VOLUME 10 Issue 1
JULY 2025

E-ISSN: 2622-3384
P-ISSN: 2527-9939

JURNAL MEdu Mangifera Edu

# Development of test instruments to assess high school students' critical thinking on the digestive system

Zaenal Abidin*, Aghniyah Mawaddah Mahar Azizi Soewondo
Biology Education, Postgraduate Program, Kuningan University, Kuningan, West Java, Indonesia
*Corresponding author: zaenalabidin@uniku.ac.id

## ABSTRACT

*Critical thinking is one of the vital 21st-century skills, but many Indonesian students still show low proficiency in this area. To address this gap, it is essential to develop valid and reliable assessment instruments, particularly in science subjects like biology. This study aimed to develop and validate a test instrument to measure high school students' critical thinking skills on the topic of the digestive system. The research followed the ADDIE development model (Analyze, Design, Develop, Implement, Evaluate) and involved 74 eleventh-grade students. The instrument consisted of 30 multiple-choice items based on six critical thinking indicators by Facione: interpretation, analysis, evaluation, inference, explanation, and self-regulation. Validity and reliability were analyzed using the Rasch model with QUEST software. All items were found to be valid and fit the Rasch model, indicating strong construct validity. The item reliability was categorized as excellent (0.93), while student answer consistency was low (0.59), suggesting limited accuracy or care in responses. Most items (70%) were of moderate difficulty. The students' average critical thinking score was 47.25%, with self-regulation being the strongest indicator (63.51%) and analysis and inference being the weakest (35.13% and 38.64%). These results indicate that, although the instrument is effective in measuring critical thinking, students still encounter challenges, particularly in analytical reasoning. Enhancing learning strategies and assessment tools is crucial for improving students' cognitive development in science education.*

## ARTICLE INFO

**Keywords**
ADDIE, Critical thinking skills, Digestive system, Test instrument

**Received**
February 16, 2025

**Revised**
March 17, 2025

**Accepted**
June 30, 2025

**Published**
July 31, 2025

**How to cite**
Abidin Z., & Soewondo, A. M. M. A. (2025). Development of test instruments to assess high school students' critical thinking on the digestive system. *Jurnal Mangifera Edu*, 10(1), 12-22. https://doi.org/10.31943/mangiferaedu.v10i1.219.

## INTRODUCTION

Nowadays, learning is oriented towards the skills and competencies possessed by teachers and students, where critical thinking skills become one of the four main skills that need to be trained, in addition to collaboration, creativity, and communication (Halimah et al., 2023). Critical thinking skills are classified as cognitive abilities, which are essential indicators of achievement for students. Critical thinking skills enable students to overcome problems they encounter, make decisions, and face challenges in various domains of their lives (Kurniahtunnisa et al., 2024). Kurniahtunnisa et al. (2024) stated that with critical thinking skills, students can provide appropriate responses to various problems, analyze them, organize solution strategies, and apply and evaluate the techniques used.

Critical thinking is a comprehensive approach to thinking that examines topics, content, and problems in depth to enhance the quality of thought. Critical thinking is also used to select and analyze whether something is right or wrong (Wood, 2002). According to the Ministry of Education and Culture (2016), learners in the 21st century must be able to meet the demands of life, including developing critical thinking skills. It is crucial to train critical thinking through learning activities so that students can make decisions in their future lives. Education plays a vital role in developing critical thinking skills. Through education, individuals are equipped to analyze information, evaluate arguments, and make informed, rational decisions. The primary goal of education is to equip students with critical thinking skills, enabling them to become lifelong learners and address various challenges encountered in everyday life (Duron et al., 2006). Students' critical thinking skills can be assessed through indicators of critical thinking. According to Facione (2015), there are six aspects of critical thinking skills, which include: (1) *interpretation*; (2) *analysis*; (3) *evaluation*; (4) *inference*; (5) *explanation*; and (6) *self-regulation*.

Critical thinking is one of the basic skills for facing the complexities of life in the modern era. Hartono et al. (2023) stated that critical thinking skills in the 21st century are fundamental because they help develop students' cognitive abilities needed to analyze and evaluate information to solve problems and make informed decisions. Students' critical thinking skills in Indonesia are still relatively low. This can be seen from the results of PISA 2022, although the PISA results are closely related to students' literacy skills, critical thinking skills can also be analyzed. Rahayuni (2016) stated that there is a relatively strong positive relationship between critical thinking skills and scientific literacy, where the higher a student's critical thinking skills, the higher their scientific literacy scores. Therefore, when literacy skills are low, one of the contributing factors is the students' low critical thinking skills.

The implementation of the learning process is not only related to the design, model, and method during the teaching process but also the assessment process, learning assessment is defined as an assessment activity that occurs in learning to find out / measure whether the goals set out have been achieved or not (Arsi Prabaningtias et al., 2023). In developing assessments, it is crucial to analyze the items to determine their validity and reliability. After that, the questions need to be revised and tested on students to see their further validity (Nawawi & Wijayanti, 2018). Additionally, when developing test instruments, there are challenges in ensuring that the questions developed are valid and meet established criteria. The process of measuring critical thinking is not easy, as it requires tools that can test students' understanding of basic concepts as well as their ability to apply knowledge to complex situations. The critical thinking disposition test in biology, which has to be developed with contextual relevance to biology, is considered more objective in assessing critical thinking tendencies. Moreover, it comprises questions designed to guide individuals in exploring their critical thinking dispositions (Syafitriet et al., 2019). The critical thinking assessment instrument is designed to measure their skills in analyzing, evaluating, and synthesizing related information, as well as their ability to recognize assumptions, link concepts, make evidence-based decisions, and solve problems critically (Bhakti et al, 2023).

The instruments used must meet the criteria of validity and reliability, so that the evaluation results can be relied upon to support meaningful and holistic learning (Huber & Kuncel, 2016). This

research focuses on developing valid and reliable question instruments to obtain accurate data to determine students' ability to think critically.

## METHOD

This research implements the RnD development method by adapting the ADDIE model**.** ADDIE stands for *Analyze*, *Design*, *Develop*, *Implement,* and *Evaluate* (Waruwu, 2024). According to Mariam and Nam (2019), this model is commonly used in the context of developing performance-based learning products. **First**, the *analysis* stage. This stage involves analyzing needs and determining the purpose and scope of developing test instruments. **Second**, the *design* stage. This stage consists of designing the product to be developed. The product design is still conceptual, underpinning the process, such as compiling a grid and test instrument format. In this study, the test instrument was developed in the form of a multiple-choice questionnaire, consisting of 30 items that measure six indicators of critical thinking. **Third**, the *development* stage. This stage involves developing test instruments that are ready for application or testing. At this stage, an instrument was made to measure critical thinking skills. **Fourth**, the *implementation* stage. This stage involves applying the test instrument that has been developed, where it is administered to all XI MIPA class students at MAN 2 Kuningan, totaling 74 students. At this stage, the researcher obtained feedback on the products that were developed and implemented. **Fifth**, the *evaluation* stage. This stage involves evaluating the test instrument by analyzing the test result data. Data analysis techniques are employed to process data into information, facilitating the interpretation of results that are easy to understand. The analysis carried out in this study aims to determine the quality of the instrument used to measure the critical thinking skills of students. The analysis was carried out using Quest Software, which produces results in the form of reliability, validity, difficulty level, and analysis of critical thinking skills. Almizi et al. (2023) stated that Quest provides a comprehensive platform for test and questionnaire analysis, offering data analysts access to the latest advancements in Rasch measurement theory, as well as various traditional analytical methods.

## RESULTS AND DISCUSSION

The critical thinking skills test instrument on digestive system material for class XI SMA consists of 30 items. Based on the items developed, it measures six indicators of critical thinking according to Facione (2015), including: (1) *interpretation*; (2) *analysis*; (3) *evaluation*; (4) *inference*; (5) *explanation*; and (6) *self-regulation*. So that each indicator consists of 5 different items, each item is a complex multiple-choice question that has a score of 1 if answered correctly and 0 if answered incorrectly. The instrument was tested on all XI-grade high school students of the MIPA Department, comprising 74 students. The results of the analysis obtained based on research on the development of critical thinking test instruments at the implementation stage are as follows:

### 1. Analysis of Item *Validity* Estimation / Instrument Item Fit (*Goodness of Fit Test)*

Estimation of item validity or item fit *(goodness-of-fit test)* aims to evaluate the extent to which the items on the test instrument align with the measurement model used. Some categories that assess the model's fit according to Nur *et al.* (2022) include examining the average *Mean* INFIT *Mean Square* (*Mean INFIT MNSQ*) value and its standard deviation. In addition, it can also be done by

looking at the average value of INFIT t (*Mean INFIT t*) and its standard deviation. The value limit on the overall item that is declared fit is if the *Mean of Square* INFIT value is ±1.00 and the standard deviation is 0.00. Based on the analysis results obtained, it is evident that the average/mean value of *INFT MNSQ* falls within the range of 0.85 to 0.121. It can be said that all items are considered valid and suitable for the Rasch model. These results are also in line with the results of research (Hanna & Retnawati, 2022) which states that items that are relevant to the Rasch model and meet the requirements are in the *INFIT MNSQ* value range between 0.77-1.33, if the *INFIT MNSQ* result value is 1.4 then it is said to be invalid. Setyawarno (2017) also noted that *INFIT MNSQ* can be used to compare the determination of each item with the criterion model, where a value in the range of 0.77-1.33 indicates that the item is relevant to the Rasch model. A question is considered valid if it successfully performs its intended function, which is to assess the specific construct it is meant to measure accurately.

On the other hand, an item may be deemed invalid due to several contributing factors. This aligns with the view that test validity is influenced by three main aspects: the quality of the test instrument itself, the procedures used during administration and scoring, and the responses provided by students (Arifin,2016). Anwar (2020) said that test-related factors that can reduce validity include: (1) unclear instructions; (2) the use of complex vocabulary and sentence structures; (3) ambiguous wording that allows for multiple interpretations; (4) an overemphasis on specific question types, making the correct answers easier to guess; (5) inconsistencies in the construction of test items; and (6) predictable answer patterns.

Analisis Butir Soal PG Critical Thinking Sistem Pencernaan

Item Estimates (Thresholds) In input Order
all on all (N = 74 L = 30 Probability Level= .50)

| | ITEM NAME | SCORE | MAXSCR | THRSH 1 | INFT MNSQ | OUTFT MNSQ | INFT t | OUTFT t |
|---|---|---|---|---|---|---|---|---|
| 1 | item 1 | 59 | 74 | -1.60 .30 | 1.02 | 1.10 | .2 | .4 |
| 2 | item 2 | 14 | 74 | 1.37 .31 | 1.03 | 1.12 | .2 | .5 |
| 3 | item 3 | 12 | 74 | 1.56 .33 | 1.14 | 1.34 | .7 | 1.1 |
| 4 | item 4 | 71 | 74 | -3.39 .59 | .92 | .48 | .0 | -.7 |
| 5 | item 5 | 41 | 74 | -.40 .25 | 1.00 | .99 | .0 | .0 |
| 6 | item 6 | 35 | 74 | -.05 .24 | .99 | 1.01 | -.2 | .1 |
| 7 | item 7 | 16 | 74 | 1.20 .29 | .94 | .88 | -.3 | -.4 |
| 8 | item 8 | 22 | 74 | .75 .27 | .95 | 1.02 | -.4 | .1 |
| 9 | item 9 | 21 | 74 | .82 .27 | 1.01 | 1.03 | .1 | .2 |
| 10 | item 10 | 42 | 74 | -.45 .25 | .95 | .94 | -.7 | -.4 |
| 11 | item 11 | 30 | 74 | .24 .25 | 1.06 | 1.06 | .7 | .4 |
| 12 | item 12 | 43 | 74 | -.51 .25 | 1.01 | .99 | .2 | .0 |
| 13 | item 13 | 35 | 74 | -.05 .24 | 1.01 | 1.01 | .2 | .1 |
| 14 | item 14 | 31 | 74 | .18 .25 | .99 | .98 | -.1 | -.1 |
| 15 | item 15 | 4 | 74 | 2.80 .52 | .93 | .64 | .0 | -.5 |
| 16 | item 16 | 48 | 74 | -.81 .25 | .97 | .93 | -.3 | -.4 |
| 17 | item 17 | 26 | 74 | .49 .25 | .94 | .93 | -.6 | -.4 |
| 18 | item 18 | 31 | 74 | .18 .25 | .99 | .97 | -.1 | -.1 |
| 19 | item 19 | 27 | 74 | .42 .25 | 1.08 | 1.07 | .9 | .5 |
| 20 | item 20 | 37 | 74 | -.17 .24 | .88 | .87 | -1.7 | -1.0 |
| 21 | item 21 | 39 | 74 | -.28 .24 | 1.01 | 1.02 | .1 | .2 |
| 22 | item 22 | 15 | 74 | 1.28 .30 | 1.08 | 1.10 | .5 | .5 |
| 23 | item 23 | 18 | 74 | 1.04 .28 | 1.21 | 1.42 | 1.4 | 1.8 |
| 24 | item 24 | 39 | 74 | -.28 .24 | .91 | .89 | -1.3 | -.8 |
| 25 | item 25 | 44 | 74 | -.57 .25 | 1.00 | .97 | .0 | -.1 |
| 26 | item 26 | 57 | 74 | -1.43 .28 | 1.01 | .95 | .1 | -.1 |
| 27 | item 27 | 41 | 74 | -.40 .25 | .94 | .92 | -.8 | -.5 |
| 28 | item 28 | 43 | 74 | -.51 .25 | 1.05 | 1.11 | .8 | .8 |
| 29 | item 29 | 45 | 74 | -.63 .25 | .85 | .85 | -1.9 | -1.0 |
| 30 | item 30 | 48 | 74 | -.81 .25 | 1.10 | 1.14 | 1.1 | .9 |
| Mean | | | | .00 | 1.00 | .99 | .0 | .0 |
| SD | | | | 1.14 | .08 | .17 | .8 | .6 |

**Figure 1**. Item recapitulation results

Based on the data contained in Figure 1, it is clear that all items are relevant and follow the Rasch model analysis. The level of fit of the items can also be seen from the mean value of *INFIT MNSQ* and its standard deviation value. When viewed based on this value, the results of the analysis of critical thinking instrument items on digestive system material *are fit* for the Rasch model, as evidenced by the average *INFIT MNSQ* value and its standard deviation of 1.00 and 0.08, respectively. The results of the analysis are following the results of the study (Satria Mukti & Istiyono, 2018) which states that the results of the data analysis obtained show that the *INFIT Mean of Square* value is 1.00. The standard deviation value is 0.04, indicating that all items on the test are suitable for the Rasch model. The item fit map for each item can be seen in Figure 2, which can be used to determine whether the item fits the Rasch model.

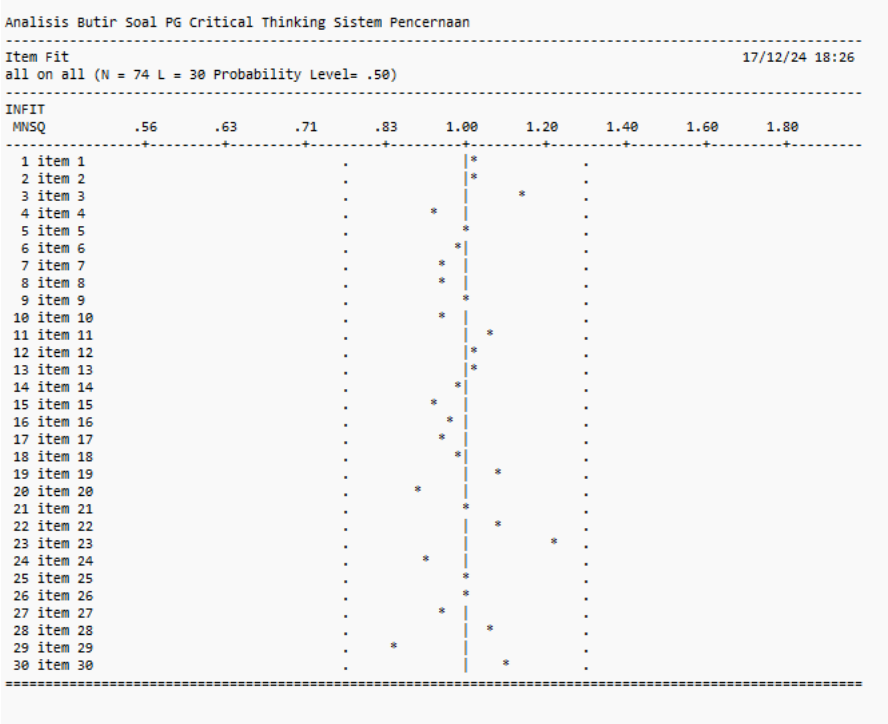


**Figure 2.** Rasch model fit map

Based on the *Fit map* Figure 2, it is evident that the value occupied by the leftmost boundary point is 0.83, while the value of the point on the far right is 1.21. Therefore, all items from 1 to 30 fit the Rasch model because the *INFIT MNSQ* value falls within the range of 0.77-1.33.

**2. Reliability Estimation Analysis**

The value of the Rasch model, as implemented in the QUEST program, is evident in two types of values: the *reliability of item estimates* and the *reliability of case estimates*. Pratama (2020) states that the *reliability of item estimate* values is related to the analysis of the number of question items that fit the model, while the *reliability of case estimate* values shows how consistent the answers given by test takers or students are. Referring to the opinion of Susdelina et al. (2018), the following criteria are listed for the reliability value of the Rasch model: <0.67 weak, 0.67-0.80 moderate, 0.81-0.90 good, 0.91-0.94 excellent, and>0.94 ideal.

Based on the analysis results obtained, it is evident that the *reliability* value of *the item estimate* listed in Figure 3 on the left is 0.93. Based on these results, if interpreted according to the criteria

used, it is classified as having an extraordinary/excellent reliability value because the value falls within the range of 0.91-0.94. Therefore, these results impact the items that *fit* the model, where the higher the reliability, the more items *fit* the model (Hanna & Retnawati, 2022).

| Summary of item Estimates | |
|---|---|
| Mean | .00 |
| SD | 1.14 |
| SD (adjusted) | 1.10 |
| Reliability of estimate | .93 |

| Summary of case Estimates | |
|---|---|
| Mean | -.16 |
| SD | .65 |
| SD (adjusted) | .50 |
| Reliability of estimate | .59 |

**Figure 3.** Reliability value results

Furthermore, the second result is the analysis of *the reliability of case estimate* results, as shown on the right side of Figure 3. The *reliability of the case estimate* result obtained is 0.59, which, when compared to the existing categories, is included in the weak/low category, because the value is <0.67. This *reliability* value relates to the consistency of the answers given by students, so that the *reliability of case estimate* value of 0.59 indicates the inconsistency of students' answers and also implies that they are less careful and careful when answering questions, participants' careless answers can be one of the factors that affect the reliability value to be low (Pratama, 2020). In addition, the number of participants or students can also affect the *reliability of the case estimate* value. This is in line with Purba (2018), who analyzed achievement test instruments using the Rasch model, proving that the number of test participants (<100) affects the reliability value of the test. Whereas in this study, only 74 students were test participants. In addition, based on the research findings of Hakiki et al. (2018), which show that the number of items does not correlate with the test reliability value.

### 3. Analysis of Level of Difficulty Estimation

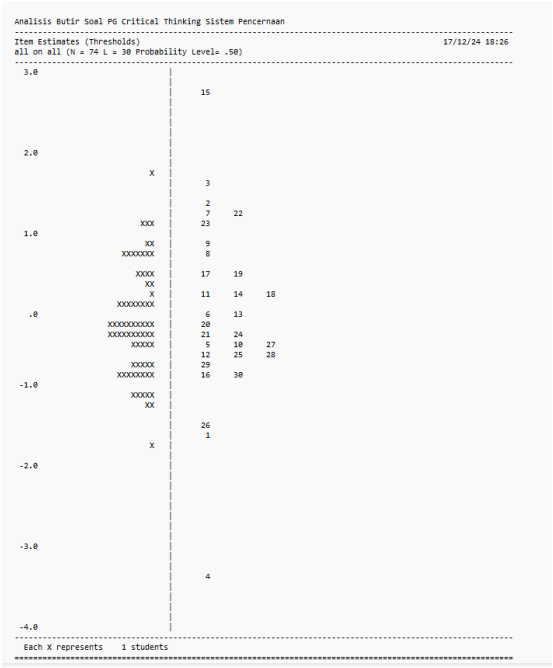


**Figure 4.** Distribution of item difficulty levels

The level of difficulty of the items can be analyzed by examining the results of the item estimation analysis *(Threshold)*. The criteria for categorization, according to Setyawarno (2017), range

from -2.0 to 2.0. The detailed description is as follows: b > 2 (Very Difficult), 1 < b ≤ 2 (Difficult), -1 < b ≤ 1 (Medium), -1 < b ≥ -2 (Easy), b < -2 (Very Easy). The distribution of item difficulty levels is shown in Figure 4 below.

Based on the data distribution of the level of difficulty of the items in Figure 4 above, it can be analyzed that the lower the position of the item number, the easier the level of difficulty, and vice versa, the higher the position of the item number, the more difficult the level of difficulty. Therefore, it can be concluded that question number 4 has a level of difficulty classified as very easy, as it has a *threshold* value of <-3.0, which, when compared to the category used, is <-2.0. Additionally, its position is also at the bottom. Question number 15 has the highest level of difficulty because it has a value *(threshold)* greater than 2, and its position is also at the very top. Table 1 contains a description of the level of difficulty for each item, categorized by the used categories.

**Table1** . The results of the recapitulation of the level of difficulty of the Rasch model items

| Item | *Threshold* value | Difficulty Level | Item | *Threshold* value | Difficulty Level |
|---|---|---|---|---|---|
| 1 | -1.60 | Easy | 16 | -0,81 | Medium |
| 2 | 1.37 | Difficult | 17 | 0,49 | Medium |
| 3 | 1.56 | Difficult | 18 | 0,18 | Medium |
| 4 | -3.39 | Very Easy | 19 | 0,42 | Medium |
| 5 | -0,40 | Medium | 20 | -0,17 | Medium |
| 6 | -0,05 | Medium | 21 | -0,28 | Medium |
| 7 | 1.20 | Difficult | 22 | 1,28 | Difficult |
| 8 | 0,75 | Medium | 23 | 1,04 | Difficult |
| 9 | 0,82 | Medium | 24 | -0,28 | Medium |
| 10 | -0,45 | Medium | 25 | -0,57 | Medium |
| 11 | 0,24 | Medium | 26 | -1,43 | Easy |
| 12 | -0,51 | Medium | 27 | -0,40 | Medium |
| 13 | -0,05 | Medium | 28 | -0,51 | Medium |
| 14 | 0,18 | Medium | 29 | -0,63 | Medium |
| 15 | 2,80 | Very Difficult | 30 | -0,81 | Medium |

Based on the results of the recapitulation of the level of difficulty in Table 1, it can be seen that the questions developed are less varied in terms of difficulty. This is because the questions developed are predominantly in the medium category, with a percentage of 70%. Questions that fall into the difficult category account for 16.67%. Furthermore, the rate of items in the easy category is only 6%, while the percentage in the very difficult and easy categories is only 3.33%. Setyawarno (2017) said that through analysis using the QUEST program, it can analyze the ability of test takers seen in the *Summary of Case Estimate* on the reliability of estimate with the following criteria: > 1.00 (High Ability), -1.00 - 1.00 (Medium Ability), and < -1.00 (Low Ability). Based on the *Summary of Case Estimate, the* value is 0.57, indicating that students' abilities are classified as moderate.

**4. Item Estimation Analysis Passed (*Fit*)**

Based on Rasch analysis using QUEST software, it can determine which items fall or pass based on the OUTFIT t-value in the QUEST program. If the OUTFIT t-value ≤ 2.00, then the item passes; if the OUTFIT t-value≥ 2.00, the item fails (Pratama, 2020). This principle aligns with Setyawarno's (2017) opinion, which states that a question item is successful if the OUTFIT t value is less than or

equal to 2.00, and fails if the OUTFIT t value is greater than or equal to 2.00. In Figure 1, the OUTFIT t value has been displayed. Based on the results of the fit analysis and estimation of passed items, it can be concluded that all items developed have passed and can be used. The fit components of the OUTFIT t values are summarized in Table 2.

**Table2** . Item fit recapitulation

| Item | *OUTFIT t-value* | Description | Item | *OUTFIT t-value* | Description |
|---|---|---|---|---|---|
| 1 | 0,4 | Qualified | 16 | -0,4 | Qualified |
| 2 | 0,5 | Qualified | 17 | -0,4 | Qualified |
| 3 | 1,1 | Qualified | 18 | -0,1 | Qualified |
| 4 | -0,7 | Qualified | 19 | 0,5 | Qualified |
| 5 | 0,0 | Qualified | 20 | -1,0 | Qualified |
| 6 | 0,1 | Qualified | 21 | 0,2 | Qualified |
| 7 | -0,4 | Qualified | 22 | 0,5 | Qualified |
| 8 | 0,1 | Qualified | 23 | 1,8 | Qualified |
| 9 | 0,2 | Qualified | 24 | -0,8 | Qualified |
| 10 | -0,4 | Qualified | 25 | -0,1 | Qualified |
| 11 | 0,4 | Qualified | 26 | -0,1 | Qualified |
| 12 | 0,0 | Qualified | 27 | -0,5 | Qualified |
| 13 | 0,1 | Qualified | 28 | 0,8 | Qualified |
| 14 | -0,1 | Qualified | 29 | -1,0 | Qualified |
| 15 | -0,5 | Qualified | 30 | 0,9 | Qualified |

## 5. Analysis of Critical Thinking Skills Results

Critical thinking, according to Fisher (2008), is the ability to actively interpret and evaluate the results of observation, communication, information, and argumentation. According to Facione (2015), there are six aspects of critical thinking skills, namely: (1) interpretation, (2) analysis, (3) evaluation, (4) inference, (5) explanation, and (6) self-regulation. Based on the analysis of test scores, a figure of 47.25% of students' ability to master critical thinking skills was obtained. Furthermore, it is further analyzed according to the six indicators proposed by Facione (2015), where the percentage interpretations are as follows: analysis, 35.13%; inference, 38.64%; explanation, 46.75%; evaluation, 46.21%; and self-regulation, 63.51%.

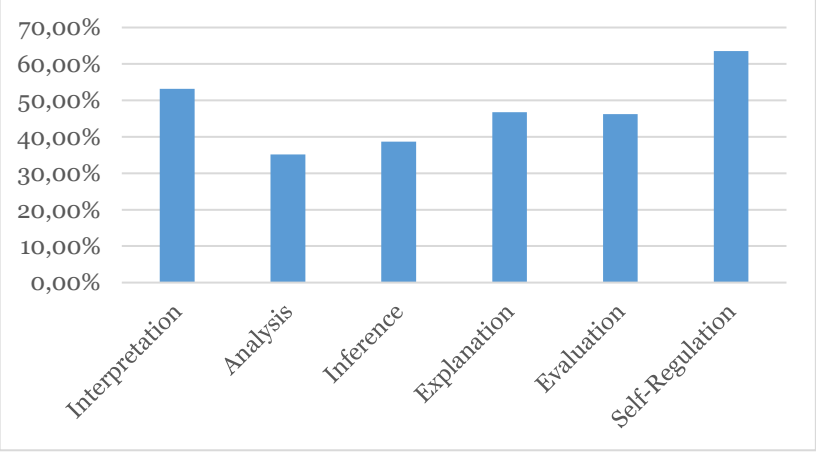


**Figure 5.** Critical thinking skills profile diagram

Based on the analysis of each critical thinking skills indicator, it was found that the highest mastery was observed in the aspect of self-regulation. *Self-regulation* ability refers to the personal awareness that enables one to evaluate their abilities when presenting arguments, drawing conclusions, or making decisions for future self-improvement (Aston, 2023). The sub-skills of the self-regulation indicator include self-assessment and self-evaluation. Students can evaluate a problem and determine actions relevant to the situation at hand (Halimah et al., 2023). The ability of self-regulation is closely related to indicators of evaluation and explanation. This is because Saputri et al. (2018) describe the definition of the explanation indicator as a critical thinking skill that involves the ability to present information convincingly and coherently. This includes explaining methods and results, justifying procedures, and defending concepts or points of view with well-reasoned arguments. Meanwhile, evaluation is the ability to assess the credibility of statements, descriptions, and questions.

The interpretation indicator gets a relatively large number, namely 53.20%. The interpretation indicator is the ability to understand and express the meaning of situations, data, events, or procedures (Halimah et al., 2023). While the indicator values are relatively low, they are related to analysis and inference. Skills involve examining and breaking down information into parts to understand it better, and inference is the ability to draw conclusions based on evidence and reasoning (Halimah et al., 2023).

Mulyani (2022) stated that factors influencing critical thinking skills encompass psychological, sociological, and educational aspects. Psychological factors are related to human thinking modes that require independence in thinking. Sociological factors include pressure to conform, group affiliation, and mapping of surrounding social life experiences. Educational factors encompass the learning methods employed in schools, learning objectives, motivation, willingness to learn, emotions, and parental attitudes and habits.

## CONCLUSION

This study examines the development and evaluation of test instruments to assess the critical thinking skills of high school students concerning material on the digestive system. This study employed the ADDIE model and involved 74 grade XI students, using 30 items based on six critical thinking indicators, as outlined by Facione. The analysis results showed that all items were valid and in accordance with the Rasch model, although the reliability of student answer consistency was low (0.59). Most of the questions had a medium level of difficulty, and students' critical thinking ability, as a whole, was recorded at 47.25%. The self-regulation aspect showed the highest mastery (63.51%), while analysis (35.13%) and inference (38.64%) were the weakest. This study highlights the need for improved teaching methods to enhance students' critical thinking skills. This research makes a significant contribution to the field of science education by providing valuable insights into students' strengths and weaknesses in critical thinking. The findings reveal that self-regulation is the most developed aspect, while analysis and inference remain areas of concern. These findings help educators tailor instructional strategies more effectively. Furthermore, this research emphasizes the importance of assessment not only as a means of evaluation but also as a tool to drive meaningful learning, encouraging the integration of thinking skill measurements in science classrooms to support deeper cognitive development

**REFERENCES**

Almizi, M., Hidayatullah, H., Ikhsanudin, I., & Novaliah, N. (2023). A Practical Using Of The Quest Program To Analyze The Characteristics Of The Test Items In Educational Measurement. *JISAE: Journal of Indonesian Student Assessment and Evaluation*. https://doi.org/10.21009/jisae.v9i1.31163.

Anwar, Yenny., Djunaidah, Zen., Adinda Tiara. (2020). Developing Critical Thinking Skills Assessment of Digestive System for Senior High Schools. *4th Sriwijaya University Learning and Education International Conference (SULE-IC 2020)*. https://doi.org/10.2991/assehr.k.201230.078

Arifin, Zainal. (2016). Evaluasi Pembelajaran Prinsip, Teknik dan Prosedur. Bandung: PT Remaja Rosdakarya. https://books.google.co.id/books?id=V15NMwEACAAJ

Arsi Prabaningtias, D., Arman Madrasah Tsanawiyah Negeri, D., & Article, G. (2023). Analysis of Learners' Ability in End-of-Semester Summative (SAS) Using Item Response Theory (IRT) Approach Assisted by Modular Object-Oriented Dynamic Learning Environment (MOODLE). Journal of Information Technology Education, 1(1), 27-32. http://dx.doi.org/10.30812/upgrade.v1i1.3155

Aston, K. J. (2023). 'Why is this hard, to have critical thinking?' Exploring the factors affecting critical thinking with international higher education students. Active Learning in Higher Education, 1-14. https://doi.org/10.1177/14697874231168341

Bhakti, Y., Arthur, R., & Supriyati, Y. (2023). Development of an assessment instrument for critical thinking skills in Physics: a systematic literature review. Journal of Physics: Conference Series, 2596. https://doi.org/10.1088/1742-6596/2596/1/012067

Duron, R., Limbach, B., & Waugh, W. (2006). Critical thinking framework for any discipline. International Journal of Teaching and Learning in Higher Education 2006, 17(2), https://www.scirp.org/reference/referencespapers?referenceid=2698438

Facione, P. a. (2015). Critical Thinking: What It Is and Why It Counts. In Insight assessment (Issue ISBN 13: 978-1-891557-07-1.). https://www.law.uh.edu/blakely/advocacy-survey/Critical%20Thinking%20Skills.pdf

Fisher, A. (2008). Critical thinking. Jakarta: Erlangga. https://lib.ui.ac.id/detail.jsp?id=20401661

Hakiki, A. W., Fitri, A. R., & Agung, I. M. (2018). Analysis of Psychometric Properties of Merkaufgaben (ME) Subtests with Rasch Model. Journal of Psychology, 14(1), 40. https://doi.org/10.24014/jp.v14i1.4900

Halimah, F., Nurrizki, F., Fadmala, E. I., & Dewi, N. K. (2023). Profile analysis of critical thinking skills of high school students in Biology subject. *National Seminar on Social Science, Education, Humanities (SENASSDRA)*, *2*, 89-96. https://prosiding.unipma.ac.id/index.php/SENASSDRA/article/view/4110

Hanna, W. F., & Retnawati, H. (2022). QUALITY ANALYSIS OF MATHEMATICS ITEMS USING THE RASCH MODEL WITH THE HELP OF QUEST SOFTWARE. AKSIOMA: Journal of Mathematics Education Study Program, 11(4), 3695. https://doi.org/10.24127/ajpm.v11i4.5908

Huber, C.H and Kuncel, N.R. (2016). Does College Teach Critical Thinking? A Meta-Analysis. Review of Educational Research June, Vol. 86, No. 2, pp. 431 - 468. https://doi.org/10.3102/0034654315605917

Kasse, F., & Atmojo, I. R. (2022). Analysis of 21st Century Skills through Science Literacy in Elementary School Students. Journal of Education and Development, 10(1). https://jurnal.arkainstitute.co.id/index.php/educenter/article/download/464/405

Ministry of Education and Culture. (2016). Senior high school/madrasah aliyah (SMA/MA) subject syllabus: biology subject. Retrieved from http://www.syaiflash.com/ rpp2016

Kurniahtunnisa, Wiesje Merry Warouw, Z., & Rukmana, M. (2024). Development Of Critical Thinking Ability Test Instruments On Breathing Systems Science Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Manado 3). Eduproxima: Scientific Journal of Science Education, 448-456. https://doi.org/10.29100/.v6i2.5224

Mulyani, A. Y. (2022). Development of Critical Thinking in Improving the Quality of Education in Indonesia. DIAJAR: Journal of Education and Learning, 1(1), 100-105. https://doi.org/10.54259/diajar.v1i1.226

Nawawi, S., & Wijayanti, T. F. (2018). Development of biology assessment based on critical thinking skills integrated with Islamic values. *Journal of Science Education Innovation*, *4*(2), 136-148. https://doi.org/10.21831/jipi.v4i2.21265

Nur, R. M.A. (2022). Validity and Reliability Analysis Using the Rasch Model to Measure the Quality of Mathematics Test Items of Vocational High Schools. Journal of Educational Research and Evaluation, 11 (1), 103-113. http://dx.doi.org/10.15294/jere.v11i2.58835

Pratama, D. (2020). Analysis of Teacher-made Test Quality Through the Rasch Model Item Response Theory (IRT) Approach. Tarbawy: Journal of Islamic Education, 7(1), 61-70. https://doi.org/10.32923/tarbawy.v7i1.1187

Purba, S. E. D. (2018). Rasch model analysis of achievement test instruments on basic subjects and electrical measurements. Wiyata Dharma: Journal of Educational Research and Evaluation, 6(2), 142. http://dx.doi.org/10.30738/wd.v6i2.3393

Rahayuni, Galuh. (2016). Hubungan Keterampilan Berpikir Kritis Dan Literasi Sains Pada Pembelajaran IPA Terpadu Dengan Model PBM dan STM. Jurnal Penelitian dan Pembelajaran IPA V. https://dx.doi.org/10.30870/jppi.v2i2.926

Saputri, A. C., Sajidan, S., & Rinanto, Y. (2018). Critical thinking skills profile of senior high school students in Biology learning. Journal of Physics: Conference Series, 1006(1). https://doi.org/10.1088/1742-6596/1006/1/012002

Satria Mukti, T., & Istiyono, E. (2018). Instrument for Assessing the Critical Thinking Ability of X Grade High School Students on Biology Learning. BIOEDUKASI: Journal of Biology Education, 11. https://jurnal.uns.ac.id/bioedukasi/article/view/21624

Setyawarno, D. (2017). Use of Application of Software Iteman (Item and Test Analysis) to Analyze Multiple Choice Items Based on Classical Test Theory. State University of Yogyakarta, 1(May). http://dx.doi.org/10.19109/jifp.v1i1.866

Susdelina, Perdana, S. A., & Febrian. (2018). Quality analysis of measurement instruments for understanding the concept of quadratic equations through classical test theory and Rasch model. Journal of Kiprah, VI(1), 41-48. https://doi.org/10.31629/kiprah.v6i1.574

Syahfitri Jayanti, Harry Firman, Sri Redjeki, Siti Sriyati. (2019). Development and Validation of Critical Thinking Disposition Test in Biology. International Journal of Instruction Vol 12, No 4, 381-392. http://dx.doi.org/10.29333/iji.2019.12425a

Waruwu, M. (2024). Research and Development (R&D) Methods: Concepts, Types, Stages and Advantages. Scientific Journal of Education Profession, 9(2), 1220-1230. https://doi.org/10.29303/jipp.v9i2.2141

Wood, R. (2002). Critical thinking. Retrieved https://www.robinwood.com/Democracy/GeneralEssays/CriticalThinking.pdf