Jurnal MATH-UMB.EDU
Vol 13 (2), 2026

DEVELOPMENT OF ESSAY INSTRUMENT TO MEASURE MATHEMATICAL CONCEPTUAL UNDERSTANDING ON SIMILARITY AND CONGRUENCE

Muhammad Faruq Wahyu Utomo[1], Novaliyosi[2], Aan Hendrayana[3], Abdul Fatah[4]
Universitas Sultan Ageng Tirtayasa, Banten, Indonesia
muhammadfaruq.wu@gmail.com

Recived :10-12-2025 ***Revised :07-02-2026*** ***Accepted :01-03-2026*** ***Published : 02-03-2026***

Abstract

This study aims to develop a valid and reliable essay test instrument to assess mathematical conceptual understanding in ninth-grade junior high school students, focusing on similarity and congruence. Employing a modified 4D model (Define, Design, Develop, Disseminate), the research involved designing a test blueprint based on seven conceptual understanding indicators, expert validation using the Content Validity Ratio (CVR), and field testing to evaluate construct validity, reliability, item difficulty, and discrimination indices. The results indicated that all test items were valid, achieved a high reliability coefficient ($\alpha = 0.83$), and exhibited varied difficulty and discrimination levels meeting established criteria. Consequently, this instrument is capable of comprehensively measuring students' conceptual understanding and serves as an effective assessment tool in mathematics education.

Keywords: Essay Instrument, Mathematical Conceptual Understanding, Similarity and Congruence.

INTRODUCTION

Conceptual understanding is fundamental to mathematics education, enabling students to construct meaning, establish connections between ideas, and apply procedures logically to solve problems (Saputra, 2022). Unlike rote memorization, deep comprehension empowers learners to justify their reasoning and adapt strategies to various problem contexts. This capability is particularly vital in geometry topics such as similarity and congruence, which demand both visual and logical reasoning to interpret relationships between figures. However, research indicates that students often struggle to grasp these concepts profoundly, frequently relying on formula memorization without true conceptual insight (Permana, 2021). Consequently, a more robust evaluation approach is required to accurately assess the depth of student comprehension.

In this context, developing a valid and reliable instrument specifically designed to assess conceptual understanding is critical. Current assessments predominantly utilize multiple-choice formats, which fail to fully capture students' cognitive processes (Iskandar et al., 2022). Furthermore, while previous studies utilizing the ADDIE model have focused on developing teaching materials to improve general learning outcomes, they have not yet established a structured measurement tool grounded in specific indicators of conceptual understanding (Loi, 2023). This highlights an urgent need for a systematically designed essay-based instrument focusing on similarity and congruence to provide a comprehensive and accurate evaluation of students' conceptual proficiency.

Concept mastery is a prerequisite for achieving the primary goals of mathematics instruction, moving beyond content memorization to genuine understanding. The theoretical framework for this study relies on the National Council of Teachers of Mathematics (NCTM, 2000), which defines conceptual understanding as the ability to: define concepts verbally and in writing; identify examples and non-examples; represent concepts through models, diagrams, and symbols; translate between different modes of representation; recognize defining characteristics and conditions; and compare or contrast related concepts. Based on this framework, this study adapts these standards into a curriculum-aligned instrument featuring seven key indicators, as follows:

Table 1. Indicators of Students' Conceptual Comprehension Skill

No	Indicator Description
1	Restating a concept
2	Classifying objects according to specific properties based on the concept
3	Providing examples and non-examples of the concept
4	Presenting concepts in various forms of mathematical representation
5	Developing necessary or sufficient conditions for a concept
6	Applying, utilizing, and selecting specific procedures or operations
7	Applying concepts or algorithms to problem-solving

Sources: (NCTM, 2000; Pasha & Aini, 2022; Saputra, 2022)

Previous studies have explored the development of evaluation instruments for conceptual understanding. However, these efforts have predominantly focused on multiple-choice formats, which are often insufficient for deeply exploring students' cognitive processes. (Iskandar et al., 2022) support this observation, noting that most existing conceptual comprehension tests rely heavily on selected-response items. Furthermore, (Retno Kuncoro & Martila Ruli, 2022) highlighted low student mastery of conceptual indicators, particularly in essay questions, although their study was limited to elementary levels or different topics such as sets and algebra. Consequently, developing an essay-based instrument is essential to assess the seven indicators of conceptual understanding holistically and validly.

The development of this instrument is underpinned by three theoretical frameworks. First, Skemp's theory distinguishes between instrumental and relational understanding, emphasizing that students must not only master procedures but also grasp the underlying reasons for their application. This establishes the foundation for measuring indicators such as mathematical representation and the formulation of necessary and sufficient conditions (Erawati et al., 2025). Second, the principle of authentic assessment Brown & Hudson, (1998) argues that effective evaluation must assess not only the final result but also the thinking process through meaningful tasks, such as essay questions (Noordin & Darmi, 2022). Third, constructivist theory serves as the pedagogical basis, viewing knowledge as actively constructed through contextual and reflective experiences. (Erawati et al., 2025) suggest that this approach enhances conceptual comprehension through active exploration. Synergistically, these theories reinforce the principle that evaluation instruments must reflect the profound cognitive processes underlying mathematical understanding.

In the context of geometry learning—specifically similarity and congruence—there is a critical need for an instrument capable of comprehensively assessing conceptual understanding. Existing tools generally fail to thoroughly evaluate students based on established conceptual indicators. Therefore, the central research question of this study is: "How can a valid, reliable, and standardized essay instrument be developed to measure the mathematical concept comprehension of ninth-grade students on the topic of similarity and congruence?" Accordingly, this study aims to design an essay-based assessment that is not only content-valid but also empirically tested for its psychometric properties, including validity, reliability, discrimination power, and difficulty index.

This study aims to provide both theoretical and practical contributions. Theoretically, the findings are expected to enrich the literature regarding instrument development based on conceptual understanding indicators. Practically, the instrument can serve as a diagnostic tool for teachers to identify students' comprehension levels, while for researchers and curriculum developers, it offers a reference for designing more meaningful evaluations. However, this study has limitations: the scope is restricted to the topic of similarity and congruence for ninth-grade students. Additionally, as the instrument was tested on a limited research subject within a specific school, the development outcomes cannot be directly generalized to the broader population without further empirical study.

RESEARCH METHOD

Research Design

This study employs a Research and Development (R&D) approach aimed at constructing a valid and reliable essay test instrument to assess the mathematical conceptual understanding of ninth-grade students on the topic of similarity and congruence. Specifically, the development process adopts the 4D model—comprising Define, Design, Develop, and Disseminate phases—as the systematic framework to guide the creation of the instrument (Thiagarajan et al., 1974**.** This model was selected to ensure a structured progression from needs analysis to final dissemination. Consequently, this approach emphasizes not only product creation but also rigorous validation and quality evaluation to meet practical and academic standards (Creswell & Creswell, 2018).

Participants

The participants comprised 43 ninth-grade students from Mutiara Bangsa 2 Junior High School, Tangerang, during the 2024/2025 academic year. The subjects were selected using purposive sampling to ensure representation of diverse academic abilities and backgrounds. The sample consisted of 23 students from Class 9B and 20 students from Class 9C, all of whom had previously studied similarity and congruence, ensuring they possessed sufficient prior knowledge to participate in the instrument trial (Nyimbili & Nyimbili, 2024).

Instrument

The primary instrument developed is an essay test comprising seven items, each specifically designed to measure distinct dimensions of mathematical conceptual understanding. The essay format was chosen for its capacity to capture students' cognitive processes more comprehensively than selected-response formats. It allows students to explain their reasoning, connect concepts, and apply knowledge in real-world contexts—nuances often missed by multiple-choice tests (Eldakhakhny & Elsamanoudy, 2023). This format enables students to explicitly demonstrate their logical thinking, providing richer diagnostic information for educators.

The items were constructed based on a blueprint mapping relevant conceptual competencies, including: restating concepts, classifying objects, providing examples and non-examples, utilizing representations, and applying procedures or algorithms. This structure aligns with psychometric development principles (Maric et al., 2023), prioritizing content validity to ensure the assessment is representative and contextually appropriate. By integrating these indicators into essay questions, the instrument is intended to capture the depth of student understanding accurately.

Instrument Development Procedure

The study adapted the 4D model (Define, Design, Develop, Disseminate) for educational instrument development. The Define phase involved a needs analysis and a review of conceptual understanding indicators to establish the test blueprint. The Design phase focused on prototyping the essay test and conducting initial expert validations, followed by revisions based on feedback. The Develop phase involved field testing the instrument on ninth-grade students to evaluate its psychometric properties quantitatively, including item validity, internal reliability, discrimination power, and difficulty index. Finally, the Disseminate phase entailed finalizing the instrument based on trial data and preparing it for broader utilization by teachers and researchers (Johan et al., 2023).

Data Analysis Technique

Data analysis focused on evaluating four key psychometric properties: item validity, internal consistency reliability, discrimination power, and difficulty index. Item validity was tested using Pearson Product-Moment correlation between item scores and total scores, while reliability was measured using Cronbach's Alpha. Discrimination power was determined by the score difference between high- and low-achieving groups, and the difficulty index was calculated based on the proportion of correct responses. The analysis procedures and threshold values followed recommendations by (Sivakumar & Thirumoorthy, 2019) and utilized Microsoft Excel for efficient calculation (Amelia & Erita, 2024).

Content Validity Analysis Before field testing, the instrument underwent a rigorous content validity assessment. This process involved a panel of three experts, consisting of mathematics education lecturers and experienced junior high school mathematics teachers. These experts evaluated the relevance of each item against the theoretical indicators. The analysis utilized the Content Validity Ratio (CVR) method proposed by Lawshe, as refined by (Obilor & Miwari, 2022). The CVR formula is as follows:

$$\boldsymbol{CVR} = \frac{\boldsymbol{n_e} - \left(\frac{\boldsymbol{N}}{\boldsymbol{2}}\right)}{\frac{\boldsymbol{N}}{\boldsymbol{2}}}$$

where n_e represents the number of experts rating an item as "essential," and N denotes the total number of experts. The calculated CVR value is then compared against the critical minimum threshold established for the specific number of panelists. According to (Obilor & Miwari, 2022), for a panel of five or fewer experts, an item is considered valid if the CVR is ≥ 0.99. Confirming content validity ensures that the instrument fulfills criteria for clarity and relevance regarding content, face, and language. Once validated by experts, the instrument proceeds to empirical testing for construct validity (Amelia & Erita, 2024).

Subsequently, construct validity testing is conducted to determine the extent to which the items accurately measure the theoretical construct of mathematical conceptual understanding. This step is crucial to ensure that the instrument demonstrates theoretical consistency with the measured indicators in an empirical setting, beyond mere content relevance. Construct validity was assessed using the *Pearson Product-Moment* Correlation coefficient to measure the relationship between individual item scores and the total instrument score. A high correlation coefficient indicates strong consistency between the item and the overall construct. The *Pearson* Correlation formula is as follows:

$$\boldsymbol{r_{xy}} = \frac{\boldsymbol{N\sum XY - (\sum X)(\sum Y)}}{\sqrt{[\boldsymbol{N\sum X^2 - (\sum X)^2}][\boldsymbol{N\sum Y^2 - (\sum Y)^2}]}}$$

where X represents the individual item score, Y denotes the total score, and N refers to the number of respondents. The calculated correlation coefficient (r_{count}) is compared against the critical value (r_{table}) at a 5% significance level ($\alpha = 0.05$) with degrees of freedom $df = N - 2$. If $r_{count} > r_{table}$, the item is deemed to possess construct validity (Amelia & Erita, 2024).

Reliability refers to the consistency of test outcomes when administered under comparable conditions. A reliable instrument yields stable and reproducible scores (Saputri et al., 2023). To measure internal consistency, this study utilized Cronbach's Alpha coefficient, which evaluates the correlation among items within the instrument. A high Alpha value indicates strong reliability. The formula is defined as follows:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_T^2}\right)$$

where k represents the number of items, σ_i^2 denotes the variance of each item, and σ_T^2 refers to the total variance. A coefficient of $\alpha \geq 0.70$ indicates acceptable reliability (Amelia & Erita, 2024). Since Microsoft Excel lacks a built-in function for Cronbach's Alpha, the value was calculated manually using variance components derived from the dataset. The interpretation of the Alpha value is presented in Table 2.

Table 2. Reliability Criteria

Alpha Score	**Criteria**
$\geq 0{,}80$	High
0,70 – 0,79	Moderate
< 0,70	Low

Sources: (Amelia & Erita, 2024; Sivakumar & Thirumoorthy, 2019)

Discrimination power measures an item's ability to differentiate between high-ability and low-ability students. A high discrimination index implies that the item effectively distinguishes students based on their mastery of the material (Qomariyah et al., 2022). To calculate this, participants were categorized into two groups based on their total scores: the upper group (top 27%) and the lower group (bottom 27%). The discrimination power (D) is determined using the following formula:

$$\boldsymbol{D = \frac{B - S}{N}}$$

Keterangan:
D : Discriminative Power
B : Upper Group
S : Lower Group
N : Number of students of each group

This calculation was performed in Microsoft Excel by computing the difference in the proportion of correct answers between the upper and lower groups. The criteria for interpreting discrimination power are detailed in Table 3.

Table 3. Discrimination Power Criteria

Score	**Criteria**
> 0.70	Very Good
$0{,}40 - 0.70$	Good
0.30 – 0.39	Moderate
< 0.30	Low

Source: (Khumaira et al., 2024).

The difficulty index (P) indicates the proportion of students who answered a specific item correctly, serving as a metric to balance the test's complexity. It is calculated by dividing the number of correct responses by the total number of participants (Saputri et al., 2023). The formula is as follows:

$$P = \frac{N_p}{N}$$

where P represents the difficulty index, N_p is the number of students answering correctly, and N is the total number of test-takers (Saputri et al., 2023). This analysis was conducted using standard frequency functions in Microsoft Excel. The interpretation of the difficulty index levels is provided in Table 4.

Table 4. Difficulty Index Criteria

Difficulty Score	**Criteria**
$0{,}00 - 0{,}30$	Very Difficult
0,31 – 0,50	Difficult
$0{,}51 - 0{,}70$	Moderate
$0{,}71 - 1{,}00$	Easy

Source: (Qomariyah et al., 2022)

The results from the validity, reliability, discrimination power, and difficulty index analyses serve as the empirical basis for refining and finalizing the essay test instrument. This rigorous evaluation ensures that the instrument fulfills the psychometric standards required for a standardized assessment tool suitable for broader application. By employing these comprehensive data analysis techniques, this study guarantees that the final instrument is not only valid and reliable but also possesses appropriate discrimination and difficulty levels. Consequently, it effectively and accurately measures the mathematical conceptual understanding of ninth-grade students regarding similarity and congruence.

RESULT

The development of the essay test instrument followed the systematic phases of the 4D model: Define, Design, Develop, and Disseminate (Johan et al., 2023; Thiagarajan et al., 1974). This section presents the empirical findings obtained from each stage, with a primary focus on the psychometric evaluation of the instrument. The analysis covers content and construct validity, internal consistency reliability, as well as item characteristics, specifically difficulty indices and discrimination power. The following subsections detail the specific outcomes of each development phase.

1) Define Phase

The Define phase aimed to identify the instructional problems and specify the requirements for developing the assessment instrument. This stage primarily involved front-end analysis and concept analysis based on preliminary research conducted at Mutiara Bangsa 2 Junior High School. The needs analysis revealed that ninth-grade students frequently experienced difficulties in comprehending conceptual problems related to similarity and congruence. Observations indicated that students often failed to recognize the underlying geometric rules or postulates (e.g., Side-Side-Side, Side-Angle-Side) required to determine similarity. Furthermore, students struggled to connect these geometric concepts with related mathematical topics, leading to fragmented understanding. Consequently, there was a critical need for an instrument capable of diagnosing these conceptual gaps. Following this analysis, the key concepts were formulated to align with the seven indicators of mathematical conceptual understanding adapted from the National Council of Teachers of Mathematics (NCTM, 2000).

2) Design Phase

The Design phase focused on constructing the instrument prototype based on the conceptual framework established in the Define phase. This stage began with the development of a criterion-referenced test blueprint to ensure content validity by explicitly mapping the subject matter—similarity and congruence—against the seven indicators of conceptual understanding. This blueprint served as a strategic guide to guarantee that the instrument comprehensively measured the intended cognitive skills, as detailed in the item distribution presented in Table 5.

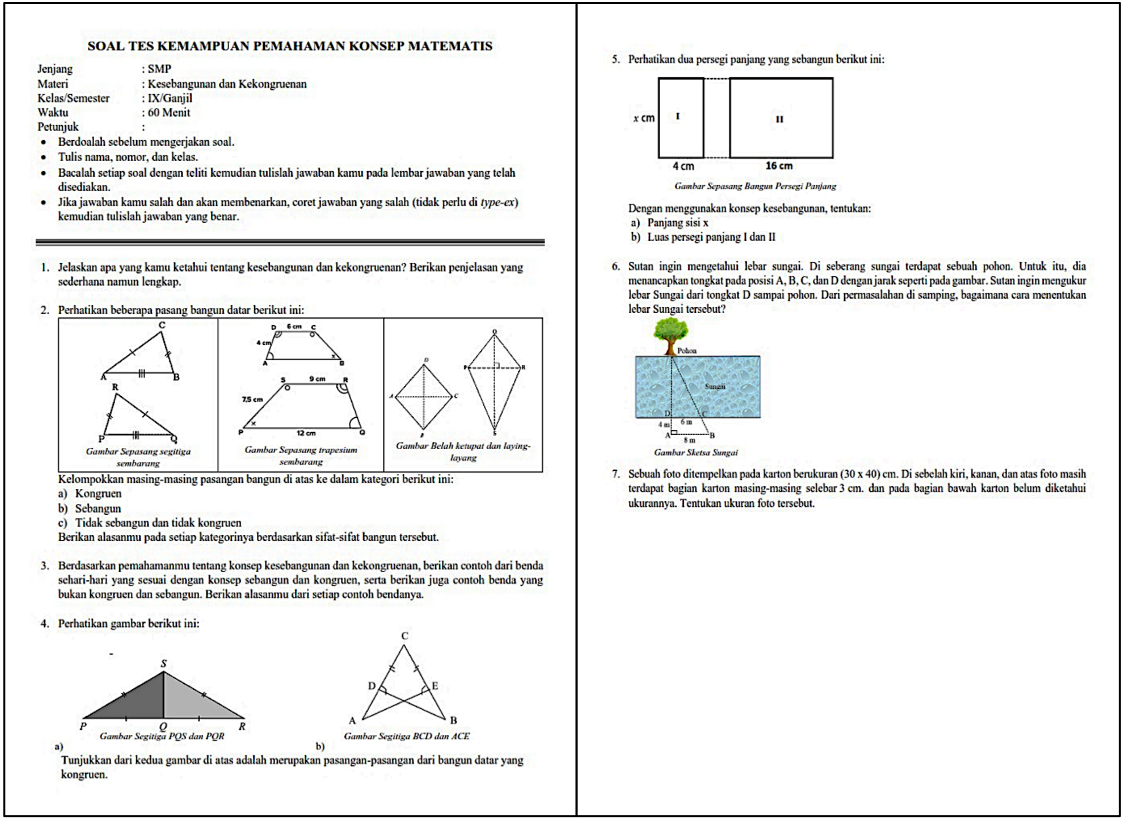
Table 5. Blueprint of Mathematical Conceptual Understanding Instrument

Indicator of Mathematical Concept Understanding	Topic Indicator	Item Number
Restating a concept	Explaining the definition of similarity and congruence.	1
Classifying objects according to specific properties based on the concept	Distinguishing and classifying plane figures that exhibit properties of similarity, congruence, or neither.	2
Providing examples and non-examples of the concept	Giving examples of everyday objects that exhibit similarity, congruence, or neither.	3
Presenting concepts in various forms of mathematical representation	Determining the length of sides or magnitude of angles in similar or congruent figures using visual models.	4a,4b
Developing necessary or sufficient conditions for a concept	Determining the necessary or sufficient conditions for two triangles to be considered congruent.	5a, 5b
Applying, utilizing, and selecting specific procedures or operations	Calculating the length of sides or angles using the concept of similarity/congruence to solve geometric problems.	6

Indicator of Mathematical Concept Understanding	Topic Indicator	Item Number
Applying concepts or algorithms to problem-solving	Solving contextual problems related to similarity and congruence in real-life scenarios.	7

Source: Adaptable from (Giriansyah et al., 2023)

Based on the blueprint above, the researcher designed the initial draft of the instrument in the form of an essay test. The essay format was selected to allow students to explicitly demonstrate their reasoning processes and ability to connect concepts, which is often limited in multiple-choice formats. This draft included seven items, a scoring rubric, and an answer key, which were subsequently prepared for expert validation.



SOAL TES KEMAMPUAN PEMAHAMAN KONSEP MATEMATIS

Jenjang : SMP
Materi : Kesebangunan dan Kekongruenan
Kelas/Semester : IX/Ganjil
Waktu : 60 Menit
Petunjuk :

- Berdoalah sebelum mengerjakan soal.
- Tulis nama, nomor, dan kelas.
- Bacalah setiap soal dengan teliti kemudian tulislah jawaban kamu pada lembar jawaban yang telah disediakan.
- Jika jawaban kamu salah dan akan membenarkan, coret jawaban yang salah (tidak perlu di *type-ex*) kemudian tulislah jawaban yang benar.

1. Jelaskan apa yang kamu ketahui tentang kesebangunan dan kekongruenan? Berikan penjelasan yang sederhana namun lengkap.

2. Perhatikan beberapa pasang bangun datar berikut ini:

Gambar Sepasang segitiga sembarang

Gambar Sepasang trapesium sembarang

Gambar Belah ketupat dan laying-layang

Kelompokkan masing-masing pasangan bangun di atas ke dalam kategori berikut ini:
a) Kongruen
b) Sebangun
c) Tidak sebangun dan tidak kongruen
Berikan alasanmu pada setiap kategorinya berdasarkan sifat-sifat bangun tersebut.

3. Berdasarkan pemahamanmu tentang konsep kesebangunan dan kekongruenan, berikan contoh dari benda sehari-hari yang sesuai dengan konsep sebangun dan kongruen, serta berikan juga contoh benda yang bukan kongruen dan sebangun. Berikan alasanmu dari setiap contoh bendanya.

4. Perhatikan gambar berikut ini:

a) *Gambar Segitiga PQS dan PQR*

b) *Gambar Segitiga BCD dan ACE*

Tunjukkan dari kedua gambar di atas adalah merupakan pasangan-pasangan dari bangun datar yang kongruen.

5. Perhatikan dua persegi panjang yang sebangun berikut ini:

Gambar Sepasang Bangun Persegi Panjang

Dengan menggunakan konsep kesebangunan, tentukan:
a) Panjang sisi x
b) Luas persegi panjang I dan II

6. Sutan ingin mengetahui lebar sungai. Di seberang sungai terdapat sebuah pohon. Untuk itu, dia menancapkan tongkat pada posisi A, B, C, dan D dengan jarak seperti pada gambar. Sutan ingin mengukur lebar Sungai dari tongkat D sampai pohon. Dari permasalahan di samping, bagaimana cara menentukan lebar Sungai tersebut?

Gambar Sketsa Sungai

7. Sebuah foto ditempelkan pada karton berukuran (30 x 40) cm. Di sebelah kiri, kanan, dan atas foto masih terdapat bagian karton masing-masing selebar 3 cm. dan pada bagian bawah karton belum diketahui ukurannya. Tentukan ukuran foto tersebut.

Figure 1. Preliminary Design of Mathematical Concept Understanding Ability Test Questions

3) Development Phase

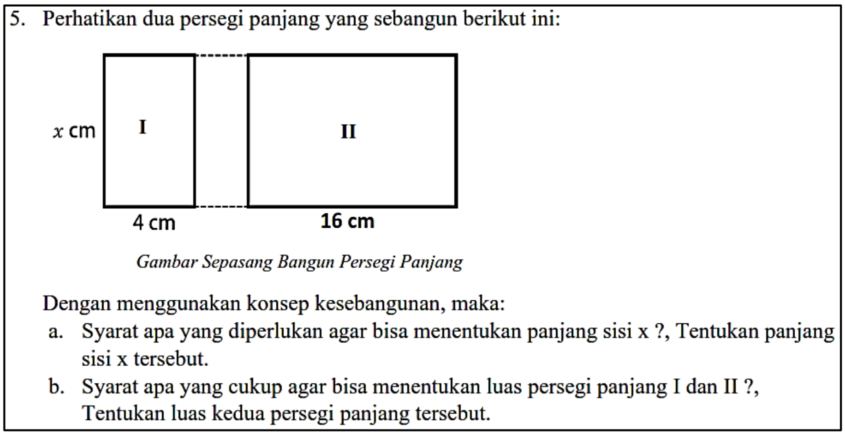
The Develop phase focused on validating and refining the drafted instrument. This stage comprised two primary steps: expert validation (content validity) and empirical field testing. The initial draft was evaluated by a panel of three experts, consisting of mathematics teachers from Mutiara Bangsa 2 Junior High School. The experts assessed the relevance of each item against the conceptual understanding indicators using the Content Validity Ratio (CVR) method. The results of this assessment are presented in Table 6.

Table 6. Content Validity Ratio (CVR) Results

Item No.	Essential (n_e)	Experts (N)	CVR Score	Category
1	3	3	1	Valid

Item No.	Essential (n_e)	Experts (N)	CVR Score	Category
2	3	3	1	Valid
3	3	3	1	Valid
4a	3	3	1	Valid
4b	3	3	1	Valid
5a	1	3	0.33	Revision
5b	1	3	0.33	Revision
6	3	3	1	Valid
7	3	3	1	Valid

Based on Table 6, items 5a and 5b achieved a CVR score of 0.33, indicating the need for revision. The experts noted that these items, which aimed to measure the indicator *"developing necessary and sufficient conditions,"* implicitly required students to use conditions (e.g., side lengths in rectangles) but did not explicitly ask them to identify or explain those conditions. Consequently, the items were revised to explicitly prompt students to state the necessary requirements before performing calculations. Following this revision, the instrument was declared content-valid and ready for field testing.



5. Perhatikan dua persegi panjang yang sebangun berikut ini:

Gambar Sepasang Bangun Persegi Panjang

Dengan menggunakan konsep kesebangunan, maka:

a. Syarat apa yang diperlukan agar bisa menentukan panjang sisi x ?, Tentukan panjang sisi x tersebut.

b. Syarat apa yang cukup agar bisa menentukan luas persegi panjang I dan II ?, Tentukan luas kedua persegi panjang tersebut.

Figure 2. Revised Question Item Number 5

The revised instrument was tested on May 26, 2025, involving 43 ninth-grade students (24 boys and 19 girls) from classes 9B and 9C at Mutiara Bangsa 2 Junior High School. The trial aimed to evaluate construct validity, reliability, difficulty index, and discrimination power. Validity was tested using the Pearson Product-Moment Correlation between item scores and the total score. The results are summarized in Table 7.

Table 7. Construct Validity Results

Item No	Correlation (r_{count})	Critical Value (r_{table})	Note
1	0.45	0.36	Valid
2	0.49	0.36	Valid
3	0.57	0.36	Valid
4a	0.83	0.36	Valid
4b	0.71	0.36	Valid
5a	0.62	0.36	Valid

Item No	Correlation (r_{count})	Critical Value (r_{table})	Note
5b	0.65	0.36	Valid
6	0.83	0.36	Valid
7	0.70	0.36	Valid

As shown in Table 7, all items yielded correlation coefficients (r_{count}) ranging from 0.45 to 0.83, surpassing the critical table value of 0.36 (at 5% significance level). Thus, all items are considered valid.

Internal consistency was measured using Cronbach's Alpha. The analysis yielded an Alpha coefficient of 0.83, which exceeds the minimum threshold of 0.60. This indicates that the instrument has **High Reliability** and can produce consistent measurements of students' conceptual understanding.

The difficulty index analysis classifies items into easy, medium, and difficult categories. The results are presented in Table 8.

Table 8. Dificulty Index Result

Item No	Average	Max Score	Difficulty Index	Description
1	2.79	3	0.93	Easy
2	2.63	3	0.88	Easy
3	2.00	3	0.67	Moderate
4a	1.63	3	0.54	Moderate
4b	0.84	3	0.28	Difficult
5a	2.44	4	0.61	Moderate
5b	1.91	3	0.64	Moderate
6	1.19	4	0.30	Difficult
7	2.09	4	0.52	Moderate

The analysis shows a balanced distribution: two items are easy (1, 2), five are moderate (3, 4a, 5a, 5b, 7), and two are difficult (4b, 6). This variety allows the instrument to accommodate students with varying ability levels.

Discrimination power measures the item's ability to distinguish between high and low-performing students. The results are detailed in Table 9.

Table 9. Discrimination Power Result

Item No.	Discrimination Index	Description
1	0.17	Low
2	0.31	Moderate
3	0.36	Moderate
4a	0.92	Very Good
4b	0.61	Good
5a	0.69	Good
5b	0.47	Good
6	0.92	Very Good
7	0.61	Good

Most items demonstrated effective discrimination, with four items classified as "Good" and two as "Very Good." Items 2 and 3 were "Moderate." Item 1 showed low discrimination (0.17), which is expected given its high difficulty index (0.93/Easy); since most students could answer it correctly, it did not significantly differentiate between the upper and lower groups.

4) Dissemination Phase

The Disseminate phase constitutes the final stage of the development process, aiming to promote the adoption and utilization of the developed product. Following the rigorous empirical testing in the Develop phase, the instrument was finalized into a complete assessment package. This package consists of the seven validated essay items, a standardized scoring rubric, and a comprehensive answer key, ensuring that the instrument is ready for practical implementation.

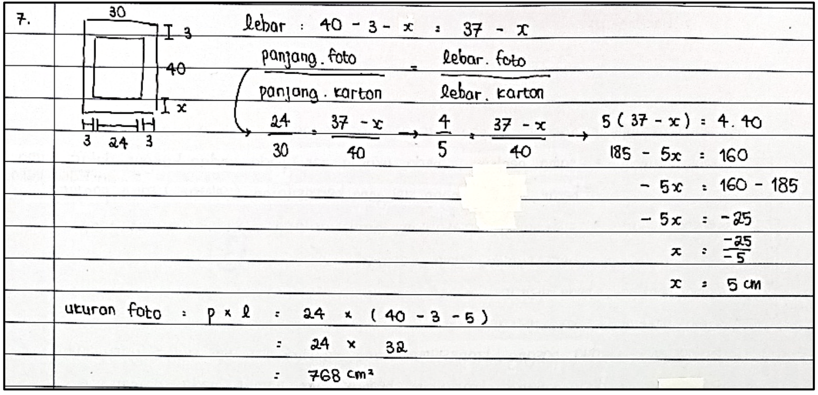
The dissemination process was conducted through two primary mechanisms. First, the final instrument was distributed to mathematics teachers at Mutiara Bangsa 2 Junior High School to be utilized as a diagnostic tool for measuring students' conceptual understanding in future learning sessions. Second, the dissemination was expanded to the broader academic community through the publication of this research article. By making the instrument accessible via this publication, it allows other researchers and practitioners to adopt, adapt, or further develop the instrument for various educational contexts. Consequently, this phase confirms that the product has transitioned from a developmental prototype into a standardized assessment tool ready for wider application.

DISCUSSION

The successful development of this essay-based instrument underscores the growing imperative of qualitative assessment in mathematics education. The findings of this study align with (Tadhkiroh et al., 2023), who demonstrated that instruments developed using the 4D model are capable of generating robust qualitative and quantitative data to validate constructs in educational assessments. Furthermore, these results reinforce the work of (Sriyanti et al., 2019), which highlighted that essay-based diagnostic tools offer a more accurate identification of students' conceptual understanding compared to selected-response formats. This study confirms that such instruments can meet rigorous psychometric criteria, achieving high content validity (CVR = 1.00) and high reliability ($\alpha = 0.83$), supported by a score distribution that facilitates in-depth analysis of difficulty and discrimination.

A specific psychometric observation was noted regarding Item 1, which exhibited low discrimination power $(D = 0.17)$ despite being constructively valid. This occurrence is attributed to the item's low difficulty level (index > 0.90), categorized as "easy," meaning nearly all respondents from both upper and lower groups answered correctly. This phenomenon is consistent with findings by (Mustaqim & Sulisti, 2024), who state that extremely easy items tend to have poor discrimination because they fail to effectively differentiate between varying ability levels. Similarly, (Saputri et al., 2023) note that easy items often produce low discrimination indices statistically. However, despite its low discrimination power, Item 1 was retained in the final instrument. This decision is justified because the item remains valid and serves an essential instructional function: measuring the basic level of conceptual comprehension required by the indicators and building student confidence at the beginning of the test.

The novelty and significance of this research lie in its comprehensive focus on similarity and congruence—a topic that has rarely been assessed using a fully validated essay instrument covering all conceptual indicators. This addresses a gap identified in previous studies, such as those by (Sutriani et al., 2021) and (Sriyanti et al., 2019), which often relied on instruments that did not explicitly measure all dimensions of mathematical concept comprehension or were limited to multiple-choice formats lacking cognitive depth. By utilizing a rigorous essay format, this instrument provides a more accurate and holistic depiction of students' conceptual mastery. To illustrate the instrument's capacity to capture deep reasoning, the following analysis presents a subject's response to Item 7, which corresponds to the indicator of "applying concepts or algorithms to problem-solving".


30
3
40
x
3
24
3

7.

lebar : 40 - 3 - x : 37 - x

$\frac{\text{panjang . foto}}{\text{panjang . karton}} = \frac{\text{lebar . foto}}{\text{lebar . karton}}$

$\frac{24}{30} : \frac{37-x}{40} \rightarrow \frac{4}{5} : \frac{37-x}{40} \rightarrow$ 5 (37 - x) : 4 . 40

185 - 5x : 160

- 5x : 160 - 185

- 5x : -25

x : $\frac{-25}{-5}$

x : 5 cm

ukuran foto : p × l : 24 × (40 - 3 - 5)

: 24 × 32

: 768 cm²

Figure 3. Sample Student Response to Problem Number 7

As depicted in Figure 3, the subject successfully answered Item 7 by accurately applying the concept of similarity with a comprehensive problem-solving approach. The response demonstrates that the student did not merely apply a rote comparison formula; instead, they began by systematically identifying the given information (frame size and distances) and translating it into a visual representation to aid understanding. Subsequently, the student devised a logical solution strategy, calculating the unknown distance of the lower side of the frame to determine the photo's width, ultimately finding the actual dimensions of the photo. This structured response provides empirical evidence that the essay instrument effectively measures the depth of students' mathematical conceptual understanding, confirming the instrument's validity beyond statistical metrics.

The development of this valid and reliable instrument for ninth-grade students regarding similarity and congruence offers significant practical implications for the field of mathematics education. Teachers and researchers can utilize this instrument as a robust diagnostic tool to identify specific gaps in students' conceptual understanding, enabling the design of targeted and effective pedagogical interventions. Furthermore, this instrument serves as a valuable reference for curriculum developers and educators in designing evaluation assessments, particularly for geometry topics. Ultimately, this study contributes to the improvement of mathematics education quality, supporting the optimization of student learning outcomes in mastering the concepts of similarity and congruence.

CONCLUSION

This study successfully developed a valid and reliable essay-based instrument to assess mathematical conceptual understanding of similarity and congruence for ninth-grade students. The development process, guided by the systematic 4D model (Define, Design, Develop, and Disseminate), ensured that the instrument met rigorous psychometric standards. Statistical analysis confirmed that the instrument possesses high content and construct validity, as well as strong internal consistency reliability. Furthermore, the analysis of item characteristics—specifically discrimination power and difficulty indices—demonstrates the instrument's capability to effectively differentiate between varying levels of student ability while maintaining an appropriate difficulty distribution. Consequently, this instrument is deemed suitable as a standardized diagnostic tool for evaluating mathematical learning outcomes in junior high schools.

Despite the positive findings, this study acknowledges several limitations. First, the empirical trial was conducted with a relatively small sample size ($N = 43$) from a single junior high school, which may limit the generalizability of the results to a broader population. Second, the scope of the instrument is restricted solely to the topic of similarity and congruence, meaning it cannot be used to assess conceptual understanding in other geometry topics without further adaptation. Therefore, future research is recommended to test this instrument on a larger and more diverse demographic to enhance its external validity. Additionally, further studies could expand the development of similar essay-based instruments to cover other critical mathematical topics, thereby enriching the repository of quality assessment tools available for mathematics educators.

REFERENCE

Amelia, N., & Erita, S. (2024). Eksplorasi Validitas dan Reliabilitas Soal Pemahaman Konsep dalam Asesmen Pembelajaran. *Jurnal BIMA: Pusat Publikasi Ilmu Pendidikan Bahasa Dan Sastra*, *2*(1), 222–232.

Creswell, J. W., & Creswell, J. D. (2018). Research Design: Fifth Edition. In SAGE (Ed.), *Writing Center Talk over Time* (5th ed.). SAGE Publications. https://doi.org/10.4324/9780429469237-3

Eldakhakhny, B., & Elsamanoudy, A. Z. (2023). Discrimination Power of Short Essay Questions Versus Multiple Choice Questions as an Assessment Tool in Clinical Biochemistry. *Cureus*, *15*(2). https://doi.org/10.7759/cureus.35427

Erawati, N. K., Suastra, W., Atmaja, A. W. T., & Tika, I. N. (2025). Peran Konstruktivisme Dalam Mengembangkan Pemahaman Konseptual Matematika: Perspektif Filsafat Ilmu. *Emasains: Jurnal Edukasi Matematika Dan Sains*, *14*(1), 105–114.

Giriansyah, F. E., Pujiastuti, H., & Ihsanudin, I. (2023). Kemampuan Pemahaman Matematis Siswa Berdasarkan Teori Skemp Ditinjau dari Gaya Belajar. *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *7*(1), 751–765. https://doi.org/10.31004/cendekia.v7i1.1515

Iskandar, R. S. F., Karjanto, N., Kusumah, Y. S., & Ihsan, I. R. (2022). *A Systematic Literature Review : Ethnomathematics in Geometry. 1990*. http://arxiv.org/abs/2212.11788

Johan, J. R., Iriani, T., Maulana, A., & Negeri, U. (2023). Penerapan Model Four-D dalam Pengembangan Media Video Keterampilan Mengajar Kelompok Kecil dan Perorangan. *Jurnal Pendidikan West Science*, *01*(06), 372–378.

Khumaira, R., Ramadan, R. R., Khairunnisa, S., Miftahul Jannah, & Hendri Marhadi. (2024). Analisis Daya Beda pada Soal tes Mata Pelajaran Matematika di Kelas IV SDN 136 Pekanbaru. *Jurnal Penelitian Ilmiah Multidisipin*, *8*(5), 533–540.

Loi, A. (2023). Pengembangan Modul Kekongruenan dan Kesebangunan untuk Meningkatkan Kemampuan Pemahaman Konsep Matematika. *FAGURU: Jurnal Ilmiah Mahasiswa Keguruan*, *2*(2), 99–23.

Maric, D., Fore, G. A., Nyarko, S. C., & Varma-Nelson, P. (2023). Measurement in STEM education research: a systematic literature review of trends in the psychometric evidence of scales. *International Journal of STEM Education*, *10*(1). https://doi.org/10.1186/s40594-023-00430-x

Mustaqim, M., & Sulisti, H. (2024). Analisis Butir Soal Pas Matematika Peminatan: Daya Pembeda, Tingkat Kesukaran, Dan Kualitas Pengecoh. *Al-'Adad : Jurnal Tadris Matematika*, *3*(1), 44–56. https://doi.org/10.24260/add.v3i1.3011

NCTM. (2000). Principal Standards for School Mathematics. In *National Council of Teachers of Mathematics* (Issue 1). NCTM. http://scioteca.caf.com/bitstream/handle/123456789/1091/RED2017-Eng-8ene.pdf?sequence=12&isAllowed=y%0Ahttp://dx.doi.org/10.1016/j.regsciurbeco.2008.06.005%0Ahttps://www.researchgate.net/publication/30532048

4_SISTEM_PEMBETUNGAN_TERPUSAT_STRATEGI_MELESTARI

Noordin, M., & Darmi, N. (2022). Exploring ESL teachers' alternative assessment strategies and practices in the classroom. *Journal of Language and Linguistic Studies*, *18*(1), 411–426. https://doi.org/10.52462/jlls.191

Nyimbili, F., & Nyimbili, L. (2024). Types of Purposive Sampling Techniques with Their Examples and Application in Qualitative Research Studies. *British Journal of Multidisciplinary and Advanced Studies*, *5*(1), 90–99. https://doi.org/10.37745/bjmas.2022.0419

Obilor, E. I., & Miwari, G. U. (2022). Content Validity in Educational Assessment. *International Journal of Innovative Education Research*, *10*(2), 57–69. www.seahipaj.org

Pasha, V. F., & Aini, I. N. (2022). Deskripsi Kemampuan Pemahaman Konsep Matematis Ditinjau dari Self-Regulated Learning. *Teorema: Teori Dan Riset Matematika*, *7*(2), 235. https://doi.org/10.25157/teorema.v7i2.7217

Permana, F. A. (2021). Upaya Meningkatkan Pemahaman Materi Kesebangunan dan Kekongruenan melalui Metode Praktek Langsung. *Jurnal Serambi PTK*, *VIII*(5), 466–478.

Qomariyah, R. S., Putri, G. A., Putri, D. R., Putri, D. S., & P, M. R. T. (2022). Analisis Tingkat Kesukaran Dan Daya Pembeda Pada Butir Soal Pilihan Ganda Mata Pelajaran Bahasa Indonesia Kelas V Semester 1 SDN Kedungdalem 2. *Jurnal Pendidikan, Sains, Dan Teknologi*, *1*(2), 74–80.

Retno Kuncoro, A., & Martila Ruli, R. (2022). Analisis Kemampuan Pemahaman Konsep Matematis Siswa SMP Pada Materi Relasi dan Fungsi Berdasarkan Teori Honey Mumford. *Jurnal Ilmiah Dikdaya*, *12*(1), 39. https://doi.org/10.33087/dikdaya.v12i1.271

Saputra, H. (2022). Kemampuan Pemahaman Matematis. *PHI: Jurnal Pendidikan Matematika*, *6*(1), 1. https://doi.org/10.33087/phi.v6i1.180

Saputri, H. A., Zulhijrah, Larasati, N. J., & Shaleh. (2023). Analisis Instrumen Assesmen : Validitas, Reliabilitas, Tingkat Kesukaran, dan Daya Beda Butir Soal. *Didaktik : Jurnal Ilmiah PGSD FKIP Universitas Mandiri*, *09*(05), 2986–2995.

Sivakumar, A., & Thirumoorthy, G. (2019). Measurement and Evaluation in Education and Psychology. In *Ideal Publishing Solutions* (1st ed., Vol. 47, Issue 1). A.P.H. Publishing Corporation. https://doi.org/10.2307/1403221

Sriyanti, A., Mania, S., & A, N. H. (2019). Pengembangan Instrumen Tes Diagnostik Berbentuk Uraian Untuk Mengidentifikasi Pemahaman Konsep Matematika Wajib Siswa Man 1 Makassar. *De Fermat : Jurnal Pendidikan Matematika*, *2*(1), 57–69. https://doi.org/10.36277/defermat.v2i1.40

Sutriani, S., Sukmawati, S., & Rukli, R. (2021). Pengembangan instrumen tes hasil belajar matematika berbasis pendekatan kontekstual siswa kelas IV sekolah dasar wilayah II marioriwawo kabupaten soppeng. *Delta-Pi: Jurnal Matematika Dan Pendidikan Matematika*, *10*(1), 1–20. https://doi.org/10.33387/dpi.v10i1.2559

Tadhkiroh, T., Akbar, B., & Hartini, T. I. (2023). Pengembangan Instrumen Penilaian Kinerja pada Muatan IPA Kurikulum 2013 Tingkat Sekolah Dasar. *Jurnal Basicedu*, *7*(1), 631–644. https://doi.org/10.31004/basicedu.v7i1.4720

Thiagarajan, Sivasailam, & Others. (1974). *Instructional Development for Training Teachers of Exceptional Children: A Sourcebook* (I. U. Bloomington. (ed.); 1st ed.). National Center for Improvement of Educational System.