level of 5% with the following criteria: if the correlation coefficient  $r_{xy} > r_{table}$ , then the test item is considered valid (Syofian, 2015). In the Rasch Model, validity testing is referred to as fit and misfit testing (valid and non-valid items), which can be conducted by analyzing the output from the Item fit order in the Winsteps program (Muntazhimah, 2020). According to Sumintono & Widhiarso (2015), the acceptance criteria for validity testing in the Rasch Model are as follows:

- a) Outfit MNSQ (Mean Square):  $0,5 < \text{Outfit} - \text{MNSQ} < 1,5$
- b) Outfit ZSTD (Z – Standard):  $-2,0 < \text{ZSTD} < +2,0$
- c) Pt Measure Corr (Point Measure Correlation):  $0,4 < \text{Pt Measure Corr} < 0,85$

The results of the validity analysis of the SPLDV concept mastery test items based on classical test theory and the Rasch model are presented in Table 1.

**Table 1. Comparison of Validity Analysis Results of Test Items using Classical Test Theory and Rasch Model**

No	Result	Items Number	
		Classical Test Theory	Rasch Model
1	Valid	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15	5, 6, 9, 11
2	Invalid	-	1, 2, 3, 4, 7, 8, 10, 12, 13, 14,

The results of the validity analysis of the SPLDV concept mastery test items according to classical test theory indicate that all test items are considered valid. However, only 4 test items are declared valid when using the Rasch model. 11 other test items are deemed invalid

as they do not meet the criteria for Outfit MNSQ, Outfit ZSTD, and Point Measure Correlation (Pt Measure Corr).

In analyzing the quality of the SPLDV concept mastery items, four items are considered valid or acceptable in both the classical test theory (CTT) approach and the Rasch model. These four items are items 5, 6, 9, and 11. According to Embretson and Reise (2013), these results can differ because Classical Test Theory assumes that undesirable respondent characteristics, such as item bias and Differential Item Functioning (DIF), can be addressed using specific statistical methods. However, the Rasch model takes a stricter approach to undesirable respondent characteristics. The Rasch model demands that items exhibiting bias or DIF be revised or removed from the test.

### Reliability Test

To evaluate the quality of test items in terms of reliability using the classical test theory approach, Cronbach's Alpha (KR-20) formula is commonly employed. According to Guilford (1956), as cited in Lestari and Yudhanegara (2017), test items are considered reliable if they meet the criteria for the coefficient of instrument reliability correlation, as listed in Table 2.

**Table 2. Criteria the coefficient of reliability correlation in the classical test theory**

Coefficient Correlation	Correlation	Interpretation/Reliability
$0,90 \leq r \leq 1,00$	Very High	Very high reliability
$0,70 \leq r < 0,90$	High	High reliability
$0,40 \leq r < 0,70$	Moderate	Moderate reliability
$0,20 \leq r < 0,40$	Low	Low reliability
$0 \leq r < 0,20$	Very Low	Very low reliability

In the Rasch model, the criteria used to determine Item Reliability and Person Reliability values are based on several factors described by Sumintono and Widhiarso (2015), as shown in Table 3.

**Table 3. Criteria Reliability in the Rasch model**

Reliability Value (Person/Item)	Interpretation
$> 0,94$	Special
$0,91 - 0,94$	Excellent

0,81 – 0,90	Good
0,67 – 0,80	Fair
< 0,67	Weak

The information regarding the comparison of the reliability analysis results for the SPLDV concept mastery items using the classical test theory and Rasch model approaches is documented in Table 4.

**Table 4. Comparison of Item Reliability Analysis Results through Classical Test Theory and Rasch Model Approaches**

Teori Tes Klasik		Model Rasch			
Alpha Cronbach	Intrepretation	Person Reliability	Intrepretation	Item Reliability	Intrepretation
0,862	High	0,44	Weak	0,77	Fair

Table 4 presents the results of the reliability analysis of the SPLDV concept mastery instrument. The Classical Test Theory (CTT) approach, using the Alpha Cronbach coefficient, indicates a reliability coefficient of 0.862, which is interpreted as high. On the other hand, using the Rasch model, the test reliability coefficient is obtained as 0.77 (sufficient). This suggests that the reliability of the items, as assessed by both approaches, is sufficiently reliable for use with the same subjects, even at different times, places, or by different individuals.

In addition to test reliability, the analysis using the Rasch model also provides information about respondent reliability, in this case, the students, referred to as person reliability. Based on the analysis results, the person reliability value obtained is 0.44, indicating low consistency in student responses. This can be observed through analyzing abilities, where students' abilities are evaluated based on their response patterns or answers in the scale map table. The results are presented in the following output:

## GUTTMAN SCALOGRAM OF RESPONSES:

```

Person |Item
| 111 111
|784011225639354
|-----
5 +111111111111111 05
12 +111111111111111 12
17 +111111111111111 17
22 +111111111111111 22
23 +111111111111111 23
28 +111111111111111 28
29 +111111111111111 29
31 +111111111111111 31
32 +111111111111111 32
34 +111111111111111 34
35 +111111111111111 35
38 +111111111111111 38
39 +111111111111111 39
43 +111111111111111 43
50 +111111111111111 50
51 +111111111111111 51
60 +111111111111111 60
64 +111111111111111 64
1 +111111101111111 01
4 +111111111110111 04

```

**Figure 1. Guttman Scalogram of Responses**

In this case, only 18 out of 64 students could answer the items consistently. Students from the 19th position onwards showed patterns of answers that could have been better.

The analysis results using both approaches show differences in the values and categories of item reliability. The Classical Test Theory approach and the Rasch model yield different analyses with different criteria. According to Embretson and Reise (2013), these differences in results can occur because Classical Test Theory uses reliability coefficients such as Cronbach's alpha coefficient or Split-Half coefficient to measure test reliability. On the other hand, the Rasch model uses reliability indices such as the Rasch reliability index or the person separation index (PSI).

### Item Difficulty

According to Lestari and Yudhanegara (2017), in classical test theory, the difficulty index of an item can be interpreted based on the following criteria:

**Table 5. Criteria for Difficulty Index of Test Items in Classical Test Theory**

Difficulty Index Value	Interpretation
DI = 0,00	Very Difficult
$0,00 < DI \leq 0,30$	Difficult
$0,30 < DI \leq 0,70$	Moderate
$0,70 < DI < 1,00$	Easy
DI = 1	Very Easy

In the Rasch model, the difficulty level of test items is categorized based on the Measure logit and the Standard Deviation (SD) logit of the item. The categories for item difficulty in the Rasch model are as follows, according to Sumintono and Widhiarso (2015):

**Table 6. Criteria for Difficulty Level of Test Items in the Rasch Model**

Difficulty Level	Criteria
0,0 logit + 1SD	Difficult
> + 1SD	Very Difficult
0,0 logit – 1SD	Easy
< – 1SD	Very Easy

**Table 7. Comparison of item difficulty analysis results using the classical test theory (CTT) approach and the Rasch model.**

Classical Theory Test			Rasch Model		
No	DI	Interpretation	No	Measure Value (logit)	Interpretation
1	0,94	Easy	1	-0,13	Easy
2	0,83	Easy	2	0,03	Difficult
3	0,80	Easy	3	0,32	Difficult
4	0,94	Easy	4	-1,15	Very Easy
5	0,83	Easy	5	0,03	Difficult
6	0,83	Easy	6	0,03	Difficult
7	0,94	Easy	7	-1,45	Very Easy
8	0,94	Easy	8	-1,45	Very Easy



9	0,72	Easy	9	0,96	Difficult
10	0,89	Easy	10	0,68	Difficult
11	0,88	Easy	11	-0,48	Mudah
12	0,84	Easy	12	-0,13	Mudah
13	0,67	Moderate	13	1,30	Very Difficult
14	0,64	Moderate	14	1,52	Very Difficult
15	0,67	Moderate	15	1,30	Very Difficult

Table 7 compares the analysis results of item difficulty using both approaches. The classical test theory (CTT) approach found that 12 items in the SPLDV concept mastery test instrument were easy, while three items were moderate.

In the analysis of item difficulty based on the Rasch model, the difficulty level of each item is evaluated based on the measured value in logits. In this case, three items were categorized as very easy (items 4, 7, and 8), three items were categorized as easy (items 1, 11, and 12), six items were categorized as difficult (items 2, 3, 5, 6, 9, and 10), and three items were categorized as very difficult (items 13, 14, and 15).

The analysis results of item difficulty using the classical test theory and Rasch model approaches show differences in categorizing item difficulty levels. The classical test theory approach resulted in two categories, easy and moderate, while the Rasch model categorized items into four categories: very easy, easy, difficult, and very difficult.

However, three items (numbers 1, 11, and 12) fall into the same category in both the classical test theory and Rasch model analysis, which is the easy category. Additionally, there is a difference in the analysis of item difficulty for six items (numbers 2, 3, 5, 6, 9, and 10). According to the classical test theory analysis, these six items are categorized as easy, while according to the Rasch model analysis, they are categorized as difficult. According to Ayala (2013), these differences can occur because classical test theory generally interprets item difficulty based on descriptive statistics such as the percentage of correct answers. On the other hand, the Rasch model uses the difficulty level measurement provided by the item difficulty parameter in the Rasch model (e.g., in logits).

### Item Discrimination

Item discrimination refers to the ability of an item to differentiate between students who can answer the item correctly and those who have a low level of ability (Arikunto, 2015). In other words, item discrimination indicates the extent to which an item can distinguish between students with high ability and those with low ability in answering the item.

The results of calculating the item discrimination index based on classical test theory are generally categorized into four categories, as shown in Table 8.

**Table 8. Categories of Item Discrimination Index in Classical Test Theory**

Item Discrimination (ID)	Interpretation	Items Number	Number of Items
$ID \geq 0,70$	Very Good	9, 10, 12	3
$0,40 \leq ID < 0,70$	Good	1, 2, 4, 5, 6, 7, 8, 11, 15	9
$0,20 \leq ID < 0,40$	Adequate	3, 13, 14	3
$ID < 0,20$	Poor	-	0

Based on the information provided in Table 8, it can be concluded that the analysis of item discrimination through the classical test theory approach shows that most items fall into the "good" category. Three items have very good item discrimination, while three others have adequate item discrimination.

In the Rasch model, the analysis at the individual ability level is used to determine the item's ability to differentiate between students who can answer the item correctly and those who are not. This method can also be used to identify groups of respondents based on the person separation index. The higher the item separation value, the better the overall instrument quality in identifying groups of respondents and groups of items (Sumintono & Widhiarso, 2014). The strata equation (H) can be used for more detailed grouping and obtaining more detailed information (Sumintono & Widhiarso, 2014).

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

Based on the analysis results, the item separation value is 1.83, and the strata equation (H) is calculated to be 2.77, rounded to 3. This indicates that there are three groups of items that can be identified. As for the person separation value of 0.89, the strata equation (H) is

calculated to be 1.52, rounded to 2. This shows that the respondents can be differentiated into two groups based on the person separation value. The analysis results are presented in the following output.

Person	64 INPUT	64 MEASURED	INFIT		OUTFIT			
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	12.2	15.0	2.25	1.08	.99	.0	1.03	.1
P.SD	2.9	.0	1.59	5.0	.28	.9	.58	1.0
REAL RMSE	1.19	TRUE SD	1.06	SEPARATION	.89	Person RELIABILITY	.44	

  

Item	15 INPUT	15 MEASURED	INFIT		OUTFIT			
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	52.2	64.0	.00	.44	.96	-.1	1.03	.1
P.SD	6.0	.0	.93	.08	.34	1.5	.72	1.8
REAL RMSE	.45	TRUE SD	.81	SEPARATION	1.83	Item RELIABILITY	.77	

**Figure 2. Separation Value in Rasch Model**

Based on the analysis results of both approaches, there is an agreement in the analysis of item discrimination. The classical test theory approach and the Rasch model categorize most items as having poor discrimination. This indicates that those items could be more effective in distinguishing students with high and low abilities.

In the analysis using the Rasch model, a person strata equation (H) value of 2 was found, indicating that the measurement of SPLDV concept mastery needs to be sufficiently effective in identifying respondents into groups based on their level of conceptual understanding. This suggests that the instrument needs improvement to depict students' abilities accurately.

Based on the analysis results, it can be concluded that although there are differences in analysis methods and parameters used, the classical test theory approach and the Rasch model share the same principles as tools for measuring the psychometric properties of test items. The main difference lies in the analysis of item discrimination. The classical test theory approach analyzes item discrimination at the item level, while analysis through the Rasch model uses separation values to analyze respondents. In the Rasch model analysis, the separation value of respondents is used to identify groups of students based on their ability levels. This indicates that both approaches have slightly different analytical focuses but still provide helpful information in understanding the quality of the measurement instrument. The choice of the appropriate approach will depend on the research objectives and the needs of instrument use. The classical test theory approach can provide more detailed information at the item level, while the Rasch model provides more specific information regarding individual students' ability levels.

One limitation of the classical test theory is that it can introduce bias in the information about test quality. According to Widhiarso (2016), some limitations can affect the information provided by the test and result in bias in the test's quality. One significant limitation is its dependency on the characteristics of the sample used in the test. In this case, the analysis results can only be generalized to samples with similar characteristics to the analyzed data. As a result, the test is considered valid only when applied to individuals with limited characteristics.

## CONCLUSION

The research findings indicate differences in the quality of the measurement instrument for concept understanding using the classical test theory approach and the Rasch model. In terms of validity, the analysis through the classical test theory approach shows good quality, while the analysis through the Rasch model still needs improvement. Regarding reliability analysis, the classical test theory approach yields high reliability with a value of 0.862, while the Rasch model approach has sufficient reliability of 0.77. In analyzing item difficulty, the classical test theory approach still needs to show better quality. In contrast, the analysis through the Rasch model shows variation in item difficulty, ranging from very easy, easy, difficult, to very difficult. Most item discriminations in the SPLDV concept mastery instrument using the classical test theory approach are good for item discrimination analysis. However, the analysis through the Rasch model shows the quality that still needs to be satisfactory, with a respondent separation value of 2. It should be noted that the quality of the instrument can differ depending on the approach used. In this case, there are differences in the analysis results between the classical test theory approach and the Rasch model.

## REFERENCES

- Arikunto, S. (2015). *Dasar-dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara.
- Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.
- Hayati, S., & Lailatussaadah, L. (2016). Validitas dan Reliabilitas Instrumen Pengetahuan Pembelajaran Aktif, Kreatif dan Menyenangkan (Pakem) Menggunakan Model Rasch. *Jurnal Ilmiah Didaktika*, 16(2), 169– 179. doi: 10.22373/jid.v16i2.593
- Lestari, K., & Yudhanegara, M. (2017). *Penelitian Pendidikan Matematika*. Bandung: PT Refika Aditama.

- Marjiastuti, K., & Wahyuni, S. 2014. Analisis Kemampuan Peserta Didik Dengan Model RASCH. Seminar Nasional Evaluasi Pendidikan, 94–102.
- Misbah, I.H & Sumintono, B. (2014). Pengembangan dan validasi instrumen "persepsi siswa terhadap karakter moral guru" di Indonesia dengan model rasch, dipresentasikan dalam Seminar Nasional "Pengembangan Instrumen Penilaian Karakter yang Valid" di Fakultas Psikologi, Universitas Muhammadiyah Surakarta
- Pisca, C. C. (2014). Analisis Perbandingan Kualitas Butir Soal Ujian Sekolah Bahasa Prancis SMAN 10 Yogyakarta Tahun Ajaran 2013/2014 Dilihat dari Paradigma Klasik dan Modern (PhD Thesis). Universitas Negeri Yogyakarta.
- Sugiyono. 2018. Metode Penelitian Kuantitatif, Kualitatif, dan R&D, Cetakan Ke28, Alfabeta, Bandung.
- Sumintono , B. & Widhiarso, W. (2014). Aplikasi Model Rasch Untuk Penelitian Ilmu-Ilmu Sosial. Cimahi: Trim Komunikata Publishing House.
- Sumintono, B., & Widhiarso, W. (2015). Aplikasi model Rrasch pada Assessment Pendidikan. Cimahi: Trim Komunikata Publishing House.
- Syofian. (2015). Metode Penelitian Kuantitatif dilengkapi Perbandingan Perhitungan Manual & SPSS. Jakarta: PRENAMEDIA GROUP.
- Tri Wahyuningsih, E. (2015). Analisis Butir Soal Tes Objektif Buatan Guru Ulangan Semester Ganjil Mata Pelajaran Ekonomi Kelas X di Sma Negeri 1 Mlati Tahun Ajaran 2013/2014 (PhD Thesis). Fakultas Ekonomi.
- Widhiarso, W. (2016). Penerapan Model Rasch untuk Mengevaluasi Tes UKKS dan UKPS. Tenaga Kependidikan, 1 (1). pp. 50-51.