

Development of a Critical Thinking Reading Assessment Instrument for Rural Indonesian Senior High School Students

Semi Sukarni¹, Abdul Ngafif², Iqbal Salman Abdurrohman³, Alek Andika⁴

¹ Universitas Muhammadiyah Purworejo, Indonesia; semisukarni@umpwr.ac.id

² Universitas Muhammadiyah Purworejo, Indonesia; abdulngafif@umpwr.ac.id

³ Universitas Muhammadiyah Purworejo, Indonesia ; iqbalsalman100@gmail.com

⁴ Universitas Ma'arif Nahdlatul Ulama Kebumen, Indonesia; mr12011989@gmail.com

ARTICLE INFO

Keywords:

assessment;
critical thinking;
reading;
rural senior high school

Article history:

Received 2026-04-06

Revised 2026-06-13

Accepted 2026-09-29

ABSTRACT

Critical thinking is essential for EFL reading comprehension; however, classroom assessments often provide limited opportunities to evaluate students' analytical and evaluative reasoning, particularly in rural educational contexts. This study aimed to develop and evaluate a critical-thinking-oriented reading assessment instrument for rural Indonesian senior high school students. The study employed a Research and Development design using the ADDIE model. A needs analysis was conducted with 15 senior high school English teachers. The developed instrument, based on Facione's critical thinking framework, comprised 20 multiple-choice and five essay items. Its development involved expert judgment, a limited trial, item analysis, reliability testing, revision, and field implementation with 107 eleventh-grade students at a rural senior high school. The needs analysis indicated a strong demand for a critical-thinking-oriented reading assessment ($M = 4.56$), while only 20% of the participating teachers reported incorporating critical thinking indicators into their assessment practices. Expert judgment supported the instrument's content appropriateness. Empirical item analysis showed that 20 of 25 items met the applied validity criterion, while five items required revision. The multiple-choice and essay sections yielded Cronbach's alpha coefficients of .723 and .693, respectively. Field implementation produced a mean student score of 72.19 ($SD = 12.45$). The findings provide initial evidence that the developed instrument is applicable for assessing critical thinking through EFL reading in a rural senior high school context. The mixed item format may also support teachers in translating critical thinking constructs into classroom assessment practices. Further multi-school validation is required to establish broader validity, reliability, and applicability.

This is an open access article under the [CC BY-NC-SA](https://creativecommons.org/licenses/by-nc-sa/4.0/) license.



Corresponding Author:

Semi Sukarni

Universitas Muhammadiyah Purworejo, Indonesia; semisukarni@umpwr.ac.id

1. INTRODUCTION

Critical thinking has become an essential competence in contemporary education because learners are increasingly required to interpret, analyze, evaluate, and make reasoned judgments about information rather than merely reproduce knowledge. Within language education, critical thinking is particularly relevant because language learning involves not only linguistic competence but also the ability to construct meaning, evaluate information, and communicate reasoned interpretations. Facione (1990) conceptualizes critical thinking through six interrelated cognitive skills: interpretation, analysis, inference, evaluation, explanation, and self-regulation. These dimensions provide a useful theoretical basis for educational assessment because they specify cognitive processes that can be translated into observable task demands. More recent work on language assessment similarly emphasizes the importance of clearly defining critical thinking as a construct before incorporating it into language assessment (Yin et al., 2024).

The relationship between critical thinking and reading comprehension is particularly important in English as a Foreign Language (EFL) education. Effective reading extends beyond identifying vocabulary, retrieving factual information, or recognizing explicitly stated ideas. Students are also expected to infer implicit meanings, examine relationships among ideas, evaluate claims and evidence, distinguish relevant from irrelevant information, and justify interpretations. Research has consequently emphasized the pedagogical value of integrating critical thinking into EFL reading activities (Indah et al., 2022; Oktaviah et al., 2021). Empirical evidence also indicates meaningful connections between reading comprehension and critical thinking among EFL learners, although individual critical-thinking dimensions may contribute differently to reading performance (Alshehri, 2024). Thus, reading provides an appropriate context in which students can demonstrate critical thinking through their engagement with textual information.

Despite this conceptual alignment, assessing critical thinking through reading remains challenging. Classroom reading assessments often emphasize outcomes that can be efficiently scored, such as identifying main ideas, locating factual details, recognizing vocabulary, and selecting correct answers. Such tasks are useful for assessing particular dimensions of comprehension but may provide insufficient evidence of how students analyze information, evaluate alternatives, construct inferences, or justify their responses. Brookhart (2010) argues that higher-order thinking assessment should require students to process and use information rather than simply recall it. Similarly, critical-thinking assessment requires careful alignment between the construct to be measured, the tasks used to elicit it, and the criteria used to evaluate students' responses. Research in language assessment has therefore increasingly emphasized construct-based approaches that explicitly connect critical-thinking dimensions with observable language performance (Yin et al., 2024).

Assessment format is also important in capturing different aspects of students' thinking. Multiple-choice items can provide efficient and relatively objective assessment across a range of content, but selected responses alone may offer limited information about the reasoning underlying students' choices. Open-ended tasks, in contrast, can provide opportunities for students to explain interpretations, evaluate information, justify conclusions, and support arguments with textual evidence. Fajari (2021) demonstrates the potential of open-ended questions for assessing critical-thinking skills, while research on authentic assessment in Indonesian English classrooms suggests that assessment activities can create opportunities for learners to evaluate evidence, reflect on issues, formulate logical opinions, and draw conclusions (Umam et al., 2024). A combination of selected-response and constructed-response items may therefore provide complementary evidence of students' reading comprehension and critical-thinking processes.

These assessment issues are particularly relevant in Indonesian EFL education. Previous studies have examined higher-order thinking, critical literacy, and critical-thinking assessment in various educational contexts (Arini et al., 2023; Indah et al., 2022; Kurniawati et al., 2020; Rustam & Priyanto, 2022; Sukarni et al., 2022). Nevertheless, translating theoretical critical-thinking frameworks into practical classroom assessment remains challenging. Teachers must determine which dimensions of critical thinking should be assessed, how these dimensions can be represented through reading tasks, and how students' responses can be evaluated consistently. The challenge is therefore not simply to increase the

number of difficult questions but to establish explicit alignment among the critical-thinking construct, reading objectives, item formats, and scoring criteria. A theoretically grounded yet practically manageable instrument may consequently support both student assessment and teachers' assessment literacy.

The need for context-sensitive assessment becomes more important when considering rural schools. Educational contexts can differ in access to learning resources, technology, English exposure, and opportunities for professional development. Research comparing rural and urban learning contexts has documented differences in learners' experiences and approaches to English learning (Fu & Liu, 2024), while studies of rural education have highlighted additional challenges associated with resources and teacher support. Such contextual conditions do not imply that rural students inherently possess lower critical-thinking ability; rather, they indicate that assessment instruments should be designed and evaluated with attention to the environments in which they will be used. Consequently, an instrument developed for rural EFL learners should employ accessible language, age-appropriate reading materials, relevant contexts, clear instructions, and feasible scoring procedures instead of assuming that instruments developed for other populations can be transferred without contextual evaluation.

Although previous research has provided critical-thinking frameworks and developed instruments for secondary-school students, a more specific gap remains at the intersection of critical thinking, EFL reading assessment, mixed item formats, and rural Indonesian senior high school contexts. Orhan and Ay (2022), for example, developed and validated a critical-thinking skills test for high school students, demonstrating the importance of systematic validity and reliability evidence in instrument development. However, general critical-thinking measures are not necessarily designed to elicit thinking through EFL reading tasks. Likewise, Indonesian research has examined critical thinking and higher-order assessment, but there remains a need for context-sensitive instruments that operationalize an established critical-thinking framework specifically through reading comprehension tasks while remaining practical for classroom use. Addressing this gap requires an instrument-development process that begins with teachers' needs, translates theoretically defined critical-thinking dimensions into reading tasks, incorporates complementary item formats, and evaluates the resulting instrument empirically in its intended educational setting.

Accordingly, the present study develops a critical-thinking-oriented reading comprehension assessment instrument for rural Indonesian senior high school students using the ADDIE model. The instrument operationalizes Facione's (1990) critical-thinking dimensions through multiple-choice and open-ended reading items, thereby combining efficient selected-response assessment with opportunities for students to demonstrate reasoning through constructed responses. The study contributes practically by providing a context-sensitive assessment resource for EFL teachers and methodologically by illustrating how critical-thinking dimensions can be translated into reading objectives, item specifications, and scoring criteria. Rather than assuming broad applicability, the study provides initial evidence from the instrument's intended rural context as a basis for subsequent validation across more diverse populations.

Based on these objectives, the study addresses three research questions: (1) How is a critical-thinking-oriented reading comprehension assessment instrument developed for rural Indonesian senior high school students using the ADDIE model? (2) What evidence of item validity and reliability is obtained from the limited trial of the developed instrument? and (3) How does the developed instrument perform when implemented with students in its intended rural senior high school context?

2. METHODS

2.1 Design

This research was designed using a Research and Development (R&D) approach with the ADDIE model for an instrument development based on assessment theory and deep learning (Branch, 2010). The

problem-solving strategy focuses on the following five main steps, namely analyze, design, develop, implement, and evaluate. The figure of this process is shown in Figure 1.

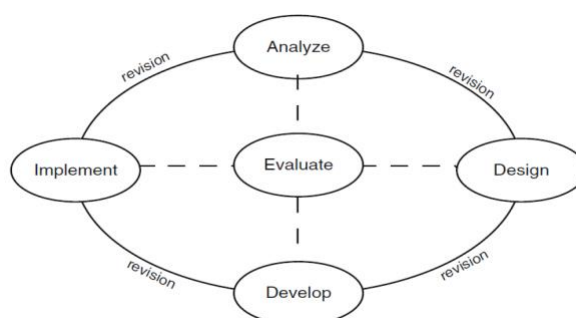


Figure 1. ADDIE Concept

2.2 Research Setting and Participants

This research was conducted in a senior high school in Purworejo Regency. The data were collected for three months, that is July to October 2025, starting from the need analysis of the critical thinking skill assessment up to the implementation of the instrument to assess the students' critical thinking. In the process of analysing the need of CTS instrument, it involves senior high school teachers in Purworejo regency. The need analysis the CTS instrument involved 15 senior high school teachers in Purworejo regency. Furthermore, in implementation, the researchers involved 107 of the eleventh-grade students at senior high school which is located in a rural area of Purworejo regency.

2.3 Research Procedure

Based on Branch (2010) cited in Adeoye et al. (2024); Feng and Sangsawang (2023); Li and Cheong (2023); Sial et al. (2024), there are five stages in the ADDIE model, namely:

(1) Analysis (Needs and Context Analysis)

The activities in this stage are a literature review of definitions and indicators of critical thinking. The researchers had decided to adopt Facione's indicators: curriculum and competency standards analysis (target grade/target classes); audience analysis: participant characteristics (grade, age, language proficiency level); instrument user needs (teachers, students).

(2) Design (Instrument Design)

The activities in this stage includes: creating a detailed blueprint: domain (Facione), operational indicators, cognitive levels (C1–C6), item format, and weighting of each item; determine the final format namely 20 multiple choice items + 5 essays; design a scoring rubric for essay items (4 levels: excellent, good, fair, poor); preparing administration guidelines and instructions for respondents and raters (scoring guide).

(3) Development

The activities in this stage are writing draft items according to the grid, internal review and revision of the draft, and checking content validation by content experts. The first expert is a senior lecturer who has a lot of experience in teaching and assessing reading comprehension, while the second expert is a senior high school teacher who has taught English for more than 25 years. The two experts assessed relevance, appropriateness, and representativeness; used the Content Validity Index (CVI) or a scale of 1–4; collected qualitative comments; and revised items based on expert input.

(4) Implementation (Field Test)

The instrument was conducted in two stages: a limited tryout and a field test. The limited tryout was conducted in one class of 32 students. This stage aimed to obtain initial data on text readability, instruction clarity, item understandability, and the empirical quality of each item.

After revisions were made, the instrument was retested in the field with three classes totaling 107 students. This field test aimed to obtain an overview of the instrument's performance in the target population.

(5) Evaluation (Analysis and Finalization)

A comprehensive evaluation of the instrument's implementation results. At this stage, researchers analyze findings from limited tryouts and field tests to determine the instrument's final quality. The evaluation is conducted by reviewing weak items, revising problematic components, and refining the instrument's format and scoring rubric.

2.4 Technique of Data Analysis

(1) Qualitative Analysis

This technique was used for needs analysis (interview data, observations, open-ended questionnaires); expert validation (narrative input from experts); teacher/student feedback on the product

(2) Quantitative Analysis

This technique was used to test content and construct validity; test instrument reliability; test product practicality and effectiveness.

3. FINDINGS AND DISCUSSION

3.1 Finding

The findings relate to the research questions, which were intended to give explanations on the need analysis of English teachers on the Critical Thinking skill (CTS) instrument; developing the CTS instruments, and providing implementation results of the CTS for assessing rural senior high school students' CTS reading performance. The detail descriptions are provided below.

3.1.1 The result of the need analysis of the Critical Thinking skill (CTS) instrument

In this stage, the researchers draw the information from English teachers related to the need of Critical Thinking skill (CTS) instrument. The data of the need analysis of the critical thinking skill instrument were obtained from a questionnaire which was distributed to 15 (fifteen) English teachers of Purworejo, Central Java. The questionnaire includes 5 Likert scale items and 3 open-ended items. The result of the need analysis is presented in Table 1.

Table 1. The result of the need analysis in the critical thinking skill instrument in reading

No	Statement	Response					Mean	Interpretation
		SA	A	N	D	SD		
1	I need an instrument that can assess students' critical thinking skills when reading texts.	10	5	0	0	0	4.7	Very High
2	The assessment instruments currently used do not adequately measure critical thinking.	4	11	0	0	0	3.6	High
3	It is necessary to develop reading instruction that encourages text analysis and evaluation.	12	3	0	0	0	4.8	Very High
4	There needs to be an instrument based on critical thinking indicators.	13	2	0	0	0	4.9	Very High
5	Teachers need practical guidance in assessing students' critical thinking skills.	12	3	0	0	0	4.8	Very High
Mean							4.56	Very High

Table 1 shows the need analysis of the critical thinking skill instrument in reading obtained from 15 English teachers. From the 5 items of the questionnaire referring to the need, the mean is 4.56. It can be concluded that the need of critical thinking skill instrument is extremely needed.

Three open-ended questions to enrich the data on the closed-ended questions. The result is presented below.

(1) *What instruments are typically used to assess reading comprehension?*

Table 2. Teachers' Current Reading Assessment Practices

No	Instrument type	Frequency
1	Multiple choice	10
2	Essay item	5
3	Retell	4
4	Combined	6

Based on the questionnaire results, ten teachers commonly use multiple-choice type questions for assessing reading comprehension. Five teachers use essay-type questions; four teachers use retelling, and six teachers use combinations of multiple-choice and essay questions.

(2) *Are there any critical thinking (CT) indicators that have been used? Please list them.*

The questionnaire result is reported in Figure 2.

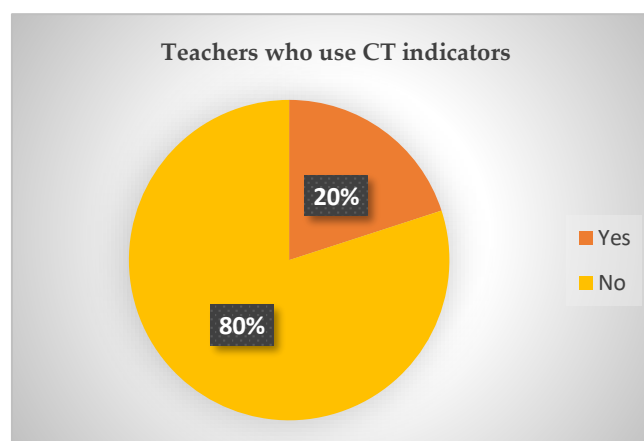


Figure 2. Teachers who use CT indicators

Based on Figure 2, only 20% of teachers who participated in this study use CT indicators in constructing test items; meanwhile, 80% of teachers do not include the CT indicators.

(3) *What are the obstacles faced in measuring critical thinking skills?*

The questionnaire result obtained from teacher respondents is reported in Table 3.

Table 3. Obstacles faced in measuring critical thinking skills

Theme	Code
Construct	(1) Limitations of measurement instruments, differences in competence, and complexity. (2) Instrument limitations and bias in assessment (3) The instrument used is not yet fixed.
Time	(4) Because of the open-ended questions, the answer variations are very wide. (1) Time, measurement tools. (2) Difficulty in creating valid and reliable questions.

Recourses	(1) Resources, practice opportunities. (2) Difficulty in creating valid questions. (3) Composing questions.
Students	(1) The large number of students requires a long time. (2) The variety of students' English competencies

Table 3 shows teachers' difficulties in measuring critical thinking skills, which were put into four themes. The theme "construct" seemed to be the theme with the most problems.

3.1.2 The validity and reliability of the developed assessment instrument through limited trials and empirical analysis

a. Instrument blueprint

The design phase is used to verify the desired performances and determine the appropriate testing methods. Based on the results of the needs analysis above, the researchers developed an instrument draft by adopting critical thinking skills based on Facione's indicators.

Based on Facione's (1990) framework, the following are six key indicators of critical thinking skills used in developing questions and assessment rubrics, interpretation, analysis, inference, evaluation, explanation and self-regulation. Based on the result of the need analysis, the researchers made an assessment blueprint which consists of 20 multiple-choice items and 5 essay items to measure students' critical thinking performance in reading descriptive texts. Each indicator was operationalized into specific item objective. The blueprint is presented in Table 4.

Table 4. Distribution of Reading Test Items based on Facione's Critical Thinking Indicators

No.	Indicator	Item Objective	Item Numbers
1	Interpretation	Students can identify the main ideas and supporting details in descriptive texts	1, 8, 15, 16, 21
2	Analysis	Students can identify comparisons, classifications of information, cause-and-effect relationships presented descriptive texts.	2, 6, 9, 13, 17, 22
3	Inference	Students can make conclusions, infer implicit meaning, and make prediction of descriptive text.	3, 10, 18, 23
4	Evaluation	Students can evaluate the writer's arguments or information presented in descriptive text.	4, 19, 24
5	Explanation	Students are able to give explanation in their answer logically based on the evidence in the text	5, 12, 20, 25
6	Self-Regulation	Students can reflect their reasons and reconsider their assumptions when it is necessary	7, 14

b. Expert validation

After the development of the instrument was completed, the researchers conducted instrument validation through expert judgment. The first is a lecturer with expertise in ELT and assessment, and the second is a senior English teacher as a practitioner. The experts filled in the table checklist, which is constructed on a Likert scale with Very appropriate (VA), Appropriate (A), Somewhat Appropriate (SA), Inappropriate (IA), Very Inappropriate (VIA). The result is shown in Table 5.

Table 5. Expert validation

No	Assessed Aspects	Assessment					Mean	Interpretation
		VA	A	SA	IA	VIA		
1	Indicators' suitability to critical thinking dimensions.	2					5	Very High
2	Clarity of language in questions	2					5	Very High
3	Text context's suitability for high school students' level	1	1				4.5	Very High
4	Clarity of assessment rubric	2					5	Very High
5	Potential of the instrument to encourage students' critical thinking processes	2					5	Very High

To interpret the result, the writers adopted Silaen & Widiyono's mean interpretation (2013:128), i.e., Very High (4.20 – 5.00), High (3.40 – 4.19), Medium (2.60 – 3.39), Low 1.80 – 2.59, and Very Low (1.00 - 1.79). Based on the expert validation, the result is very high, which means that the instrument was very appropriate for assessing students' critical reading skills.

Scale note

1 = Very Inappropriate,

2 = Inappropriate,

3 = Somewhat Appropriate,

4 = Appropriate,

5 = Very Appropriate

c. Limited trial and item analysis

The test instrument was also tried out on 34 students of the eleventh grade of Senior High School. To check the instrument's validity, the researchers applied it in SPSS version 22. The result of the validity test is presented in Table 6.

Table 6. The result of the item validity test

No Item	Type of Item	r-count	r-table	Interpretation
1	Multiple choice	.783	0.275	Valid
2	Multiple choice	.661	0.275	Valid
3	Multiple choice	.578	0.275	Valid
4	Multiple choice	.213	0.275	Invalid
5	Multiple choice	.495	0.275	Valid
6	Multiple choice	.677	0.275	Valid
7	Multiple choice	.093	0.275	Invalid
8	Multiple choice	.255	0.275	Invalid
9	Multiple choice	.500	0.275	Valid
10	Multiple choice	.678	0.275	Valid
11	Multiple choice	.578	0.275	Valid

12	Multiple choice	.490	0.275	Valid
13	Multiple choice	.766	0.275	Valid
14	Multiple choice	.400	0.275	Valid
15	Multiple choice	.231	0.275	Invalid
16	Multiple choice	.495	0.275	Valid
17	Multiple choice	.193	0.275	Invalid
18	Multiple choice	.485	0.275	Valid
19	Multiple choice	.333	0.275	Valid
20	Multiple choice	.561	0.275	Valid
21	Essay	.314	0.275	Valid
22	Essay	.501	0.275	Valid
23	Essay	.428	0.275	Valid
24	Essay	.568	0.275	Valid
25	Essay	.658	0.275	Valid

Based on the result of validity testing, 80% of items are valid, while 20% are not valid. The valid items can be used for further testing, while the invalid items, namely number 4, 7, 8, 15, and 17 needed to be revised.

Based on the findings of item analysis and expert validation, invalid test items were corrected by determining each item's shortcomings. Enhancing the quality of distractions in multiple-choice questions, aligning the items with Facione's critical thinking indicators, and improving the instructions' clarity were the main goals of the revision process. To make sure the items more accurately assessed higher-order thinking abilities, ambiguous phrasing and extremely simple or implausible answer options were changed.

Here is a sample of a multiple-choice valid instrument of "analysis"

(1) *How is the Tengerese culture connected to Mount Bromo?*

- a. *They use Mount Bromo as a farming area.*
- b. *They built a temple and conduct ceremonies there.*
- c. *They believe the crater produces fertile soil for farming.*
- d. *They migrated because of volcanic activity.*

The sample of the essay question of "Evaluation".

(2) *Based on the text, evaluate whether Labuan Bajo should prioritize tourism growth or environmental protection. Support your answer with evidence from the text.*

For the scoring rubric, the multiple-choice section consisted of 20 items. Each correct answer was scored 1, while wrong answers were scored 0. Meanwhile, five questions in the essay part were designed to evaluate higher-order critical thinking abilities. An analytical scoring rubric that was modified from Facione's critical thinking indicators was used to assess each essay item. The following criteria were used to assess responses on a scale of 0 to 4: Score 4= The answer is extremely well-developed, logical, and supported by pertinent data or logic; score 3 = Although not fully developed, the response is reasonable and straightforward with sufficient supporting information; score 2= There is insufficient justification and explanation in the response; score 1 The answer is ambiguous, nonsensical, or hardly pertinent to the query; score 0= No response or an unrelated response.

d. *Reliability testing*

The reliability of the test item was also tested with SPSS version 22. The result of the reliability test is shown in Table 7.

Table 7. The result of the reliability test of items

Type of instrument	Cronbach's Alpha	CA level	Interpretation
Multiple-choice	.723	High	Reliable
Essay	.693	High	Reliable

Table 6 presents the internal consistency reliability of the multiple-choice and essay sections of the developed instrument. The Cronbach's alpha coefficients were .723 for the multiple-choice items and .693 for the essay items. Based on the criterion proposed by Sugiyono (2013), in which a coefficient of .60 or higher indicates acceptable reliability, both sections demonstrated adequate internal consistency for use in this study. In addition to internal consistency analysis, two raters independently scored the essay responses (Items 21–25) to enhance scoring objectivity. Both raters applied the same analytical scoring rubric derived from Facione's critical thinking framework, particularly the dimensions of interpretation, analysis, inference, evaluation, and explanation. The use of common scoring criteria was intended to promote consistency in evaluating students' constructed responses. However, the degree of agreement between the two raters should be reported using an appropriate inter-rater reliability coefficient to provide stronger evidence of scoring reliability.

3.1.3 The implementation of the Critical Thinking Skill instrument in reading comprehension of rural senior high school

After the invalid test items were revised, the instrument was implemented to test the students' critical thinking skills in three classes of the eleventh grade of a rural senior high school in Purworejo, with a total of 107 students. The result of the students' achievement is presented in Table 8.

Table 8. The result of students' critical thinking skills (CTS)

Value	Grade	Level of Achievement	Frequency	Percentage
80 – 100	A	Excellent	37	34.5%
66 – 79	B	Good	37	34.5%
56 – 65	C	Sufficient	23	21.4
40 – 55	D	Fairly Sufficient	10	9.3%
<39	E	Low	0	0
			107	

Table 8 shows the test results of students' critical thinking skills (CTS). It spreads from the fairly sufficient to the excellent category, while there is no student in the low category. The percentage of students in the good and very good category are the same (34.5%). Roughly, the students' CTS is good. The result is also represented in figure 3 below.

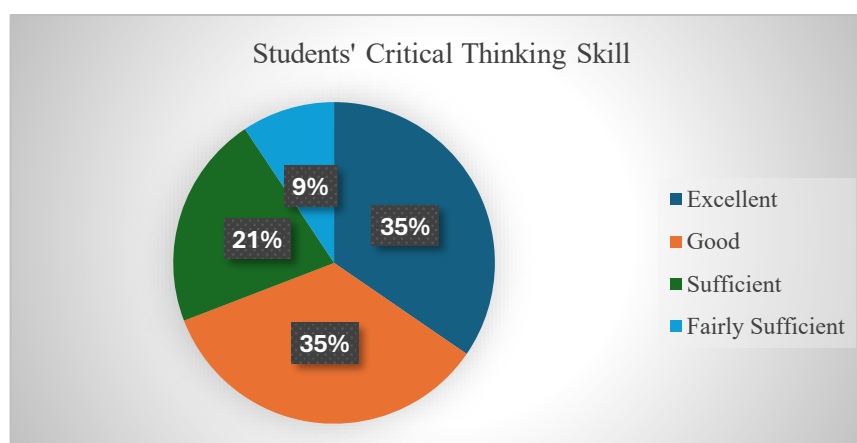


Figure 3. The result of students' critical thinking skill (CTS)

The students' CTS spreads from excellent to fairly sufficient category. The excellent and good category have the same slice, while fairly sufficient category has the smallest. To make the analysis of CTS easier, the researchers used SPSS to do the statistical analysis. The result is shown in the table 8.

Table 9. The result of descriptive analysis from SPSS

Statistics		
Critical Thinking		
N	Valid	107
	Missing	3
Mean		72.1869
Median		74.0000
Mode		84.00
Std. Deviation		12.44621
Variance		154.908
Range		52.00
Minimum		44.00
Maximum		96.00
Sum		7724.00

Table 9 shows there are 107 students involved in field implementation. The highest score is 96, while the lowest is 44. The mean score of is 72.2, and it is categorized in 'Good level, although some students were still in the sufficient and fairly sufficient level.

3.2 Discussion

The findings demonstrate a substantial need for a reading assessment instrument that explicitly incorporates critical thinking into EFL assessment practices in rural senior high schools. The high overall needs-analysis score ($M = 4.56$) indicates that teachers perceive existing reading assessments as insufficient for capturing students' higher-order thinking processes. More importantly, only 20% of the participating teachers reported using critical thinking indicators when constructing reading assessment

items. This finding suggests a gap between the increasing emphasis on higher-order thinking in language education and teachers' practical assessment literacy. Previous research has similarly indicated that teachers recognize the importance of higher-order thinking skills while experiencing difficulties related to their complexity and classroom implementation (Sukarni et al., 2022). Authentic assessment research in Indonesian ELT has also highlighted challenges in translating higher-order learning objectives into feasible classroom assessment practices (Suwartono & Riyani, 2019). Therefore, the need identified in this study should not be interpreted merely as a demand for an additional testing instrument; rather, it reflects a broader need to support teachers in operationalizing critical thinking constructs into measurable reading tasks.

The teachers' preference for combining multiple-choice and open-ended items further highlights the importance of using complementary assessment formats. Multiple-choice items provide efficient and relatively objective measurement across several aspects of reading comprehension, whereas open-ended tasks provide students with opportunities to explain reasoning, evaluate information, and support judgments with textual evidence. This combination is particularly relevant to critical thinking because such competence cannot be adequately represented by the selection of correct answers alone. Brookhart (2010) argues that higher-order thinking assessment should require learners to process, evaluate, and use information rather than merely reproduce previously learned knowledge. Similarly, Fajari (2021) found that open-ended questions can provide opportunities to assess critical-thinking processes that are less visible in conventional selected-response tasks. The mixed-format design adopted in this study—20 multiple-choice and five essay items—can therefore be understood as an attempt to balance scoring efficiency with opportunities to observe students' reasoning. In this respect, the instrument may function not only as a measurement tool but also as a practical model for teachers seeking to align reading assessment with higher-order learning objectives.

The development process also demonstrates the importance of grounding assessment tasks in an explicit theoretical framework. The instrument blueprint operationalized six critical thinking dimensions adapted from Facione (1990)—interpretation, analysis, inference, evaluation, explanation, and self-regulation—into reading-related objectives and item formats. This theoretically informed approach addresses concerns that critical thinking assessment may become overly general unless the targeted cognitive processes are clearly defined. Existing research has shown that Facione's dimensions provide a useful framework for examining learners' critical thinking profiles (Arini et al., 2023), while Orhan and Ay (2022) demonstrated that critical thinking constructs can be systematically translated into assessment instruments for secondary-school learners. In the present study, expert judgment indicated strong perceived appropriateness of the instrument in terms of indicator alignment, language clarity, contextual suitability, scoring rubrics, and its potential to elicit critical thinking. Nevertheless, expert judgment should be interpreted as evidence of content appropriateness rather than as definitive evidence of construct validity. Further validation with larger and more diverse samples would therefore be required to establish more robust evidence regarding the instrument's underlying construct structure.

The empirical item analysis provides additional evidence concerning the performance of the developed instrument. Twenty of the 25 items met the validity criterion applied in the limited trial, whereas five multiple-choice items required revision. Rather than being interpreted as a weakness of the development process, the identification of problematic items illustrates the importance of iterative evaluation within instrument development. Revising ambiguous wording, implausible distractors, and insufficient alignment with critical thinking indicators is consistent with the purpose of developmental testing: to identify weaknesses before an instrument is implemented with the target population. The reliability coefficients of .723 for the multiple-choice section and .693 for the essay section indicate that the two components achieved usable levels of internal consistency within this developmental context. These findings are broadly consistent with Orhan and Ay (2022), who emphasize that evidence of both validity and reliability is essential before a critical thinking assessment is used to interpret students' performance. However, because essay responses involve scorer judgment, internal consistency alone

does not provide complete reliability evidence. Future validation should therefore report inter-rater agreement explicitly and examine whether the scoring rubric produces consistent judgments across different raters and school contexts.

The implementation results provide an initial picture of how the instrument functions among its intended users rather than definitive evidence about the critical thinking ability of rural Indonesian students as a population. The 107 students obtained a mean score of 72.19 (SD = 12.45), with considerable variation in individual performance. This result is important because rural educational contexts are often discussed primarily in terms of resource constraints and differences in educational opportunities. Previous research has documented differences between rural and urban learners in educational conditions and learning outcomes (Fu & Liu, 2024; Ikhlas, 2020). However, the present findings should not be interpreted as demonstrating that rural students generally possess either high or low critical thinking ability, since the participants were drawn from only one school and no urban comparison group was included. Instead, the results demonstrate that the developed instrument was implementable in a rural EFL setting and was capable of differentiating students across a range of performance levels. This distinction is important because the principal contribution of an instrument-development study concerns the quality and usability of the assessment rather than broad generalizations about the population tested.

Taken together, the findings suggest that the main contribution of this study lies in connecting critical thinking theory with feasible classroom assessment practice in a context that remains underrepresented in EFL assessment research. The combination of theoretically mapped indicators, selected-response and constructed-response tasks, expert review, empirical item analysis, and implementation in a rural senior high school provides an initial framework for context-sensitive critical-thinking reading assessment. The instrument may also contribute to teachers' assessment literacy by illustrating how abstract critical thinking dimensions can be transformed into reading objectives, test items, and scoring criteria. Nevertheless, its current evidence base remains context-specific. Further studies should examine the instrument with larger multi-school samples, different reading genres, and students with varying English proficiency levels, while employing stronger evidence of construct validity and inter-rater reliability. Such validation would determine whether the instrument can be extended beyond its original rural context and establish its broader usefulness for critical-thinking-oriented EFL reading assessment in Indonesian secondary education.

4. CONCLUSION

This research demonstrates the importance of developing a critical thinking-based reading comprehension assessment instrument for high school students, particularly in rural areas. The research findings confirm that English teachers need an instrument that combines multiple-choice and essay questions, as this combination can provide a more comprehensive picture of students' critical thinking skills in reading comprehension.

Furthermore, the developed instrument met adequate quality criteria based on validity and reliability, making it suitable for use as a critical thinking assessment tool in descriptive reading texts. Implementation results also indicated that students' critical thinking skills were quite good, with an average score of 72.2, although some students still scored lower. Therefore, this instrument can be utilized by teachers in other schools, coupled with more intensive learning efforts to improve students' critical thinking skills, particularly in rural schools.

This study has several limitations. First, it only involved students from one rural school, so the results cannot be broadly generalized to other school contexts with different characteristics. Second, the instrument developed focused on critical thinking skills in reading comprehension of descriptive texts only. It does not cover other text types or other language skills. Third, the study sample size was relatively limited, to one school only. Therefore, further research is recommended involving a wider

sample, various text types, and implementation over a longer period to obtain more comprehensive results.

REFERENCES

- Adeoye, M. A., Wirawan, K. A. S. I., Pradnyani, M. S. S., & Septiarini, N. I. (2024). Revolutionizing education: Unleashing the power of the ADDIE model for effective teaching and learning. *JPI (Jurnal Pendidikan Indonesia)*, 13(1), 202–209.
- Arini, R., Rahayu, Y. S., & Erman, E. (2023). Profile of critical thinking results analyzed from facione indicators and gender of learners. *IJORER: International Journal of Recent Educational Research*, 4(4), 434–446.
- Brookhart, S. M. (2010). *Assess Higher-Order Thinking Skills in Your Classroom*. ASCD.
- Butler, H. A. (2024). Predicting everyday critical thinking: A review of critical thinking assessments. *Journal of Intelligence*, 12(2), 16.
- Chusni, M. M., Saputro, S., Suranto, S., & Rahardjo, S. B. (2020). Review of critical thinking skill in indonesia: Preparation of the 21st century learner. *Journal of Critical Reviews*, 7(9), 1230–1235.
- Fajari, L. E. W. (2021). Open-Ended Questions to Assess Critical-Thinking Skills in Indonesian Elementary School. *International Journal of Instruction*, 14(1), 615–630.
- Feng, J., & Sangsawang, T. (2023). Information technology, according to the ADDIE model on English subject teaching, enhances the learning achievement of Shunde Polytechnic students in China. *Turkish Online Journal of Educational Technology-TOJET*, 22(4), 121–131.
- Fu, H., & Liu, H. (2024). A comparative study of learners' conceptions of and approaches to learning English between high school students in urban and rural areas of China. *Frontiers in Psychology*, 15, 1324366.
- Fullan, M., Quinn, J., & McEachen, J. (2018). Praise for Deep Learning: Engage the World Change the World. *Corwin. Ontario Principals Council*.
- Ikhlas, M. (2020). Do a science process skills affect on critical thinking in science? Differences in urban and rural. *International Journal of Evaluation and Research in Education (IJERE)*.
- Indah, R. N., Toyyibah, T., Budhiningrum, A. S., & Afifi, N. (2022). The research competence, critical thinking skills and digital literacy of Indonesian EFL students. *Journal of Language Teaching and Research*, 13(2), 315–324.
- Jiang, R. (2022). Understanding, Investigating, and promoting deep learning in language education: A survey on chinese college students' deep learning in the online EFL teaching context. *Frontiers in Psychology*, 13, 955565.
- Kurniawati, N., Sugaryamah, D., & Hasanah, A. (2020). Proposing a model of critical literacy program for fostering Indonesian EFL students' critical thinking skills. *Journal of Education and Learning (EduLearn)*, 14(2), 234–247.
- Li, H., & Cheong, J. P. G. (2023). Using the ADDIE model to design and develop physical education lessons incorporated with a functional training component. *Frontiers in Public Health*, 11, 1201228.
- Liaw, S.-S., Huang, H.-M., & Chen, G.-D. (2007). Surveying instructor and learner attitudes toward e-learning. *Computers & Education*, 49(4), 1066–1080.
- Lisdawati, L., & Umam, A. (2022). Promoting HOTS and critical thinking to English students of higher education through authentic assessment. *Erudita: Journal of English Language Teaching*, 2(1), 15–25.
- Liu, D., & Wang, X. (2025). Design and Application of English Smart Classroom Teaching Based on Deep Learning. *International Journal of High-Speed Electronics and Systems*, 2540399.

- Mohamed, R., & Lebar, O. (2017). Authentic assessment in assessing higher order thinking skills. *International Journal of Academic Research in Business and Social Sciences*, 7(2), 466–476.
- Muthohharoh, S. R., Bharati, D. A. L., & Rozi, F. (2020). The Implementation of Authentic Assessment to Assess Students' Higher Order Thinking Skills in Writing at MAN 2 Tulungagung. *English Education Journal*, 10(3), 374–386.
- Oktaviah, R., Rokhman, N. M., & Maulana, A. R. (2021). Inquiry-based teaching to develop efl students' critical thinking in reading comprehension. *Education of English as Foreign Language*, 4(1), 49–56.
- Orhan, A., & Ay, Ş. Ç. (2022). Developing the critical thinking skill test for high school students: A validity and reliability study. *International Journal of Psychology and Educational Studies*, 9(1), 130–142.
- Paul, R., & Elder, L. (2006). Critical thinking: The nature of critical and creative thought. *Journal of Developmental Education*, 30(2), 34.
- Paulsen, V. H., & Kolstø, S. D. (2022). Students' reasoning when faced with test items of challenging aspects of critical thinking. *Thinking Skills and Creativity*, 43, 100969.
- Rustam, R., & Priyanto, P. (2022). Critical thinking assessment in the teaching of writing Indonesian scientific texts in high school. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 26(1), 12–25.
- Said, A. (2025). A Comparative Analysis of Assessment Practices and Their Impact on Student Achievement in Rural vs. Urban Government Secondary Schools in Islamabad. *Sjesr*, 8(2), 19–27.
- Saputri, I., Nurkamto, J., & Wahyuni, D. S. (2019). The implementation of authentic assessment in English language teaching. *English Education*, 6(3), 270–277.
- Sial, Z. A., Fatima, Z., & Fatima, S. Z. (2024). Addie model of instructional effectiveness: analyzing the impact on students learning. *Journal of Law Social Studies (JLSS)*, 6(1), 64–72.
- Sukarni, S., Sudar, S., & Tarwiyah, S. (2022). English Teachers' Perception of Higher Order Thinking Skills Viewed from Importance and Complexity. *Jurnal Pendidikan Progresif*, 12(1), 254–266.
- Supriyatno, T., Susilawati, S., & Hassan, A. (2020). E-learning development in improving students' critical thinking ability. *Cypriot Journal of Educational Sciences*, 15(5), 1099–1106.
- Suwartono, T., & Riyani, C. (2019). Authentic assessment in ELT: Hopes, challenges, and practices. *REFLEKSI EDUKATIKA: Jurnal Ilmiah Kependidikan*, 9(2).
- Yazdanmehr, E., & Akbari, R. (2015). Cognitive aspects of teacher expertise in ELT. *Journal of English Language Teaching and Learning*, 7(15), 157–180.