Bulletin of Electrical Engineering and Informatics
Vol. 14, No. 4, August 2025, pp. 2947~2961
ISSN: 2302-9285, DOI: 10.11591/eei.v14i4.8962

Combination of item response theory and k-means for adaptive assessment

Wargijono Utomo, Waras Kamdi, Eddy Sutadji, Dwi Agus Sudjimat
Department of Vocational Education, Graduate School, State University of Malang, Malang, Indonesia

Article Info

Article history:

Received Jul 5, 2024
Revised Apr 14, 2025
Accepted May 27, 2025

Keywords:

Adaptive assessment
Basic programming
Clustering
Item response theory
K-means

ABSTRACT

This study focuses on developing an adaptive assessment system for basic programming courses using a combination of item response theory (IRT) and the K-mean. The main objective is to enhance the precision of assessments by adapting the difficulty of questions to students' cognitive levels while grouping them based on both cognitive and affective characteristics. The key contribution is the creation of a more personalized assessment framework, addressing the shortcomings of traditional assessments, which often fail to accommodate varying student abilities. Methodologically, the study employs IRT to dynamically assess students' abilities, and students are categorized into different groups based on their answer patterns using K-means. The research design involves a student motivation survey and a programming skills test. Data is collected through the Google Quiz platform and analyzed using R Studio Software to apply the algorithms. The results demonstrate that combining IRT and K-means successfully adjusts the difficulty of questions and more accurately clusters students, providing more relevant feedback. In conclusion, this method enhances adaptive assessments' effectiveness and fosters personalized learning experiences. The findings have implications for broader application in courses with diverse student competencies.

Corresponding Author:

Wargijono Utomo
Department of Vocational Education, Graduate School, State University of Malang
St. Semarang 5, Malang 65145, East Java, Indonesia
Email: wargiono@unkris.ac.id

1. INTRODUCTION

The era of digital education, adaptive assessment has become one of the important tools in evaluating student performance. This assessment is structured in such a way that the degree of difficulty of the questions can be adjusted to the abilities of each individual, thus producing a more precise and personal evaluation [1]-[3]. Courses such as basic programming need this kind of approach because of the variation in students' abilities in understanding concepts and completing practical tasks [4], [5]. However, most of the assessments used today are still conventional and uniform, which are unable to accommodate differences in student ability levels. Item response theory (IRT) has been widely applied in educational assessment to provide more dynamic assessments [6].

IRT focuses on the relationship between students' abilities and the characteristics of the questions they answer, allowing adjustments to the level of difficulty based on individual performance. A study by [7]-[9] showed that IRT can significantly improve the accuracy of student ability evaluation. Conversely, clustering methods like K-mean clustering are commonly employed to group students in accordance with similarities in their cognitive or emotional capabilities, enabling educators to develop more targeted instructional strategies [10]-[12].

However, research combining IRT and K-means in the context of adaptive assessment is still limited, especially for technical courses such as basic programming. Many existing assessment systems have not fully utilized the strengths of both methods simultaneously. For example, research by [13], [14] indicates the great potential of this combination in more personalized and relevant assessments, but its application in the context of technical education still needs further research.

This gap suggests that there is room for further research to explore the effectiveness of this combination approach. To address these issues, this study proposes a combination of IRT and K-means as a solution to create a more comprehensive adaptive assessment. IRT will be utilized to tailor the question difficulty according to students' cognitive abilities, whereas K-means will categorize students based on their cognitive and affective traits [15], [16]. The combination of these two approaches is expected to provide a clearer picture of students' abilities, as well as allow for the provision of questions that are more appropriate to their needs.

The suggested approach not only addresses the need for more adaptive assessment methods but also provides new insights into the use of analytical technologies within the educational field. By combining the strengths of both techniques, this study seeks to create a system that more precisely evaluates students' academic performance while also uncovering the affective factors that impact their learning experience. This is important because students' motivation and emotional engagement in learning play a significant role in their success, especially in demanding courses such as programming.

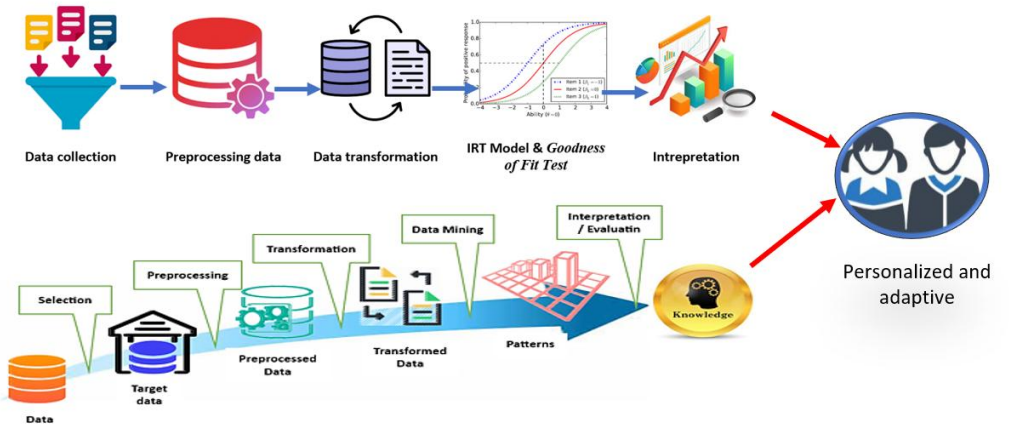
The success of this combination can be seen from the results of initial research which shows that the approach is able to increase the relevance of questions and accelerate the process of identifying student needs. Several studies also show that grouping students based on their cognitive and affective characteristics can help in designing more targeted interventions, which ultimately improve overall learning outcomes [17]. Therefore, the solution presented in this research not only enhances the adaptive assessment framework but also lays the foundation for the creation of more efficient evaluation systems in the future. With a strong theoretical foundation and empirical evidence, this study seeks to introduce innovation in adaptive assessment systems, especially in the context of basic programming. The process of developing, implementing, and evaluating this system will be described in detail in the following sections, which include the methodology, analysis results, and implications for future education.

2. MATERIALS AND METHOD

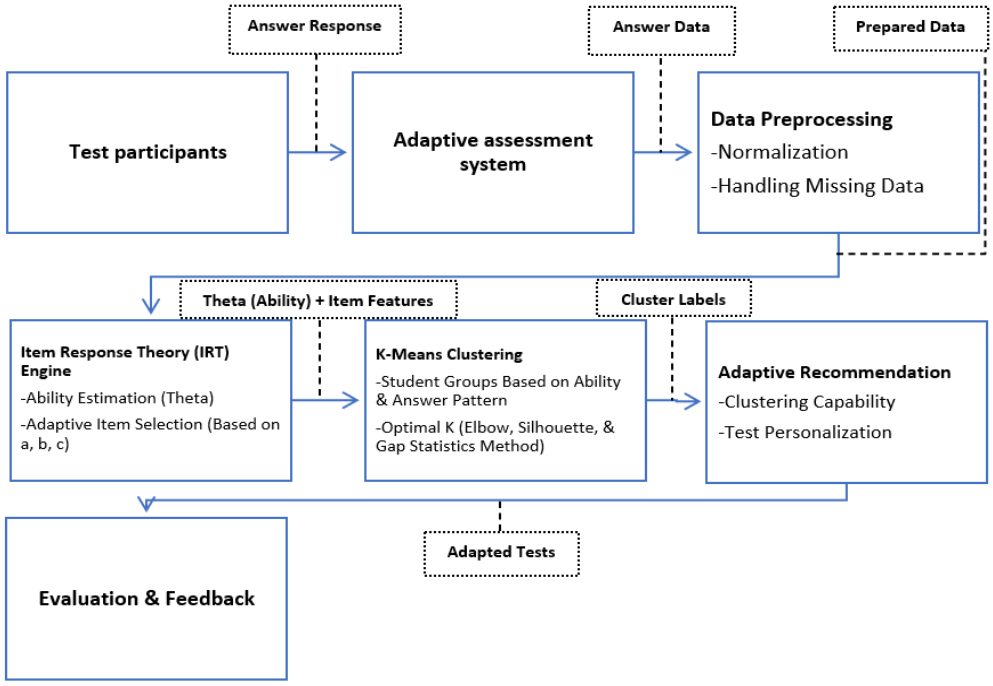
The research materials of this paper consist of several important elements, namely data collection, instrumentation, procedures, and measurements, which can be explained as follows: i) data collection: data is obtained from the results of cognitive and affective assessments of students taking basic programming courses. Data collection is carried out through exam questions and questionnaires distributed using an online platform; ii) instrumentation: exam questions are compiled using the IRT principle with various levels of difficulty, while the questionnaire designed measures affective aspects such as student motivation and involvement in learning. The learning assessment instrument includes affective and cognitive aspects accompanied by criteria, scoring rubrics, and thinking levels which can be seen at the link: https://tau.id/5oray; iii) procedure: procedural steps include distributing exam questions and questionnaires to students, collecting answers, and processing data using statistical software (R Studio) to apply IRT and the K-means algorithm; and iv) measurement: the main measurements in this study are students' cognitive abilities calculated based on exam results, as well as affective factors evaluated through questionnaires. The K-means algorithm is employed to classify students according to their cognitive and affective patterns.

2.1. Adaptive assessment system

This research aims to develop and test an adaptive assessment system in basic programming courses by combining IRT and K-means. The proposed system framework combines two models, namely: first, the IRT model which includes data collection, pre-processing, transformation, IRT model, and goodness of fit test, as well as interpretation; secondly, clustering model using the K-mean algorithm involves several phases: preprocessing, data mining, transformation, interpretation, and evaluation, which can be seen in Figure 1(a). The explanation of the components of the adaptive assessment system architecture is as follows, in Figure 1(b): i) user (student): answers the test items presented; ii) adaptive assessment system: manages the test and selects items adaptively; iii) data preprocessing: cleans and prepares answer data for analysis, iv) IRT engine: calculates ability (theta) and selects items based on maximum information; v) K-means clustering: groups students based on theta and answer patterns with the optimal number of clusters; vi) adaptive recommendation: provides items according to clusters (high/medium/low) to improve assessment accuracy; and vii) evaluation and feedback: presents test results and personalized learning suggestions.



(a)



(b)

Figure 1. The framework and system architecture; (a) the framework of the proposed system and (b) adaptive assessment system architecture

The combination of IRT and K-means method for adaptive assessment offers advantages over traditional adaptive systems. Traditional adaptive systems usually use only IRT to adjust questions based on students' abilities. However, this approach integrates the K-means for student classification according to their abilities levels and response patterns. This approach allows for more accurate question adjustments, as students are grouped into clusters such as high, medium, or low. Each group gets appropriate questions, for example, easy questions for the low group and challenging questions for the high group. The combination of IRT and K-means makes the system more adaptive and helps improve student learning outcomes more effectively.

2.2. Item response theory

The psychometric method known as IRT is used to evaluate how an individual's ability affects the likelihood of giving an accurate answer to a specific item. IRT consists of various models, including 1PL (which focuses on item difficulty), 2PL (which adds item discrimination), and 3PL (which incorporates guessing factors) [18], [19]. IRT is superior to classical test theory because it can take into account individual

and item characteristics more accurately. IRT is applied in education, psychology, and the social sciences to create more precise and adaptive assessments. This study uses one logistic parameter because it adjusts to the small amount of data. Rasch model or 1-parameter logistic (1PL) is used especially to evaluate the level of difficulty of an item in a test or questionnaire (1). The item characteristic curve (ICC) in this model is represented by an equation that illustrates how the likelihood of a respondent answering an item correctly varies with their level of ability [20], [21]. The adaptive assessment process with IRT includes: i) data collection: data was collected from the results of affective and cognitive tests for basic programming courses; ii) data preprocessing: checking data quality, such as missing data or anomalies; iii) data transformation: after the data has been successfully collected and checked, the data is transformed from Excel format into delimited format; iv) IRT model and goodness of fit test: next, estimate parameters using 1PL or Rasch model and test the suitability of the IRT model with R Studio Software; and v) interpretation, as the final step, involves making interpretations and decisions based on the results of data analysis [22].

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1+e^{(\theta - b_i)}} \quad (1)$$

where: $P_i(\theta)$ indicates the likelihood that a test-taker with a certain level of ability θ answers item i correctly, θ indicates the test-taker's degree of ability, b_i is item i's difficulty parameter, a_i represents the discrimination coefficient for item i, e refers to the natural logarithm's base (approximately 2.718), and i range from 1 to n.

2.3. K-means clustering process with knowledge discovery in database

Knowledge discovery in database (KDD) is a process aimed at uncovering valuable and insightful knowledge from large datasets. KDD includes various stages that systematically transform raw data into useful information and knowledge [23], [24]. The following are the main stages in the KDD process: i) selection, the first stage in KDD is selecting relevant data from various data sources. This involves identifying and extracting relevant subsets of data for analysis purposes; ii) preprocessing: after selecting the data, the subsequent step involves preparing and cleaning the data through preprocessing. This stage involves cleaning the data to address problems such as missing data, duplication, or inconsistent data. Preprocessing also includes data normalization and transformation to ensure the data is properly formatted for effective analysis; iii) transformation: data transformation entails converting the data into a structure more appropriate for the data mining procedure. This includes dimensionality reduction, data aggregation, or creating new features. This transformation aims to simplify and improve data quality so that analysis can be carried out more effectively; iv) data mining: this is the central phase of the KDD process, where methods like machine learning and statistical analysis, or data mining are used to uncover patterns, relationships, or meaningful insights from the processed data. In this study, clustering techniques were utilized with the K-means method, supported by the R Studio Software; and v) interpretation/evaluation, once patterns or knowledge are discovered through data mining, the next step is interpretation and evaluation of the results. The results found must be evaluated to ensure that they are valid, useful, and can be interpreted correctly.

2.4. Algorithm K-means

K-mean is a clustering or partitioning technique initially introduced by J. B. MacQueen [25]. It is widely utilized in data mining clustering algorithm, K-mean and pattern recognition due to its simplicity. Known as one of the simplest methods, it primarily relies on the Euclidean distance metric. It is valued for its speed, simplicity, and scalability, making it particularly effective for adaptive systems like cluster-based student assessments [26]. However, choosing the right algorithm depends on the particular characteristics of the dataset. K-means functions by grouping data into a set quantity of clusters based on feature similarities. The algorithm follows these steps: first, define the quantity of clusters (k) and randomly select initial centers of clusters. Second, calculates the distance from the cluster center and each data point. Third, assign the closest cluster is indicated by each data point. Fourth, recalculate the cluster centers and repeat the process from step two to step four until the data points no longer shift between clusters [27]. In this clustering process, identification of data to be grouped is carried out using the Euclidean distance (2):

$$d(\times, y) = \sqrt{\sum_{i=1}^{n} (\times_i - y_i)^2} \quad (2)$$

where $d(\times, y)$ represents the separation between the data instances located at x and y points, $\times_i$ denotes the $\times$ value at the i-th record, y_i refers to the center value of y to the i, th entry, n reflects the overall count of entries.

2.5. Cluster optimization

Clustering outcomes are affected by the chosen quantity of clusters. One of the key challenges is estimating the optimal quantity of clusters beforehand. This can be assessed using techniques like the Elbow technique, Silhouete evaluation, and gap statistics to identify the most appropriate clustering count [28]. The following is a brief explanation of the method:

2.5.1. Elbow

Elbow is an approach to finding the ideal quantity of clusters by analyzing the relationship graph between the quantity of clusters and the resulting variance. In this graph, we look for "Elbow" points where the decrease in variance between clusters is very significant before becoming flatter. This point shows the transition from a steep drop to a flatter drop, indicating the correct quantity of clusters. This method calculates the sum square error (SSE) for each cluster value (k), assisting in determining the ideal number of clusters (3) [29]:

$$SSE(k) = \sum_{i=1}^{n} \sum_{j=1}^{k} \left\| \times_i - \mu_j \right\|^2 \quad (3)$$

where: indicates the overall quantity of data instances, k signifies the quantity of clusters being analyzed, $\times_i$ alludes to the case of data at position i, and μ_j represents the j-th cluster's centroid.

2.5.2. Silhouette

The Silhouette approach was introduced by Rousseeuw *et al.* in 1990 [30], designed to evaluate whether an item i has been appropriately assigned to its cluster. The score for the silhouette every item or point of data i is computed separately, using (4) [31], [32]:

$$S(i) = \frac{b(i) - a(i)}{max\ (a(i), b(i))} \quad (4)$$

whereas $b(i)$ indicates the mean separation between item i and every item in the closest adjacent cluster, $a(i)$ reflects the mean separation between item i and all remaining items within the identical group.

2.5.3. Gap statistics

Gap statistics compare the internal cluster variability of the real data with that of a reference set created from a random distribution. After clustering both the observed and reference datasets using various values of k, the within-cluster dispersion is calculated, and the gap statistic is subsequently derived from these results, using (5) [33], [34]:

$$Gap_n(k) = E_n^* \{log\ (W(k)\} - log\ \big(W(k)\big) \quad (5)$$

where, $W(k)$ is the total variation inside the cluster, E_n^* {.} signifies the anticipated value for a size n dataset extracted from the baseline setup. Gap statistics measure the disparity between the actual $W(k)$ results and their corresponding expected values under the assumption of no distinct cluster structure.

2.6. Cluster evaluation with the Dunn index

The Dunn index serves as a measure that captures the minimum separation between distinct clusters and the maximum compactness within an individual cluster [35]. In simpler terms, better clustering results are indicated by larger distances between clusters and smaller sizes within each cluster [36]. In this context, a higher Dunn value indicates a more effective partitioning or organization of the clusters (6):

$$D = \frac{\{d_c(C_i, C_j)\}\, i \neq j}{\{\Delta(C_l)\} 1 \leq l \leq k} \quad (6)$$

where $d_c\big(C_i, C_j\big) = min\ \{d(x, y)\}$ represents the minimum distance between two clusters, where $x\ is\ a\ member\ of\ cluster\ C_i, and\ y\ is\ a\ member\ of\ cluster\ C_j$:

$$\Delta(C_l) = \{d\ (x, y)\}\ or\ diameter\ of\ C_l x, y\ \in\ C_l$$

$d\ (x, Y)$ indicates distance in Euclidean terms between two data point, while k signifies the overall number of groups or clusters.

3. RESULTS AND DISCUSSION

Data was collected base on the results of affective and cognitive formative tests of basic programming courses totaling 168 informatics students at the Wahana Mandiri Computer Academy in Bekasi, West Java, Indonesia, even semester of the 2022-2023 academic year using a Google form quiz with 20 polytomous questions using a Likert scale (1, 2, 3, and 4) for affective and 40 dichotomous type questions using binary (0, and 1) for cognitive, as seen in Tables 1 and 2.

Table 1. Affective domain polytomous

Id	Name	Item1	Item2	Item3	Item4	...	Item20	Amount	Score
1	Khairil Aslam	3	3	3	3	...	4	34	3.4
2	Dinda Kamelia	3	2	2	2	...	1	34	3.4
...	...	...	...	...	...	...	...	...	...
168	Ahmad Sutisna	2	1	1	2	...	1	26	2.6

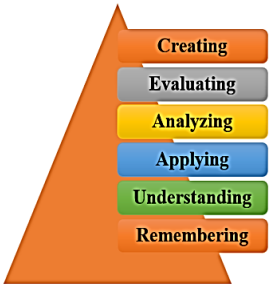
Table 2. Dichotomous of the cognitive domain

Id	Name	Item1	Item2	Item3	Item4	...	Item40	Correct	Score
1	Khairil Aslam	1	1	1	0	...	1	17	8.5
2	Dinda Kamelia	0	1	1	1	...	1	15	7.5
...	...	...	...	...	...	...	...	...	...
168	Ahmad Sutisna	1	0	1	1	...	0	12	6.5

The test question instrument is compiled using the principles of IRT and Bloom's taxonomy in the affective and cognitive domains with various levels of difficulty. The basic programming test competencies include basic concepts of computer programming, such as algorithms, data types, variables, operations, input/output, control structures, functions, procedures, arrays, searching, sorting [37], [38]. Meanwhile, the affective domain includes five levels of expertise according to Krathwohl *et al.* (1964) [39] Figure 2 [40], [41], and the cognitive domain includes six levels of expertise from the revised results of Lorin Anderson and David Krathwohl (2001) [42] can be seen in Figures 3 [43]-[45].

Level	Description	Action Verbs
Receiving	The student shows willingness to listen or engage.	Listen, Attend, Discuss, Follow, Read
Responding	The student reacts and actively participates in activities.	Interpret, Clarify, Contribute, Present, Perform
Valuing	The student attaches personal values and expresses opinions.	Dispute, Reject, Challenge, Justify, Criticize
Organizing	The student reconciles internal conflicts and develops a new value system.	Develop, Formulate, Modify, Reconcile, Compare
Characterizing	The student adopts a consistent belief system and personal philosophy.	Act, Demonstrate, Influence, Resolve, Practice

Figure 2. Affective level competencies



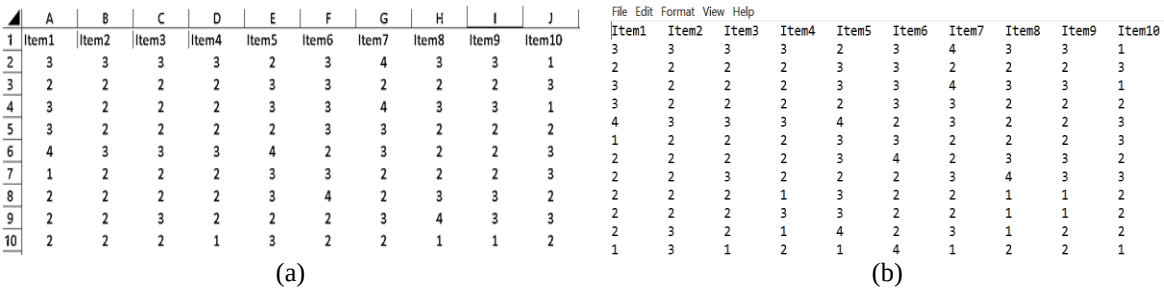
Level	Description	Action Verbs
Remembering	The student recalls facts or information from memory.	State, Define, List, Identify, Recognize, Recall
Understanding	The student explains the meaning or interpretation of a fact in their own words.	Summarize, Rewrite, Illustrate, Describe, Explain, Interpret
Applying	The student uses knowledge to solve real-life problems.	Estimate, Calculate, Construct, Demonstrate, Solve, Prove
Analyzing	The student critically analyzes real-life issues, breaking down complex problems based on known facts and theories.	Analyze, Distinguish, Predict, Investigate, Criticize, Evaluate
Evaluating	The student compares events, critically analyzes parameters, weighs pros and cons, and makes judgments.	Evaluate, Compare, Justify, Defend, Assess, Review
Creating	The student creates new ideas or directions, developing new structures or patterns.	Create, Synthesize, Develop, Generate, Build

Figure 3. Cognitive level competencies based on the 2000 revision

3.1. Item response theory

3.1.1. Estimation of item response theory in the affective domain

The data was collected from the results of filling out the questionnaire on the affective domain of students using the polytomous format utilizes four levels on a Likert scale: not at all (1), occasionally (2), frequently (3), consistently (4) in the first stage, then in the second stage data processing and data transformation were carried out, from excel format into delimited format with the help of R Studio Software, as seen in Figure 4(a) format excel and Figure 4(b) format delimited.



	A	B	C	D	E	F	G	H	I	J
1	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10
2	3	3	3	3	2	3	4	3	3	1
3	2	2	2	2	3	3	2	2	2	3
4	3	2	2	2	3	3	4	3	3	1
5	3	2	2	2	2	3	3	2	2	2
6	4	3	3	3	4	2	3	2	2	3
7	1	2	2	2	3	3	2	2	2	3
8	2	2	2	2	3	4	2	3	3	2
9	2	2	3	2	2	2	3	4	3	3
10	2	2	2	1	3	2	2	1	1	2

(a)

File Edit Format View Help

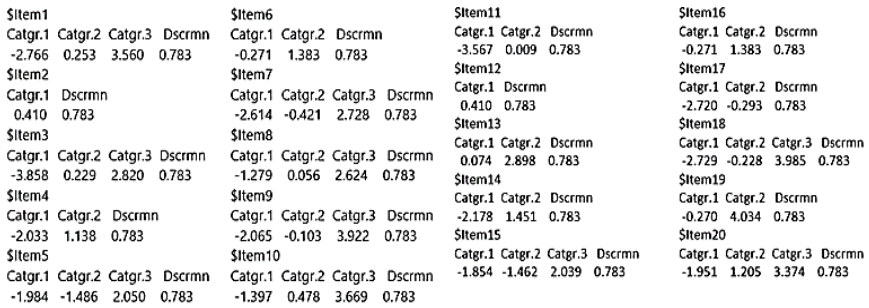
Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10
3	3	3	3	2	3	4	3	3	1
2	2	2	2	3	3	2	2	2	3
3	2	2	2	3	3	4	3	3	1
3	2	2	2	2	3	3	2	2	2
4	3	3	3	4	2	3	2	2	3
1	2	2	2	3	3	2	2	2	3
2	2	2	2	3	4	2	3	3	2
2	2	3	2	2	2	3	4	3	3
2	2	2	1	3	2	2	1	1	2
2	2	2	3	3	2	2	1	1	2
2	3	2	1	4	2	3	1	2	2
1	3	1	2	1	4	1	2	2	1

(b)

Figure 4. Processing and transformation of data; (a) excel format and (b) delimited format

The third stage is calculating IRT: analysis using the logistic 1 parameter model (IRT 1PL) focuses on estimating the difficulty level parameters and discriminating for each test item. In this model, parameter estimation is carried out using the constrained method. The estimation results show that the level of difficulty (values a and b) of each test item varies greatly.

From Figure 5, it is evident that the constrained model's difficulty level parameter ranges from -0.270 to 0.410 (item 19 and item 12), which is in the medium category. However, to be considered a good item, the difficulty level (b) should ideally be in the -2 to +2 range [46]. Meanwhile, the level of discrimination (a) with the constrained model ranges from 0.783, which is included in the discrimination classification index in the very good or excellent category (0.70-1.00).



```
$Item1
Catgr.1 Catgr.2 Catgr.3  Dscrmn
 -2.766   0.253   3.560   0.783
$Item2
Catgr.1  Dscrmn
  0.410   0.783
$Item3
Catgr.1 Catgr.2 Catgr.3 Dscrmn
 -3.858   0.229   2.820   0.783
$Item4
Catgr.1 Catgr.2  Dscrmn
 -2.033   1.138   0.783
$Item5
Catgr.1 Catgr.2 Catgr.3  Dscrmn
 -1.984  -1.486   2.050   0.783
$Item6
Catgr.1 Catgr.2  Dscrmn
 -0.271   1.383   0.783
$Item7
Catgr.1 Catgr.2 Catgr.3  Dscrmn
 -2.614  -0.421   2.728   0.783
$Item8
Catgr.1 Catgr.2 Catgr.3  Dscrmn
 -1.279   0.056   2.624   0.783
$Item9
Catgr.1 Catgr.2 Catgr.3  Dscrmn
 -2.065  -0.103   3.922   0.783
$Item10
Catgr.1 Catgr.2 Catgr.3  Dscrmn
 -1.397   0.478   3.669   0.783
$Item11
Catgr.1 Catgr.2  Dscrmn
 -3.567   0.009   0.783
$Item12
Catgr.1  Dscrmn
  0.410   0.783
$Item13
Catgr.1 Catgr.2  Dscrmn
  0.074   2.898   0.783
$Item14
Catgr.1 Catgr.2  Dscrmn
 -2.178   1.451   0.783
$Item15
Catgr.1 Catgr.2 Catgr.3  Dscrmn
 -1.854  -1.462   2.039   0.783
$Item16
Catgr.1 Catgr.2  Dscrmn
 -0.271   1.383   0.783
$Item17
Catgr.1 Catgr.2  Dscrmn
 -2.720  -0.293   0.783
$Item18
Catgr.1 Catgr.2 Catgr.3  Dscrmn
 -2.729  -0.228   3.985   0.783
$Item19
Catgr.1 Catgr.2  Dscrmn
 -0.270   4.034   0.783
$Item20
Catgr.1 Catgr.2 Catgr.3  Dscrmn
 -1.951   1.205   3.374   0.783
```

Figure 5. Affective constraint estimation results (1PL)

Fourth stage goodness of fit test: choosing the right analysis model is critical to accurate estimation of individual abilities. The suitability between the model and the data is the main benchmark, because an inappropriate model can cause estimation errors. However, no model can perfectly fit the data because each model has limitations. Models are compared according to their complexity and fitting quality employing BIC and AIC [47], [48]. As the AIC and BIC values decrease, the better the model, although there is no exact limit to a "good" value.

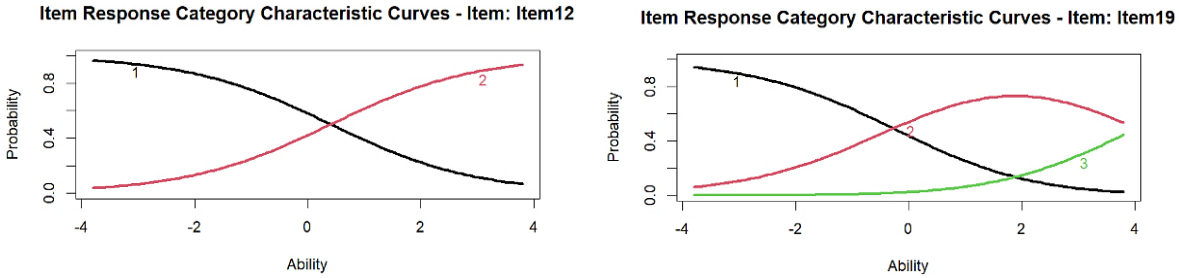
Referring to Figure 6, the outcomes of the likelihood ratio test (LRT) from the ANOVA comparison of different models indicate that the OUT2 model outperforms the OUT1 model. This is indicated by the AIC and BIC values are lower for OUT2 (AIC: 5138.69 and BIC: 5280.68) compared to OUT1 (AIC: 5146.03, and BIC: 5285.13). Smaller AIC and BIC values suggest that the OUT2 model provides more suited to the data. and reduces the amount of information loss while considering the number of parameters and observations. In addition, the significant LRT statistic (p-value=0.002) indicates a significant difference between the two models. Therefore, it can be said that the OUT2 the model is superior in terms of fitting the data than the OUT1 model, making it the best choice of the two models.

```
> anova(OUT1, OUT2)
 Likelihood Ratio Table
       AIC     BIC  log.Lik  LRT df p.value
OUT1 5146.03 5285.13 -2525.02      48
OUT2 5138.69 5280.68 -2520.34 9.35 49   0.002
```

Figure 6. Anova AIC and BIC results

Final stage–model interpretation: the ICC in the partial credit model (PCM) illustrates the link between a test taker's ability level and the likelihood of choosing a specific answer for an item [49], [50]. Interpretation of the ICC involves two main components: i) difficulty: the horizontal position of the ICC indicates the item's degree of difficulty. Items to the left of the ability axis are considered easy, while items to the right are considered difficult and ii) discrimination: the shape and slope of the curve indicate the discrimination of the item. A curve with a high slope indicates good discrimination, because it is able to differentiate participants with significantly different abilities. Conversely, a curve with a low slope indicates low discrimination.

Based on Figure 7(a) item 12 is in the difficult category and Figure 7(b) the estimated difficulty level shows that item 19 is included in the easy category. An individual with level of ability of θ=0.410 has a 50% probability of scoring 0 or above on the item, meanwhile, a person with level of ability of θ=-0.270 has a 50% likelihood of achieving a score of 0 or more on the item. This shows that item 19 is easier for test takers to answer correctly compared to item 12.

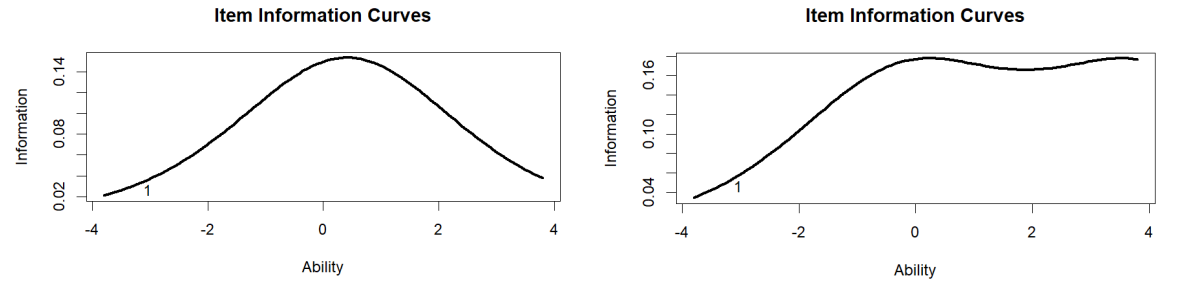


(a) (b)

Figure 7. Characteristic curves; (a) items 12 and (b) item 19

In polytomous models, the amount of information an item contributes depends on its slope parameter; the greater the slope, the more information is provided. A greater distance of location parameters (b1, b2, b3, b4) also increases the amount of information provided. Optimally informative polytomous items have large locations and broad category coverage above theta. The information function is best illustrated by the item information curve, which shows that item information is not static and depends on theta level.

Figure 8(a) item 12, with the highest slope, provides the most statistical information, while Figure 8(b) illustrates that item 19, with the lowest slope, is the least informative. Items tend to provide maximum information in the -2.5 to +1 theta range. The "wavy" curve reflects that item information is a combination of information from each category which is combined to form the item information function.



(a) (b)

Figure 8. Information function of items; (a) 12 and (b) 19

3.1.2. Item response theory estimation in the cognitive domain

The data was collected from the basic programming exam that measures the cognitive domain of students using the Dichotomous type using Binary items if the correct answer is scored 1 but if the answer is wrong the score is 0 in the first stage. The second stage involves data processing and transformation. The data that was originally in Excel format was changed to delimited format with the help of R Studio Software, as seen in Figure 9(a) format excel and Figure 9(b) format delimited.

	A	B	C	D	E	F	G	H	I	J
1	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10
2	1	1	1	0	1	0	1	0	1	1
3	0	1	1	1	1	1	1	1	1	1
4	1	1	1	0	1	0	1	0	1	0
5	1	1	1	1	1	1	1	1	1	0
6	1	1	1	0	1	0	1	0	1	1
7	0	1	1	1	1	1	1	1	1	1
8	1	1	1	0	1	0	1	1	1	1
9	0	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	0	1	0	1	0

(a)

File Edit Format View Help

Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10
1	1	1	0	1	0	1	0	1	1
0	1	1	1	1	1	1	1	1	1
1	1	1	0	1	0	1	0	1	0
1	1	1	1	1	1	1	1	1	0
1	1	1	0	1	0	1	0	1	1
0	1	1	1	1	1	1	1	1	1
1	1	1	0	1	0	1	1	1	1
0	1	1	1	1	1	1	1	1	1
1	1	1	1	1	0	1	0	1	0
1	1	1	0	1	0	1	0	1	0

(b)

Figure 9. Processing and transformation of data; (a) excel format and (b) delimited format

The third stage is calculating IRT: analysis using the logistic 1 parameter model (IRT 1PL) focuses on estimating the difficulty level parameters and discriminating for each test item [51], [52]. In this model, parameter estimation is carried out using the constrained method. The estimation results show that the level of difficulty (values a and b) of each test item varies greatly.

Figure 10 it's visible that the difficulty level parameter for the constrained model ranges from -0.132 to 1.925 (item 31 and item 2), which is in the medium category. However, to be considered a good item, the difficulty level (b) should ideally be in the -2 to +2 range [46]. Meanwhile, the level of discrimination (a) with the constrained model ranges from 1.00, which is included in the discrimination classification index in the negative (non-discriminating) and excellent category (≤0-≥0.40) [53], [54].

```
Dffclt Dscrmn
Item1  -0.3277231   1    Item11 -1.5740909   1    Item21 -2.2764782   1    Item31 -0.1323425   1
Item2   1.9253858   1    Item12 -0.8459230   1    Item22 -3.4016141   1    Item32 -0.6312226   1
Item3  -2.5496394   1    Item13 -1.9815581   1    Item23 -1.4257738   1    Item33 -2.1966599   1
Item4  -1.2444205   1    Item14 -1.2882553   1    Item24 -2.1212412   1    Item34 -1.0372833   1
Item5  -2.0497253   1    Item15 -1.9815699   1    Item25 -2.5497655   1    Item35 -0.3277231   1
Item6  -0.6661006   1    Item16 -1.3788712   1    Item26 -3.0441021   1    Item36  0.5570863   1
Item7  -2.9005737   1    Item17 -1.8540149   1    Item27 -1.6804092   1    Item37 -0.6658263   1
Item8  -1.4257738   1    Item18 -2.2764683   1    Item28 -2.1966564   1    Item38 -1.2444205   1
Item9  -1.5232846   1    Item19 -3.4000760   1    Item29 -1.6264245   1    Item39 -2.0497253   1
Item10 -0.9209828   1    Item20 -2.7718791   1    Item30 -2.1966510   1    Item40 -0.6661006   1
```

Figure 10. Constraint estimation results (1PL)

Fourth stage test of goodness of fit: predicated on Figure 11, In the ANOVA analysis, the OUT2 model shows a significant increase in data explanation compared to the OUT1 model (p-value <0.001). AIC and BIC values are lower for the OUT2 model also indicate its superiority in simplicity and explainability. Therefore, it can be concluded that the OUT2 model is better than the OUT1 model in modeling data.

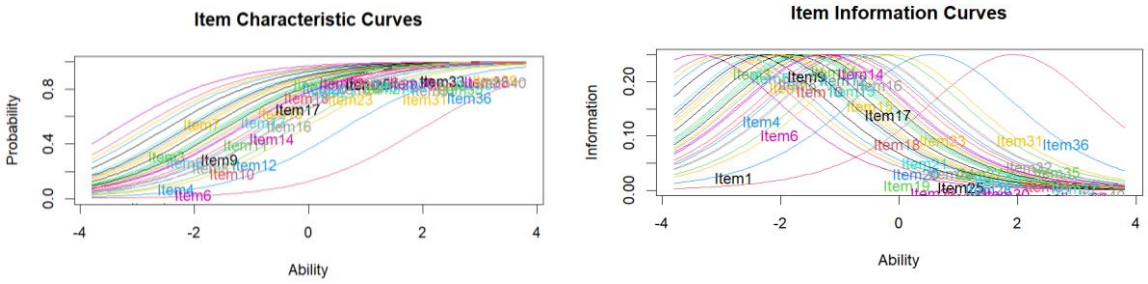
```
> #Fit model
> anova(out1,out2)
 Likelihood Ratio Table
       AIC    BIC  log.Lik    LRT df p.value
out1 5071.18 5187.09 -2495.59
out2 4315.12 4546.95 -2077.56 836.05 40  <0.001
```

Figure 11. Anova AIC and BIC test results

Last stage model interpretation: Figure 12(a) shows how the level of item difficulty affects the likelihood that test-takers will provide accurate answers. for example, the level of difficulty for item 1 is

-0.328, while the difficulty level of item 2 is 1.925. The curve shows that participants with low ability were more likely to answer easy items correctly, whereas participants with high ability were more likely to answer difficult items correctly.

Figure 12(b) shows the unique information function of each item. for example, item 2 provides maximum information when the participant's ability is at θ=1 and remains informative above average (θ=4), but does not provide information below average (θ=-4). This means that item 2 is effective in measuring high ability. In contrast, item 1 provides information on low ability (θ=-4) and is uninformative on high ability (θ=4), thus more effectively distinguishing participants with low ability.



(a) (b)

Figure 12. Curve ICC and IIF; (a) item characteristic curve and (b) item information function test

3.2. K-means clustering

K-means clustering, which was first presented by J. B. MacQueen [25]. It is commonly used in data analysis and pattern recognition because of its straightforwardness and ease of use, particularly with the use of Euclidean distance. Its advantages lie in its speed, simplicity, and ability to handle large amounts of data. Therefore, K-means is suitable for use in adaptive systems such as cluster-based student assessments [26]. The K-means clustering the data utilized in this investigation came from the results of basic programming formative test covering the affective and cognitive domains, as displayed in Tables 1 and 2. Before running the clustering process, a data preprocessing stage was carried out on several attributes used, including no. id, name, and affective and cognitive values. Details of the processed attributes is displayed in Table 3.

3.2.1. Data normalization

In K-means data mining, data standardization is an essential step that adjusts the scale of variables so that they have an average near zero and a spread approximately equal to one as shown in Figure 13 [55]. This normalization process successfully adjusted the scale of each variable, this results in the affective and cognitive data having an average close to zero and a variability approximately equal to one.

Table 3. Affective and cognitive score dataset

Id	Name	Affective	Cognitive
1	Khairil Aslam	3.4	8.5
2	Dinda Kamelia	3.4	7.5
...	...	...	...
168	Ahmad Sutisna	2.6	6.5

```
      Affective    Cognitive
[1,]  0.66318384 -0.19880056
[2,] -0.54784752  1.64094702
[3,]  0.66318384 -2.03854815
[4,] -0.54784752  0.93335180
[5,]  0.96594168 -0.19880056
```

Figure 13. Data normalization

3.2.2. Determination of the number of clusters

With three distinct techniques for determine the quantity of clusters yields various answers, which can be seen in Figure 14(a) Elbow (WSS): suggests 3 clusters, with a significant decrease in WSS after 3 clusters. Figure 14(b) Silhouette: suggests 2 clusters, with the highest silhouette value in the 2 clusters. Figure 14(c) gap statistics: shows 1 cluster, which is not informative. Conclusion: differences in results are caused by the characteristics of the data and respective methods. Researchers chose the Elbow method with 3 clusters because it shows that data variations are well explained without additional clusters.

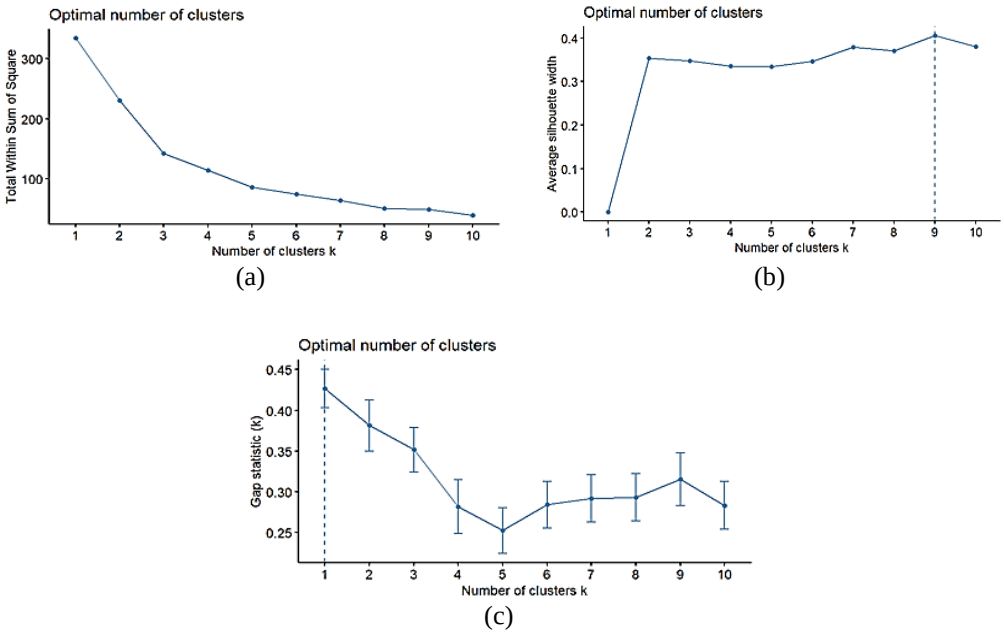


Figure 14. Method; (a) Elbow, (b) Silhouette, and (c) gap statistics

3.2.3. K-means clusterization process

In the K-means clustering analysis with 3 clusters, three groups of observations were formed with sizes of 63, 55, and 50 observations respectively Figure 15(a) visualization, and Figure 15(b) cluster 1: affective -0.4277 and cognitive 0.8525. Cognitive abilities are high, affective abilities are slightly below average. Cluster 2: affective -0.6304 and cognitive 0.7983. Challenges in affective and cognitive aspects, both below average. Cluster 3: affective 1.2324 and cognitive -0.1960. affective response is positive, cognitive understanding is slightly below average. These variations indicate differences in cognitive abilities and affective responses between groups. Educators can use this information to provide specific interventions or learning enrichment to more effectively support the needs of each group.

```
> # Application of K-Means clustering
> final <- kmeans(DATAAFKO_scale, centers= 3, nstart = 25)
> print(final)
K-means clustering with 3 clusters of sizes 63, 55, 50

Cluster means:
  Affective   Cognitive
1 -0.4277055  0.8524838
2 -0.6304178 -0.7983267
3  1.2323686 -0.1959702
```

(a)

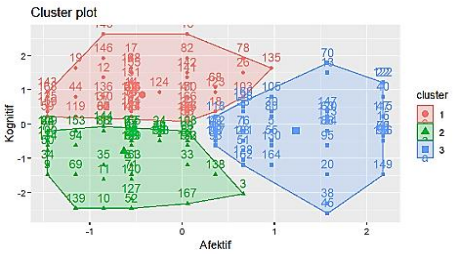


Figure 15. K-means clustering results; (a) three clusters and (b) visualization

Based on the results of student profiling [56], it can be shown in Figure 16(a) cluster 1: very high cognitive ability, fairly good affective, average cognitive 8.24, average affective 2.34, strategy additional challenges through research projects or complex assignments. Cluster 2: high cognitive ability, low affective, average cognitive 7.08, average affective 2.27, strategy: personal guidance, mentoring, interactive activities. Cluster 3: high cognitive ability, medium affective, average cognitive 7.50, average affective 2.89, strategy: motivating and emotionally involving activities, such as group discussions or collaborative projects. Educational institutions can use this information to design adaptive and effective learning strategies, ensuring each student gets the support they need.

Based on the analysis of the Dunn index values provided, it can be concluded that grouping data with three clusters (k=3) is the most optimal configuration. This is indicated in Figure 16(b) by the highest Dunn index value (0.05682544) compared to the grouping of two clusters (0.01707518) and four clusters (0.02415005) [57].

```
> # Group profiling
> DATAAFKO %>%
+   mutate(Cluster = final$cluster) %>%
+   group_by(Cluster) %>%
+   summarise_all("mean")
# A tibble: 3 × 3
  Cluster Affective Cognitive
    <int>    <dbl>     <dbl>
1       1     2.34      8.24
2       2     2.27      7.08
3       3     2.89      7.50
```

(a)

K	Dunn index
2	0.01707518
3	0.05682544
4	0.02415005

(b)

Figure 16. Profiling and Dunn index values; (a) student profiling and (b) Dunn index scores

4. CONCLUSION

This study aims to develop an adaptive assessment system in a basic programming course by combining IRT and the K-mean. The findings of the study indicate that integrating these two methods effectively enhances assessment accuracy by tailoring question difficulty to students' cognitive abilities and grouping them according to their cognitive and affective traits. This achievement is in line with the initial objectives of the study, which focus on improving the effectiveness of assessment and a more personalized learning experience for students. This study's importance stems from its capacity to overcome the shortcomings of traditional assessment systems that often do not consider variations in student abilities. Thus, the developed system not only provides more accurate assessments but also has the ability to enhance student motivation and participation. in the learning process. Future studies are encouraged to investigate the implementation of this adaptive assessment system across different subject areas and within wider educational settings, including online or distance learning environments. Future studies could focus on developing more sophisticated algorithms for student clustering and evaluating the lasting impact of this system on student performance and involvement. Through these efforts, it is hoped that this study can make a greater contribution to innovation in educational assessment systems.

ACKNOWLEDGMENTS

The author would like to express his appreciation and gratitude to Krisnadwipayana University, Bekasi, West Java, for the facilities provided for this research, as well as to the Doctoral Program supervisor at the Postgraduate School, Department of Vocational Education, State University of Malang, East Java, for the support and guidance provided during this research process.

FUNDING INFORMATION

This research did not receive funding from any institution.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Wargijono Utomo	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Waras Kamdi		✓				✓		✓	✓	✓	✓	✓		
Eddy Sutadji	✓		✓	✓			✓			✓	✓		✓	✓
Dwi Agus Sudjimat					✓		✓			✓		✓		✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The author declares that there is no conflict of interest in the preparation and publication of this article, whether financial, personal, or professional.

INFORMED CONSENT

We have obtained approval from stakeholders in the publication of this article, both individuals and institutions.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] A. Brancaccio, D. de Chiusole, and L. Stefanutti, "Algorithms for the adaptive assessment of procedural knowledge and skills," *Behav. Res. Methods*, vol. 55, no. 7, pp. 3929–3951, 2023, doi: 10.3758/s13428-022-01998-y.

[2] S. A. Burr et al., "A narrative review of adaptive testing and its application to medical education," *MedEdPublish*, vol. 13, p. 221, 2023, doi: 10.12688/mep.19844.1.

[3] T. Goto, K. Kano, and T. Shiose, "Students' acceptance on computer-adaptive testing for achievement assessment in Japanese elementary and secondary school," *Front. Educ.*, vol. 8, pp. 1–12, Jul. 2023, doi: 10.3389/feduc.2023.1107341.

[4] S. A. M. Hogenboom, F. F. J. Hermans, and H. L. J. Van der Maas, "Computerized adaptive assessment of understanding of programming concepts in primary school children," *Comput. Sci. Educ.*, vol. 32, no. 4, pp. 418–448, 2022, doi: 10.1080/08993408.2021.1914461.

[5] A. C. M. Yang, B. Flanagan, and H. Ogata, "Adaptive formative assessment system based on computerized adaptive testing and the learning memory cycle for personalized learning," *Comput. Educ. Artif. Intell.*, vol. 3, pp. 1-10, Oct. 2022, doi: 10.1016/j.caeai.2022.100104.

[6] M. Boussakuk, A. Bouchboua, M. El Ghazi, M. El Bekkali, and M. Fattah, "Design of Computerized Adaptive Testing Module into Our Dynamic Adaptive Hypermedia System," *Int. J. Emerg. Technol. Learn.*, vol. 16, no. 18, pp. 113–128, 2021, doi: 10.3991/ijet.v16i18.23841.

[7] L. Xu, Z. Jiang, Y. Han, H. Liang, and J. Ouyang, "Developing Computerized Adaptive Testing for a National Health Professionals Exam: An Attempt from Psychometric Simulations," *Perspect. Med. Educ.*, vol. 12, no. 1, pp. 462–471, 2023, doi: 10.5334/pme.855.

[8] C. Demir and B. F. French, "Applicability and Efficiency of a Computerized Adaptive Test for the Washington Assessment of the Risks and Needs of Students," *Assessment*, vol. 30, no. 1, pp. 238–247, 2023, doi: 10.1177/10731911211047892.

[9] D. Xiong *et al.*, "Development of short forms for screening children's dental caries and urgent treatment needs using item response theory and machine learning methods," *PLoS One*, vol. 19, no. 3, pp. 1–18, 2024, doi: 10.1371/journal.pone.0299947.

[10] R. Liu, "Data Analysis of Educational Evaluation Using K-Means Clustering Method," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/3762431.

[11] R. G. Santosa, Y. Lukito, and A. R. Chrismanto, "Classification and Prediction of Students' GPA Using K-Means Clustering Algorithm to Assist Student Admission Process," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 7, no. 1, pp. 1-10, 2021, doi: 10.20473/jisebi.7.1.1-10.

[12] D. Yu, X. Zhou, Y. Pan, Z. Niu, and H. Sun, "Application of Statistical K-Means Algorithm for University Academic Evaluation," *Entropy*, vol. 24, no. 7, pp. 1–23, 2022, doi: 10.3390/e24071004.

[13] N. I. M. Talib, N. A. Abd Majid, and S. Sahran, "Identification of Student Behavioral Patterns in Higher Education Using K-Means Clustering and Support Vector Machine," *Appl. Sci.*, vol. 13, no. 5, 2023, doi: 10.3390/app13053267.

[14] O. Gonzalez, "Psychometric and Machine Learning Approaches to Reduce the Length of Scales," *Multivariate Behav. Res.*, Aug. 2020, doi: 10.1080/00273171.2020.1781585.

[15] S. Kim, S. Cho, J. Y. Kim, and D. J. Kim, "Statistical Assessment on Student Engagement in Asynchronous Online Learning Using the k-Means Clustering Algorithm," *Sustain.*, vol. 15, no. 3, 2023, doi: 10.3390/su15032049.

[16] G. Feng, M. Fan, and Y. Chen, "Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining," *IEEE Access*, vol. 10, pp. 19558–19571, 2022, doi: 10.1109/ACCESS.2022.3151652.

[17] J. C. A. Sami and U. Arumugam, "A descriptive analysis of students learning skills using bloom's revised taxonomy," *J. Comput. Sci.*, vol. 16, no. 2, pp. 183–193, 2020, doi: 10.3844/JCSSP.2020.183.193.

[18] P. C. Bürkner, "Bayesian Item Response Modeling in R with brms and Stan," *J. Stat. Softw.*, vol. 100, no. 5, 2021, doi: 10.18637/JSS.V100.I05.

[19] A. Gyamfi and R. Acquaye, "Parameters and Models of Item Response Theory (IRT): A Review of Literature," *Acta Educ. Gen.*, vol. 13, no. 3, pp. 68–78, 2023, doi: 10.2478/atd-2023-0022.

[20] S. E. Stemler and A. Naples, "Rasch Measurement v. Item Response Theory: Knowing When to Cross the Line," *Pract. Assessment, Res. Eval.*, vol. 26, pp. 1–16, 2021, doi: 10.7275/v2gd-4441.

[21] K. M. Sattelmayer, K. C. Jagadamma, F. Sattelmayer, R. Hilfiker, and G. Baer, "The assessment of procedural skills in physiotherapy education: a measurement study using the Rasch model," *Arch. Physiother.*, vol. 10, no. 1, pp. 1–11, 2020, doi: 10.1186/s40945-020-00080-0.

[22] Z. Y. Zhuang, C. K. Ho, P. J. B. Tan, J. M. Ying, and J. H. Chen, "The optimal setting of A/B exam papers without item pools: A hybrid approach of IRT and BGP," *Mathematics*, vol. 8, no. 8, pp. 1–29, 2020, doi: 10.3390/MATH8081290.

[23] G. H. Alshammri, A. K. Samha, E. E. D. Hemdan, M. Amoon, and W. El-Shafai, "An Efficient Intrusion Detection Framework in Software-Defined Networking for Cybersecurity Applications," *Comput. Mater. Contin.*, vol. 72, no. 2, pp. 3529–3548, 2022, doi: 10.32604/cmc.2022.025262.

[24] L. Schroeder, M. R. Veronez, E. M. de Souza, D. Brum, L. Gonzaga, and V. F. Rofatto, "Respiratory diseases, malaria and leishmaniasis: Temporal and spatial association with fire occurrences from knowledge discovery and data mining," *Int. J.*

Environ. Res. Public Health, vol. 17, no. 10, pp. 1–23, 2020, doi: 10.3390/ijerph17103718.
[25] I. Zada *et al.*, "Performance Evaluation of Simple K -Mean and Parallel K -Mean Clustering Algorithms: Big Data Business Process Management Concept," *Mob. Inf. Syst.*, vol. 2022, 2022, doi: 10.1155/2022/1277765.
[26] T. Omar, A. Alzahrani, and M. Zohdy, "Clustering Approach for Analyzing the Student's Efficiency and Performance Based on Data," *J. Data Anal. Inf. Process.*, vol. 08, no. 03, pp. 171–182, 2020, doi: 10.4236/jdaip.2020.83010.
[27] R. Vankayalapati, K. B. Ghutugade, R. Vannapuram, and B. P. S. Prasanna, "K-means algorithm for clustering of learners performance levels using machine learning techniques," *Rev. d'Intelligence Artif.,* vol. 35, no. 1, pp. 99–104, 2021, doi: 10.18280/ria.350112.
[28] P. Patel, B. Sivaiah, and R. Patel, "Approaches for finding Optimal Number of Clusters using K-Means and Agglomerative Hierarchical Clustering Techniques," in *2022 Int. Conf. Intell. Controll. Comput. Smart Power*, 2022, pp. 1–6, doi: 10.1109/ICICCSP53532.2022.9862439.
[29] A. Et-Taleby, M. Boussetta, and M. Benslimane, "Faults detection for photovoltaic field based on k-means, elbow, and average silhouette techniques through the segmentation of a thermal image," *Int. J. Photoenergy*, 2020, doi: 10.1155/2020/6617597.
[30] J. Raymaekers and P. J. Rousseeuw, "Silhouettes and Quasi R.esidual Plots for Neural Nets and Tree-based Classifiers," *J. Comput. Graph. Stat.*, vol. 31, no. 4, pp. 1332–1343, 2022, doi: 10.1080/10618600.2022.2050249.
[31] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Student Engagement Level in an e-Learning Environment: Clustering Using K-means," *Am. J. Distance Educ.*, vol. 34, no. 2, pp. 137–156, 2020, doi: 10.1080/08923647.2020.1696140.
[32] S. Özarpacl, B. Klllç, O. C. Bayrak, A. Özdemir, Y. Yilmaz, and M. Floyd, "Comparative analysis of the optimum cluster number determination algorithms in clustering GPS velocities," *Geophys. J. Int.*, vol. 232, no. 1, pp. 70–80, 2023, doi: 10.1093/gji/ggac326.
[33] C. Liu, X. Liu, and W. Li, "Design and Application of Aerospace Accelerometer Testing System using Gap Statistics Based K-Means Clustering Method," in *2024 Int. Conf. Intell. Algorithms Comput. Intell. Syst.*, 2024, pp. 1–4, doi: 10.1109/IACIS61494.2024.10721751.
[34] I. K. Khan *et al.*, "Determining the optimal number of clusters by Enhanced Gap Statistic in K-mean algorithm," *Egypt. Informatics J.*, vol. 27, pp. 1-14, Jul. 2024, doi: 10.1016/j.eij.2024.100504.
[35] R. Holloway *et al.*, "Optimal location selection for a distributed hybrid renewable energy system in rural Western Australia: A data mining approach," *Energy Strateg. Rev.*, vol. 50, pp. 1-13, Jun. 2023, doi: 10.1016/j.esr.2023.101205.
[36] H. N. Abdulrazzak, G. C. Hock, N. A. Mohamed Radzi, N. M. L. Tan, and C. F. Kwong, "Modeling and Analysis of New Hybrid Clustering Technique for Vehicular Ad Hoc Network," *Mathematics*, vol. 10, no. 24, 2022, doi: 10.3390/math10244720.
[37] N. Kiesler, "Towards a Competence Model for the Novice Programmer Using Bloom's Revised Taxonomy - An Empirical Approach," in *Proc. 2020 ACM Conf. Innov. Technol. Comput. Sci. Educ.*, 2020, doi: 10.1145/3341525.3387419.
[38] Z. Ullah, A. Lajis, M. Jamjoom, A. Altalhi, and F. Saleem, "Bloom's taxonomy: A beneficial tool for learning and assessing students' competency levels in computer programming using empirical analysis," *Comput. Appl. Eng. Educ.*, vol. 28, no. 6, pp. 1628–1640, 2020, doi: 10.1002/cae.22339.
[39] K. S. Astuti, M. Belly, R. Maulana, and A. Armini, "Differences in Affective Domain Development Music Learning between Indonesia, The Netherlands, and France," *Harmon. J. Arts Res. Educ.*, vol. 24, no. 1, pp. 62–76, 2024, doi: 10.15294/harmonia.v24i1.44034.
[40] E. Mutlu, B. B. Altunbaş, and S. Kambur, "Taxonomic Investigation of Affective Domain Objectives in the Life Science Curriculum," *Hacettepe Egit. Derg.*, vol. 37, no. 1, pp. 188–203, 2022, doi: 10.16986/HUJE.2020063462.
[41] R. G. Sable and K. D. Bhatt, "NEP 2020: Linking Emotional Intelligence and Bloom's Affective Domain Categories to New Pedagogical and Curricular Structure," *Int. J. Prof. Bus. Rev.*, vol. 8, no. 7, 2023, doi: 10.26668/businessreview/2023.v8i7.3038.
[42] H. Amin and M. S. Mirza, "Comparative study of knowledge and use of Bloom's digital taxonomy by teachers and students in virtual and conventional universities," *Asian Assoc. Open Univ. J.*, vol. 15, no. 2, pp. 223–238, 2020, doi: 10.1108/AAOUJ-01-2020-0005.
[43] Y. Cheng, Y. Cai, H. Chen, Z. Cai, G. Wu, and J. Huang, "A Cognitive Level Evaluation Method Based on a Deep Neural Network for Online Learning: From a Bloom's Taxonomy of Cognition Objectives Perspective," *Front. Psychol.*, vol. 12, pp. 1–15, Oct. 2021, doi: 10.3389/fpsyg.2021.661235.
[44] T. Muhayimana, L. Kwizera, and M. R. Nyirahabimana, "Using Bloom's taxonomy to evaluate the cognitive levels of Primary Leaving English Exam questions in Rwandan schools," *Curric. Perspect.*, vol. 42, no. 1, pp. 51–63, 2022, doi: 10.1007/s41297-021-00156-2.
[45] R. Prakash and R. Litoriya, "Pedagogical Transformation of Bloom Taxonomy's LOTs into HOTs: An Investigation in Context with IT Education," *Wirel. Pers. Commun.*, vol. 122, no. 1, pp. 725–736, 2022, doi: 10.1007/s11277-021-08921-2.
[46] S. Rakkapao, S. Prasitpong, and K. Arayathanitkul, "Analysis test of understanding of vectors with the three-parameter logistic model of item response theory and item response curves technique," *Phys. Rev. Phys. Educ. Res.*, vol. 12, no. 2, pp. 1–10, 2016, doi: 10.1103/PhysRevPhysEducRes.12.020135.
[47] J. Kasali and A. A. Adeyemi, "Model-Data Fit using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and The Sample-Size-Adjusted BIC," *Sq. J. Math. Math. Educ.*, vol. 4, no. 1, pp. 43–51, 2022, doi: 10.21580/square.2022.4.1.11297.
[48] J. Zhang, Y. Yang, and J. Ding, "Information criteria for model selection," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 15, no. 5, 2023, doi: 10.1002/wics.1607.
[49] R. M. van der Lans *et al.*, "Student Perceptions of Teaching Quality in Five Countries: A Partial Credit Model Approach to Assess Measurement Invariance," *SAGE Open*, vol. 11, no. 3, 2021, doi: 10.1177/21582440211040121.
[50] S. Salimpour, R. Tytler, B. Doig, M. T. Fitzgerald, and U. Eriksson, "Conceptualising the Cosmos: Development and Validation of the Cosmology Concept Inventory for High School," *Int. J. Sci. Math. Educ.*, vol. 21, no. 1, pp. 251–275, 2023, doi: 10.1007/s10763-022-10252-y.
[51] L. Maldonado-Murciano, H. M. Pontes, M. Barrios, J. Gómez-Benito, and G. Guilera, "Psychometric Validation of the Spanish Gaming Disorder Test (GDT): Item Response Theory and Measurement Invariance Analysis," *Int. J. Ment. Health Addict.*, vol. 21, no. 3, pp. 1973–1991, 2023, doi: 10.1007/s11469-021-00704-x.
[52] N. H. C. Lah, Z. Tasir, and N. F. Jumaat, "Applying alternative method to evaluate online problem-solving skill inventory (OPSI) using Rasch model analysis," *Educ. Stud.*, vol. 49, no. 4, pp. 644–666, 2023, doi: 10.1080/03055698.2021.1874310.
[53] S. Soeharto and B. Csapó, "Assessing Indonesian student inductive reasoning: Rasch analysis," *Think. Ski. Creat.*, vol. 46, no. Aug. 2022, doi: 10.1016/j.tsc.2022.101132.
[54] T. Orozco, E. Segal, C. Hinkamp, O. Olaoye, P. Shell, and A. M. Shukla, "Development and validation of an end stage kidney disease awareness survey: Item difficulty and discrimination indices," *PLoS One*, vol. 17, no. 9, pp. 1–12, 2022, doi:

10.1371/journal.pone.0269488.
[55] A. A. Khan, M. S. Bashir, A. Batool, M. S. Raza, and M. A. Bashir, "K-Means Centroids Initialization Based on Differentiation Between Instances Attributes," *Int. J. Intell. Syst.*, no. 1, 2024, doi: 10.1155/2024/7086878.
[56] M. Orsoni *et al.*, "Preliminary evidence on machine learning approaches for clusterizing students' cognitive profile," *Heliyon*, vol. 9, no. 3, pp. 1-11, 2023, doi: 10.1016/j.heliyon.2023.e14506.
[57] H. Mahmood, T. Mehmood, and L. A. Al-Essa, "Optimizing Clustering Algorithms for Anti-Microbial Evaluation Data: A Majority Score-Based Evaluation of K-Means, Gaussian Mixture Model, and Multivariate T-Distribution Mixtures," *IEEE Access*, vol. 11, no. August, pp. 79793–79800, 2023, doi: 10.1109/ACCESS.2023.3288344.

BIOGRAPHIES OF AUTHORS

Wargijono Utomo is a lecturer in the Information Systems Study Program, Faculty of Engineering, Krisnadwipayana University, Jakarta, Indonesia. He took a Bachelor's degree in Industrial Engineering from Persada Indonesia University YAI Jakarta in 2003, a Bachelor's degree in Informatics Engineering at STMIK Triguna in 2015, a Masters in Informatics Engineering from Pamulang University, West Java in 2016 and, in 2021, a student at the Department of Vocational Education, graduate school, State University of Malang, Malang, East Java. His areas of research interest are educational data mining, machine learning, deep learning, and smart education. He can be contacted at email: wargiono@unkris.ac.id.

Waras Kamdi is a lecturer in the Vocational Education Department at the Postgraduate School, State University of Malang and a Professor in the field of learning technology. He can be contacted at email: waras.ft@um.ac.id.

Eddy Sutadji is a lecturer in the Vocational Education Department at the Postgraduate School, State University of Malang and a Professor in the field of educational research and evaluation. He can be contacted at email: eddy.sutadji.ft@um.ac.id.

Dwi Agus Sudjimat is a lecturer in the Vocational Education Department at the Postgraduate School, State University of Malang and a Professor in the field of learning technology. He can be contacted at email: dwi.agus.ft@um.ac.id.