



Developing an Artificial Intelligence (AI)-Assisted Instrument for Assessing the Short Story Writing Skills of Grade XI Students at Senior High School

Andy Dahnial Pratama MZ¹, Ida Nur'aeni², Sukma³, Yunidar⁴, Juniati⁵

^{1,2,3,4,5}Master's Program in Indonesian Language Education, Universitas Tadulako, Palu, Indonesia

Article Info	Abstract
Received: 2026-10-12 Revised: 2026-10-16 Accepted: 2026-10-17 Keywords: 4D model; analytic rubric; artificial intelligence; assessment instrument; short story writing DOI: 10.24256/ideas.v14i2.12887 Corresponding Author: Andy Dahnial Pratama MZ andydahnial08@gmail.com Master's Program in Indonesian Language Education, Universitas Tadulako, Palu, Indonesia	<i>Assessing short story writing is prone to subjectivity when it is not supported by a valid and reliable instrument, while the use of artificial intelligence (AI) in drafting assessment instruments has rarely been documented. This study aims to describe the development process and the quality of an AI-assisted instrument for assessing the short story writing skills of grade XI students at SMA Al-Azhar Mandiri Palu. It followed the 4D development model (define, design, develop, disseminate), with AI (Claude Opus 5.5) serving as a drafting aid rather than as a rater of students' work. Data were collected through interviews, observation, document analysis, validation by five experts, the scoring of 30 short stories by two teachers, and a practicality questionnaire, and were analyzed using feasibility percentages, the intraclass correlation coefficient (ICC), Cronbach's alpha, percentage agreement, and practicality percentages. The AI generated 32 rubric descriptors from a single prompt. After the researcher's review, 12 descriptors were accepted unchanged, 17 were revised to make them observable and consistent with the quantitative requirement, and 3 were rejected because they relied on impressionistic criteria. The 12 descriptors accepted unchanged were later revised on the language validator's advice, so none of the AI-generated descriptors was used unchanged in the final product. The instrument obtained a feasibility percentage of 92.63% (very valid), a single-measures ICC of 0.758 (good; 95% CI 0.541–0.878), and a practicality percentage of 98.33% (very practical). Cronbach's alpha was low (–0.224 and 0.251) because scores clustered at the top of the scale, while exact inter-rater agreement reached 94.17%. The instrument is feasible for use, and AI is effective as a drafting aid provided that its output is reviewed by the researcher and validated by experts.</i>

1. Introduction

Writing short stories (*cerpen*) is a competence in Indonesian language learning at senior high school that combines mastery of intrinsic elements with the ability to use language and express ideas creatively. Mastery of intrinsic elements is closely related to students' ability to write short stories (Ahmad, Br Ginting, & Sidiqin, 2020). Because the outcome is a written product, this skill is assessed through performance assessment, in line with the principles of authentic assessment (Santi, Silvia, & Damaianti, 2023). Such assessment requires a valid and reliable instrument, because the assessment of written work is prone to subjectivity when it is not accompanied by clear assessment standards (Alfianika & Sitohang, 2022). Nurgiyantoro (2016) distinguishes holistic-impressionistic assessment, which is quick but prone to subjectivity, from analytic assessment, which breaks a composition down into several components and is therefore more focused and accountable.

The same problem was found at SMA Al-Azhar Mandiri Palu. A preliminary study in the form of an interview with an Indonesian language teacher showed that the teacher found it difficult to assess creativity fairly, felt burdened by the number of intrinsic elements to be assessed, especially language style, and acknowledged that not all intrinsic elements were covered by the existing assessment. Scoring consistency also relied on shared views among Indonesian language teachers rather than on a standardized instrument. Such assessment may produce unfair scores, low inter-rater consistency, and feedback that is not specific enough to support students' writing development.

Many efforts have been made to develop instruments for assessing writing skills. Santoso, Suyono, Harsiati, and Sari (2023) implemented a short story writing assessment instrument using inspirational songs for grade XI senior high school students; Alfianika and Sitohang (2022) developed rubrics for narrative and descriptive paragraph writing with validity levels of 95% and 96.25%; and Hikmah (2021) designed an instrument for assessing expository text writing based on a construct of micro- and macro-skills. The need for such instruments is also evident in training programs on developing short story writing assessment instruments for teachers (Sauhenda, Marnina, & Narahawarin, 2025).

These studies produced feasible instruments, but all of them were developed through conventional procedures without involving AI. Yet developing an analytic rubric requires care and time, particularly in formulating descriptions of performance levels. Brookhart (2018) emphasizes that such descriptions should be written in descriptive language that depicts what work at each level looks like, rather than in evaluative language such as "excellent" or "poor".

The development of artificial intelligence (AI), particularly large language models, opens new opportunities in assessment. Oktavianus, Naibaho, and Rantung (2023) and Muhajir et al. (2026) show that AI can support learning and assessment through fast and systematic processing, while Hanis and Wahyudin (2024) found that elementary school teachers have begun to use AI to plan assessments, although not yet optimally.

In a number of international studies, AI is mainly positioned as a rater or a feedback provider. Mizumoto and Eguchi (2023) used a GPT model to score 12,100 essays from the TOEFL11 corpus and found a level of accuracy and reliability that could support human evaluation, whereas Steiss et al. (2024) found that human feedback on students' writing was of higher quality than ChatGPT feedback in most categories. Kasneci et al. (2023)

stress that using large language models in education requires human oversight and the ability to understand the limitations of their output.

These studies show that the development of writing assessment instruments and the use of AI in assessment are still discussed separately. To the best of the authors' knowledge, no study has developed an instrument for assessing short story writing skills using AI as an aid in the development process. This gap matters because studies of AI as a rater have not answered a more fundamental question for teachers: how feasible are the instrument formulations produced by AI, which parts need to be changed by humans, and what is the quality of the instrument after validation and trial? Without such documentation, teachers have no basis for judging how far AI output can be relied on when developing assessment instruments.

Based on this gap, this study addresses two questions: (1) how is an AI-assisted instrument for assessing the short story writing skills of grade XI students at SMA Al-Azhar Mandiri Palu developed, and (2) what is the quality of the instrument in terms of validity, reliability, and practicality? The novelty of this study lies in combining the two strands of research through the development of an analytic rubric whose descriptors were drafted with AI assistance and whose every change was documented, from the first AI output to the final product. This documentation clearly separates the roles of AI and humans, yielding a replicable procedure for using AI in developing assessment instruments.

2. Method

Research Design

This study employed a research and development (R&D) design to produce an assessment instrument and test its quality (Waruwu, 2024). Development followed the 4D model, which consists of the define, design, develop, and disseminate stages, because its stages are systematic and widely used for developing assessment instruments (Maydiantoro, 2021; Sihombing, Zamsiswaya, & Sawaluddin, 2024). The approach was quantitative, supported by qualitative data in the form of suggestions from validators and teachers. The AI used was Claude, specifically the Claude Opus 5.5 model (Anthropic, 2026), which served as an aid for drafting the instrument rather than as a rater of students' work.

Participants, Setting, and Time

The study was conducted at SMA Al-Azhar Mandiri Palu, Palu City, Central Sulawesi, from June to October 2026. The participants were five expert validators, two Indonesian language teachers who acted as both raters and users of the instrument, and 30 students of class XI E in the 2026/2027 academic year whose short stories were scored in the trial. The five validators comprised one assessment expert (V1), two language experts (V2 and V3), and two curriculum experts (V4 and V5). Four validators were from the Faculty of Teacher Training and Education, Universitas Tadulako, while V4 was an Indonesian language teacher at the school and thus represented a practitioner's perspective. The number of raters was set at two because the school has two Indonesian language teachers who teach grade XI; this limitation was offset by the number of works scored, namely 30.

Development Procedure

The define stage comprised an interview with the teacher, a review of the teaching module and the rubric used by the teacher, observation of assessment practice, analysis of student characteristics, content analysis, and formulation of development objectives. The design stage comprised developing a blueprint of eight aspects, namely theme, plot, characters and characterization, setting, point of view, language style, moral message, and language conventions; selecting an analytic rubric format with four score levels (4 = very good, 3 = good, 2 = fair, 1 = poor); drafting descriptors with AI assistance; and assembling the complete instrument into Draft I.

The prompt to the AI had five components: the AI's role as an expert in Indonesian language assessment; the context and constraint that the AI would only help develop the instrument; the task of developing an analytic rubric for eight aspects with four score levels; descriptor requirements; and an output format in the form of a table. The descriptor requirements specified (a) one or two plain sentences for each level, (b) differentiation of levels by the degree to which the indicator is achieved and, where possible, by features observable in the text, (c) each level being clearly different from the adjacent levels, with each feature assessed in only one aspect, and (d) the use of the average number of spelling, diction, or sentence errors per paragraph for the language conventions aspect.

The develop stage comprised expert validation, Revision I based on the validators' suggestions, a limited trial, and Revision II based on the trial results, while the disseminate stage was carried out on a limited scale by introducing the final product and handing it over to the teachers. All prompts and AI outputs were documented in conversation transcripts. The activities and their timing are presented in Table 1.

Table 1. Stages, Activities, and Timeline of the Study

Stage	Activity	Date
Define	Teacher interview and observation of assessment practice in class XI E	Monday, 15 June 2026
Design	Drafting rubric descriptors with AI assistance (session 1)	Tuesday, 25 August 2026
Design	Researcher's review of the AI output and AI-assisted assembly of the complete instrument (session 2)	25–29 August 2026
Develop	Short story writing by 30 students of class XI E as trial material	Monday, 31 August – Friday, 11 September 2026
Develop	Validation by V1, V2, V3, and V5; validation by V4	Monday, 14 September and Friday, 18 September 2026
Develop	AI-assisted Revision I based on the validators' suggestions (session 3)	Sunday, 20 September 2026
Develop	Limited trial: scoring of 30 short stories by two raters	Monday–Thursday, 21–24 September 2026

Stage	Activity	Date
Develop	Completion of the practicality questionnaire by both raters	Friday, 25 September 2026
Develop	Revision II based on the trial results	After the trial, before dissemination
Disseminate	Introduction and handover of the final product to the teachers	Monday, 5 October 2026

Instruments and Data Collection

Data were collected through interviews, observation, document analysis, questionnaires, a writing performance test, and documentation. The validation sheets used a 1–5 scale and were prepared separately for each field of expertise by adapting the feasibility components of the Indonesian National Education Standards Agency (BSNP), namely content, presentation, language, and graphic feasibility. Because the product was an assessment instrument, an assessment construction aspect was added, covering descriptor clarity, gradation between score levels, scoring guidelines, and descriptor specificity, with reference to the requirements of a good instrument according to Arikunto (2018).

The assessment expert rated 10 items, while each language expert and curriculum expert rated 7 items. In the trial, the two teachers scored the same 30 works independently at separate times or places, without discussion and without knowing each other's scores, accompanied by the researcher. Practicality was measured with a teacher response questionnaire on a 1–4 Likert scale consisting of 15 items in three aspects: ease of use, scoring time efficiency, and clarity and readability of the rubric.

Data Analysis

The status of each AI-generated descriptor was determined by comparing the AI's wording with that of Draft I, using three categories: accepted unchanged (used without modification), revised (the distinguishing criterion was retained, but the wording was clarified, completed, or made more observable), and rejected (the distinguishing criterion was not used and was replaced by a new formulation based on a different criterion). Changes in Revision I and Revision II were recorded in the same way, so that the wording of each descriptor could be traced across the four versions of the instrument.

Validity was analyzed using a feasibility percentage, namely the total validator score divided by the maximum score and multiplied by 100%, with a minimum criterion of 61% or the valid category (Maulana, 2022). Inter-rater reliability was analyzed using the intraclass correlation coefficient (ICC), two-way random model, absolute agreement type, on total scores. The single-measures coefficient was used as the basis for decision because in practice students' work is usually scored by one teacher, with a minimum criterion of 0.75 (good); the 95% confidence interval was reported to show the precision of the estimate (Koo & Li, 2016).

As supporting information, internal consistency was computed with Cronbach's alpha separately for each rater, with the eight aspects as items and 0.70 as the reference (Rasyid, Hartanto, & Priyanto, 2026), while exact inter-rater agreement was computed for

each aspect to identify the aspect most prone to divergent interpretation, which served as the basis for Revision II. Practicality was analyzed using the percentage of questionnaire scores, with a minimum criterion of 61% or the practical category (Maulana, 2022). Statistical calculations were performed in Python with the assistance of Claude and checked again by the researcher. The instrument was declared feasible if it met the valid, reliable, and practical criteria.

3. Results

Define Stage: Needs for Instrument Development

The define stage yielded findings from three mutually reinforcing sources. The interview on 15 June 2026 showed that the teacher found creativity the hardest to assess fairly, felt that there were too many intrinsic elements to assess, especially language style, and stated that the intrinsic elements had not been fully assessed. The teacher had never used AI for teaching or assessment but believed that AI could help develop a rubric as long as its output was adjusted to actual needs. The review of the teaching module showed that the teacher's rubric contained five aspects with four score levels: relevance of the theme to real-life problems, building elements of the short story, plot and conflict, language and style, and completeness and independence.

Characters and characterization, setting, point of view, and moral message were merged into a single aspect, while language style was combined with language conventions. Its descriptors were short phrases relying on degree words, such as "most elements are present" (*sebagian besar unsur hadir*), and the rubric had no blueprint, scoring guidelines, or validity and reliability testing. Observation in class XI E in the 2025/2026 academic year confirmed that the assessment did not cover all seven intrinsic elements and language conventions, scores were not determined on the basis of clear descriptors, and creativity was not assessed with clear criteria.

Based on these findings and the analysis of student characteristics and content, the development objective was formulated as follows: to produce an instrument consisting of a blueprint, an analytic rubric, and scoring guidelines that assess eight aspects, namely the seven intrinsic elements according to Nurgiyantoro (2018) and a language conventions aspect covering spelling, diction, and sentence effectiveness, with observable descriptors at every performance level.

Design Stage: AI Output and the Researcher's Review

The AI generated 32 descriptors (8 aspects $\times$ 4 levels) from a single prompt on 25 August 2026. The output met the format requirements but did not fully meet the content requirements. First, several descriptors consisted only of negations or degree words without observable features, for example "The theme of the story is not very clear" (*Tema cerita kurang jelas*) and "The theme of the story is unclear" (*Tema cerita tidak jelas*) for scores 2 and 1. Second, the requirement to use the average number of errors per paragraph in the language conventions aspect was applied only to scores 4 and 2, while scores 3 and 1 still used "some errors" (*beberapa kesalahan*) and "many errors" (*banyak kesalahan*). Third, all language style descriptors relied on the degree of being "interesting" (*menarik*), for example "The use of language and figures of speech is fairly interesting" (*Penggunaan bahasa dan majas cukup menarik*), even though language style was among the elements that the teacher found hardest to assess fairly.

The researcher's review of each descriptor (Table 2) showed that only 12 of the 32 descriptors (37.5%) were accepted unchanged, while 20 descriptors (62.5%) were changed by the researcher. The average length of the descriptors, counted in the original Indonesian wording, increased from 9.50 to 13.25 words. The changes followed four patterns: (1) adding an observable consequence or feature to descriptors that consisted only of negations or degree words, for example "The theme of the story is not very clear" became "The theme of the story is not very clear so that the main idea of the story is hard to identify"; (2) clarifying the scope of the elements observed, for example the setting descriptor for score 3, "The setting is described well, but it is not quite complete", became "The setting of place, time, and atmosphere is depicted, but one of its elements is insufficiently described"; (3) applying the quantitative requirement that the AI had not followed in the language conventions aspect; and (4) replacing impressionistic criteria in the language style aspect with criteria of accuracy, variety, and the effect of language and figures of speech on the story. In the language style descriptor for score 4, the element of variety from the AI's wording was retained, so it was categorized as revised, whereas for scores 3, 2, and 1 the criterion of being "interesting" was replaced entirely, so they were categorized as rejected.

The researcher then asked the AI to assemble the complete instrument from a template prepared by the researcher, which consisted of six parts: instructions for use, blueprint, rubric, scoring and score conversion guidelines, individual scoring sheet, and class score recap sheet. The AI was asked to copy the template word for word and to fill in the rubric section with the reviewed descriptors. A review of the transcript showed that the AI changed one sentence interpreting the 71–85 score range without being asked, from *Sebagian besar aspek tergarap dengan baik dengan beberapa kekurangan kecil* ("Most aspects are developed well, with a few minor shortcomings") to *Sebagian besar aspek tergarap baik dengan beberapa kelemahan kecil* ("Most aspects are well developed, with a few minor weaknesses"), even though the AI stated that the entire text was identical to the template. The assembled document became Draft I.

Table 2. Status of AI-Generated Descriptors after the Researcher's Review

Aspect	Accepted unchanged	Revised by the researcher	Rejected	Revised on the validator's advice*
Theme	1	3	0	1
Plot	2	2	0	2
Characters and characterization	3	1	0	3
Setting	1	3	0	1
Point of view	1	3	0	1
Language style	0	1	3	0
Moral message	2	2	0	2

Aspect	Accepted unchanged	Revised by the researcher	Rejected	Revised on the validator's advice*
Language conventions	2	2	0	2
Total	12 (37.5%)	17 (53.1%)	3 (9.4%)	12 (37.5%)

*Recorded in Revision I. Percentages are calculated from 32 descriptors.

Develop Stage: Expert Validation and Revision I

Draft I was validated by the five experts on 14 and 18 September 2026. All validators gave percentages in the very valid category, ranging from 85.71% to 100%, and the overall percentage of 92.63% exceeded the minimum criterion of 61% (Table 3). By feasibility aspect, the highest percentage was for graphic feasibility (100%), followed by presentation feasibility and assessment construction feasibility (95.00% each), content feasibility (92.50%), and language feasibility (87.50%). No item received a score below 4, so no item needed improvement on the basis of its score.

Table 3. Expert Validation Results

Validator	Field of expertise	Score	Percentage	Category	Validator's conclusion
V1	Assessment	47/50	94.00%	Very valid	Feasible with revision
V2	Language	34/35	97.14%	Very valid	Feasible with revision
V3	Language	30/35	85.71%	Very valid	Feasible with revision
V4	Curriculum (practitioner)	30/35	85.71%	Very valid	Feasible with revision
V5	Curriculum	35/35	100%	Very valid	Feasible without revision
Overall		176/190	92.63%	Very valid	–

Four validators declared the instrument feasible for use with revision. V1 and V2 suggested wording improvements to the instructions for use and the introductory sentence of the rubric, V3 asked for improvements in sentence structure and word choice by underlining the parts to be revised, and V4 asked for an explicit statement that AI served only as an aid and that the short stories had to be the students' own work. Revision I was carried out by the AI at the researcher's command on 20 September 2026, with instructions to revise only the parts suggested, to copy the other parts exactly as they were, and to present a table of changes with reasons.

The result was 19 changes: 2 items in the instructions for use (including a new item on the role of AI), 1 introductory sentence of the rubric, 12 descriptors, and 4 sentences interpreting final scores. A text comparison showed no other changes outside this list, except for the addition of learning outcomes and learning objectives to the identification section of the instrument, made by the researcher.

Tracing the status of the descriptors revealed an important finding. All twelve descriptors revised on V3's advice were descriptors previously accepted unchanged from

the AI, whereas none of the 20 descriptors edited by the researcher fell within the parts underlined by V3. Some of these revisions aligned the sentence structure with the researcher-edited descriptors in the same aspect, for example by adding the word *digambarkan* ('depicted') to the characters and characterization descriptors for scores 3, 2, and 1 to match the descriptor for score 4.

Limited Trial: Reliability and Practicality

The trial was conducted on 21–24 September 2026 by scoring 30 short stories that the students had written between 31 August and 11 September 2026. The mean total scores given by rater 1 and rater 2 were 30.47 and 30.20, respectively, out of a maximum of 32, with mean final scores of 95.21 and 94.38. Scores clustered at the top of the scale. For theme, characters and characterization, setting, point of view, and moral message, more than 86% of the works received a score of 4 from both raters, whereas variation in scores appeared only in plot, language style, and language conventions (Table 5). In other words, in this sample the four-level scale effectively functioned as a two-level scale, namely scores 3 and 4. The reliability results are presented in Table 4.

Table 4. Reliability Test Results

Measure	Coefficient	95% CI	Remarks
ICC, <i>single measures</i>	0.758	0.541–0.878	Good; $F(29, 29) = 7.981$; $p < 0.001$
ICC, <i>average measures</i>	0.862	0.702–0.935	Good
Cronbach's alpha, rater 1	–0.224	–	Negative; $\Sigma\sigma_b^2 = 0.968 > \sigma_t^2 = 0.809$
Cronbach's alpha, rater 2	0.251	–	Low; $\Sigma\sigma_b^2 = 1.044$; $\sigma_t^2 = 1.338$

The ICC was computed with a two-way random model, absolute agreement type, on total scores ($n = 30$; $k = 2$); $MSR = 1.908$; $MSC = 1.067$; $MSE = 0.239$. Alpha was computed with the eight aspects as items.

The single-measures ICC of 0.758 falls in the good category and meets the minimum criterion of 0.75. However, it is only slightly above the threshold, and its confidence interval (0.541–0.878) indicates moderate to good reliability (Koo & Li, 2016). The two raters' total scores were identical for 17 works (56.7%) and differed by at most one point for 29 works (96.7%). Of the 14 aspect scores that differed, 11 showed rater 1 giving the higher score. This small systematic tendency is still taken into account by the absolute-agreement ICC.

Both Cronbach's alpha coefficients were far below the 0.70 reference, and the coefficient for rater 1 was even negative. The negative value is not a calculation error. Alpha becomes negative when the sum of the item variances exceeds the variance of the total score, which means that the average covariance between aspects is negative (–0.003), or nearly zero. For rater 1, 15 of the 30 works scored below 4 on only one aspect, and that aspect differed from work to work, so 20 of the 28 aspect pairs had negative covariances. The score variances of the five aspects that almost always received a score of 4 were also very small (0.033–0.120).

Table 5. Mean Scores, Number of Works Scoring 4, and Inter-Rater Agreement by Aspect

Aspect	Mean score		Works scoring 4		Exact agreement	
	R1	R2	R1	R2	<i>n</i>	%
Theme	3.93	3.93	28	28	30	100
Plot	3.67	3.63	20	20	29	96.67
Characters and characterization	3.93	3.87	28	26	28	93.33
Setting	3.97	3.97	29	29	30	100
Point of view	3.93	3.97	28	29	29	96.67
Language style	3.67	3.60	20	18	28	93.33
Moral message	3.97	3.97	29	29	28	93.33
Language conventions	3.40	3.27	12	8	24	80.00
Overall	30.47*	30.20*	194	187	226/240	94.17

R1 = rater 1; R2 = rater 2. *Mean total score (maximum 32). In the overall row, the columns for works scoring 4 show the number of scores of 4 among the 240 aspect scores (80.8% and 77.9%). Neither rater gave a score of 1, and the only score of 2 was given by rater 2 for plot.

Exact inter-rater agreement reached 94.17% of the 240 aspect scores. Language conventions had the lowest agreement (80.00%), with six works scored differently, and all of these differences occurred at the boundary between score 4 (an average of at most two errors per paragraph) and score 3 (an average of three to five errors per paragraph). Conversely, the 100% agreement for theme and setting occurred mostly at the highest score, because almost all works received a score of 4 from both raters.

The questionnaire completed by both raters on 25 September 2026 yielded a practicality percentage of 98.33% (very practical), with 97.92% for ease of use, 96.88% for scoring time efficiency, and 100% for clarity and readability of the rubric. A score of 3 was given for only two items: the item on using the instrument without special training (rater 1) and the item on the time needed to score one work (rater 2). Neither rater reported any obstacles or suggestions.

Revision II and Dissemination

Because the questionnaire contained no suggestions, Revision II was based on the agreement data by aspect. It targeted the language conventions aspect, taking into account two notes that the AI had made beyond its instructions in Revision I: the descriptor for score 1 did not yet contain the phrase “on average” (*rata-rata*), and the ranges “at most two” and “three to five” left a gap for fractional averages, such as 2.5 errors per paragraph. The researcher changed the range for score 4 to “fewer than three” and the range for score 3 to “three to fewer than six” errors per paragraph, added “on average” to score 1, and added an item to the instructions for use explaining how to calculate the average number of errors, namely by summing all spelling, diction, and sentence errors in the work and

dividing the total by the number of paragraphs. This revision did not change the aspects, the score scale, or the validated error limits. The final product was handed over to the Indonesian language teachers at SMA Al-Azhar Mandiri Palu on 5 October 2026 as a limited dissemination.

Tracing the Descriptors from AI Output to the Final Product

Tracing the four versions of the instrument shows that the role of AI changed from stage to stage. The AI drafted the descriptors in the first version, merely assembled the document from the researcher's template in Draft I, and carried out the changes derived from the validators' suggestions in Revision I, whereas Revision II was carried out entirely by the researcher. The wording that came from the AI was also limited to the rubric descriptors. The other parts of the instrument, namely the instructions for use, blueprint, final score formula, score interpretation sentences, and scoring sheets, came from the researcher's template. Examples of descriptor wording along each path of change are presented in Table 6, while the path of all 32 descriptors from the AI output to the final product is shown in Figure 1.

Table 6. Examples of the Descriptor Trail from AI Output to the Final Product

Aspect (score)	Version	Descriptor wording (translated)	Remarks
Theme (score 2)	AI output	The theme of the story is not very clear.	Revised in Draft I.
	Draft I	The theme of the story is not very clear so that the main idea of the story is hard to identify.	Observable consequence added.
Characters and characterization (score 3)	AI output	The characters are fairly clear and their traits are portrayed, although there is a slight inconsistency.	Accepted unchanged in Draft I.
	Revision I	The characters are depicted fairly clearly and their traits are portrayed, although there is a slight inconsistency of traits in certain parts.	V3: aligned with score 4 (<i>digambarkan</i> , 'depicted'); inconsistency specified.
Setting (score 3)	AI output	The setting is described well, but it is not quite complete.	Revised in Draft I.
	Draft I	The setting of place, time, and atmosphere is depicted, but one of its elements is insufficiently described.	Elements of setting specified; limit of the shortcoming set.
Language style (score 3)	AI output	The use of language and figures of speech is fairly interesting.	Rejected in Draft I.
	Draft I	Language and figures of speech are used fairly accurately, although their variety is still limited.	"Fairly interesting" replaced with accuracy and variety.

Aspect (score)	Version	Descriptor wording (translated)	Remarks
Language conventions (score 4)	AI output	Spelling, diction, and sentences are used accurately and effectively; on average, errors do not exceed two per paragraph and do not interfere with meaning.	Accepted unchanged in Draft I.
	Revision I	Spelling, diction, and sentences are used accurately and effectively; on average, there are at most two errors per paragraph and the meaning is easy to understand.	V3: structure aligned with score 3; error limit unchanged.
	Revision I I	Spelling, diction, and sentences are used accurately and effectively; on average, there are fewer than three errors per paragraph and the meaning is easy to understand.	Trial results: “at most two” became “fewer than three”, closing the gap for fractional averages.
Language conventions (score 3)	AI output	Spelling, diction, and sentences are fairly accurate, with some errors, but the meaning can still be understood.	Revised in Draft I.
	Draft I	Spelling, diction, and sentences are fairly accurate; on average, there are three to five errors per paragraph, but the meaning remains understandable.	Quantitative requirement applied.
	Revision I I	Spelling, diction, and sentences are fairly accurate; on average, there are three to fewer than six errors per paragraph, but the meaning remains understandable.	Trial results: “three to five” became “three to fewer than six”.

Descriptors are translated from Indonesian. Bold marks words or phrases changed from the previous version. Versions that did not change the wording are omitted; the last row of each descriptor is the wording in the final product.

Table 6 and Figure 1 show that none of the 32 AI-generated descriptors survived unchanged to the final product. Twenty descriptors were changed by the researcher in Draft I, the twelve descriptors initially accepted unchanged were changed on V3’s advice in Revision I, and three language conventions descriptors were changed again in Revision II. The average descriptor length in the original Indonesian increased from 9.50 words in the AI output to 13.25 words in Draft I, 13.75 words in Revision I, and 13.84 words in the final product, in line with the dominant direction of change, namely adding observable features or consequences.

Language style was the aspect most changed by the researcher, characters and characterization the aspect most changed on the validator’s advice, and language conventions the only aspect changed at all three stages, with seven changes in total. What survived from the AI output was the core criterion of most descriptors, such as clarity and consistency for theme and coherence and cause-and-effect relationships for plot, while

the framework of eight aspects and four score levels had already been set by the researcher in the prompt.

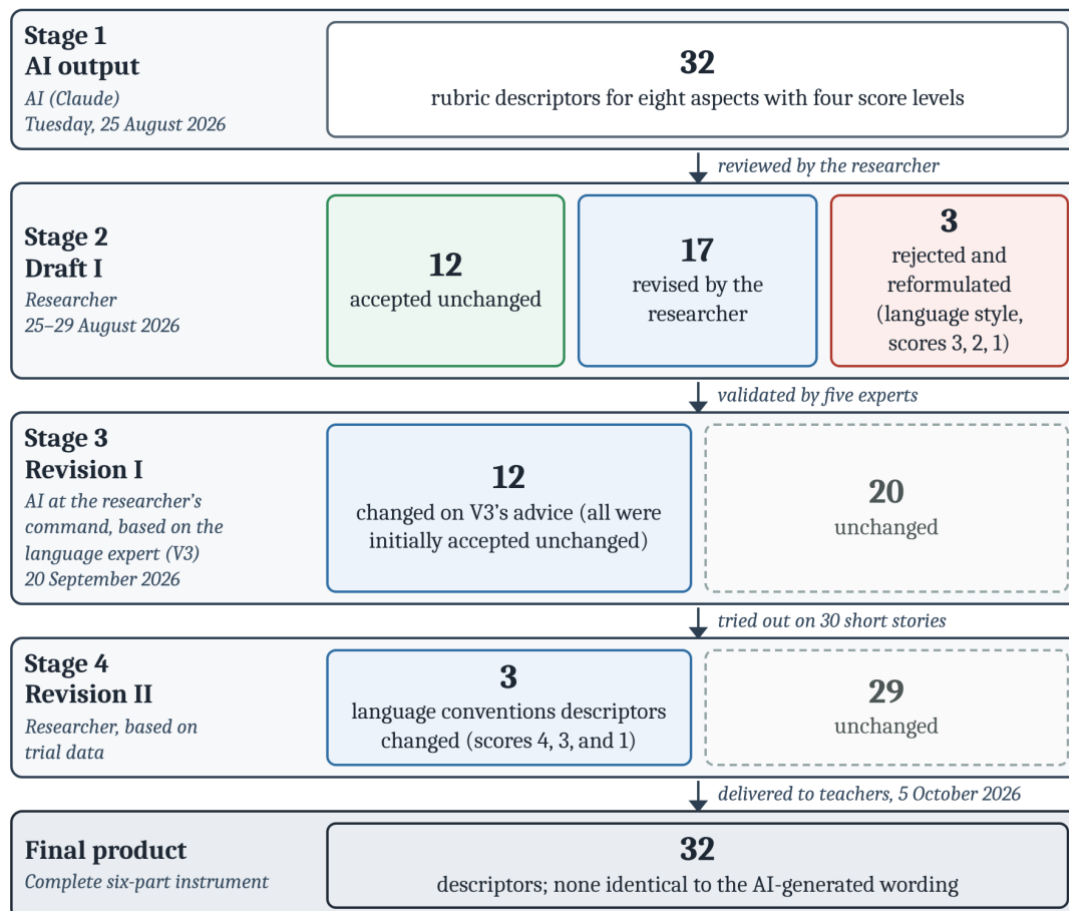


Figure 1. Changes to the Rubric Descriptors from AI Output to the Final Product

4. Discussion

The Role of AI and Human Control in Instrument Development

The findings show that AI made a real contribution in terms of speed and orderliness. A single prompt produced 32 descriptors with a uniform structure, the six-part instrument was assembled in one session, and Revision I produced a table of changes with reasons, so that every change could be traced. The AI even flagged two weaknesses in the language conventions wording beyond its instructions, which later proved relevant to the trial data. This contribution is consistent with Oktavianus et al. (2023) and Muhajir et al. (2026), who show that AI can support assessment through fast and systematic processing.

However, the data also reveal concrete limits of AI. As many as 62.5% of the AI-generated descriptors had to be changed by the researcher, and none of them was used unchanged in the final product. The AI did not fully comply with explicit instructions, either the quantitative requirement in the language conventions aspect or the instruction to copy the template, while it stated that its output was identical to the template. The most important finding is the AI's tendency to formulate impressionistic criteria. All language style descriptors relied on the degree of being "interesting", whereas descriptions of

performance levels in a rubric should depict what the work looks like at each level rather than give evaluative judgments (Brookhart, 2018). Had this output been accepted without screening, an instrument intended to reduce subjectivity would have reintroduced impression-based scoring, which, according to Nurgiyantoro (2016), is quick but prone to subjectivity. This tendency appeared precisely in the element that the teacher found hardest to assess fairly.

The findings from the validation stage reinforce the importance of human editing. All descriptors revised on the language validator’s advice were descriptors that had been accepted unchanged from the AI, and some revisions were needed because unedited AI wording sat alongside researcher-edited wording within the same aspect. This means that even AI wording that appears adequate is not necessarily ready to use and still requires linguistic review.

This finding supports the view of Hanis and Wahyudin (2024) and Kasneci et al. (2023) that the use of AI in education requires human control, while providing empirical evidence of where such control is most needed, namely the observability of criteria, compliance with requirements, and linguistic consistency across descriptors. Whereas previous studies tested the quality of AI as a rater or a feedback provider (Mizumoto & Eguchi, 2023; Steiss et al., 2024), this study shows that at an earlier stage, namely the formulation of assessment criteria, AI output also needs testing and editing. Based on this experience, the procedure for using AI in developing assessment instruments is summarized in Table 7.

Table 7. Procedure for Using AI in Developing Assessment Instruments

No.	Step	Role of AI	Role of humans
1	Setting the aspects, indicators, and descriptor requirements	–	The researcher formulates them based on the analysis of needs, documents, students, and content.
2	Writing the prompt	–	The researcher writes a prompt containing the role, context and constraints, task, descriptor requirements, and output format.
3	Generating draft descriptors	Drafts descriptors for each aspect and score level.	The researcher saves the transcript of prompts and AI outputs.
4	Reviewing the AI output	–	The researcher assigns a status to each descriptor (accepted unchanged, revised, or rejected) with reasons.
5	Assembling the complete instrument	Copies the researcher’s template and formats the document.	The researcher compares the assembled document with the template word for word.
6	Expert validation	–	Validators assess feasibility and give suggestions.

No.	Step	Role of AI	Role of humans
7	Revision based on the validators' suggestions	Revises the suggested parts and presents a table of changes with reasons.	The researcher verifies each change and ensures there are no changes beyond the suggestions.
8	Trial and quality analysis	Assists with statistical calculations.	The researcher determines the analysis techniques, checks the calculations, and interprets them.
9	Revision based on trial data	Can help formulate alternative improvements.	The researcher decides on improvements based on inter-rater agreement data and user input.

This procedure is the main contribution of this study to the identified gap. Previous studies developed writing assessment instruments through conventional procedures (Alfianika & Sitohang, 2022; Hikmah, 2021; Santoso et al., 2023), whereas studies on the use of AI in assessment have focused on assessment planning (Hanis & Wahyudin, 2024). This study combines the two and documents every decision, so that the roles of AI and humans can be clearly separated. Two practical lessons can be drawn. First, prompts to AI should always request a table of changes so that every change can be checked. Second, the AI's claim that its output complies with the instructions needs to be verified independently, even for a copying task.

Quality of the Developed Instrument

Validity

The feasibility percentage of 92.63% is comparable to the findings of Alfianika and Sitohang (2022), who obtained validity levels of 95% and 96.25% for narrative and descriptive paragraph writing rubrics. The difference is that validation in this study involved five validators from three fields of expertise, including a practitioner, so the feasibility of the instrument was examined from the perspectives of assessment, language and literature, and curriculum. This validity is content validity based on expert judgment. Content validity indicates that the design of the instrument is feasible but does not guarantee that the instrument works well in practice; hence, the trial remains necessary.

Reliability and Inter-Rater Agreement

The single-measures ICC of 0.758 indicates that two teachers scoring independently with this instrument produced relatively consistent scores. This addresses the initial problem that scoring consistency had previously relied only on shared views among teachers. Nevertheless, the conclusion that the instrument is reliable should be read with caution, because the coefficient is only slightly above 0.75 and the lower bound of its confidence interval falls in the moderate category. Koo and Li (2016) recommend interpreting reliability levels on the basis of the confidence interval rather than the point estimate alone, so the reliability of the instrument is more accurately described as moderate to good. The higher average-measures coefficient (0.862) indicates that for high-stakes decisions, the mean score of two raters will produce more consistent results.

The low Cronbach's alpha should be interpreted carefully. Alpha is a property of the scores from a particular sample and can be lowered by a small number of items, weak interrelatedness among items, or a heterogeneous construct (Tavakol & Dennick, 2011). All three conditions are present in this study: there are only eight items, the narrow score distribution weakens the interrelatedness among items, and the analytic rubric is designed to assess different elements of the short story separately (Nurgiyantoro, 2016). The score pattern also shows no halo effect, because the raters did not give uniform scores across all aspects; the language conventions aspect, for example, consistently received lower scores. Nevertheless, the low alpha means that these data cannot yet show that the eight aspects work in harmony when summed into a total score, so this needs to be retested with a sample of more varied ability.

The high agreement percentages in most aspects should also be interpreted together with their score distributions. Feinstein and Cicchetti (1990) show that percentage agreement can be high largely because of chance when scores are concentrated in one category. In contrast, the language conventions aspect, which received the fewest scores of 4, showed the lowest agreement, with all differences occurring at the boundary between scores 4 and 3. This pattern shows that quantitative descriptors do not fully align raters' interpretations when the ranges leave gaps and the method of counting errors is not explained. Revision II has closed these gaps, but its effectiveness still needs to be demonstrated in a further trial. It should also be noted that count-based criteria make it easier to align scores, but Brookhart (2018) cautions that rubrics that count elements tend to be more useful for grading than for learning. Therefore, scores for the language conventions aspect should be accompanied by feedback on the dominant types of errors.

Practicality

The practicality percentage of 98.33% indicates that the teachers perceived the instrument as easy to use, efficient, and clear and readable, consistent with Maulana's (2022) view that the practicality of a developed product is measured through user responses. The two scores of 3 suggest that the use of the instrument could still benefit from a brief introduction before scoring and that counting language errors per paragraph takes time. Because the questionnaire was completed by the two teachers who also served as raters, response bias cannot be ruled out.

Ceiling Effect and Discriminating Power of the Instrument

The very high student scores, with means of 95.21 and 94.38, have at least three possible explanations. First, the works scored were the product of problem-based learning that included outlining, drafting, peer editing, and revision, so they had already been improved several times. Second, the score 4 descriptors for some aspects, such as setting, which requires place, time, and atmosphere to be depicted, appear to be attainable by most students and therefore do not yet distinguish good work from very good work. Third, one of the raters was the homeroom teacher who knew the students, so the possibility of leniency in scoring cannot be ignored. These possibilities cannot be confirmed with the available data, but the implication is clear: the ability of the instrument to discriminate among high-achieving students has not yet been tested.

5. Conclusion

The AI-assisted instrument for assessing the short story writing skills of grade XI students at SMA Al-Azhar Mandiri Palu was developed through the four stages of the 4D model. AI served effectively as a drafter of the initial descriptor framework, an assembler of documents, and an aid in revision, but substantive decisions about the descriptors were still determined by the researcher's review, the validators' suggestions, and the trial data. Of the 32 AI-generated descriptors, 12 were accepted unchanged, 17 were revised, and 3 were rejected because they relied on impressionistic criteria. The twelve descriptors accepted unchanged were later revised on the language validator's advice, so none of the AI-generated descriptors was used unchanged in the final product. The resulting instrument is valid (92.63%; very valid), reliable (single-measures ICC of 0.758; good, with a 95% confidence interval of 0.541–0.878), and practical (98.33%; very practical), and is therefore feasible for use. The low Cronbach's alpha indicates that consistency across aspects could not be demonstrated in a sample whose scores clustered at the top of the scale.

These findings should be read with their limitations in mind. The trial involved only one class with 30 works and two raters from the same school, so the reliability results cannot yet be widely generalized, as reflected in the width of the ICC confidence interval. The ceiling effect limited the testing of internal consistency and discriminating power; the Revision II product has not been revalidated or retested, and the AI output depends on the model and prompt used, so the same results may not be obtained with a different model, version, or prompt. In addition, practicality was assessed only by the two teachers who also served as raters, and dissemination was limited to one school.

Teachers are advised to use this instrument together with the instructions for counting language errors, to calibrate with fellow raters using several sample works before scoring, and to use the mean score of two raters for high-stakes decisions. Teachers and researchers who use AI to develop instruments are advised to review every AI-generated descriptor, including those that appear adequate, to request a table of changes in every revision, and to verify AI output independently. Future studies are advised to test the instrument on larger and more diverse samples with more raters, to sharpen the score 4 descriptors and revalidate them, to test reliability at the aspect level and the discriminating power of the instrument, to test the effectiveness of Revision II, and to compare AI-assisted instruments with instruments developed without AI.

6. Acknowledgments

The authors thank the principal of SMA Al-Azhar Mandiri Palu for permission to conduct the research; Ulfah, Moh. Tahir, Syamsudin, Akhlis, and Agustan as expert validators; Ahmad Maulidi and Meylissa Putri as raters; and the students of class XI E of SMA Al-Azhar Mandiri Palu. This article is based on the first author's thesis in the Master's Program in Indonesian Language Education, Universitas Tadulako. Claude (Anthropic) was used as an aid in developing the instrument and in the statistical calculations, as described in the Method section, and assisted in preparing this manuscript from the thesis and translating it into English; all content has been checked by the authors, who take full responsibility for it.

7. References

- Ahmad, K., Br Ginting, S. U., & Sidiqin, M. A. (2020). Hubungan penguasaan unsur intrinsik cerpen dengan kemampuan menulis cerpen siswa kelas XI SMK Swasta Maju Binjai tahun pelajaran 2019/2020 [The relationship between mastery of the intrinsic elements of short stories and the short story writing ability of grade XI students of SMK Swasta Maju Binjai in the 2019/2020 academic year]. *Jurnal Serunai Bahasa Indonesia*, 17(1), 7–19. doi:10.37755/jsbi.v17i1.257
- Alfianika, N., & Sitohang, K. (2022). Validitas pengembangan rubrik penilaian menulis paragraf narasi dan deskripsi dalam pembelajaran bahasa Indonesia [The validity of developing assessment rubrics for writing narrative and descriptive paragraphs in Indonesian language learning]. *Fon: Jurnal Pendidikan Bahasa dan Sastra Indonesia*, 18(2). doi:10.25134/fon.v18i2.5592
- Anthropic. (2026). Claude (Version Opus 5.5) [Large language model]. Retrieved from <https://claude.ai>
- Arikunto, S. (2018). *Dasar-dasar evaluasi pendidikan* [Fundamentals of educational evaluation] (2nd ed.). Jakarta: Bumi Aksara.
- Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, 3, 22. doi:10.3389/feduc.2018.00022
- Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6), 543–549. doi:10.1016/0895-4356(90)90158-L
- Hanis, M., & Wahyudin, D. (2024). Pemanfaatan artificial intelligence (AI) dalam penyusunan asesmen pembelajaran bagi guru sekolah dasar [The use of artificial intelligence (AI) in developing learning assessments for elementary school teachers]. *Jurnal Ilmiah Profesi Pendidikan*, 9(2), 1199–1207. doi:10.29303/jipp.v9i2.2252
- Hikmah, S. N. A. (2021). Pengembangan instrumen asesmen keterampilan menulis teks eksposisi [Developing an assessment instrument for expository text writing skills]. *Jurnal Tarbiyatuna: Jurnal Kajian Pendidikan, Pemikiran dan Pengembangan Pendidikan Islam*, 2(1), 59–69. doi:10.30739/tarbiyatuna.v2i01.975
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. doi:10.1016/j.lindif.2023.102274
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. doi:10.1016/j.jcm.2016.02.012
- Maulana, A. (2022). Analisis validitas, reliabilitas, dan kelayakan instrumen penilaian rasa percaya diri siswa [An analysis of the validity, reliability, and feasibility of a student self-confidence assessment instrument]. *Jurnal Kualita Pendidikan*, 3(3), 133–139. Retrieved from <https://journal.kualitama.com/index.php/jkp/article/view/331>

- Maydiantoro, A. (2021). Model-model penelitian pengembangan (research and development) [Research and development models]. *Jurnal Pengembangan Profesi Pendidik Indonesia*, 1(2), 29–35.
- Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), 100050. doi:10.1016/j.rmal.2023.100050
- Muhajir, M., Rismawati, R., Marlis, M. I., Faisal, S. F. R., Vanezal, V., Raldiastari, S., & Hakil, M. (2026). Pemanfaatan kecerdasan buatan (AI) dalam asesmen pembelajaran mendalam di SDIT Darul Fikri Makassar [The use of artificial intelligence (AI) in deep learning assessment at SDIT Darul Fikri Makassar]. *Jurnal Petisi (Pendidikan Teknologi Informasi)*, 7(2), 302–310. doi:10.36232/jurnalpetisi.v7i2.4417
- Nurgiyantoro, B. (2016). *Penilaian pembelajaran bahasa berbasis kompetensi* [Competency-based assessment in language learning]. Yogyakarta: BPFE.
- Nurgiyantoro, B. (2018). *Teori pengkajian fiksi* [The theory of fiction analysis]. Yogyakarta: Gadjah Mada University Press.
- Oktavianus, A. J. E., Naibaho, L., & Rantung, D. A. (2023). Pemanfaatan artificial intelligence pada pembelajaran dan asesmen di era digitalisasi [The use of artificial intelligence in learning and assessment in the digitalization era]. *Jurnal Kridatama Sains dan Teknologi*, 5(2), 473–476. doi:10.53863/kst.v5i02.975
- Rasyid, A., Hartanto, R., & Priyanto, C. (2026). Reliabilitas alat ukur, jenis-jenis, cara pengukuran, dan faktor-faktor yang memengaruhinya [The reliability of measuring instruments: Types, measurement methods, and influencing factors]. *Tsaqofah: Jurnal Penelitian Guru Indonesia*, 6(2), 1285–1294. doi:10.58578/tsaqofah.v6i2.8511
- Santi, A., Silvia, D., & Damaianti, V. S. (2023). Penilaian autentik pembelajaran bahasa Indonesia menulis karya ilmiah: Penggunaan dan pencapaian keterampilan peserta didik [Authentic assessment in Indonesian language learning of scientific writing: Its use and students' skill attainment]. *Fon: Jurnal Pendidikan Bahasa dan Sastra Indonesia*, 19(2), 226–238. doi:10.25134/fon.v19i2.7710
- Santoso, R. S. B., Suyono, Harsiati, T., & Sari, N. P. (2023). Implementasi instrumen asesmen keterampilan menulis cerpen dengan media lagu inspiratif untuk siswa kelas XI SMA di era merdeka belajar [Implementing a short story writing assessment instrument using inspirational songs for grade XI senior high school students in the Merdeka Belajar era]. *Ghancaran: Jurnal Pendidikan Bahasa dan Sastra Indonesia*. doi:10.19105/ghancaran.vi.11772
- Sauhenda, A. F., Marnina, & Narahawarin, M. F. (2025). Pelatihan penyusunan instrumen penilaian keterampilan menulis cerita pendek di SMP Negeri Urumb Merauke [Training on developing short story writing assessment instruments at SMP Negeri Urumb Merauke]. *Jurnal Pelayanan Masyarakat*, 2(4), 283–290. doi:10.62951/jpm.v2i4.2743
- Sihombing, B., Zamsiswaya, & Sawaluddin. (2024). Model pengembangan 4D (define, design, develop, dan disseminate) dalam pembelajaran pendidikan Islam [The 4D development model (define, design, develop, and disseminate) in Islamic education

- learning]. *Journal of Islamic Education El Madani*, 4(1), 11–19. doi:10.55438/jiee.v4i1.135
- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., . . . Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, 101894. doi:10.1016/j.learninstruc.2024.101894
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53–55. doi:10.5116/ijme.4dfb.8dfd
- Waruwu, M. (2024). Metode penelitian dan pengembangan (R&D): Konsep, jenis, tahapan, dan kelebihan [Research and development (R&D) methods: Concepts, types, stages, and advantages]. *Jurnal Ilmiah Profesi Pendidikan*, 9(2), 1220–1230. doi:10.29303/jipp.v9i2.2141