Available online at BIOMA: Jurnal Ilmiah Biologi
Websites:http://journal.upgris.ac.id/index.php/bioma/index
BIOMA: Jurnal Ilmiah Biologi, 14 (1), April 2025, 121-141
DOI: https://doi.org/10.26877/bioma.v13i1.1810

CLASS EVALUATION: DEVELOPMENT OF HIGHER ORDER THINKING SKILLS TEST INSTRUMENTS IN DIGESTIVE SYSTEM LEARNING

Iis Sumiati, Neneng Windayani*, Ida Farida, Tri Cahyanto
Master of Science Education, Postgraduae Program, UIN Sunan Gunung Djati Bandung
Jl. Cimencrang, Panyileukan, Gedebage Kota Bandung - Indonesia
**Corresponding author: nenengwinda.ftk@uinsgd.ac.id*

ARTICLE INFO

Article history

Submission	2024-01-28
Revision	2025-03-01
Accepted	2025-04-30

Keywords:
Assessment instrument development
Digestive system
Higher order thinking skills

ABSTRACT

Today's educational context, where critical thinking and deep understanding are key competencies, existing assessment tools are often inadequate in capturing students' higher-order thinking skills (HOTs), especially in science learning. Many of these instruments are limited in scope, lack cognitive alignment, and fail to provide reliable data for instructional improvement. This study aims to develop and validate a test instrument that is meticulously designed to measure HOTs in the digestive system of junior high school students, ensuring a strong cognitive alignment. The instrument was developed using Resnick's (1987) Higher Order Thinking Skills framework and followed the ADDIE model (Analysis, Design, Development, Implementation, Evaluation) within a Research and Development (R&D) design. Furthermore, the instrument demonstrated high internal consistency with a Cronbach's alpha value of 0.878. These results indicate that the test is valid and reliable for assessing students' HOTs. The developed instrument not only provides a meaningful alternative for educators seeking to implement more accurate and cognitively aligned assessments in science classrooms, but also offers practical implications for curriculum development and instructional improvement in Indonesian secondary education.

INTRODUCTION

Existing global challenges have changed learning dynamics in the 21st century (Cahyanto et al., 2023). In 21st-century learning, thinking skills are necessary and the primary foundation for students to face rapid and complex changes in modern life (Rusandi et al., 2024). Higher-order thinking skills (HOTs) allow students to solve complex problems, adapt to rapid change, and face innovative challenges they may have never encountered before (Rahmi et al., 2021). These skills enable students to become

independent thinkers who can explore new ideas and solve problems unconventionally by relying on memorization and optimizing the ability to link and analyze various existing facts (Rohmah et al., 2023). In addition, higher-order thinking skills also enable students to interpret, filter, and manage the information obtained more critically (Ahmad et al., 2018). Higher-order thinking skills support students in developing collaboration, effective communication and leadership skills (Kwangmuang et al., 2021). This is a necessary aspect of working with a team and contributing to a diverse environment.

While the development of higher-order thinking skills is a key goal in education, it is equally important to have effective evaluation tools that can measure students' progress and mastery of these skills. In the ever-evolving landscape of education, it is crucial that the evaluation tools used in the classroom accurately reflect students' abilities to understand complex concepts (Setiawan et al., 2021). (Setiawan et al., 2021). Practical evaluation tools should encompass a variety of assessment forms to gauge students' overall understanding, including their higher-order thinking skills. The HOTS instrument, for instance, plays a significant role in creating a learning environment that fosters problem-solving and adaptability to diverse situations (Lu et al., 2021).

Developing effective test instruments for higher-order thinking skills is not easy. Requires deep knowledge of the material, a strong understanding of how students learn, and good test design skills. Selection of appropriate content, preparation of questions that can encourage students to think critically, as well as objective measurement of desired skills are some of the challenges faced by developers of higher-order thinking skills test instruments (Wijnen et al., 2021), especially in complex material such as the digestive system. The digestive system is a field that involves various concepts, ranging from the physical structure of the digestive organs to complex biochemical processes (Ahİ, 2017). In this context, developing test instruments that include an in-depth understanding of functions, digestive processes, system-related problems, and the student's ability to apply this knowledge in real-life scenarios is essential.

However, several studies have shown that science teachers in Indonesian high schools often face challenges in developing evaluation instruments that align with higher-order thinking skills (Nida et al., 2020; Darong et al., 2024); similar issues are observed in the digestive system topic. The instrument focuses too much on specific facts without measuring deep conceptual understanding (Suhirman, 2024). The instruments created

always fall into the trap of measuring how well students memorize information about the parts of the digestive system without assessing their understanding of the relationships between parts, the processes involved, or the health implications of the system's function (Rohmawati et al., 2025). In other cases, sometimes evaluation instruments tend to be too heavy on technical terms or medical language, making it difficult for students to understand the questions or assignments given (Ni'mah et al., 2018). Teachers can also fall into the trap of creating evaluation instruments that are too one-way, only measuring students' knowledge without providing opportunities for them to apply the knowledge in authentic contexts or to think critically about the digestive system (Utami et al., 2021).

Therefore, research was carried out to develop and validate a test instrument specifically designed to measure higher-order thinking skills on digestive system material at the secondary school level, especially middle school or junior high school. In this study, researchers used the higher-order thinking skills framework developed by Resnick, a highly respected approach in education (Ansori et al., 2019). This framework emphasizes deep understanding, not just memorizing facts or procedures, and directs students to truly understand concepts and ideas (Afandi et al., 2018). Resnick's framework encourages students to apply their knowledge and skills in real and relevant contexts (Ansori et al., 2019). In addition, Resnick's framework emphasizes the importance of critical thinking skills such as analysis, evaluation, and synthesis, which can help students become more independent thinkers (Afandi et al., 2018).

Previous research has investigated the development of higher-order thinking skills test instruments in high school, focusing on mathematics learning. The research uses the SEM (structural equation modelling) model and has produced a product that can evaluate student learning success and the quality of teacher teaching (Benidiktus Tanujaya, 2016). Another research developed a test instrument for higher-order thinking skills in junior high school on quadratic equations. This research uses the APOS model (Action, Process, Object, Schema) with Anderson and Krathwohl's taxonomy indicators C3, C4, and C6 (application, analysis, evaluation). It produces 17 HOT question items ready to be used (Kim How et al., 2022). Apart from that, similar research also focuses on junior secondary students in the digestive system material. However, this research uses the Borg and Gall model with indicators of Anderson and Krathwohl's taxonomy of higher-order thinking

abilities C4 to C6 (analysis, evaluation and creation), which produces 27 valid question items (Utami et al., 2021).

While previous studies have shown promising results in developing HOTs-based test instruments (Widyaningsih et al., 2021; Kim How et al., 2022; Rohmawati et al., 2025), there is still a pressing need to contribute new insights by adopting a different approach. Our research, which uses the Research and Development (R&D) method through the ADDIE (Analysis, Design, Development, Implementation, Evaluation) approach, aims to develop more structured and efficient test instruments. We also use Resnick's framework to measure higher-order thinking skills, as its perspective is more comprehensive and focuses on in-depth understanding and application in authentic contexts, unlike previous research that focused more on Anderson's taxonomy. We believe that our approach will provide a fresh perspective and valuable insights into the challenges, approaches, and value of effective test instruments in understanding student understanding. We hope that this article will serve as a basis for test instrument developers, teachers, and other stakeholders in the education sector to optimize the evaluation of high-order thinking skills in the digestive system.

MATERIALS AND METHODS

This research uses a Research and Development Design (R&D) approach by adapting the ADDIE model, which consists of five stages: Analysis, Design, Development, Implementation, and Evaluation (Hidayat & Nizar, 2021). Therefore, the research procedure begins with the Analysis stage. At this stage, the researcher conducts a source analysis to identify the foundations, guidelines, and standards used in developing higher-order thinking skills (HOTs) test instruments before the preparation begins. Next, in the design stage, the researcher creates a blueprint for the instrument by determining HOT indicators and designing the content of the test instrument. This process is guided by Resnick's (1987) Higher Order Thinking Skills framework, which identifies five main characteristics of higher order thinking: 1)Describe the material; 2)Make conclusions; 3) Build representations/solutions; 4)Analyze; and 5)Building relationships by involving basic mental activities. These characteristics served as a reference for formulating the

indicators and ensuring that each item developed measures the abilities to analyze, evaluate, and create in line with higher-order cognitive processes.

In the development stage, researchers construct test items based on the indicators and content previously designed. This stage also involves a crucial validation process, where two content experts ensure the relevance and clarity of each item in measuring HOTs. The Implementation stage sees the testing of the instrument on ninth-grade students, with a focus on item validity, reliability, discrimination index, and difficulty level. Finally, in the Evaluation stage, researchers assess the quality of the developed instrument based on the item analysis results, making necessary improvements before finalizing the test.

Research data was collected at *SMP Al-Qona'ah Bandung* via a Google form link distributed to students. The population in the study included all class IX students at the school, while the sample used for testing the instrument was 30 students in class IX-C who were selected using a simple random sampling technique. The simple random sampling technique was chosen to obtain a sample with heterogeneous capabilities, which in this study refers to students categorized based on their academic performance into four levels: excellent, good, fair, and poor. This diverse stratification allows the instrument to be tested across a range of student competencies, providing a broader and more accurate picture of how well the instrument functions in diverse learning contexts (Arieska & Herdiani, 2018).

The data analysis in this study focused on evaluating the quality of the test items. Item validity was measured using the Pearson Product-Moment correlation, a statistical measure of the strength and direction of the linear relationship between two variables. Reliability was assessed using Cronbach's Alpha, a measure of internal consistency. In addition, item difficulty indices and discrimination indices were calculated to determine the suitability of each question in measuring higher-order thinking skills. All analyses were conducted using IBM SPSS *Statistics 26.0 for Windows*.

RESULTS AND DISCUSSION

In this comprehensive research, a series of meticulous procedures have been conducted to reveal and analyze various relevant findings. Each methodological step has

been carefully prepared, allowing for a more precise and detailed description of the research results.

Stage 1: Analysis

At this stage, source analysis consists of curriculum analysis, which includes relevant content analysis and analysis of the HOTs framework. The HOTs framework is a set of guidelines that define and measure higher-order thinking skills. Curriculum analysis is carried out to ensure that the curriculum used by an educational institution is in line with the desired HOT capabilities/standards. If there are elements in the HOTs test instrument that are not in line with the content of the curriculum, this can question the validity or relevance of the curriculum in developing higher-level thinking skills. At this stage, the researcher identified learning outcomes/campaign *pembelajaran* (CP) contained in the digestive system material. The CP used in this research is *"Identifying life organization systems and carrying out analysis to find the relationship between organ systems and their functions as well as abnormalities or disorders that arise in certain organ systems (digestive system, circulatory system, respiratory system and reproductive system)"*.

After the CP was found, the researcher carried out an analysis by breaking down the CP into a flow of learning objectives/ *alur tujuan pembelajaran* (ATP) to determine the coverage of material studied in the digestive system at the junior high school level in the independent curriculum. This is done because the questions must reflect the learning objectives set in the curriculum so that the evaluation can accurately measure students' understanding and achievement by what is expected from the curriculum (Nasir, 2015). In the next stage, researchers carried out content analysis to ensure that the question items in the HOTs test instrument were relevant to the experiences experienced by students or at least related to situations they might have encountered in real life, both from teachers and the mass media. Researchers also consider that each element must meet most or all of the HOT characteristics recommended by the Curriculum Development Division (Table 1).

Table 1. Characteristics Higher Order Thinking (HOT) Questions

Characteristics	Description/Explanation of Each Characteristic
There is stimulation	The material or items provided must provide a strong basis for students to think critically, draw conclusions, and produce new ideas that are deeply connected to the material being studied.

Characteristics	Description/Explanation of Each Characteristic
It consists of various levels of cognitive stimulation	Variation in command wording refers to the use of different types of questions, instructions, or tasks designed to measure a learner's understanding, analysis, evaluation, and creativity.
An unusual context	The use of the situation or context in question rarely occurs or has never been questioned before, either in class or in textbooks.
Suitable for situations in real/concrete life	Questions must be designed so that students can relate learning to experiences and situations that can be identified or experienced in everyday life outside the classroom environment.
Non-recurring items	Each question or question asked is designed in such a way that the material or information used is not directly repeated or too similar to the previous one.

Source: Modification of the Malaysian Curriculum Development Division (Kim How *et al*., 2022).

Stage 2: Design

At the design stage, the researcher determines the indicators and content presented in the instrument. As explained previously, the HOT indicators used in research are a framework for high-order thinking skills developed by Resnick, 1987 with the following indicators: 1)Describe the material; 2)Make conclusions; 3)Building representations/solutions; 4)Analyze; and 5)Building relationships by involving basic mental activities (Resnick, 1987 in Tan & Halili, 2015). Each indicator was developed into two questions. The type of question created is a subjective question (essay). There are several reasons behind the choice of subjective questions in this instrument development research, including it can encourage students' critical thinking and flexibility in expressing ideas in different ways, direct students to analytical thinking, strengthen personal relevance, which increases students' interest, and it can measure understanding in more depth (Asrul et al., 2022). The content that will be asked in the instrument includes a discussion of 1)The function of the digestive organ system and its processes in humans; 2) Disturbances in the digestive system; 3)Healthy and nutritious food and its sources; and 4) Process of the digestive system of ruminant animals. Based on the series of procedures that have been carried out, 10 questions were obtained with detailed instrument grids (Table 2).

Table 2. Instrument Grids

Learning Achievements	HOTs Indicator (Resnick)	Learning Materials	Numb
Identifying the organizational systems of life and	Describe the material	Functions of the digestive system organs and how they work in humans (small intestine)	1

Learning Achievements	HOTs Indicator (Resnick)	Learning Materials	Numb
conducting analysis to find the relationship between organ systems and their functions as well as abnormalities or disorders that appear in certain organ systems (digestive system)		Disorders of the human digestive system (maag/gastric acid)	6
	Making conclusions	Disorders of the human digestive system (diarrhea)	2
		Functions of the digestive system organs and how they work in ruminant animals (abomasum)	7
	Building representations/ solutions	Healthy and nutritious food as an effort to maintain the health of human digestive organs	3
		Sources of healthy and nutritious food and their benefits for the body (milk)	8
	Analyze	Types and functions of enzymes in the human digestive system (amylase enzyme)	4
		Functions of the digestive system organs and how they work in humans (tongue)	9
	Building relationships by engaging in fundamental mental activities	Choking mechanism as one of the processes/parts in the digestive system	5
		Efforts to maintain the health of the digestive system organs	10

Stage 3. Development

The grid that has been prepared will be the basis for answering questions about the instrument's development. Researchers use this grid as a guide to develop various questions that reflect the scope of the material and the desired level of difficulty to accurately measure the desired knowledge or ability. This process involves careful consideration of the content, the level of difficulty, and the language used in the questions (Table 3).

Meanwhile, for verification and variation of the level of accuracy of HOTs and language indicators, all questions on the test instrument were reviewed and evaluated by a panel of experts. This panel comprises lecturers in charge of the Learning Evaluation course with doctoral degrees, professors in Education, and lecturers in charge of the Biology Education course with postgraduate degrees. These experts were appointed as interpreters to ensure that the HOTs domain level that the researcher has determined, the content of the digestive system material, and language are not only accurate but also applicable for further use by students, underscoring the value of their work.

Table 3. HOTs Question Item Example

Numb	Question Description
2.	Today Rani didn't go to school because she was sick. Rani had a stomachache. According to the doctor, one of Rani's digestive organs was infected by bacteria so that water absorption was not optimal and caused the feces to become runny. Based on the description of the story, which organ is experiencing this disorder? Explain the reason!
5.	Jakarta - Several people have experienced unfortunate incidents due to food. One of them was a middle-aged man who died from choking while eating my cake. He enjoyed my cake as usual until he finally choked. At first, Ji Rimei had difficulty breathing because her throat was swollen. After that, the child realized that his father had experienced serious choking. (*Source: https://food.detik.com/info-kuliner/d-6549689/tragis-pria-kehilangan-nyawa-karena-tersedak-kue-ku.)* The text above is an excerpt from a choking incident. Choking is an incident that almost everyone has experienced. Explain: a. Why do we experience choking? b. How does choking occur?
9.	The tongue is one of the organs involved in the digestive system. Inside the tongue are papillae. Papillae themselves are protrusions on the upper surface of the tongue. Papillae contain taste nerve cells, sensory cells, supporting cells, and also serous glands that produce saliva. These taste nerve cells have been programmed to have four tastes, namely sweet, sour, salty, and bitter. But in reality when eating food containing chili, we taste spicy. Based on this case, explain: a. Where does the spicy taste come from? b. How does the spicy taste mechanism occur?

Stage 4. Implementation

Upon the completion of the HOTs test instrument, a pivotal phase follows: the implementation. This involves the actual testing of the test items. The initial test instrument was comprised of 10 subjective questions designed as structured problem-solving exercises.

At this stage, the test item's results are not just reviewed, but carefully analyzed, as they are an integral part of the instrument development process. These results provide an in-depth picture of the validity, reliability, discriminatory power, and difficulty level of each test item that has been developed, engaging us in the refinement of the instrument.

Validity Test

Our validity testing is a comprehensive process, evaluating the instrument's ability to accurately, consistently, and relevantly measure the intended concept or variable (Hylton et al., 2022). This thorough testing uses Pearson product-moment on the SPSS 26.0 for Windows application (Table 4).

Table 4. Instrument Validity Analysis

Numb	Question	Sig.Hit	Interpretation	Correlation
Q1	Pearson Correlation	,840**	Valid	Very High
	Sig. (2-tailed)	,000		

Numb	Question	Sig.Hit	Interpretation	Correlation
Q2	Pearson Correlation	,569**	Valid	Enough
	Sig. (2-tailed)	,001		
Q3	Pearson Correlation	,618**	Valid	High
	Sig. (2-tailed)	,000		
Q4	Pearson Correlation	,742**	Valid	High
	Sig. (2-tailed)	,000		
Q5	Pearson Correlation	,824**	Valid	Very High
	Sig. (2-tailed)	,000		
Q6	Pearson Correlation	,682**	Valid	High
	Sig. (2-tailed)	,000		
Q7	Pearson Correlation	,418*	Valid	Enough
	Sig. (2-tailed)	,021		
Q8	Pearson Correlation	,780**	Valid	High
	Sig. (2-tailed)	,000		
Q9	Pearson Correlation	,457*	Valid	Enough
	Sig. (2-tailed)	,011		
Q10	Pearson Correlation	,906**	Valid	Very High
	Sig. (2-tailed)	,000		

**. Correlation is significant at the 0.05 level (2-tailed).*
***. Correlation is significant at the 0.01 level (2-tailed).*
(Source: SPSS 26.0 for windows)

According to Puspasari, H & Puspita (2022), a question item instrument can be declared valid if the calculated r-value, which measures the strength and direction of the relationship between two variables, is greater than the table r value (0.361 for significance 0.05 and 0.463 for significance 0.01 with a sample size of 30). Similarly, the computed sig value, which indicates the probability that the observed correlation could be due to random chance, should be less than the reference sig value (0.05 or 0.01). Based on these provisions, it can be seen that all questions that have been asked are valid with a correlation level of 30% in the sufficient category (range 0.40 - 0.60), 40% in the high category (range 0.60 - 0.80), and 30% in the very high category (range 0.80 - 1.00).

Reliability Test

The reliability test results were obtained using Alpha Cronbach on 10 questions. This method calculates the average correlation of each item with every other item, providing a measure of internal consistency. The results proved the questions were reliable if the calculated sig value > reference sig 0.60 (Maryani et al., 2021). The calculation results show a figure of 0.878, which means that the instrument has a high level of reliability for use. If the reliability of an instrument is high, the measurement error will be reduced; conversely, if the instrument's reliability is low, the measurement error will be greater (Ghozali, 2018). A high level of reliability on each item indicates that the

measurement results achieved by the same individual when retested with an identical test at different times show high consistency.

Discrimination Power and Level of Difficulty Test

The discriminatory power test, a crucial component of the evaluation process, aims to assess the instrument's ability to distinguish between respondents with high performance and those with low performance (Damayanti et al., 2021). Similarly, the difficulty level test, another key aspect, aims to measure the level of difficulty or ease of a question in the instrument for respondents (Damayanti et al., 2021). By conducting these tests, instrument developers can ensure that the instrument has the right variation in difficulty levels. Both types of tests were conducted using the SPSS 26.0 for Windows applications (Table 5).

Table 5. Analysis of Discriminatory Power and Level of Difficulty

Numb.	Corrected Item (Discriminatory Power)	Information	P (Difficulty)	Information
Q1	,778	Very Good	0,37	Currently
Q2	,517	Good	0,00	Difficult
Q3	,522	Good	0,40	Currently
Q4	,622	Good	0,40	Currently
Q5	,758	Very Good	0,43	Currently
Q6	,598	Good	0,17	Difficult
Q7	,359	Enough	0,00	Difficult
Q8	,721	Very Good	0,03	Difficult
Q9	,372	Enough	0,00	Difficult
Q10	,861	Very Good	0,40	Currently

Source: SPSS 26.0 for windows)

From Table 5, it can be seen that all questions on the HOTs test instrument have positive discriminatory power (have high discriminatory power) with a sufficient category of 20% (range 0.30 - 0.39), good of 40% (range 0.40 - 0.70), and very good of 40% (range 0.71 - 1.00). For the difficulty level test, 50% of the questions are at the difficult level (range 0.00 - 0.29), and the other 50% are at the moderate level (range 0.30 - 0.69).

Stage 5. Evaluation

The analysis of the test results is a crucial step in our evaluation process. It involves a thorough review of the students' answers to the tested questions, providing us with valuable insights into how well the test items measure students' understanding of the material being tested. This analysis also helps us determine which test items are suitable for use in the actual research sample if this instrument is to be used (Meesak et al., 2022).

Question number one develops an instrument that highlights the organs involved in the digestive system and their functions. This question shows an overview of the human digestive system along with a brief definition of the digestive process. The focus of the question is the digestive organs involved in absorbing food nutrients and how the body processes and absorbs these nutrients. The trial results showed that question number one has high validity with a significance level of 0.840, indicating a strong relationship between the question and what should be measured. A high correlation strengthens the question's ability to differentiate students based on their knowledge or skills (Elmedina Nikoçeviq-Kurti, 2022). The discriminatory power of this question is also in the outstanding category, which allows it to effectively separate high and low achievement. Although the level of difficulty is not too extreme (moderate category), this shows that the question is not too easy or difficult; the moderate level of difficulty is also suitable for measuring variations in student abilities (Pradipta & Kurniawan, 2023). Thus, the interpretation results confirm that question number one is suitable for use as an evaluation instrument.

Question number two develops an instrument that highlights disorders in the human digestive system. This question introduces the problems experienced by one of the characters written by the researcher, where this character has a stomach ache due to a bacterial infection that causes his faeces to become liquid. The focus of the problem asked is which organ is thought to be disturbed and the reasons for the supporting theory. The trial results showed that question number two has moderate validity with a significance figure of 0.569, indicating a relevant relationship with what should be measured. The correlation of the question is in the sufficient category (still within the accepted range), which indicates a sufficient relationship between the question and what should be measured. The discriminatory power of question number two is also good, which means it can separate high and low achievement well. This question's difficulty level is in the difficult category, which indicates that it may be challenging for students but can still be used to measure their abilities at a higher level. Difficult questions are considered suitable in tight selection situations because they are expected to be able to distinguish between high and low-ability participants more clearly (Nasir, 2015). Moreover, the purpose of developing this instrument is to measure high-level thinking skills. Thus, the

interpretation results state that even though question number two is not perfect, the question can still be used as an evaluation instrument.

Question number three develops the instrument by explaining the cooperation of digestive organs and glands in carrying out their processes. This question highlights the importance of healthy and nutritious food in maintaining a healthy digestive system. The problem focuses on the definition of healthy and nutritious food and the importance of consuming such food. The question also asks about other efforts that can be made to maintain a healthy digestive system.

Meanwhile, question number four discusses enzymes in the digestive system with a table regarding the location and function of each enzyme. The main problem is why the amylase enzyme is produced twice during digestion, namely in the oral cavity and the small intestine. The trial results showed that these two questions had high validity with significance figures of 0.618 and 0.742, indicating a significant relationship with what should be measured. High correlation confirms the reliability of the questions in measuring students' understanding or performance. The discriminatory power of these questions is also quite good, with moderate difficulty. Thus, the interpretation results state that questions three and four can be used as evaluation instruments.

Question five introduces the choking mechanism as an evaluation instrument. This question presents a news text that narrates an incident of a man's death due to choking on food. The question's focus is on understanding the reasons behind the choking incident and how it could have occurred. The trial results show that question number five has high validity, marked by a significance figure of 0.824, and a very high correlation. The question also demonstrates excellent discriminatory power with moderate difficulty. The interpretation results confirm that question number five is a suitable evaluation instrument, further reinforcing the effectiveness of the assessment methods.

Question number six discusses one of the disorders of the human digestive system, namely gastritis. This question begins with a story about an employee who often misses his meal times, which results in complaints of pain in the pit of the stomach and nausea. The focus of the questions developed is why the employee can suffer from gastritis and how such a disease can occur in someone. On the other hand, question number eight discusses the application of healthy and nutritious food sources in everyday life. This question begins with a list of food ingredients, often allergens, including milk. The focus

of the problem developed is a solution for people who are allergic to cow's milk but still have to consume milk because of the importance of the content in the ingredients.

The trial results showed that questions six and eight had high validity with a significance figure of 0.682 and 0.780 and a correlation that was in the high category. Both also showed good and excellent discriminatory power. Instruments with good discriminatory power tend to measure the desired concept or skill more accurately. This reflects that the instrument is well-designed and able to provide informative results. The difficulty level of both questions is in the difficult category. Questions with a high difficulty level are often considered good in measuring high-level thinking skills because they can encourage participants to think deeper, analyze information, and evaluate existing options. Complex questions can also measure participants' ability to adapt to new challenges and solve complex problems. This reflects the participants' ability to think of alternative solutions and strategies that can be applied in various contexts. The interpretation results suggest that questions number six and eight are worthy of use as evaluation instruments.

Question number seven discusses the function of the digestive system organs and how they work in ruminant animals. The question is presented with a stimulus in the form of a picture of the digestive tract in ruminant animals (one of which is a cow). The focus of the problem discussed is asking which organ in the stomach of a cow functions as the actual place of digestion, considering that ruminant animals have four types of stomachs with different functions. Students must answer the question with reasons supported by theory. Meanwhile, question nine discusses one of the human digestive organs, the tongue. The question is given a stimulus in the form of information that the tongue has papillae that can taste various food flavours, such as sweet, sour, salty, and bitter. The focus of the problem discussed is why someone can taste spicy while the tongue only has papillae with four flavours.

The trial results showed that both questions had good validity with a significance figure of 0.418 and 0.457 and a correlation in the pretty good category. The discriminatory power of the two questions was categorized as sufficient; in some contexts, using instruments with "sufficient" discriminatory power is still acceptable, depending on the instrument's purpose (Yoel et al., 2018). The level of difficulty of the two questions was in the difficult category. Thus, the interpretation results stated that questions seven and

nine could still be used. However, they required adjustments or revisions to improve the quality of the instrument. One type of revision that can be made is improving the instructions or context in the question instructions. The arrangement of the language of the question instructions must provide sufficient information to understand the context and solve the questions correctly (Widana, 2017; Nurrifqy, 2024). Clear contextual settings will help students understand the situation presented.

Question number ten discusses efforts to maintain the health of the digestive system organs. The question presents stimuli in the form of benefits and consequences of sweet foods. The focus of the questions developed is why one should not consume too much sweet food and the disorders it can cause. The trial results showed that question number one has high validity, a significance level of 0.906, and a very high correlation. The discriminatory power of this question is also in the outstanding category, and the difficulty level is in the moderate category. Thus, the interpretation results confirm that question number ten is suitable for use as an evaluation instrument.

CONCLUSION

Based on the research, the developed test instrument for assessing higher-order thinking skills (HOTs) on the digestive system topic demonstrated strong content validity, with Content Validity Index (CVI) values ranging from 0.42 to 0.91. The instrument also showed high internal consistency, as indicated by a Cronbach's Alpha of 0.878. Eight out of ten items met the criteria for high validity and good to excellent discriminatory power, effectively measuring students' conceptual understanding and cognitive skills. Although items 7 and 9 yielded relatively lower validity and moderate discrimination indices, they remain potentially usable with appropriate revision—such as improving item clarity and context—and further empirical testing. In conclusion, the instrument is considered valid and reliable for evaluating HOTs in junior high school science education. Its role in promoting student learning and understanding makes it a valuable resource for teachers seeking to implement competency-based assessments aligned with current curriculum standards.

SUGGESTION

In future research, it is suggested to create instruments with a more diverse set of questions that encompass a broader range of science topics, with a focus on the digestive system. This expansion is expected to deepen students' understanding and knowledge.

REFERENCES

Afandi, A., Sajidan, S., Akhyar, M., & Suryani, N. (2018). Pre-Service Science Teachers' Perception About High Order Thinking Skills (HOTS) in the 21st Century. *International Journal of Pedagogy and Teacher Education*, *2*(1), 107. https://doi.org/10.20961/ijpte.v2i1.18254

Ahİ, B. (2017). Thinking about digestive system in early childhood: A comparative study about biological knowledge. *Cogent Education*, *4*(1), 1–16. https://doi.org/10.1080/2331186X.2017.1278650

Ahmad, S., Prahmana, R. C. I., Kenedi, A. K., Helsa, Y., Arianil, Y., & Zainil, M. (2018). The instruments of higher order thinking skills. *Journal of Physics: Conference Series*, *943*(1). https://doi.org/10.1088/1742-6596/943/1/012053

Ansori, M., Nurkamto, J., & Suparno, S. (2019). Teacher's Beliefs and Practices in the Integration of Higher Order Thinking Skills in Teaching Reading. *ELS Journal on Interdisciplinary Studies in Humanities*, *2*(4), 541–555. https://doi.org/10.34050/els-jish.v2i4.8164

Arieska, P. K., & Herdiani, N. (2018). Pemilihan Teknik Sampling Berdasarkan Perhitungan Efisiensi Relatif. *Jurnal Statistika*, *6*(2), 166–171. https://jurnal.unimus.ac.id/index.php/statistik/article/view/4322/4001

Asrul, Saragih, A. H., & Mukhtar. (2022). *Evaluasi Pembelajaran*.

Benidiktus Tanujaya. (2016). Development of an instrument to measure higher order thinking skills in senior high school mathematics instruction. *Journal of Education and Practice*, *7*(21), 144–148.

Cahyanto, T., Al Zahro, I. R., & Windayani, N. (2023). Konsep dan Implementasi Literasi Halal pada Pembelajaran IPA. *Jurnal Penelitian Sains Dan Pendidikan (JPSP)*, *3*(2), 158–172. https://doi.org/10.23971/jpsp.v3i2.7237

Damayanti, Wi. D., Halidjah, S., & Pranata, R. (2021). Analisis tingkat kesukaran butir soal pilihan gandapada penilaian tengah semester kelas iv. *Jurnal Pendidikan Dan Pengajaran Khatulistiwa*, *10*(11), 1–10. https://jurnal.untan.ac.id/index.php/jpdpb/article/view/50458/75676591120

Darong, H. C., Erna Mena Niman, & Fransiskus Nendi. (2024). Pendampingan Berbasis Hots: Strategi Peningkatan Kemampuan Guru. *Qardhul Hasan: Media Pengabdian Kepada Masyarakat*, *10*(2), 153–161. https://doi.org/10.30997/qh.v10i2.13539

Elmedina Nikoçeviq-Kurti. (2022). European Journal of Educational Research. *European Journal of Educational Research*, *11*(3), 1245–1257. https://www.researchgate.net/profile/Suntonrapot-Damrongpanit/publication/356662582_Effects_of_Mindset_Democratic_Parenting_Teaching_and_School_Environment_on_Global_Citizenship_of_Ninth-grade_Students/links/61a6dda685c5ea51abc0f7b6/Effects-of-Mindset-Dem

Ghozali, I. (2018). Aplikasi Analisis Multivariate dengan Program IBM SPSS 25 Edisi 9. *Semarang: Badan Penerbit Universitas Diponegoro. Variabel Pemoderasi. E-Jurnal Akuntansi Universitas Udayana*, *23 (2)*(1470), 1494.

Hidayat, F., & Nizar, M. (2021). Model Addie (Analysis, Design, Development, Implementation and Evaluation) Dalam Pembelajaran Pendidikan Agama Islam Addie (Analysis, Design, Development, Implementation and Evaluation) Model in Islamic Education Learning. *Jurnal UIN*, *1*(1), 28–37.

Hylton, S. P., Joseph, J. D., Ward, T. J., & Gareis, C. R. (2022). Examining the Validity of a Student Teaching Evaluation Instrument. *Teacher Educators' Journal*, *15*(1), 77–101.

Kim How, R. P. T., Zulnaidi, H., & Rahim, S. S. B. A. (2022). Development of Higher-Order Thinking Skills test instrument on Quadratic Equations (HOTS-QE) for Secondary School Students. *Pegem Egitim ve Ogretim Dergisi*, *13*(1), 379–394. https://doi.org/10.47750/pegegog.13.01.41

Kwangmuang, P., Jarutkamolpong, S., Sangboonraung, W., & Daungtod, S. (2021). The development of learning innovation to enhance higher order thinking skills for students in Thailand junior high schools. *Heliyon*, *7*(6), e07309. https://doi.org/10.1016/j.heliyon.2021.e07309

Lu, K., Yang, H. H., Shi, Y., & Wang, X. (2021). Examining the key influencing factors on college students' higher-order thinking skills in the smart classroom environment. *International Journal of Educational Technology in Higher Education*, *18*(1), 1–13. https://doi.org/10.1186/s41239-020-00238-7

Maryani, I., Prasetyo, Z. K., Wilujeng, I., Purwanti, S., & Fitrianawati, M. (2021). HOTs Multiple Choice and Essay Questions: A Validated Instrument to Measure Higher-order Thinking Skills of Prospective Teachers. *Journal of Turkish Science Education*, *18*(4), 674–690. https://doi.org/10.36681/tused.2021.97

Meesak, A. M., Rozgonjuk, D., Õun, T., & Kikas, E. (2022). Validation of an e-instrument for assessing five-year-old children's development in Estonia: a comparison of children's skills and teachers' evaluations. *Education 3-13*, 1–16. https://doi.org/10.1080/03004279.2022.2137378

Nasir, M. (2015). Analisis Empirik Program Analisis Butir Soal Dalam Rangka Menghasilkan Soal Yang Baik dan Bermutu Sebagai Alat Evaluasi Pembelajaran Fisika. *Prosiding Semirata*, 336–347. jurnal.untan.ac.id

Ni'mah, S., Lestari, N. C., & Adawiyah, R. (2018). Pengembangan dan Uji Validitas Perangkat Pembelajaran SMA Berbasis Kurikulum 2013 pada Konsep Sistem Pencernaan. *Jurnal Pendidikan Hayati*, *4*(1), 22–30.

Nida, S., Rahayu, S., & Eilks, I. (2020). A survey of Indonesian science teachers' experience and perceptions toward socio-scientific issues-based science education. *Education Sciences*, *10*(2), 1–15. https://doi.org/10.3390/educsci10020039

Nurrifqy, Z. F. (2024). Peran Evaluasi Pembelajaran Dalam Meningkatkan Kualitas Pendidikan Di Sekolah Dasar. *Jurnal Pendidikan Sosial Dan Humaniora*, *3*(3), 2068–2080.

Pradipta, A. W., & Kurniawan, R. (2023). Tingkat Kesulitan dan Daya Beda Butir Soal Ujian Akhir Semester Matakuliah Penelitian Pendidikan. *Jurnal Pendidikan*, *11*(02), 234–241.

Purba, Y., O. (2018). Teknik Uji Instrumen Penelitian Pendidikan. *Widini Bhakti Persada Bandung*, *01*(02), 3–26.

Puspasari, H & Puspita, W. (2022). Uji Validitas dan Reliabilitas Instrumen Penelitian Tingkat Pengetahuan dan Sikap Mahasiswa terhadap Pemilihan Suplemen Kesehatan dalam Menghadapi Covid-19 Validity Test and Reliability Instrument Research Level Knowledge and Attitude of Students Towards. *Jurnal Kesehatan*, *13*, 65–71.

Rahmi, Y. L., Habibah, I. N., Zulyusri, Z., & Darussyamsu, R. (2021). HOTS assessment in circulatory system learning: Validity, reliability, and item quality. *JPBI (Jurnal Pendidikan Biologi Indonesia)*, *7*(2), 171–178. https://doi.org/10.22219/jpbi.v7i2.15513

Rohmah, I. G., Supeno, S., & Hariani, S. A. (2023). The Development of AKSI (Aktualisasi Siswa) Learning Model to Improve Higher Order Thinking Skills of Natural Science in Junior High School. *Bioma : Jurnal Ilmiah Biologi*, *12*(2), 77–89. https://doi.org/10.26877/bioma.v12i2.16716

Rohmawati, W., Nulhakim, L., Suryani, D. I., Sultan, U., Tirtayasa, A., & Serang, K. (2025). Pengembangan Soal HOTS Berbasis *Four Tier Multiple Choice* (4TMC) Materi Sistem Pencernaan pada Manusia untuk Siswa Kelas VIII SMP. *Eduproxima : Jurnal Ilmiah Pendidikan IPA*, Vol. 7 (2) : 623 -632.

Rusandi, D., Pramono, H., Herlangga, A. T. F., Carsiwan, C., & Priyono, D. (2024). Pendekatan Pedagogis dalam Desain Kurikulum: Studi "Menjawab Tantangan Abad 21." *JIIP - Jurnal Ilmiah Ilmu Pendidikan*, *7*(8), 8671–8676. https://doi.org/10.54371/jiip.v7i8.5076

Setiawan, J., Sudrajat, A., Aman, & Kumalasari, D. (2021). Development of higher order thinking skill assessment instruments in learning Indonesian history. *International Journal of Evaluation and Research in Education*, *10*(2), 545–552. https://doi.org/10.11591/ijere.v10i2.20796

Suhirman, S. (2024). Peningkatan Kapasitas Guru Biologi Madrasah Aliyah dalam Menyusun Tes Berbasis *Higher-Order Thinking Skills* (HOTS) Di kabupaten Lombok Tengah. *AL-HAYAT: Jurnal Pengabdian Masyarakat*, *2*(1), 1–12. https://doi.org/10.62588/ahjpm.2024.v2i1.0002

Tan, S. Y., & Halili, S. H. (2015). Effective Teaching of Higher-Order Thinking (HOT) in Education. *The Online Journal of Distance Education and E-Learning*, *3*(2), 41–47.

Utami, T. P., Sjaifuddin, S., & Berlian, L. (2021). Pengembangan Soal Uraian Berbasis Indikator Kemampuan Berpikir Tingkat Tinggi pada Konsep Sistem Pencernaan pada Manusia untuk Siswa Kelas VIII SMP/Mts. *PENDIPA Journal of Science Education*, *6*(1), 128–134. https://doi.org/10.33369/pendipa.6.1.128-134

Widana, I. W. (2017). *Modul penyusunan soal HOTS.*

Widyaningsih, S. W., Yusuf, I., Prasetyo, Z. K., & Istiyono, E. (2021). The development of the hots test of physics based on modern test theory: Question modeling through e-learning of moodle lms. *International Journal of Instruction*, *14*(4), 51–68. https://doi.org/10.29333/iji.2021.1444a

Wijnen, F., Walma van der Molen, J., & Voogt, J. (2021). Measuring primary school teachers' attitudes towards stimulating higher-order thinking (SHOT) in students: Development and validation of the SHOT questionnaire. *Thinking Skills and Creativity*, *42*(September), 100954. https://doi.org/10.1016/j.tsc.2021.100954