International Journal of Innovative Research and Scientific Studies, 8(9) 2025, pages: 113-126

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Assessing university student's Van Hiele level of abstract thinking in-two variable linear equations using IRT: Implication for Economic Education

Jesi Irwanto[1], Dian Kurniati[1*], Erfan Yudianto[1], Kiswara Agung Santoso[2]

[1]*Department of Mathematics Education, Faculty of Teacher Training and Education, University of Jember, Indonesia.*
[2]*Department of Mathematics, Faculty of Mathematics and Science, University of Jember, Indonesia.*

Corresponding author: Dian Kurniati (*Email: dian.kurniati@unej.ac.id*)

Abstract

This study aimed to identify and analyze the level of students' abstract thinking related to systems of two-variable linear equations using van Hiele's theory within the Item Response Theory (IRT) framework. This quantitative study involved 139 Indonesian students selected through random sampling. The research instrument, which was compiled based on the five van Hiele levels, was analyzed using SPSS and R Studio. The results showed that there were 8 students at level 0, 35 at level 1, 12 at level 2, 29 at level 3, and 15 at level 4. In addition, 23 students were in the transition stage, and 16 students could not be classified. The research findings indicate a significant gap in students' abstract thinking abilities, which suggests the need for a learning approach that encourages higher-level reasoning. The conclusion is that most students are at the analysis level; therefore, an appropriate learning approach is needed to develop students' abstract thinking abilities in systems of two-variable linear equations. The implication of an appropriate approach is that it will better prepare students, especially in the field of economics, to apply mathematical models in research.

Keywords: Abstract thinking, Item Response Theory (IRT), Student Understanding, Systems of linear equations in two variables.

DOI: 10.53894/ijirss.v8i9.10633
Funding: This research was funded by internal funds from the University of Jember in 2025 through publication incentive funding in reputable international journals.
History: Received: 22 August 2025 / **Revised:** 24 September 2025 / **Accepted:** 26 September 2025 / **Published:** 13 October 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

Mathematics is an abstract science that requires students to understand its basic concepts in order to develop logical, analytical, and abstract thinking skills. The system of linear equations with two variables plays a central role among the fundamental topics in mathematics. This system is introduced at the high school level and further explored in college as foundational material for advanced courses such as linear algebra, discrete mathematics, mathematical economics, and numerical analysis. A strong conceptual understanding of linear equation systems enables students to solve various

problems, including optimization cases and economic data analysis. The 2024-2025 academic experience at the Widya Gama Lumajang Institute of Technology and Business shows that many first-year students have difficulty solving applied problems. Students encounter these difficulties because they lack a deep understanding of the fundamental concepts of linear equations. These challenges are typically linked to the conventional teaching approach at the high school level, which emphasizes procedures over building conceptual understanding and abstract thinking skills.

A number of studies emphasize the importance of strengthening conceptual understanding in mathematics education. Students often have difficulty understanding systems of linear equations with two variables and require appropriate learning strategies to reduce learning barriers and foster perseverance in problem-solving [1]. Conceptual understanding is not merely about memorizing formulas but also involves students' ability to integrate various mathematical elements into a coherent structure [2]. Pre-service teachers should be able to generate connections across representations, among concepts, and to real-world situations related to the given mathematical content [3]. Students' conceptual understanding can be optimally developed through a deeper comprehension of the subject matter rather than merely memorizing procedures [4]. In addition framework for describing students' levels of understanding, ranging from visual recognition to formal abstraction [5]. Although van Hiele's theory has been widely applied to geometry, its potential for analyzing students' thinking on algebraic topics, particularly systems of linear equations with two variables, has rarely been explored.

To assess students' level of understanding, a reliable measurement method is needed. Item Response Theory (IRT) has emerged as a powerful psychometric framework for evaluating test items while analyzing students' latent abilities. Unlike Classical Test Theory, IRT considers parameters at the item level, including difficulty, discriminative power, and guessing probability, thereby providing more precise and diagnostic results [6]. Several studies have demonstrated the effectiveness of IRT in developing valid instruments and evaluating learning outcomes. Item response theory can be applied to evaluate classification consistency in large-scale assessments, demonstrating reliability and highlighting practical conditions that influence classification indices [7]. Item Response Theory can validate the structure and quality of test items and help identify the strengths of an instrument in measuring learning outcomes [8]. Several studies have also demonstrated the effectiveness of Item Response Theory in the analysis of research instruments. However, its integration with the Van Hiele framework in analyzing algebraic understanding remains limited. This limitation indicates a research gap that needs to be addressed.

This study presents three main novelties. First, it integrates Van Hiele's theory with the 3PL Item Response Theory (IRT) model to analyze students' abstract thinking levels in the context of two-variable linear equations, a topic that has been rarely explored [9]. Second, it offers a more comprehensive diagnostic mapping by not only classifying students' thinking levels but also validating the instrument through analyses of item difficulty, discrimination, and guessing parameters [10]. Third, it extends the application of Van Hiele's theory and IRT to the field of economics education, whereas most previous studies have primarily focused on secondary school students or mathematics majors.

This study aims to analyze the level of abstract thinking of first-year economics students in Indonesia, specifically at the Widya Gama Institute of Technology and Business in Lumajang. The analysis will focus on their understanding of two-variable linear equation systems, using the Van Hiele theoretical framework supported by Item Response Theory (IRT) analysis. By combining these two approaches, this study is expected to provide a more valid and reliable diagnostic framework for identifying students' levels of abstract thinking. This research will also contribute both theoretically and practically to the teaching of economics and mathematics.

2. Literature Review

2.1. Van Hiele Theory in Mathematics Education

Van Hiele theory explains the development of mathematical thinking through five hierarchical levels: visualization, analysis, informal deduction, formal deduction, and rigor [5]. At the initial stage, individuals recognize objects based on their overall appearance. They then progress to the analysis stage, where they begin to identify specific properties. Subsequently, learners start to connect these properties intuitively before advancing to the ability to construct structured logical arguments and, ultimately, to understand abstract axiomatic systems. In line with the theory, previous studies found that most university students tend to operate at the informal deduction level when solving geometry problems [11]. Highlighted the effectiveness of guided discovery learning in fostering higher levels of reasoning [2]. These findings indicate that well-designed instructional approaches can facilitate students' progression to more advanced levels of mathematical thinking.

Although the van Hiele theory has been widely applied in geometry, it is also relevant to other mathematical topics, such as systems of linear equations in two variables. In this context, students' learning can be analyzed in terms of their transition from graphical representations toward more abstract symbolic understanding. Thus, the van Hiele theory not only accounts for students' problem-solving abilities but also provides a conceptual framework for understanding the development of abstract thinking in mathematics learning.

2.2. Item Response Theory (IRT) as a Measurement Model

Item Response Theory (IRT) is one of the most widely applied frameworks in educational measurement, particularly for evaluating the quality of test items and assessing individual abilities with greater accuracy. Unlike Classical Test Theory, which focuses on total test scores, IRT emphasizes the relationship between a learner's latent ability and the probability of answering each item correctly [12]. Meanwhile among its various models the three-parameter logistic (3PL) model is frequently employed because it estimates item difficulty, discrimination, and the probability of guessing [13].

These parameters provide detailed information on how test items function across different ability levels, thereby ensuring more precise and equitable measurement.

In line with this theory, a number of studies have demonstrated the effectiveness of item response theory in the development and validation of assessment instruments. Item response theory provides a more sophisticated and flexible framework than CTT in assessing the quality of test instruments [14]. Item response theory represents a robust methodological framework for the evaluation and development of language assessments [15]. Highlighted the usefulness of the IRT model in educational measurement using the ltm package in R [16]. Model fit testing in IRT using R is effective in detecting inconsistent response patterns, thereby enhancing the quality of test score interpretation [17]. The results are further clarified with provides guidance on item response theory analysis using R packages, enabling researchers to conduct more in-depth analyses of their test data [18]. Consistent with the findings of previous research the application of graphical approaches such as the plot bin method to evaluate model fit [19]. These findings reaffirm the potential of IRT to generate valid and reliable instruments across diverse educational context.

In the present study, IRT is employed not only to evaluate the quality of the instrument but also to support the analysis of students' abstract thinking levels in learning systems of linear Equations in two variables. Thus, IRT provides a methodological foundation that enhances both the validity and the reliability of the research finding.

2.3. Two-Variable Linear Equation System

A linear equation is a form of open sentence characterized by the symbol "=". In a linear equation, all variables have an exponent of one [20]. Such equations, which express open sentences, can generally be represented in the following form.

$$ax + by = c \qquad \forall\, a, b, c, x, y \in R$$

The statement regarding systems of linear equations in two variables is supported by findings from several previous studies. A system of linear equations is defined as a system in which two or more linear equations are combined to form a set of simultaneous equations. In other words, a system of linear equations refers to a structure composed of two or more linear equations [1]. System of Linear Equations in Two Variables is defined as a set comprising two linear equations with two variables, whose solution is an ordered pair (x, y) that simultaneously satisfies both equations [21]. A system of linear equations is concerned with finding an ordered pair of numbers that represents the point of intersection of the two linear equations, either algebraically or geometrically through the intersection of two straight lines on the coordinate plane [22].

3. Research Methodology

3.1. Type of Research

This research is a type of quantitative research is the collection and analysis of numerical data to answer scientific questions; it is employed to summarize information, identify patterns, make predictions, and test causal relationships between variables, as well as to generalize findings to a broader population [23]. This research uses certain populations and samples. Data collection in this study used test package instruments (question items) and quantitative data analysis (descriptive statistics) with the aim of obtaining an initial description of the level of abstract thinking of mathematics students.

In this study, the selection of van Hiele's theory and item response theory (IRT) model is based on the main objective to analyze the level of student understanding in a comprehensive and measurable manner. A similar situation was conducted by Yudianto, et al. [24] who investigated the identification of students' geometric thinking levels based on the Van Hiele theory in the topic of analytic space geometry.

Van Hiele's theory provides a clear conceptual framework regarding the stages of abstract thinking development in mathematics, starting from level 0 (visualization) to level 4 (rigor). This theory is relevant because students do not automatically move to higher levels of thinking without gradually going through the previous stages. Therefore, classifying students' abilities using Van Hiele's theory allows researchers to identify specifically at which level comprehension difficulties occur. Meanwhile, IRT with Model 3 PL was chosen because it is able to provide detailed information about the characteristics of each item, such as difficulty level, differentiation, and chance of guessing. This model also accommodates variations in individual abilities, making it very suitable for analyzing complex Van Hiele level- based instruments. Combining the two makes this research not only able to map the cognitive position of students in Van hiele's thinking structure, but also ensure that each item used is valid and reliable in measuring the ability according to the level of thinking. This approach provides a strong basis for evaluation and improvement of learning.

3.2. Object of Research

The object of this study is the level of abstract thinking of mathematics students on the material of the system of linear equations of two variables.

3.3. Types and Sources of Data

This study consisted of primary and secondary data. The primary data were collected from item instruments completed by students. The secondary data included relevant theories and previous studies that supported this research. The data source for this study was internal, consisting of the student-completed question instrument data. The population for this study was first-year students in the management education program. A sample of 139 students was used, and the sampling technique was random sampling.

3.4. Data Analysis Technique

The validity test is carried out to determine the extent to which the instrument prepared can explore the required data or information. The instrument in this study is valid because it has a validity score above 0.3. The minimum requirement for data is considered to meet the validity requirements if r is at least 0.3. So if the correlation between the items and the total score is less than 0.3, the items in the instrument are declared invalid [25].

Reliability, or the consistency of a measurement, is carried out to determine the extent to which a data collection instrument can provide consistent results. This means that if the measurement is repeated on the same subjects at different times, the results should not be significantly different. The reliability test can be done by looking at the Cronbach Alpha coefficient [25]. The reliability criteria index is distinguished in the table as follows:

Table 1.
Reliability Criteria Index.

No.	Cronbach's Alpha	Reliability Level
1	$0.00 \leq a \leq 0.20$	Not Reliable
2	$0.21 \leq a \leq 0.40$	Slightly Reliable
3	$0.41 \leq a \leq 0.60$	Moderately Reliable
4	$0.61 \leq a \leq 0.80$	Reliable
5	$0.81 \leq a \leq 1.00$	Highly Reliable

In this study, the IRT model item test used the R studio application [26]. It is described as follows:

This test is conducted to assess the extent to which the selected IRT model (e.g. 1PL, 2PL, or 3PL) fits the participant response data. This is essential for the interpretation of item parameters (power, difficulty, and guessing) to be valid and accurate.

This test is conducted to estimate three item parameters which include Discrimination, Difficulty, and Guessing from the data obtained from respondents. To be able to estimate this, researchers compare the coefficient values that appear as a result of data processing in R studio. The research flow is shown in Figure 1.

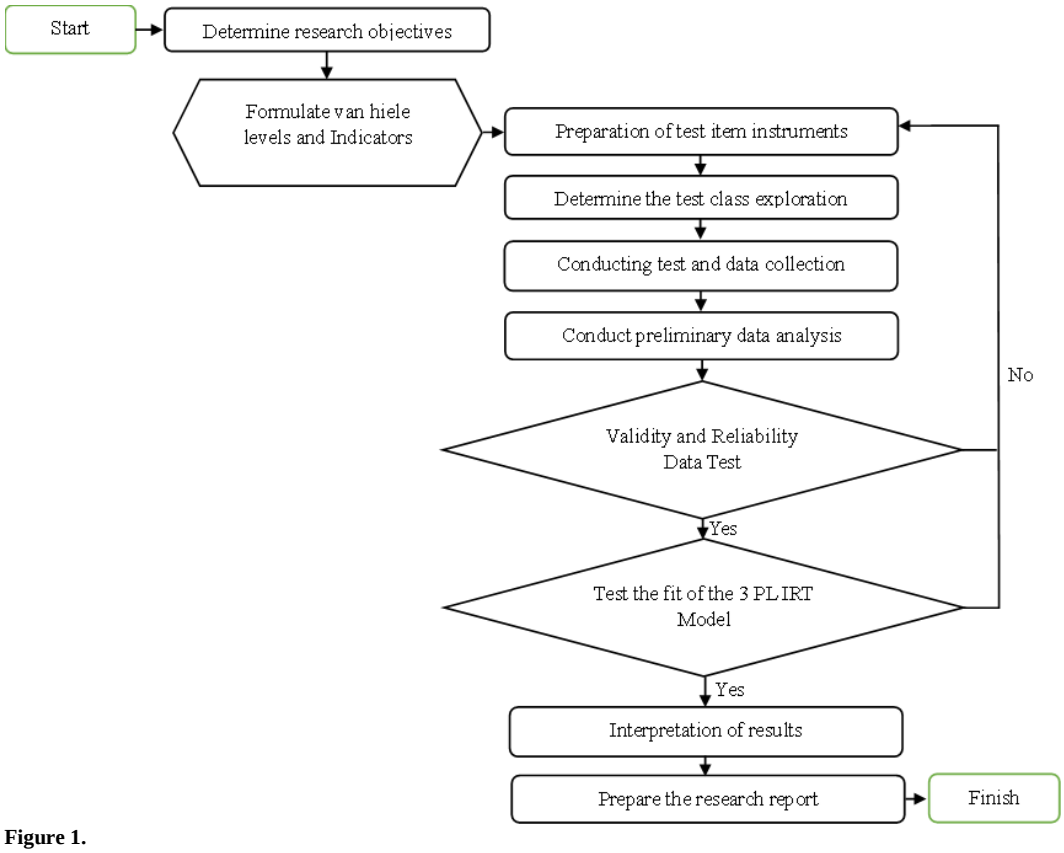

Figure 1.
Flowchart of Research Data.

4. Results

Based on the results of data collection, which involved test package tests, and data analysis of 139 respondents, the following research results are described.

4.1. Research Results

In this study, the validity test of the question instrument was carried out on each item made based on the Van Hiele level which included level 0 (visualization) there were 6 items, level 1 (analysis) 7 items, level 2 (informal deductive) 7 items, level 3 (formal deductive) 6 items, and level 4 (rigor) 6 items. The validity test results show at Table 2 until Table 6.

Table 2.
The validity test results level visualization.

Item	Corelations with total	Sig. (2-tailed)	N	Validity	Informations
B01	0.762	< 0.001	139	Valid	Highly significant
B02	0.714	< 0.001	139	Valid	Highly significant
B03	0.750	< 0.001	139	Valid	Highly significant
B04	0.496	< 0.001	139	Valid	Significant
B05	0.683	< 0.001	139	Valid	Significant
B06	0.328	< 0.001	139	Valid	Fairly significant

The validity test results show that all items (B01-B06) have a significant correlation to the total score, with a correlation value above 0.30 and a significance value (Sig.2-tailed) below 0.05. This indicates that all items are valid in measuring the intended construct. Items B01, B02, and B03 show very high correlations, signaling a strong contribution to the overall test. Although B06 has the lowest correlation (0.328), it is still within the valid category. Thus, all items are suitable for further analysis in measurement.

Table 3.
The validity test results level analysis.

Item	Corelations with total	Sig. (2-tailed)	N	Validity	Informations
B07	0.708	< 0.001	139	Valid	Highly significant
B08	0.641	< 0.001	139	Valid	Highly significant
B09	0.573	< 0.001	139	Valid	Significant
B10	0.554	< 0.001	139	Valid	Significant
B11	0.657	< 0.001	139	Valid	Highly significant
B12	0.481	< 0.001	139	Valid	Fairly significant
B13	0.464	< 0.001	139	Valid	Fairly significant

The validity test results show that all items (B07-B13) have a significant correlation to the total score, with a correlation value above 0.30 and a significance value (Sig.2-tailed) below 0.05. This indicates that all items are valid in measuring the intended construct. Items B07, B08, and B09 show a very high correlation. Thus, all items are worth using for further analysis in the measurement.

Table 4.
The validity test results level Informal deductive.

Item	Corelations with total	Sig. (2-tailed)	N	Validity	Information's
B14	0.746	< 0.001	139	Valid	Highly significant
B15	0.607	< 0.001	139	Valid	Highly significant
B16	0.767	< 0.001	139	Valid	Highly significant
B17	0.459	< 0.001	139	Valid	Fairly significant
B18	0.484	< 0.001	139	Valid	Fairly significant
B19	0.478	< 0.001	139	Valid	Fairly significant
B20	0.513	< 0.001	139	Valid	Significant

The validity test results show that all items (B14-B20) have a significant correlation to the total score, with a correlation value above 0.30 and a significance value (Sig.2-tailed) below 0.05. This indicates that all items are valid in measuring the intended construct. Items B14, B15, and B16 show a very high correlation

Table 5.
The validity test results level formal deductive.

Item	Corelations with total	Sig. (2-tailed)	N	Validity	Information's
B21	0.656	< 0.001	139	Valid	Highly significant
B22	0.572	< 0.001	139	Valid	Significant
B23	0.525	< 0.001	139	Valid	Significant
B24	0.569	< 0.001	139	Valid	Significant
B25	0.670	< 0.001	139	Valid	Highly significant
B26	0.645	< 0.001	139	Valid	Highly significant

The validity test results show that all items (B21-B26) have a significant correlation to the total score, with a correlation value above 0.30 and a significance value (Sig.2-tailed) below 0.05. This indicates that all items are valid in measuring the intended construct.

Table 6.
The validity test results level rigor

Item	Corelations with total	Sig. (2-tailed)	N	Validity	Information's
B27	0.683	< 0.001	139	Valid	Highly significant
B28	0.755	< 0.001	139	Valid	Highly significant
B29	0.320	< 0.001	139	Valid	Significant
B30	0.680	< 0.001	139	Valid	Highly significant
B31	0.714	< 0.001	139	Valid	Highly significant
B32	0.358	< 0.001	139	Valid	Fairly significant

The validity test results show that all items (B27-B32) have a significant correlation with the total score, with a correlation value above 0.30 and a significance value (Sig. 2-tailed) below 0.05. This indicates that all items are valid in measuring the intended construct.

The reliability test of the question instrument was carried out on each item made based on the Van Hiele level which included level 0 (visualization) there were 6 items, level 1 (analysis) 7 items, level 2 (informal deductive) 7 items, level 3 (formal deductive) 6 items, and level 4 (rigor) 6 items. The validity test results show at Table 7 until Table 11.

Table 7.
The reliability test results level visualization.

Cronbach's Alpha	N of items
0.683	6

Based on the reliability test results, the research instrument showed a Cronbach's Alpha value of 0.683 with a total of six items. This value is classified as sufficient and is still acceptable in early stage or exploratory research. The reliability test results also showed that students could recognize the shape or pattern of a system of linear equations with two variables visually, but had not yet understood the symbolic meaning. The sufficient reliability indicates that students at this stage have relatively consistent responses. Thus, this instrument is reliable enough to measure the intended Level 0 (Visualization).

Table 8.
The reliability test results level analysis.

Cronbach's Alpha	N of items
0.677	7

Based on the reliability test results, the research instrument showed a Cronbach's Alpha value of 0.677 with a total of seven items. This value is classified in the sufficient category and is still acceptable in early stage or exploratory research. The reliability test results also showed that they were able to identify and categorize information, but not yet build formal arguments. The reliability test results also showed that the instrument was able to stably distinguish students who were still at the "feature-aware procedural" stage but had not yet made a deep understanding. Thus, the instrument is reliable enough to measure the intended level 1 (analysis).

Table 9.
The reliability test results level informal deductive.

Cronbach's Alpha	N of items
0.651	7

The results of the reliability test, level 2 (informal deductive) instrument showed a Cronbach's Alpha value of 0.651 with a total of seven items. This value is classified in the sufficient category and is still acceptable in early stage or exploratory research. The test results show that students at this level begin to build relationships between system of linear equations with two variables concepts but are not yet systematic. Moderate reliability indicates that the items a r e able to reveal initial consistency in informal reasoning. Thus, this instrument is reliable enough to measure level 2 (informal deductive).

Table 10.
The reliability test results level formal deductive.

Cronbach's Alpha	N of items
0.656	6

The results of the reliability test, level 3 (formal deductive) instrument showed a Cronbach's Alpha value of 0.656 with a total of seven items. This value is classified in the sufficient category and is still acceptable in early stage or exploratory

research. The reliability test results also showed that students began to develop systematic logical arguments and could formalize the relationship between system of linear equations with two variables elements. Thus, this instrument is reliable enough to measure level 3 (formal deductive).

Table 11.
The reliability test results level rigor.

Cronbach's Alpha	N of items
0.621	6

The results of the reliability test, level 4 (rigor) instrument showed a Cronbach's Alpha value of 0.621 with a total of seven items. This value is classified in the sufficient category and is still acceptable in early stage or exploratory research. On the other hand, this level is the highest level, which reflects the ability to think fully abstract and translate concepts to new situations (e.g. in macro/micro economics, regression, statistical graphs). Only 15 students were at this level. The reliability is sufficient to indicate that although the instrument can be used, the rigor achievement is still very low because students have not been trained to reflect and generalize. Thus, the instrument is reliable enough to measure level 4 (rigor).

The results of the IRT model fit analysis for each item, which range from the visualization to the rigor level, indicate the accuracy of the parameters and the validity of the instrument. This is based on Van Hiele's hierarchy, as shown in the table below.

Table 12.
Fit model test level visualization.

Likelihood Ratio						
	AIC	**BIC**	**log.Lik**	**LRT**	**df**	**p.value**
Out1	1010.95	1028.51	-499.47	-	-	-
Out2	921.65	956.77	-448.82	101.3	6	<0.001
	AIC	BIC	log.Lik	LRT	df	p.value
Out2	921.65	956.77	-448.82	-	-	-
Out3	914.15	966.84	-439.07	19.5	6	<0.001
	AIC	BIC	log.Lik	LRT	df	p.value
Out1	1010.95	1028.51	-499.47	-	-	-
Out3	914.15	966.84	-439.07	120.8	12	<0.001

Based on the results of the IRT model fit test, the 3 Parameter Logistic (3PL) model shows the best fit compared to the 1PL and 2PL models. This can be seen from the lowest AIC and BIC values and the highest log likelihood, which indicates that the model fits the data better. Comparison tests between (LRT) models showed a significant improvement in fit from 1PL and 2PL. The fit significantly improved from 1PL to 2PL, and from 2PL to 3PL. Thus, the 3PL model was deemed the most suitable for use in this study because it was able to accommodate three important parameters: discrepancy, difficulty, and chance of guessing, thus providing more accurate and informative analysis results.

Table 13.
Fit model test level analysis.

Likelihood Ratio						
	AIC	**BIC**	**log.Lik**	**LRT**	**df**	**p.value**
Out1	1067.47	1087.91	-526.73	-	-	-
Out2	1045.90	1086.78	-508.95	35.57	7	<0.001
	AIC	BIC	log.Lik	LRT	df	p.value
Out2	1045.90	1086.78	-508.95	-	-	-
Out3	1020.42	1081.74	-489.21	39.48	7	<0.001
	AIC	BIC	log.Lik	LRT	df	p.value
Out1	1067.47	1087.91	-526.73	-	-	-
Out3	1020.42	1081.74	-489.21	75.04	14	<0.001

The results of the IRT model fit test at level 1 show that the 3PL (out3) model has the best fit compared to the 1PL (out1) and 2PL (out2) models. This is evidenced by the best AIC, BIC, and log-likelihood values, as well as significant Likelihood Ratio Test (LRT) results ($p < 0.001$) in each model comparison. The 2PL model is better than 1PL (LRT = 35.57), and the 3PL model is significantly superior to 2PL (LRT = 39.48) and 1PL (LRT = 75.04). Therefore, the 3PL model is most feasible to use because it is able to accommodate the guessing, difficulty, and discrimination parameters more accurately.

Table 14.
Fit model test level informal deductive.

Likelihood Ratio						
	AIC	**BIC**	**log.Lik**	**LRT**	**df**	**p.value**
Out1	1155.78	1176.01	-570.89	-	-	-
Out2	1130.59	1171.05	-551.29	39.19	7	<0.001
	AIC	BIC	log.Lik	LRT	df	p.value
Out2	1130.59	1171.05	-551.29	-	-	-
Out3	1120.26	1165.96	-547.13	39.13	7	<0.001
	AIC	BIC	log.Lik	LRT	df	p.value
Out1	1155.78	1176.01	-570.89	-	-	-
Out3	1120.26	1165.96	-547.13	47.52	14	<0.001

Based on the model results, out3 shows lower AIC and BIC than out1 and out2, as well as higher log-likelihood, indicating a better model fit. The LRT test between out2 and out3 showed a significant difference ($p<0.001$), indicating that out3 provided significant model improvement over out2. Likewise, the comparison of out1 with out3 showed significant improvement ($p < 0.001$). Therefore, out3 can be considered as the best model among the three models tested.

Table 15.
Fit model test level formal deductive.

Likelihood Ratio						
	AIC	**BIC**	**log.Lik**	**LRT**	**df**	**p.value**
Out1	1031.22	1048.57	-509.61	-	-	-
Out2	1033.98	1068.66	-504.99	9.25	6	0.16
	AIC	BIC	log.Lik	LRT	df	p.value
Out2	1033.98	1068.66	-504.99	-	-	-
Out3	1033.35	1085.37	-489.67	12.63	6	0.049
	AIC	BIC	log.Lik	LRT	df	p.value
Out1	1031.22	1048.57	-509.61	-	-	-
Out3	1033.35	1085.37	-489.67	21.88	12	0.039

The level 3 fit test results show that the out3 model has the highest log-likelihood value (-498.67) compared to out1 and out2, although its AIC and BIC values are slightly higher than out1. However, the Likelihood Ratio Test (LRT) between out2 and out3 showed a significant improvement ($p= 0.049$), as well as the comparison of out1 and out3 ($p = 0.039$), both of which are below the significance limit of 0.05. This indicates that the out3 model is statistically better than the other models in explaining the data. Thus, the 3PL model deserves to be chosen as the best model in this level 3 test.

Table 16.
Fit model test level rigor.

Likelihood Ratio						
	AIC	**BIC**	**log.Lik**	**LRT**	**df**	**p.value**
Out1	992.24	1009.58	-490.12	-	-	-
Out2	939.82	974.51	-457.91	64.42	6	<0.001
	AIC	BIC	log.Lik	LRT	df	p.value
Out2	939.82	974.51	-457.91	-	-	-
Out3	930.23	982.26	-452.12	22.6	6	<0.001
	AIC	BIC	log.Lik	LRT	df	p.value
Out1	992.24	1009.58	-490.12	-	-	-
Out3	930.23	982.26	-452.12	76.00	12	<0.001

The model fit test results show that the out2 model is significantly better than out1, with a decrease in AIC from 992.24 to 939.82 and LRT of 64.42 ($p < 0.001$). Furthermore, the comparison between out2 and out3 showed a significant improvement in fit, with AIC dropping to 930.23 and LRT of 22.6 ($p < 0.001$). In addition, the direct comparison between out1 and out3 also showed a very significant improvement, with an LRT of 76.00 ($p< 0.001$). Overall, out3 was declared the best model as it had the highest fit and better model efficiency.

The IRT analysis results for each item, from the visualization level to the rigor level, are presented through the estimation of the 3PL model coefficients, which include the discrimination, difficulty, and guessing parameters. In addition, these results are visualized using item characteristic curve graphs to provide a more comprehensive depiction of instrument performance.

Table 17.
Coefficient results – visualization level.

Item	Gussng	Dffclt	Dscrmn
B1	0.0000656248	-0.62401579	32.97454
B2	0.1562896032	-0.07869836	17.09544
B3	0.1170951598	-0.08872806	14.22420
B4	0.6449221982	0.37147204	232.11938
B5	0.3524042079	0.33373998	39.20027
B6	0.4945829892	0.71262693	12.15906

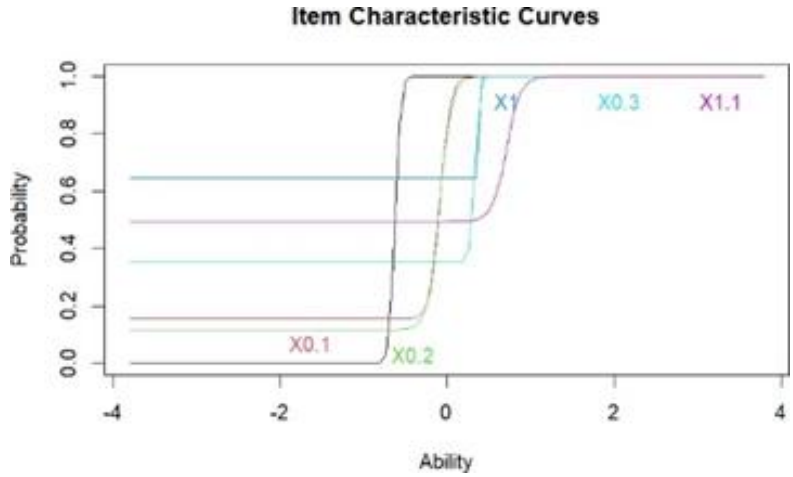

Figure 2.
Visualization level curve characteristics.

The parameter estimation results in the 3PL IRT model show significant item variation in guessing scores, difficulty, and discrimination between items. (discrimination between items. Item No. 1 had very low guessing (0.00006) and high discrimination (32.97), indicating that this item was very good at differentiating participants' abilities without the tendency to be guessed. In contrast, item No. 5 had a high guessing value (0.64) and extreme discriminating power (232.12), which may indicate overfitting or problems in the item. Item No. 6 also showed high guessing (0.49) and more moderate power differential (12.15). Overall, some items have the potential to be guessed, so they need to be reviewed for quality.

Table 18.
Coefficient results – analysis level.

Item	Gussng	Dffclt	Dscrmn
B7	3.166185e-01	0.00746738	31.7388814
B8	7.168493e-11	-0.96663863	2.3006116
B9	1.418187e-01	-0.70292466	25.4768444
B10	3.604688e-01	-0.23437068	40.7627712
B11	3.042357e-01	0.40782174	52.5387906
B12	3.500028e-06	-0.88768754	0.9609408
B13	3.701336e-01	0.68581352	32.7316287

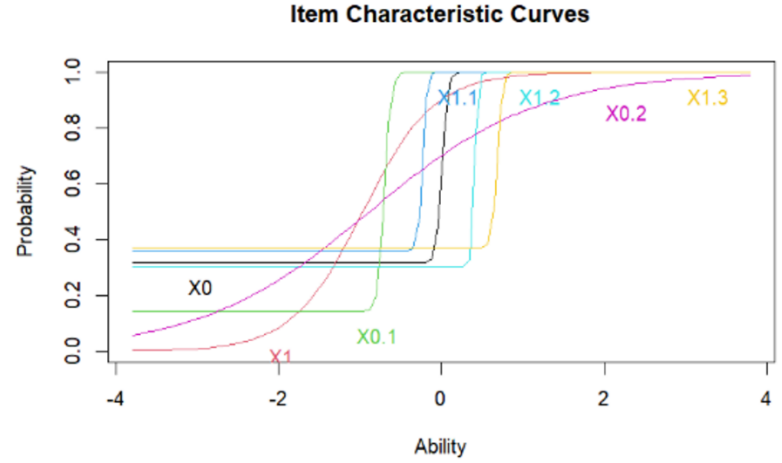

Figure 3.
Analysis level curve characteristics.

The 3PL IRT parameter estimation results show significant variations in the guessing, difficulty, and discrimination aspects between items. Some items such as item no.11 and item no.10 have a very high discrimination power of 52.54 and 40.76, indicating a very good ability to distinguish between participants with different ability levels. However, item no.12 had a very low power of difference (0.96), indicating less effectiveness in measuring ability. The high guessing values in item no.7 (0.316), item no.10 (0.360), and item no.13 (0.370) indicate a high chance of participants answering correctly due to guessing. In contrast, the guessing scores in question 8 and question 12 indicate highly accurate responses.

Table 19.
Coefficient results – informal deductive

Item	Gussng	Dffclt	Dscrmn
B14	0.01075668	0.04954356	20.7110318
B15	0.25939282	0.56938357	14.2025444
B16	0.23510853	0.34652470	6.8036510
B17	0.08509650	-1.23773625	1.0759887
B18	0.03648763	1.02404246	0.9256262
B19	0.03038188	0.90975561	0.4039755
B20	0.01107491	-1.04418810	0.5425780

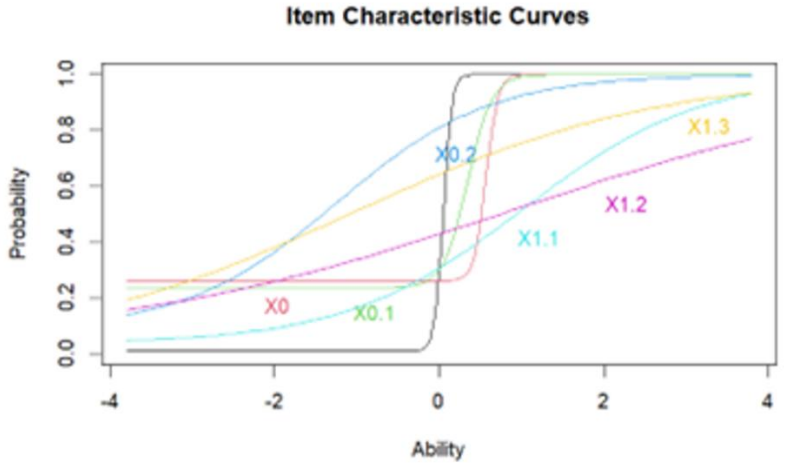

Figure 4.
Informal deductive level curve characteristics.

As a result of the model parameters, item 14 showed relatively high difficulty and discrimination scores, with significant discrimination power (20.71), indicating excellent discrimination ability. In contrast, question item 19 and question 20 had very low cut-off scores, indicating that they had less effective discrimination ability in distinguishing respondents' level of understanding. Item 17 with a negative power difference value (-1.24) also showed inconsistent difficulty. Overall, some items such as question 14 performed well, while other items need to be improved or removed to improve the accuracy of the model.

Table 20.
Coefficient results – formal deductive.

Item	Gussng	Dffclt	Dscrmn
B21	3.017442e-05	-0.19958480	1.645583
B22	2.118699e-01	0.07395137	1.375929
B23	4.633109e-01	0.62273700	14.259567
B24	2.486595e-01	0.38843884	2.323652
B25	2.930987e-01	0.50971896	13.246103
B26	5.487588e-04	-0.21450284	1.868236

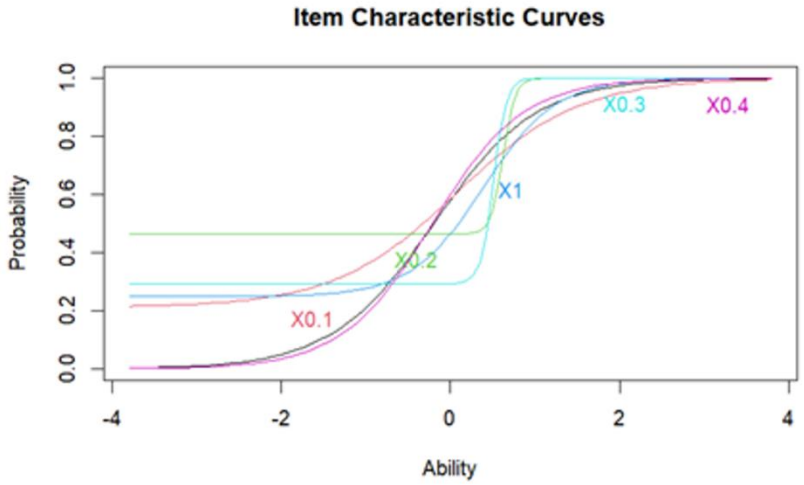

Figure 5.
Formal deductive level curve characteristics.

The results of the IRT analysis show that at level 3, item 23 and item 25 have very high discrimination values (Dscrmn), equal to 14.26 and 13.25, indicating that both items are very effective in differentiating respondents' abilities. In contrast, items such as question 22 and question 26 had lower discrimination scores but were still in the good enough category. Item difficulty scores ranged from negative to positive, indicating a good variety of difficulty levels. Guessing scores were mostly low, indicating that the probability of answering correctly due to guessing was quite small. Overall, these items do a good job of measuring respondents' abilities.

Table 21.
Coefficient results – rigor.

Item	Gussng	Dffclt	Dscrmn
B27	1.154806e-03	-0.58136394	2.15893510
B28	3.361336e-04	-0.38908127	3.58404167
B29	1.445113e-01	17.47947419	0.03330621
B30	1.296425e-01	0.24530802	3.92074766
B31	3.362394e-06	-0.09211068	2.90336199
B32	3.362394e-06	1.91643761	3.11798951

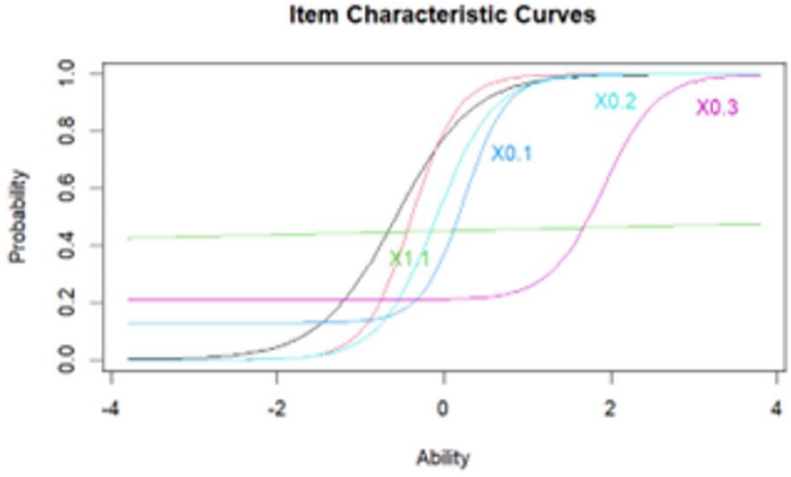

Figure 6.
Rigor level curve characteristics.

Analysis of the item parameters at the rigor level shows that most items have a low guessing rate, indicating the instrument's ability to reduce the chance of correct answers due to guessing. Difficulty scores varied, with item No. 29 being particularly high (17.48), indicating it was too difficult or disproportionate. Meanwhile, the discrimination scores were mostly high, especially in item no. 28 (3.58) and item no. 30 (3.92), indicating that these items were effective in distinguishing between high and low ability respondents. Overall, most of the items performed reasonably well.

5. Discussion

The study initially used 235 respondents from eight classes: 2MAI, 2MA2, 2MA3, 2MA4, 2MA5, 2MA6, 2MB1, 2MB2. These classes were chosen to represent the variations in students' ability to understand the material on a system of

linear equations with two variables. During data tabulation and validation, it was found that most of the initial data didn't meet the criteria for validity and reliability due to incomplete answers, inconsistent response patterns, and a failure to conform to the assumptions of the Item Response Theory (IRT) model used to analyze ability levels.

To strengthen the database and maintain representativeness, the researcher added data from 78 new respondents from classes 2MB3 and 2MB4, bringing the total to 313 data points. After a rigorous screening process to ensure validity and reliability, 139 respondents' data were found suitable for further analysis. This data reduction process is important in IRT model-based research because the accuracy of parameter estimation depends heavily on the quality of the participant responses.

The analysis of the 139 students revealed the distribution of understanding levels based on Van Hiele's theory. The results were as follows: 8 students were at Level 0 (Visualization), 35 students were at Level 1 (Analysis), 12 students were at Level 2 (Informal Deduction), 29 students were at Level 3 (Formal Deduction), 15 students were at Level 4 (Rigor). Additionally, 23 students were in a transition phase between levels, and 16 students couldn't be classified into any level. The fact that the majority of students were at Level 1 shows that most were only able to identify and classify mathematical information based on common characteristics. They weren't yet able to use deductive reasoning or build formal arguments. This suggests that the students' abstract thinking skills were still at an early stage and hadn't developed optimally. Meanwhile, the relatively small number of students at higher levels (formal deduction and rigor) indicates that only a small portion were able to think logically and abstractly in the context of a system of linear equations with two variables.

In addition, there were 23 students who were in a transition phase between levels. This indicates that the process of cognitive development isn't always linear, and some students may be in the process of adapting from one level to the next. Meanwhile, 16 students who couldn't be classified demonstrated unique challenges in their learning process, possibly due to a lack of conceptual understanding, low motivation, or inappropriate answering strategies. Overall, these results highlight the importance of using Van Hiele theory-based diagnostic instruments to accurately identify students' initial abilities. The findings also underscore the urgent need for learning approaches that can gradually develop abstract thinking. These strategies shouldn't just emphasize procedures, but also hone students' conceptual analysis, deduction, and reflection skills. It's also indicated that most freshmen haven't yet achieved the higher-order abstract thinking skills needed to deeply understand the system of linear equations with two variables. This material is crucial as a foundation for understanding advanced topics like linear algebra and economic mathematics. Validity and reliability tests of the items from each level showed that most of the instrument items were valid and reliable enough for use. This indicates that the developed instrument is capable of accurately measuring student understanding.

The three-parameter logistic (3PL) IRT analysis showed that the three-parameter model (discrimination, difficulty, and guessing) was best suited to describe the characteristics of the item. From the fit model test, the 3PL model consistently had the lowest Akaike information criterion (AIC) and Bayesian information criterion (BIC) values and the highest log-likelihood at all levels. The Likelihood Ratio Test (LRT) also confirmed that the differences between the one-parameter (1PL), two-parameter (2PL), and 3PL models were significant ($p < 0.001$), indicating that the 3PL was better able to capture the complexity of student responses.

Item parameter analysis showed high variation among items. Some items had very high discriminating power, such as at the visualization (B4) and analysis (B11) levels, potentially indicating overfitting. Conversely, some other questions, such as B29 and B19, had very low differentiating power, which indicates that they are less effective in distinguishing between students' ability levels. Therefore, improvements and redevelopment of problematic items need to be made so that the instrument can provide more representative and accurate results. This finding supports the need for a learning approach that is able to encourage the gradual development of abstract thinking.

On the other hand, this finding is in line with several studies used as supporting facts in this study. First, the results of this study support article which examines student errors in solving the system of linear equations of two variables [1]. The study showed that students often made conceptual errors, especially when correlating the solution procedure with the right mathematical representation. The same thing is also seen from the results of this study, where most students have not yet reached the formal deductive level and rigor level which represent deeper conceptual understanding and abstraction. Secondly, the results of this study corroborate the view that deep understanding of mathematics does not only depend on procedures, but is also influenced by motivation and structural understanding of concepts [27]. In line with previous study, intrinsic and extrinsic motivation affect students’ mathematics performance, and contextual and sustainable motivational strategies are needed to improve academic achievement in secondary schools [28]. While the results of this study show that students who are at the analysis level tend to only understand basic properties and procedures without being able to develop logical arguments or formal deductions as required in van Hiele's theory which is used as a guideline in the preparation of this research instrument.

The finding that only a few students are at the rigor level is also in accordance with study which emphasizes the importance of conceptual understanding in mathematics learning [2]. They emphasized that conceptual understanding involves students' ability to integrate many mathematical elements in a coherent structure. While the results of validation and reliability of items at each level show that the instrument developed is good enough to measure the ability according to van Hiele's level. This is in line with which states that the 3PL IRT model is very suitable for use in evaluating the ability of students to understand concepts cognitive, because it can map three important aspects, namely discrimination, difficulty, and guessing [19].

The implication is that the consistency between the findings in this study and the literature used as supporting evidence suggests that the challenges in understanding the system of linear equations of two variables are systemic, meaning that

they are not limited to the local context. More constructive pedagogical interventions are needed, such as thinking level-based learning and the use of technology and visual media to support students to move from lower to higher levels.

6. Conclusion

This study shows that most students' understanding of the System of Linear Equations with Two Variables is at the analysis level, according to Van Hiele's theory. This suggests that the abstract thinking ability of early-level students is still developing and has not yet reached the formal or rigorous deductive level. Students still have limitations in understanding the variables, constants, and coefficients within the formulas and linking them to a graph's visualization. The diagnostic instrument developed for this study was valid and reliable, and the 3PL IRT model provided the best analysis results. These findings indicate a need for learning that emphasizes the gradual development of abstract thinking. This will help students understand concepts more deeply, systematically, and applicatively in the context of economics and advanced mathematics. Based on these findings, there are several practical suggestions that can be applied in teaching mathematics. First, lecturers need to use a diagnostic test based on Van Hiele's levels of geometric thought to identify students' initial abilities. Second, learning should be designed gradually to encourage the development of thinking levels, for example, through the Problem-Based Learning approach and the use of visual media. Third, it is important to provide cognitive feedback that guides students in constructing deductive arguments. In addition, practice questions should vary according to the level of thinking to facilitate the transition from concrete to abstract thinking. The implementation of this strategy is expected to improve students' conceptual understanding and their readiness to face advanced mathematics material and its applications in economics.

References

[1] T. Santoso, H. L. H. Nafis, and M. Y. Oktama, "Analyzing students' error in problem solving of two-variable linear equation system: A case study of grade eight students of Indonesian junior high school," *International Journal of Learning, Teaching and Educational Research*, vol. 18, no. 11, pp. 283-296, 2019.

[2] F. A. Hidajat, L. D. Haeruman, E. D. Wiraningsih, and D. S. Pambudi, "The effect of digital technology learning based on guided discovery and self-regulated learning strategy on mathematical creativity," *International Journal of Information and Education Technology*, vol. 13, no. 3, pp. 535-543, 2023.

[3] A. L. del Carmen, W. L. del Carmen, J. García-García, and G. Salgado-Beltrán, "Mathematical connections made by preservice mathematics teachers when solving tasks about systems of linear equations," *International Electronic Journal of Mathematics Education*, vol. 19, no. 4, p. em0799, 2024. https://doi.org/10.29333/iejme/15590

[4] I. L. K. Dewi, "Identification of mathematics prospective teachers' conceptual understanding in determining solutions of linear equation systems," *European Journal of Educational Research*, vol. 10, no. 3, pp. 1157-1170, 2021. https://doi.org/10.12973/eu-jer.10.3.1157

[5] V. Hiele, "The van hiele model of thinking in geometry among adolescents," *National Council of Teachers of Mathematics*, 1988.

[6] J. Brzezińska, "Item response theory models in the measurement theory with the use of LTM package in R," *Econometrics. Ekonometria. Advances in Applied Data Analytics*, vol. 22, no. 1, pp. 11-25, 2018. https://doi.org/10.15611/eada.2018.1.01

[7] S. Zhang, J. Du, P. Chen, T. Xin, and F. Chen, "Using procedure based on item response theory to evaluate classification consistency indices in the practice of large-scale assessment," *Frontiers in Psychology*, vol. 8, p. 1676, 2017. https://doi.org/10.3389/fpsyg.2017.01676

[8] K.-L. Yang, S.-C. Fang, and S.-C. Fan, "Development and validation of an instrument for assessing secondary students' transdisciplinary STEM practices," *International Journal of STEM Education*, vol. 12, no. 1, p. 5, 2025. https://doi.org/10.1186/s40594-025-00529-3

[9] S. Senk, D. Thompson, Y. Chen, K. Voogt, and Z. Usiskin, "The van hiele geometry test: History, use, and suggestions for revisions," *University of Chicago School Mathematics Project*, 2022.

[10] Y. H. Chen, S. L. Senk, D. R. Thompson, and K. Voogt, "Examining psychometric properties and level classification of the van hiele geometry test using CTT and CDM frameworks," *Journal of Educational Measurement*, vol. 56, no. 4, pp. 733-756, 2019. https://doi.org/10.1111/jedm.12235

[11] M. Mahfut, E. Yudianto, R. Purnomo, and F. Firmansyah, "Deduction level of undergraduate students' imagination in solving geometrical problem," *Journal of Physics: Conference Series*, vol. 1465, no. 1, p. 012062, 2020. https://doi.org/10.1088/1742-6596/1465/1/012062

[12] R. P. Chalmers, "Mirt: A multidimensional item response theory package for the R environment," *Journal of Statistical Software*, vol. 48, pp. 1-29, 2012. https://doi.org/10.18637/jss.v048.i06

[13] B. Reeve, *Item response theory*. Dordrecht: Springer Netherlands, 2014, pp. 3415-3423. https://doi.org/10.1007/978-94-007-0753-5_1556

[14] J. Brzezińska, "Item response theory models in the measurement theory," *Communications in Statistics - Simulation and Computation*, vol. 49, no. 12, pp. 3299-3313, 2020. https://doi.org/10.1080/03610918.2018.1546399

[15] S. Min and V. Aryadoust, "A systematic review of item response theory in language assessment: Implications for the dimensionality of language ability," *Studies in Educational Evaluation*, vol. 68, p. 100963, 2021. https://doi.org/10.1016/j.stueduc.2020.100963

[16] C. Na, J. Clarke-Midura, J. Shumway, W. van Dijk, and V. R. Lee, "Validating a performance assessment of computational thinking for early childhood using item response theory," *International Journal of Child-Computer Interaction*, vol. 40, p. 100650, 2024. https://doi.org/10.1016/j.ijcci.2024.100650

[17] J. N. Tendeiro, R. R. Meijer, and A. S. M. Niessen, "PerFit: An R package for person-fit analysis in IRT," *Journal of Statistical Software*, vol. 74, pp. 1-27, 2016. https://doi.org/10.18637/jss.v074.i05

[18] Y.-J. Choi and A. Asilkalkan, "R packages for item response theory analysis: Descriptions and features," *Measurement: Interdisciplinary Research and Perspectives,* vol. 17, no. 3, pp. 168-175, 2019. https://doi.org/10.1080/15366367.2019.1586404

[19] S. T. Kalinowski, "A graphical method for displaying the model fit of item response theory trace lines," *Educational and Psychological Measurement,* vol. 79, no. 6, pp. 1064-1074, 2019. https://doi.org/10.1177/0013164419846234

[20] J. C. Martín, A. Maz-Machado, M. J. Madrid, and M. J. Rodríguez-Baiget, "Comprehension of linear systems with two unknowns in secondary education," *International Journal of Evaluation and Research in Education,* 2024. http://doi.org/10.11591/ijere.v13i6.28206

[21] T. I. Sari, N. Aisyah, C. Hiltrimartin, and A. Maria, "Learning design of linear equation system two variables using MEAs approach in eight grade," *AIP Conference Proceedings,* vol. 2468, no. 1, p. 070061, 2022. https://doi.org/10.1063/5.0102863

[22] A. Jupri, D. Usdiyana, and S. M. Gozali, "Teaching and learning processes for pre-service mathematics teachers: The case of systems of equations," *Eurasia Journal of Mathematics, Science and Technology Education,* vol. 20, no. 8, p. em2482, 2024. https://doi.org/10.29333/ejmste/14858

[23] A. Farazmand, *Global encyclopedia of public administration, public policy, and governance*. Cham, Switzerland: Springer Nature, 2023.

[24] E. Yudianto, T. Sugiarti, and D. Trapsilasiwi, "The identification of van Hiele level students on the topic of space analytic geometry," *Journal of Physics: Conference Series,* vol. 983, no. 1, p. 012078, 2018. https://doi.org/10.1088/1742-6596/983/1/012078

[25] N. H. M. Noor and A. M. Fuzi, "Assessment of validity, reliability, and normality in quantitative study: A survey instrument analysis with IBM SPSS," presented at the 8th ASNet International Multidisciplinary Academic Conference, 2025.

[26] A. Gunawan, M. L. F. Cheong, and J. Poh, "An essential applied statistical analysis course using rstudio with project-based learning for data science," presented at the IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), 2018.

[27] J. G. Walter and J. Hart, "Understanding the complexities of student motivations in mathematics learning," *The Journal of Mathematical Behavior,* vol. 28, no. 2-3, pp. 162-170, 2009. https://doi.org/10.1016/j.jmathb.2009.07.001

[28] A. Walker, N. R. Aguiar, R. N. Soicher, Y.-C. Kuo, and J. Resig, "Exploring the relationship between motivation and academic performance among online and blended learners: A meta-analytic review," *Online Learning,* vol. 28, no. 4, pp. 76-116, 2024. https://doi.org/10.24059/olj.v28i4.4602