ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Vol 12, Issue 02, February 2025

Comparison of Classical Item Characteristics with Several Software in Developing a Mathematics Learning Outcomes Test Instrument

[1] Zulfa Safina Ibrahim*, [2] Jumriani Sultan, [3] Nurul Wahdah

[1] [2] Student, Department of Educational Research and Evaluation, Yogyakarta State University, Yogyakarta, Indonesia
[3] Student, Department of Physic Education, Yogyakarta State University, Yogyakarta, Indonesia
Orchid Id: [1] 0009-0004-0539-4771, [2] 0009-0008-0690-1535, [3] 0000-0002-2173-6063
Corresponding Author Email: [1] zulfasafina.2022@student.uny.ac.id, [2] jumrianisultan.2022@student.uny.ac.id, [3] nurulwahdah.2022@student.uny.ac.id

Abstract*— This study outlines the development steps of a mathematics learning outcomes test and compares item characteristics using AnBuSo, ITEMAN, and RStudio software. Employing the 4D model (define, design, develop, disseminate) in Research & Development (R&D) methodology, the research took place in an East Java 8th-grade junior high school during the 2022/2023 academic year's second semester, focusing on circle material. A small-scale trial involved 33 students. Content validity assessment by 3 experts showed high validity for all items. Exploratory Factor Analysis (EFA) for construct validity revealed 3 factors comprising 7 items. Factor 1 contained 5 variables (items 1, 7, 9, 13, and 15), factor 2 had 1 variable (item 14), and factor 3 had 1 variable (item 12). Comparative analysis of item characteristics among AnBuSo, ITEMAN, and RStudio demonstrated similar and nearly identical values in differentiation and difficulty level, indicating good differentiation and moderate difficulty across all items. Thus, it's inferred that all three software options are equally effective for item characteristic analysis.

Keywords: Anbuso, Iteman, Rstudio, Classical test theory, Mathematics Instrument.

I. INTRODUCTION

The instrument can also be referred to as a tool. In a general sense, a tool is an object or device used to assist someone in performing a task or achieving a goal more effectively and efficiently [1]. Research instruments refer to various tools, objects, devices, signs, or objects used in collecting data. Instruments also encompass various methods, devices, or tools used to gather data, both in qualitative and quantitative research. Furthermore, instruments are used as tools to collect data in social research, such as through the use of questionnaires, interviews, or observations [2], [3], [4], [5].

Mathematics is one of the subjects that plays a significant role in education. According to [6], Mathematics is a discipline focused on abstract concepts and methods used to describe and explain patterns, structures, and relationships in the real world. Proficiency in mathematics is not only essential in everyday life but also serves as a strong foundation for understanding more complex scientific and technical concepts. One of the frequently taught topics in mathematics is the circle. The circle is a highly important geometric shape with numerous real-world applications. Understanding the concept of a circle involves comprehension of radius, diameter, circumference, and area of a circle. A good understanding of this material is crucial in solving mathematical, physical, and engineering problems. [7] in their book provide insights into authentic assessment approaches that can be used in developing more effective mathematics tests, including for circle-related topics. Additionally, [8] in their book offers guidance on contextual mathematics teaching, aiding teachers in designing more effective teaching strategies for circle-related topics.

In the process of learning mathematics, tests become an essential tool for measuring students' understanding of the taught material. Tests in an educational context aim to gain a better understanding of students' abilities, identify their strengths and weaknesses, and support relevant decision-making processes to enhance learning. Mathematics tests on circle-related topics aim to assess students' abilities in applying concepts and formulas related to circles. However, experiences show that students often face difficulties in understanding circle-related materials. Many of them struggle with grasping basic concepts, recognizing the appropriate formulas, and applying them to given problems. Factors such as inadequate concept comprehension, ineffective teaching methods, and insufficient practice can be primary causes of these difficulties.

To address these challenges, it's important to develop mathematics test outcomes in the context of circle-related materials. Through well-constructed test outcomes, accurate information about students' understanding of the material can be obtained. Research on the development of realistic mathematics test instruments for circle-related topics for junior high school students has been conducted by [9], discussing the development of realistic mathematics test instruments focusing on circle-related materials for junior

ISSN (Online) 2456 -1304
International Journal of Science, Engineering and Management (IJSEM)
Vol 12, Issue 02, February 2025

high school students, offering inspiration in designing more engaging and relevant mathematics tests. Research on developing contextually loaded mathematics problems to enhance high-level thinking abilities among senior high school students has also been carried out by [10]. This article offers insights into developing contextually loaded mathematics problems as an alternative method to test students' understanding of circle-related materials.

With a deep understanding of students' difficulties and weaknesses, teachers can devise more effective teaching strategies and design appropriate exercises. This article will discuss the development of mathematics test outcomes related to circle-related materials. Research and development in this regard will provide insights into how to design questions suitable for students' comprehension levels and evaluate their abilities comprehensively. Consequently, it is expected that students' understanding of circle-related materials can be enhanced, leading to improved test outcomes. This research aims not only to explain the steps in developing mathematics learning outcome test instruments but also to provide results related to comparative analysis of item characteristics using software such as AnBuSo, ITEMAN, and RStudio.

II. METHODS AND METHODOLOGY

This research is a developmental study that refers to the steps in composing questions according to [11], namely (1) determining the test objectives, (2) constructing the test blueprint, (3) writing the questions, (4) scrutinizing the questions (review and revision of questions), (5) field-testing the questions, including analysis and improvement, and (6) assembling the questions into a test instrument.

The study was conducted in one of the 8th-grade junior high schools in East Java. A small-scale instrument trial was performed with 33 students. This research took place in the second semester of the academic year 2022/2023, focusing on circle-related materials, which were also part of the curriculum for the second semester of the academic year 2022/2023. The research method employed was the R&D (Research and Development) 4D model (define, design, develop, disseminate).

A reliable instrument is one that consistently measures what is intended to be measured over time [12]. The reliability calculation for this instrument was carried out using Cronbach's Alpha, and the obtained values were based on the following reliability coefficient table.

Table 1. Reliability Coefficient

Reliability Coefficient	Reliability Level
0,80 – 1,00	Very High
0,60 – 0,80	High
0,40 – 0,60	Sufficient
0,20 – 0,40	Low
0,00 – 0,20	Very Low

One of the calculations for the content validity coefficient is by using the V Aiken index to determine whether the developed items are valid or not. Aiken (1985) formulated the Aiken's V formula to calculate the content-validity coefficient based on assessments by a group of experts, n individuals, regarding the extent to which an item represents the measured construct. The V index ranges from 0 to 1, with the formula used as follows.

$$V = \frac{\sum s}{n(c-1)} \quad (1)$$

Where,

V : Expert agreement index regarding item validity

r : The numbers given by the expert

s : $r - l_0$

l_0 : The lowest validity assessment rate (in this case is 1)

n : Number of experts

c : The highest validity assessment rate (in this case is 5)

After being calculated and obtained index v then given conclusions related to the results obtained. The following is a category of validity in the aiken validation index.

Table 2. Category of validity in the aiken validation index

Value	Description
≤ 0.4	Low validity
0.4-0.8	Medium validity
≥ 0.8	High validity

In the test trial results conducted with 33 eighth-grade students from a school in East Java, item analysis was performed on the test instrument to determine the characteristics of item discrimination, item difficulty level, and the effectiveness of distractors. Item analysis calculations were conducted using three software tools, namely AnBuSo, ITEMAN, and RStudio, to compare the results of item analysis among these three software programs.

III. RESULTS AND DISCUSSION

Content Validity

The content validity test was carried out using Equation (1) with 3 experts and obtained results as in the following table.

Table 3. Aiken V Index Calculation Results

Question	Expert 1	Expert 2	Expert 3	S1	S2	S3	Sigma S	V	Info
1	5	5	4	4	4	3	11	0,92	High
2	5	4	4	4	3	3	10	0,83	High
3	5	5	4	4	4	3	11	0,92	High
4	5	5	4	4	4	3	11	0,92	High

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Vol 12, Issue 02, February 2025

Question	Expert 1	Expert 2	Expert 3	S1	S2	S3	Sigma S	V	Info
5	5	5	4	4	4	3	11	0,92	High
6	5	5	4	4	4	3	11	0,92	High
7	5	5	4	4	4	3	11	0,92	High
8	5	4	4	4	3	3	10	0,83	High
9	5	5	4	4	4	3	11	0,92	High
10	5	5	4	4	4	3	11	0,92	High
11	5	4	4	4	3	3	10	0,83	High
12	5	5	4	4	4	3	11	0,92	High
13	5	4	4	4	3	3	10	0,83	High
14	5	4	4	4	3	3	10	0,83	High
15	5	5	4	4	4	3	11	0,92	High

From the 15 items in this instrument, the results indicated that according to the assessments of 3 experts, all items fell into the category of items with high validity because their V values were greater than 0.80, as referenced in the Validity Categories table in the Aiken Validation Index. This demonstrates that the items in this instrument can be used for test trials, confirming their validity for such purposes.

Construct Validity

The construct validity test in this study was conducted using Exploratory Factor Analysis (EFA) following prior tests of correlation and variable adequacy using Bartlett's Test and KMO. Using RStudio software, the Bartlett's Test resulted in a significance value of 0.0003 ≤ Alpha 0.05. This outcome indicates that the variables are correlated with each other and are suitable for further analysis. The KMO analysis yielded a value of 0.6 ≥ 0.5, indicating homogeneity among the variables and suitability for further analysis. Below are the correlation matrix analysis results for the 15 items as shown in Table 4. Several items with KMO values ≤ 0.5 cannot proceed in the process. Items 2, 3, 4, and 5 were excluded from further analysis according to Table 4 due to their KMO values being ≤ 0.5.

Table 4. KMO MSA value for each item.

Item	MSA
1	0.68
2	0.35
3	0.33
4	0.46
5	0.49
6	0.63
7	0.81
8	0.57
9	0.67
10	0.79
11	0.54
12	0.57
13	0.59
14	0.53
15	0.65

Determining the number of factors can be done by looking at the Scree Plot as in Figure 1. The scree plot shows that of the 11 factors that have a correlation, after extraction 3 factors are formed with eigenvalue requirements ≥1.

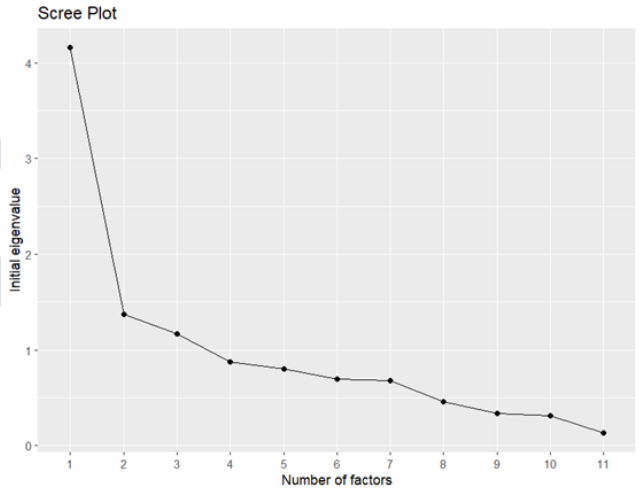

Figure 1. Scree Plot

Following is results EFA analysis with RStudio help before done rotation.

Table 5. Before Rotation

Item	ML1	ML2	ML3
V1	0.49	0.78	0.28
V6	-0.11	-0.4	0.08
V7	0.23	0.49	0.28
V8	0.4	0.22	-0.21
V9	0.59	0.4	0.17
V10	0.44	0.36	0.17
V11	0.06	0.17	-0.11
V12	0.84	-0.03	0.53
V13	0.42	0.45	-0.26
V14	0.83	-0.01	-0.55
V15	0.41	0.5	-0.18

In the EFA analysis prior to rotation, many items couldn't qualify for further analysis as they did not meet the criteria to be included in any component due to their values being < 0.5.

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Vol 12, Issue 02, February 2025

If continued, only 5 items remained: Item 1 in Component 2, Item 9 in Component 1, Item 12 in Component 1, Item 14 in Component 1, and Item 15 in Component 2. None of the items were included in Component 3. Therefore, rotation was necessary, and the results are obtained as presented in Table 6.

Table 6. After Rotation

Item	Component		
	1	2	3
V1	0.93	-0.08	0.11
V6	-0.44	-0.06	0.21
V7	0.59	-0.19	0.1
V8	0.24	0.38	-0.03
V9	0.5	0.14	0.28
V10	0.45	0.05	0.21
V11	0.18	0.11	-0.14
V12	0.06	0.07	0.95
V13	0.5	0.4	-0.18
V14	-0.04	0.98	0.09
V15	0.56	0.31	-0.15

Based on the rotated matrix component results as shown in Table 6, each variable has been grouped into respective factors. Items 6 and 11 exhibit correlation values ≤ 0.5, indicating their low correlation with the formed factors. From this rotation outcome, it is observed that Factor 1 comprises 5 variables (items 1, 7, 9, 13, and 15), Factor 2 includes 1 variable (item 14), and Factor 3 consists of 1 variable (item 12).

Analysis Characteristics Item

After the formation of three factors, an analysis of the characteristics of the items was conducted using a classical approach by comparing three software tools, namely AnBuSo, ITEMAN, and RStudio. A comparison of the discrimination values obtained from the three software tools is presented in Table 7.

Table 7. Comparison of item discrimination parameter values for AnBuSo, ITEMAN, and RStudio

Question	item discrimination					
	AnBuSo		Iteman		RStudio	
	Coefficient	Information	Coefficient	Information	Coefficient	Information
1	0.757	Good	0.769	Good	0.757	Good
7	0.436	Good	0.443	Good	0.436	Good
9	0.685	Good	0.695	Good	0.685	Good
12	0.543	Good	0.552	Good	0.543	Good
13	0.494	Good	0.502	Good	0.494	Good
14	0.477	Good	0.485	Good	0.477	Good
15	0.512	Good	0.520	Good	0.512	Good

From the analysis of item characteristics using AnBuSo, ITEMAN, and RStudio software, it was found that all items exhibited good discrimination ability. This indicates that among the seven items analyzed, all of them were able to effectively differentiate between students with lower abilities and those with higher abilities. The calculated results obtained from the three software programs showed minimal differences. The overall item discrimination for all items ranged from 0.30 to 1.00. The lowest discrimination value was 0.436 for AnBuSo and RStudio, and 0.443 for ITEMAN. On the other hand, the highest discrimination value was 0.757 for both AnBuSo and RStudio, and 0.769 for ITEMAN.

Table 8. Comparison of difficulty level values for AnBuSo, ITEMAN, and RStudio

Question	Difficulty Level					
	AnBuSo		Iteman		RStudio	
	Coefficient	Information	Coefficient	Information	Coefficient	Information
1	0.424	Medium	0.424	Medium	0.424	Medium
7	0.697	Medium	0.697	Medium	0.697	Medium
9	0.636	Medium	0.636	Medium	0.636	Medium

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Vol 12, Issue 02, February 2025

Question	Difficulty Level					
	AnBuSo		Iteman		RStudio	
	Coefficient	Information	Coefficient	Information	Coefficient	Information
12	0.545	Medium	0.545	Medium	0.545	Medium
13	0.576	Medium	0.576	Medium	0.576	Medium
14	0.424	Medium	0.424	Medium	0.424	Medium
15	0.394	Medium	0.394	Medium	0.394	Medium

In addition to obtaining the discrimination values, the analysis of item characteristics resulted in obtaining the difficulty level for each item, as shown in Table 8. In the comparison among the three software tools, all items in this instrument fall into the moderate category as they range between 0.30 and 0.70. Among the three software tools used to analyze the difficulty level of the items, the results from AnBuSo and RStudio showed a remarkable similarity up to three decimal places. Meanwhile, the analysis results from the ITEMAN software exhibited a slight difference of less than 0.1.

Effectiveness of Distractors

Table 9. Effectiveness of Distractors

Answer Distribution					
A	B	C	D	Alternative Answers Are Ineffective	Information
9.1	15.2	33.3	42.4*	-	Good
9.1	12.1	9.1	69.7*	-	Good
63.6*	27.3	3.0	6.1	-	Good
54.5*	6.1	18.2	21.2	-	Good
6.1	21.2	57.6*	15.2	-	Good
27.3	42.4*	21.2	9.1	-	Good
39.4*	27.3	27.3	6.1	-	Good

The distribution of participants' answers can be observed in Table 9. It is evident that all seven items fall into the category of being good. Each selected answer alternative represents at least 2% of the sample used in this instrument's trial. Among the seven items, the answer alternatives offered are effective and do not require any revision as distractors. Hence, it can be concluded that the distractors are working effectively.

Reliability

The reliability calculation for this instrument was performed using RStudio software, yielding a result of 0.771. Therefore, it can be concluded that the reliability of this instrument falls into the high category since the generated value is > 0.70, as per the reliability category in Table 1.

IV. CONCLUSION

The Content Validity Test was conducted with 3 experts, and the result showed that all 15 items were considered as items with high validity according to the experts (V > 0.80). Therefore, the items in this instrument can be used for the instrument's trial. The construct validity test in this study was carried out using Exploratory Factor Analysis (EFA), preceded by testing the correlation and variable suitability using Bartlett's Test and KMO. The Bartlett's Test resulted in a significance value of 0.0003 ≤ Alpha 0.05, and the KMO analysis yielded a value of 0.6 ≥ 0.5, indicating that the variables were correlated and suitable for further analysis. Some items with KMO values ≤ 0.5 could not continue the process; hence, items 2, 3, 4, and 5 were excluded from further analysis.

From the factor analysis, three factors were formed with the condition of eigenvalues ≥ 1. Based on the rotated component matrix results, three factors were formed, each with its members. Items 6 and 11 had correlation values ≤ 0.5, indicating their weak correlation with the formed factors. Factor analysis revealed that Factor 1 comprises 5 variables (items 1, 7, 9, 13, and 15), Factor 2 includes 1 variable (item 14), and Factor 3 consists of 1 variable (Item 12).

The item characteristic analysis using AnBuSo, ITEMAN, and RStudio indicated that all items had good discrimination since the discrimination values ranged between 0.30 and 1.00. In the comparison among the three software tools, the difficulty level for all items in this instrument fell into the moderate category (0.30-0.70). When comparing the results of both discrimination and difficulty level characteristics among the three software tools, all three showed similar and almost identical values. Therefore, it can be concluded that all three software tools are equally suitable for selecting for item characteristic analysis.

Observing that all seven items were categorized as good, it can be concluded that the distractors worked effectively. The reliability calculation for this instrument, done using RStudio software, resulted in a value of 0.771, classifying the instrument's reliability as high due to the generated value exceeding 0.70.This research focused more on comparing item characteristic analyses using three software tools. A suggestion for future research could be emphasizing the refinement of learning objectives (KD) down to the created items.

Acknowledgement

We would like to express our sincere gratitude to everyone who contributed to the completion of this research, especially Zulfa Safina Ibrahim, Jumriani Sultan, and Nurul Wahdah for their hard work on this research. Our deepest appreciation goes to the faculty and staff of the Educational Research and Evaluation program at Universitas Negeri Yogyakarta for their support and encouragement throughout this research. Special thanks also go to our colleagues for their insightful comments and constructive feedback, which greatly enriched this work.

REFERENCES

[1] S. Arikunto, "Dasar-Dasar Evaluasi Pendidikan," Jakarta: Bumi Aksara, 2014.
[2] M. B. Miles and A. M. Huberman, "Qualitative Data Analysis: An Expanded Sourcebook," 2nd ed. Sage Publications, 1994.
[3] L. R. Gay and P. Airasian, "Educational Research: Competencies for Analysis and Applications," 7th ed. Prentice Hall, 2003.
[4] J. W. Creswell, "Research Design: Qualitative, Quantitative, and Mixed Methods Approaches," 4th ed. Sage Publications, 2014.
[5] D. A. De Vaus, "Surveys in Social Research," 6th ed. Routledge, 2014.
[6] L. H. Loomis, "Basic Mathematics for College Students," 4th ed. Cengage Learning, 2014.
[7] R. K. Sembiring and D. Suryadi, "Penilaian Autentik dalam Pembelajaran Matematika," Alfabeta, 2017.
[8] U. T. Munandar, "Pembelajaran Matematika Kontekstual," PT Prestasi Pustakaraya, 2019.
[9] Y. W. Purnomo, "Pengembangan Instrumen Tes Matematika Realistik Materi Lingkaran untuk Siswa SMP," Jurnal Elemen, vol. 4, no. 1, pp. 37-54, 2018.
[10] H. Salim, "Mengembangkan Soal Matematika Bermuatan Kontekstual untuk Meningkatkan Kemampuan Berpikir Tingkat Tinggi Siswa SMA," Jurnal Pendidikan Matematika, vol. 10, no. 2, pp. 29-43, 2016.
[11] Kartowagiran, "Pengukuran dalam Penelitian Pendidikan," Ar-Ruzz Media, 2012.
[12] Sunarti and S. Rahmawati, "Penilaian dalam Kurikulum 2013 Membantu Guru Mengetahui Langkah-langkah Penilaian Pembelajaran," Yogyakarta: CV Andi Offset, 2014.