JURNAL EDUSCIENCE (JES)
p-ISSN : 2303 - 355X | e-ISSN : 2685 - 2217
PUBLISHED BY : LPPM of UNIVERSITAS LABUHANBATU

# Development of A Multiple-Choice Test Instrument for Clutch System Learning

Moh Zainul Muttaqin[1], Ranu Iskandar[2]

[1,2]Automotive Engineering Education, Faculty of Engineering, Universitas Negeri Semarang, Indonesia
Email: qinmuttaqin290503@students.unnes.ac.id

**ARTICLE INFO**

*Keywords:*
ADDIE
Clutch System
Multiple-Choice Test

**ABSTRACT**

**Purpose –** This research aims to develop a daily test instrument for clutch system material that is valid, reliable, and has good question item characteristics for students of the Automotive Engineering Education Study Program, Universitas Negeri Semarang.

**Methodology –** The research method used is R&D with the ADDIE model. The test subjects in this study comprised 38 students selected through purposive sampling. Data were collected through a feasibility validation sheet completed by an expert validator and the results of student responses from the instrument trial, consisting of 10 multiple-choice questions. The analysis of empirical data included validity and reliability tests, level of difficulty, discriminative power, and distractor effectiveness.

**Findings –** The results of the analysis showed that all 10 questions were declared empirically valid with values ranging from 0.594 to 0.877 (r_table = 0.329). The reliability coefficient of the instrument was obtained as (r_11) = 0.912 (very high category). The characteristics of the question items showed the proportion of the ideal difficulty level (50% easy, 40% medium, and 10% difficult), the effectiveness of the deceitful function was 83.33% (25 out of 30 deceitful), and the discriminating power was categorized as sufficient to very good (3 adequate, 4 good, and 3 very good).

**Contribution –** This study contributes to the development of a daily test instrument for clutch system material that has been proven to be valid, reliable, and to have good item quality for students of the Automotive Engineering Education Study Program, Universitas Negeri Semarang. The developed instrument can be used as an objective learning evaluation tool.

Received 29 May 2026; Received in revised form 10 June 2026; Accepted 30 July 2026
Jurnal Eduscience (JES) Volume 13 No. 4 (2026)
Available online 30 August 2026

## INTRODUCTION

Higher education aims to prepare students to master the science, technology, and skills required in the workplace, as outlined in Law Number 20 of 2003 on the National Education System (Republik Indonesia, 2003). Achieving this goal depends heavily on quality learning evaluation — a systematic process of collecting, analyzing, and interpreting information to determine the achievement of learning objectives and to inform decision-making (Fuadiy, 2021; Sukoco et al., 2025). Formative quizzes and tests play a strategic role in monitoring student progress, providing feedback to lecturers, and improving the learning process (Arikunto, 2018; Iskandar, 2019).

One of the basic competencies that must be mastered by students of the Automotive Engineering Education Study Program is understanding the clutch system, which includes working principles, components, how it works, disturbances, and maintenance and repair procedures (Okta et al., 2025). Mastery of this competence is crucial, as the clutch system is a critical component in motor vehicles, transmitting power from the engine to the drive wheels (Sugeng, 2013). Students are required not only to understand the theoretical aspects but also to apply this knowledge in practical situations in the workshop (Widianingrum, 2025).

A deep understanding of the clutch system is a crucial foundation for students in the Automotive Engineering Education Study Program to carry out professional vehicle diagnosis, maintenance, and repair. Technically, this mechanism operates on the principle of friction between the clutch discs (clutch disks) and the flywheel, supported by main components such as the diaphragm spring, release bearing, and the drive system. Therefore, mastery of these competencies needs to be measured accurately using standardized assessment instruments that meet validity and reliability criteria (Retnawati, 2016).

Although several studies have examined instrument development in vocational education, most remain limited to general pedagogical evaluation approaches (Fitriyani et al., 2021) or focus on other automotive competencies such as motorcycle injection servicing (Holik et al., 2021) and technician competency testing (Hadiyanto, 2023). To date, no published study has specifically developed and empirically validated a daily test instrument for clutch system competency using a complete Classical Test Theory analysis that simultaneously covers content validity, empirical validity, reliability, difficulty level, discriminating power, and distractor effectiveness. This gap indicates the need for an instrument that is not only theoretically grounded but also empirically proven, so that lecturers can objectively and accurately measure students' mastery of this critical competency. The novelty of this study lies in its integration of expert-based content validation with a comprehensive empirical item analysis tailored to the clutch system material, an approach not addressed in prior vocational assessment research.

The instrument functions to reveal a fact into data, so that if the instrument used in the research has good quality, valid, reliable, with a good level of difficulty, discriminating power and distractor then the data obtained will be in accordance with the facts and the actual situation in the field (Iskandar & Nashiroh, 2026; Zainal, 2017). Various research findings show that the evaluation instruments prepared by teachers still often face weaknesses in the aspects of indicator suitability, item construction, content validity, and empirical analysis of question items (Iskandar et al., 2026; Ramadhani, 2023).

Based on the description above, this study aims to: (1) determine the feasibility of the coupling system test instrument based on expert assessment (content validity) through a questionnaire, (2) determine the level of difficulty of the test instrument question that is developed, (3) determine the discriminating power of the test instrument question that is developed, (4) determine the effectiveness of the distractor in the test instrument developed, (5) determine the empirical validity of the clutch system test instrument, and (6) to determine the reliability of the clutch system test instrument in students of the Automotive Engineering Education Study Program.

This study is expected to contribute both theoretically, by providing an empirical example of comprehensive Classical Test Theory-based instrument development for technical vocational competencies, and practically, by providing lecturers and study programs with a ready-to-adapt model for developing valid and reliable assessment instruments for other automotive competencies.

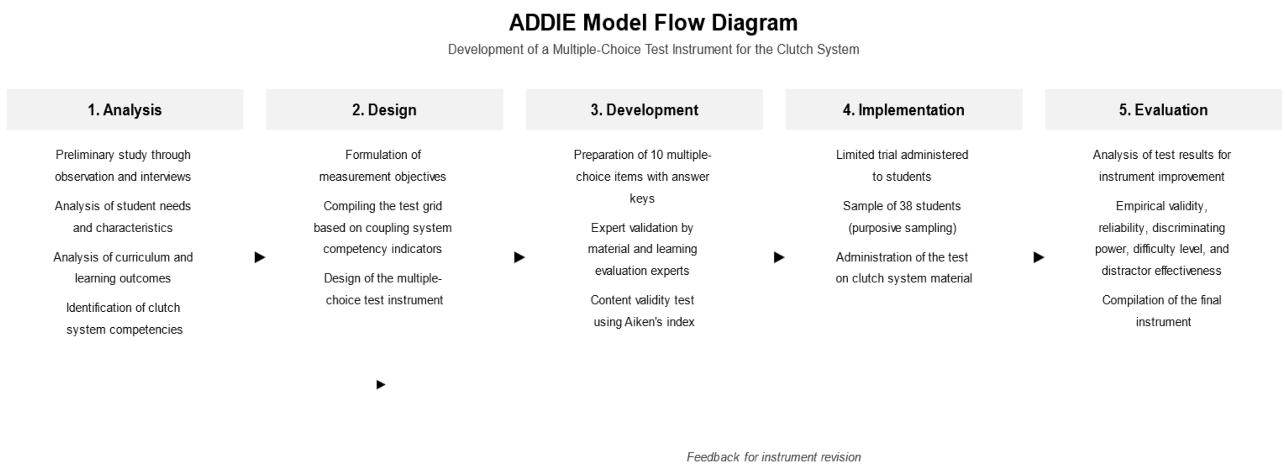
## METHODOLOGY

### Research Design

This study used a Research and Development (R&D) method with a quantitative approach to develop test instruments to validate and ensure the reliability of clutch system materials for use by students in the Automotive Engineering Education Study Program. The development model used follows the stages of formulating measurement objectives, compiling grids, writing question items, expert validation (content validity) through questionnaires, field trials, and empirical analysis of question items using the Aiken validity index for content validity, followed by a test of difficulty and variability (Handayani & Iba, 2020).

The research procedure consists of six sequential steps: (1) compiling the question grid based on learning outcomes, (2) writing question items with answer keys, (3) expert validation of instrument feasibility, (4) administering the trial to students, (5) analyzing test results, and (6) assembling the final instrument, as summarized in Table 3. This structured procedure ensures that each ADDIE stage is operationalized in a clear and replicable manner (Sugiyono, 2019; Nahdliyah, 2025).

The development of instruments in this study used the ADDIE model, which comprises five stages: Analysis, Design, Development, Implementation, and Evaluation. (Maulana et al., 2025; Khoiruddin & Iskandar, 2024). The instrument development process followed the stages of the ADDIE model, which includes: (1) analysis of student needs and characteristics, (2) design of test instruments based on coupling system competency indicators, (3) development of question items and validation by experts using questionnaires, (4) implementation of instruments in trials limited to students, and (5) evaluation of test results for instrument improvement. (Huda et al., 2024; Putra et al., 2025).




### Participant

The population in this study comprised all students of the Automotive Engineering Education Study Program, Faculty of Engineering, Universitas Negeri Semarang, totaling 88 students and spread across 2 classes. The sample comprised one class of 38 students, selected through purposive sampling, because the class had completed the coupling system material and could serve as an instrument test subject (Yuliana & Mahrizal, 2023).

### Data Collection

The data in this study consisted of qualitative and quantitative data. Qualitative data included suggestions and input from automotive lecturers as experts/validators on the suitability of indicators, construction, and language in the question items (Iskandar, 2024; Malik et al., 2024). Quantitative Data in the form of the results of the assessment of the feasibility of the instrument by experts through a questionnaire, and the results of students' answers to the test instruments that were tested. The data sources were obtained from two subject groups: the coupling system material expert and the learning evaluation expert, who served as validators. Students of the Automotive Engineering Education Study Program who have received clutch system material as test subjects.

**Instrument**

The data collection tool was carried out using two types of instruments, namely: (1) Content validation questionnaire in the form of a question item feasibility assessment questionnaire with an assessment scale, used to obtain data from validators/experts. (2) The test instrument in the form of multiple-choice questions, consisting of up to 10 items of coupling system material, is used to obtain empirical data from students. Question items compiled from a grid are first validated by experts using questionnaires, then tested with students (Nahdliyah, 2025).

**Data Analysis**

The data analysis techniques used in this study include content validity (Aiken index), the difficulty level of the question items, the discriminating power, the effectiveness of the distractor (trick), the empirical validity, and the reliability of the instrument. This study uses a classical test theory approach to examine each item using empirical data from students' responses (Fatimah & Alfath, 2019).

***Validity of Content***

The feasibility of instruments is an essential stage in educational research and development (R&D), as it aims to ensure that the instruments meet quality standards that can be scientifically justified prior to testing on students (Sugiyono, 2019). Accordingly, a feasibility assessment is conducted using a validation sheet completed by competent experts (expert judgment) in the field being developed, who are required to hold at least a Master's degree (S2) in education so that their assessments carry adequate academic weight (Widoyoko, 2012). Furthermore, the validation sheet encompasses three main aspects, namely: (1) content feasibility, (2) construction feasibility, and (3) language feasibility (Saputri et al., 2023). The resulting assessment data are then analyzed both quantitatively and qualitatively to determine the instrument's feasibility (Mardapi, 2017). The instrument is declared suitable for use if it obtains the category of "feasible" or "highly feasible" from all validators. If there are comments or suggestions, revisions are made accordingly prior to testing. Thus, expert validation serves as a guarantee that the question instruments possess adequate content feasibility and are capable of measuring what is intended in accordance with the established learning objectives (Arikunto, 2021).

***Difficulty Level***

The difficulty level of the question item is the proportion of test participants who answered a question correctly analysis of the difficulty level is carried out to find out whether the question item is considered easy, medium, or difficult, judging from the student's ability to answer, not from the perspective of the lecturer as the question maker (Susanto et al., 2015). A test instrument is considered good if it contains questions that are proportionally easy, medium, and difficult. The recommended ideal proportion is 3:5:2, with 30% easy, 50% medium, and 20% difficult questions (Fatimah & Alfath, 2019). The formula used is as follows.

Description: P = difficulty index; B = the number of students who answered correctly; JS = the total number of test takers. The difficulty index interpretation criteria are as presented by Thorndike and Hagen in the following Table (Susanto et al., 2015).

**Table 1.** Difficulty Level Criteria for Question Items

| Magnitude P | Interpretation |
|---|---|
| 0.00 - 0.30 | Difficult |
| 0.30 - 0.70 | Medium |
| 0.70 - 1.00 | Easy |

***Discriminating Power***

Differentiating power is the ability of question items to distinguish between students with high ability (upper group) and those with low ability (lower group) in their mastery of the material being tested. Question items are considered good if upper-group students can answer them but lower-group students cannot (Arifin, 2017). The formula of discriminating power is as follows, (Dewi et al., 2019).

**Table 2.** Criteria for Differentiating the Ability of Question Items

| Discriminating Power (D) | Interpretation |
|---|---|
| D < 0.0 (negative) | Very bad |
| D < 0.20 | Bad |
| 0.20 <= D < 0.40 | Enough |
| 0.40 <= D < 0.70 | Good |
| D >= 0.70 | Very good |

***Analysis of the Function of the Distractor***

A distraction, or distractor, is an incorrect answer choice in multiple-choice questions designed to outwit test-takers who do not understand the material well (Gierl et al., 2023). Distractor analysis aims to determine whether each wrong answer choice has effectively deceived test takers (Farida, 2017). A distractor is considered effective if chosen by at least 5% of test takers (Abdul, 2015). The formula for calculating the distractor's effectiveness is as follows.

*ED = (number of test takers who chose the option/number of all test takers) x 100%*

A good test instrument, as a whole, has a percentage of distractors that ranges from 80% to 100% across all items. (Ipin, 2018; Yani et al., 2014).

***Empirical Validity***

The empirical validity of the question item was tested using the biserial point correlation, which is the correlation between the item score (dichotomy: true = 1, false = 0) and the total test score of participants. The biserial point correlation formula is as follows. (Zainal, 2017).

$$rpbis = [(Mp - Mt) / St] \times \sqrt{(p/q)}$$

Remarks: rpbis = biserial point correlation coefficient; Mp = the average score of participants who answered correctly; Mt = average total score; St = the standard deviation of the total score; p = proportion of participants who answered correctly; q = 1 – p. The question item is declared empirically valid if the RPBIS value ≥ the table is at a significance level of 5% (Solichin, 2017).

***Reliability***

The reliability of the instrument indicates the consistency of measurement results when the test is administered to the same group at different times (Nahdliyah, 2025). For multiple-choice test instruments with dichotomy scoring, reliability is calculated using the Kuder-Richardson formula of 20 (KR-20) as follows.

$$r11 = [n / (n - 1)] \times [1 - (\sum pq/St^2)]$$

Remarks: r11 = reliability coefficient; n = number of questions; p = proportion of participants who answered correctly each item; q = 1 – p; $St^2$ = the variance of the total score, the interpretation of the reliability coefficient refers to the criteria: r11 < 0.20 (very low), 0.20–0.40 (low), 0.40–0.70 (adequate), 0.70–0.90 (high), and 0.90–1.00 (very high), the instrument is said to be reliable if the r11 coefficient ≥ 0.70, (Zainal, 2017).

**Ethical Considerations**

This research involves human participants, namely students of the Automotive Engineering Education Study Program at Universitas Negeri Semarang, so in its implementation, it is necessary to adhere to research ethics. Before the instrument trial, the researcher provided students with clear and complete information about the research objectives, procedures to be implemented, and the expected benefits. Approval from the study program was obtained by coordinating with the study program head and the lecturer in charge of the relevant course before the research began. All data obtained from students are kept confidential and used solely for scientific analysis in this study. The student's identity is not included in the report or publication of the research results. Student participation in this study is voluntary and free from coercion by any party. Students are informed that they have the right to refuse to participate or withdraw at any time without negative consequences for grades.

## FINDINGS

The results of the study, structured around the ADDIE model, are presented in the following four stages.

### Analysis

The analysis stage begins with a preliminary study using observations and interviews with lecturers in the Power Transfer System course at the Automotive Engineering Education Study Program, Universitas Negeri Semarang, to identify the learning conditions for the clutch system material. Through student responses, which included small group trials, an average score of 3.43 was obtained by 5 students, and large group trials obtained an average score of 3.13, (Yusuf & Permatadi, 2020). This is the basis for the argument that low learning outcomes should be attributed first to the quality of the evaluation instruments, not solely to student ability. Additionally, a pre-study interview with the course lecturer confirmed that no standardized or empirically validated test instrument had previously been used to assess clutch systems in this study program; evaluations had been conducted using ad hoc questions without systematic item analysis. This dual finding provides the empirical rationale for the present instrument development study.

An analysis of the characteristics of students in the Automotive Engineering Education Study Program was also carried out, given that students in this program tend to be more familiar with workshop practice, so the question items should be prepared with illustrations that closely reflect the real-world operation of clutch system components. Curriculum analysis is conducted on the learning outcomes of the coupling system in the study program's curriculum to ensure alignment between the question indicators and the determined competencies, in line with the importance of needs analysis as the basis for designing evaluation instruments. (Ulva & Fitri, 2022).

### Design

Research procedure steps that are used as a tool to collect data and solve problems in research :

**Table 3.** Research Procedure

| No | Stages | Activities |
|---|---|---|
| 1 | Feeding Question Grid | Compile a grid of questions aligned with the learning outcomes for the coupling system material. |
| 2 | Question Item Writing | Write 10 multiple-choice questions along with answer keys and instructions for work. |
| 3 | Validation of the feasibility of the instrument by experts | Validate question items by experts (automotive lecturers) to assess the suitability of indicators, construction, and language. |
| 4 | Testing for Students | Testing test instruments for students of the Automotive Engineering Education Study Program who have received clutch system materials. |

| No | Stages | Activities |
|---|---|---|
| 5 | Analysis of Test Results | Analyze the test results to obtain data on content validity, difficulty, discriminating power, trick effectiveness, empirical validity, and reliability. |
| 6 | Instrument Assembly | Assemble the clutch system test instrument. |
| Final | - | A good quality instrument to be used effectively in learning evaluation of students in the Automotive Engineering Education Study Program. |

At this stage, a grid of coupling system test instruments is compiled, as shown in Table 4, which includes learning outcomes, question indicators, cognitive levels, and question forms.

**Table 1.** Coupling System Test Instrument Grid (Multiple Choice Questions)

| No | LO | Indicator | Question Indicator | Level | Difficulty Index |
|---|---|---|---|---|---|
| 1 | Carry out maintenance, damage diagnosis, and overhaul of The clutch system, according to standard operational procedures | Determining the main functions of the clutch system in a motor vehicle | Determining the main functions of the clutch system | C1 | Easy |
| 2 | | Identifying the name and function of clutch disc components | Identifying coupling components that serve as friction planes | C1 | Easy |
| 3 | | Understanding the principle of self-adjusting work on hydraulic couplings | Explaining why the hydraulic clutch does not need periodic adjustment | C2 | Medium |
| 4 | | Apply an understanding of the working sequence of clutch components when the pedal is stepped on | Determining the correct working order of clutch components | C3 | Medium |
| 5 | | Apply the concept of clutch pedal free-motion in analyzing statements | Determining the correct statement about the clutch pedal free-motion | C3 | Medium |
| 6 | | Apply the concept of a slip clutch to explain its impact on the vehicle | Explain the meaning and impact of a slip clutch | C3 | Medium |
| 7 | | Apply knowledge of coupling types to identify commonly used ones | Identify the most common types of clutches in light vehicles | C3 | Easy |
| 8 | | Analyzing the symptoms of clutch component damage based on diagnosis | Diagnosing faulty components based on the sound symptoms and the weight of the pedal | C4 | Difficult |
| 9 | | Analyze clutch lining thickness measurement data to determine service actions. | Determining service actions based on clutch lining measurement data | C4 | Difficult |
| 10 | | Evaluate the symptoms of a dragging clutch and determine a comprehensive solution. | Evaluate the causes and solutions of a dragging clutch based on observed symptoms. | C5 | Difficult |

This clutch system test instrument is in multiple-choice format, with 10 questions and answer keys.

**Table 2.** Clutch System Test Instruments

| No | Level | Question Items & Answer Options | Key & Brief Discussion | Difficulty Index |
|---|---|---|---|---|
| 1 | C1 (Remember) | The main function of the clutch system in motor vehicles with manual transmission is...<br>A. Disconnecting and connecting the engine rotation to the transmission<br>B. Converting the engine rotational speed into large torque<br>C. Regulating the fuel pressure to the combustion chamber<br>D. Slowing down the speed of the vehicle with the brake friction<br>E. Stabilizes crankshaft rotation when a heavy load is applied | Key: A. The clutch disconnects and reconnects the power flow from the engine to the transmission, allowing the gears to be shifted. | Easy |
| 2 | C1 (Remember) | The component in the friction clutch that functions as the friction plane between the flywheel and the pressure plate is...<br>A. Release bearing<br>B. Diaphragm spring<br>C. Clutch cover<br>D. Pilot bearing<br>E. Clutch disk | Key: E. The clutch disc (clutch plate) is sandwiched between the flywheel and the pressure plate to transmit the torque. | Easy |

| No | Level | Question Items & Answer Options | Key & Brief Discussion | Difficulty Index |
|---|---|---|---|---|
| 3 | C2 (Understand) | Hydraulic clutches do not require periodic pedal-free motion adjustment as in mechanical clutches because...<br>A. The fluid pressure is always constant, so that the lining cannot wear out<br>B. The hydraulic system uses additional springs that compensate for the wear of the lining<br>C. The hydraulic system is self-adjusting; The fluid volume adjusts automatically to position changes due to lining wear<br>D. Hydraulic clutch uses a hard-to-wear ceramic clutch disc<br>E. The coupling master cylinder detects wear and signals to the ECU for automatic adjustment | Key: C. The hydraulic system is self-adjusting because the fluid automatically fills the space left by the pad's thinning. | Medium |
| 4 | C3 (Apply) | When the clutch pedal is fully pressed, the correct operating order of the clutch components is...<br>A. Pedal -> pressure plate -> release bearing -> diaphragm spring -> free clutch plate<br>B. Pedal -> release bearing -> cable/master cylinder -> release fork -> raised pressure plate<br>C. Pedal -> diaphragm spring -> release bearing -> release fork -> free clutch plate<br>D. Pedal -> cable/master cylinder -> release fork -> release bearing -> diaphragm spring -> pressure plate lifted -> free clutch plate<br>E. Pedal -> clutch disc -> release fork -> release bearing -> pressure plate raised | Key: D. Flow of force flows from pedal - master/cable - release fork - bearing - diaphragm spring - pressure plate lifted. | Medium |
| 5 | C3 (Apply) | Pay attention to the statements: (1) Too little free motion - slip clutch; (2) The free motion is too large - the clutch does not break fully; (3) Ideal free motion: 20-30 mm; (4) Adjustment does not affect pad life. The correct statements are...<br>A. (1), (2), and (4)<br>B. (2), (4), and (3)<br>C. (1), (3), and (4)<br>D. (4), (3), and (1)<br>E. (1), (2), and (3) | Key: E. Statements (1), (2), and (3) are true. Statement (4) is incorrect because the free-motion adjustment greatly affects the life of the pad. | Medium |
| 6 | C3 (Apply) | When a slip clutch results in...<br>A. Excessive friction occurs between the lining and the flywheel when the clutch is fully connected, causing engine power not to be perfectly channeled to the drive wheel<br>B. The clutch connects fully suddenly, causing a shock when accelerating<br>C. The clutch lining is too thick, so the pressure plate cannot press properly<br>D. Wear-release bearings so that they cannot optimally press the diaphragm spring<br>E. The clutch cannot disconnect engine power, making it difficult for the vehicle to change gears | Key: A. Slip clutch causes the lining to be pinched imperfectly, so engine power is lost (weak acceleration & wasteful fuel). | Medium |
| 7 | C3 (Apply) | The most common type of clutch used in light vehicles (passenger cars) with manual transmissions is...<br>A. Fluid coupling<br>B. Dry-type friction clutch<br>C. Electromagnetic coupling<br>D. Wet multi-plate clutch<br>E. Centrifugal coupling | Key: B. Single plate dry-type friction couplings are most common because of their simple construction and high power distribution efficiency. | Easy |
| 8 | C4 (Analyze) | Symptoms: (a) heavy pedal when stepped on, (b) creaking sound when pedal is stepped on, (c) clutch disconnecting/connecting power normally. The faulty components are...<br>A. Worn clutch pads<br>B. Diaphragm springs weaken<br>C. Release bearing damaged/ dry wear<br>D. Master cylinder clutch leaking<br>E. Molor/Ducur coupling cable | Key: C. The creaking sound and heavy pedal when stepped on indicate that the release bearing is dry/worn, but the basic function is still normal. | Difficult |
| 9 | C4 (Analyze) | Lining data: standard thickness 8.0 mm; min service limit 5.0 mm; inspection results 5.2 mm; riveting depth 1.5 mm. The right action is...<br>A. Still suitable for use; Another inspection 10,000 km ahead<br>B. Adjust pedal free motion without the need for replacement<br>C. Suitable for use because the riveting depth is well above the minimum limit | Key: D. The remaining thickness is only 0.2 mm from the minimum limit. The lining must be replaced, as it will wear out before the next service. | Difficult |

| No | Level | Question Items & Answer Options | Key & Brief Discussion | Difficulty Index |
|---|---|---|---|---|
| | | D. Replace the lining immediately due to a very thin service limit difference (0.2 mm)<br>E. Clean the surface of the lining with sandpaper | | |
| 10 | C5 (Evaluate) | Symptoms: (1) difficulty entering gear when the engine is running, (2) the car keeps moving even when the pedal is fully pressed, (3) the gear sounds when passing. When the engine is off, shift gears normally. The solution is...<br>A. Worn clutch lining - replace clutch lining<br>B. Dragging clutch due to free motion is too big - set pedal free motion & fix the drive mechanism<br>C. Broken damper spring - replace clutch disc<br>D. Release Bearing Stuck - Replace Release Bearing<br>E. Cracked flywheel - replace flywheel | Key: B. The car keeps moving when the pedal is fully pressed, indicating that the clutch is not fully disengaging (dragging) due to excessive free play. | Difficult |

Since the assessment format is in the form of a dichotomy score (1 for the correct answer, 0 for the wrong answer), not a multi-level analytical rubric, the preparation of the grid refers to the principle that a good instrument must be formulated based on a clear measurement objective so that the data produced truly represent the measured competencies (Zainal, 2017).

**Development**

The development stage begins with preparing 10 multiple-choice questions based on valid grids and answer keys. Furthermore, the question items are validated by the coupling system materials expert and the learning evaluation expert using a scale-validation questionnaire and the assessment instructions.

Content validation was conducted by two expert validators using a questionnaire-based rating scale (1 to 4), covering three main aspects: (1) material content feasibility, (2) construction feasibility, and (3) language feasibility (Saputri et al., 2023). Both validators awarded a score of 4 (Very Worthy) for all 10 question items across all three aspects. These scores were then analyzed using the Aiken validity index to assess content validity. The results confirm that all question items are content-valid and ready for empirical testing, with minor revisions to language and construction in accordance with the validator's suggestions (Mardapi, 2017; Arikunto, 2021)

The feasibility assessment of the question instrument for content validity uses a rating scale ranging from 1 to 4. A score of 4 indicates that the question item is very worthy of use, and a score of 3 indicates that the question item is eligible for use. A score of 2 indicates that the question item requires minor improvement. A score of 1 indicates that the question item needs to be replaced or deleted. Based on these results, all/part of the question items were declared valid in terms of content and worthy of testing, with revisions to the language and question construction in accordance with the validator's suggestions.

**Implementation**

The revised instrument was then tested with students in the Automotive Engineering Education Study Program, for a total of 38 students. The results of the trial were analyzed from the following four aspects.

***Validity Test***

The empirical validity of the question items was tested using the point-biserial correlation (r_pbis) to assess the correlation between each item's score (true = 1, false = 0) and students' total scores. An item is considered valid if the calculated r_pbis value is greater than or equal to the r_table value. With N = 38 (df = 36, significance level 5%), the r_table value is 0.329. The results of the validity analysis for all 10 question items are presented in Table 6.

Based on the table bellow, all 10 questions of the clutch system test instrument were declared empirically valid with r_pbis values ranging from 0.594 to 0.877, all of which exceeded the value of r_tabel = 0.329. This shows that each question item has a significant positive correlation with the total score, indicating that it accurately measures student competence in the coupling system material.

**Table 6.** Results of Empirical Validity Analysis of Question Items

| Questions | MB | Mt | St. John's | p | q | r_pbis | Remarks |
|---|---|---|---|---|---|---|---|
| 1 | 74.55 | 66.84 | 31.12 | 0.868 | 0.132 | 0.635 | Valid |
| 2 | 73.24 | 66.84 | 31.12 | 0.895 | 0.105 | 0.600 | Valid |
| 3 | 78.39 | 66.84 | 31.12 | 0.816 | 0.184 | 0.782 | Valid |
| 4 | 80.00 | 66.84 | 31.12 | 0.789 | 0.211 | 0.818 | Valid |
| 5 | 82.86 | 66.84 | 31.12 | 0.737 | 0.263 | 0.862 | Valid |
| 6 | 85.38 | 66.84 | 31.12 | 0.684 | 0.316 | 0.877 | Valid |
| 7 | 88.70 | 66.84 | 31.12 | 0.605 | 0.395 | 0.870 | Valid |
| 8 | 92.63 | 66.84 | 31.12 | 0.500 | 0.500 | 0.829 | Valid |
| 9 | 96.00 | 66.84 | 31.12 | 0.395 | 0.605 | 0.757 | Valid |
| 10 | 100.00 | 66.84 | 31.12 | 0.237 | 0.763 | 0.594 | Valid |
| **r_tabel** | **0.329** | **df=36** | **a=5%** | | | **10/10 Valid** | |

***Reliability Test***

The reliability of the instrument was tested using the Kuder-Richardson formula of 20 (KR-20), which is used for multiple-choice tests with dichotomous scoring. An instrument is considered reliable if r11 >= 0.70. Interpretation: < 0.20 (very low), 0.20-0.40 (low), 0.40-0.70 (adequate), 0.70-0.90 (high), 0.90-1.00 (very high).

**Table 7.** Results of KR-20 Reliability Coefficient Calculation

| Components | Value | Remarks |
|---|---|---|
| Number of question items (n) | 10 | |
| Total Spq | 1.7336 | |
| Total score variance (St2) | 9.684 | Scale 0-10 items |
| Coefficient KR-20 (r11) | 0.912 | Very High |
| **Conclusion** | **Reliable** | **r11 >= 0.70** |

Based on the table above, the reliability coefficient (KR-20 clutch system test instrument) is r11 = 0.912, which falls in the Very High category (0.90-1.00). This value exceeds the minimum reliability threshold (r >= 0.70), so the instrument is considered reliable and consistent in measuring students' ability in the clutch system material.

***Difficulty Index***

The difficulty level analysis aims to determine the proportion of students who answered each question item correctly. Items are categorized as Difficult (P < 0.30), Moderate (0.30 <= P <= 0.70), or Easy (P > 0.70). The difficulty level analysis results for all 10 items are presented in Table 8.

**Table 8.** Results of Analysis of the Level of Difficulty of Question Items

| Questions | B | P | Categories | Remarks |
|---|---|---|---|---|
| 1 | 37 | 0.974 | Easy | It is very simple, almost right |
| 2 | 35 | 0.921 | Easy | |
| 3 | 31 | 0.816 | Easy | |
| 4 | 30 | 0.789 | Easy | |
| 5 | 28 | 0.737 | Easy | Medium upper limit |
| 6 | 26 | 0.684 | Medium | |
| 7 | 23 | 0.605 | Medium | |
| 8 | 19 | 0.500 | Medium | Right in the middle |
| 9 | 15 | 0.395 | Medium | |
| 10 | 9 | 0.237 | Difficult | |
| **Quantity** | **253** | **-** | **5 Easy \| 4 Medium \| 1 Difficult** | |

Based on the above, of the 10 question items, there are 5 items (questions 1-5) in the easy category with a P index of > 0.70, 4 items (questions 6-9) in the medium category with P between 0.30-0.70, and 1 item (question

10) in the difficult category with P = 0.237. This distribution shows the dominance of easy questions, indicating the need to add questions at moderate and difficult levels to better reflect students' ability. The recommended ideal proportions are 30% difficult, 50% medium, and 20% easy (Fatimah & Alfath, 2019).

***Distractor***

The distractor effectiveness analysis aims to determine whether each incorrect answer option (distractor) effectively misleads students who do not master the material. A distractor is considered functional if it is chosen by at least 5% of all test participants (>= 2 students out of 38). Each question has 3 distractor options, for a total of 30 distractors. The distractor effectiveness analysis results are presented in Table 9.

**Table 9.** Results of Analysis of the Function of Question Item Distractor

| Q | Key | Option A | Option B | Option C | Option D | Functional Distractors |
|---|---|---|---|---|---|---|
| 1 | A | Key | 1 (2.6%)* | 0 (0.0%)* | 0 (0.0%)* | 0/3 (0.0%) |
| 2 | B | 2 (5.3%) | Key | 1 (2.6%)* | 0 (0.0%)* | 1/3 (33.3%) |
| 3 | C | 3 (7.9%) | 2 (5.3%) | Key | 2 (5.3%) | 3/3 (100%) |
| 4 | A | Key | 3 (7.9%) | 3 (7.9%) | 2 (5.3%) | 3/3 (100%) |
| 5 | D | 4 (10.5%) | 3 (7.9%) | 3 (7.9%) | Key | 3/3 (100%) |
| 6 | B | 5 (13.2%) | Key | 4 (10.5%) | 3 (7.9%) | 3/3 (100%) |
| 7 | C | 6 (15.8%) | 5 (13.2%) | Key | 4 (10.5%) | 3/3 (100%) |
| 8 | D | 7 (18.4%) | 7 (18.4%) | 5 (13.2%) | Key | 3/3 (100%) |
| 9 | A | Key | 9 (23.7%) | 7 (18.4%) | 7 (18.4%) | 3/3 (100%) |
| 10 | C | 10 (26.3%) | 10 (26.3%) | Key | 9 (23.7%) | 3/3 (100%) |
| Total | | | | | | 25/30 functional (83.3%) |

Description: (*) = the distractor does not work (chosen < 5% of participants).

Based on Table 4, of the 30 deceptives analyzed, 25 deceitful (83.33%) were declared to be effective, and 5 deceitful (16.67%) were not functional. Tricks that do not work are all found in question 1 (3 distractors) and question 2 (2 distractors), which are very difficult items, so that almost all students can immediately recognize the correct answer. The 83.33% functional distractor percentage falls within the good category, as it is in the required range of 80%-100% (Ipin, 2018).

***Discriminating Power***

The discriminating power was analyzed to assess each question item's ability to distinguish high-ability students (upper group) from low-ability students (lower group). Students were divided into two groups based on total scores: the upper group (the 50% highest) and the lower group (the lowest 50%), each with 19 students (JA = JB = 19). Interpretation criteria: D < 0.00 (very poor), D < 0.20 (poor), 0.20 <= D < 0.40 (adequate), 0.40 <= D < 0.70 (good), D >= 0.70 (very good), (Erfan et al., 2020).

**Table 10.** Results of Discriminating Power Analysis of Question Items

| Questions | TO | BB | YES | JB | D | Categories |
|---|---|---|---|---|---|---|
| 1 | 19 | 18 | 19 | 19 | 0.053 | Bad |
| 2 | 19 | 16 | 19 | 19 | 0.158 | Bad |
| 3 | 19 | 12 | 19 | 19 | 0.368 | Enough |
| 4 | 19 | 11 | 19 | 19 | 0.421 | Good |
| 5 | 19 | 9 | 19 | 19 | 0.526 | Good |
| 6 | 19 | 7 | 19 | 19 | 0.632 | Good |
| 7 | 19 | 4 | 19 | 19 | 0.789 | Very good |
| 8 | 19 | 0 | 19 | 19 | 1.000 | Very good |
| 9 | 15 | 0 | 19 | 19 | 0.789 | Very good |
| 10 | 9 | 0 | 19 | 19 | 0.474 | Good |
| | | | | | | **2 Bad \| 1 Enough \| 4 Good \| 3 Very Good** |

Based on the table above, the results of the discriminant power analysis indicate that 2 questions (questions 1 and 2) are categorized as weak because they do not effectively distinguish students' abilities across groups. A total of 1 question (question 3) is categorized as sufficient, 4 questions (questions 4, 5, 6, 10) are categorized as good, and 3 questions (questions 7, 8, 9) are categorized as very good. Overall, 8 out of 10 items (80%) had sufficient to very good category differentiation, indicating that the test instrument can distinguish between high- and low-ability students well. Items with poor discriminative power (items 1 and 2) need to be revised or replaced to improve the instrument's quality.

## DISCUSSION

This study successfully developed a valid and reliable clutch system test instrument using the ADDIE development model, directly addressing the research gap identified in the introduction. All 10 items met the empirical validity criterion, as detailed in Table 7. Rather than being incidental, this result reflects the strength of the development process itself: because item construction was preceded by a structured grid (Table 4) and two rounds of expert content validation before empirical trial, weak or ambiguous items were filtered out early, reducing the likelihood of items with poor correlation with scores. This supports the assessment principle that a test instrument's empirical quality is largely determined by the rigor of its earlier design and validation stages, not solely by statistical calculation at the end (Retnawati, 2016; Mardapi, 2017).

The very high KR-20 reliability coefficient further indicates that the instrument produces consistent measurement results regardless of when it is administered. This high consistency is plausible given that all items were empirically valid; in Classical Test Theory, item validity and total-score reliability are interdependent, since removing or weakening a correlated item typically lowers internal consistency (Arikunto, 2018; Syahroni, 2022). Compared to similar instrument-development studies in vocational and technical education, such as the development of a performance-assessment instrument for motorcycle injection servicing (Holik et al., 2021) and a technician competency-assessment instrument (Hadiyanto, 2023), the reliability coefficient obtained in this study is relatively higher, suggesting that combining structured expert validation with full CTT-based empirical analysis — rather than relying on reliability testing alone — yields a more robust instrument. Nevertheless, the difficulty level distribution (50% easy, 40% medium, 10% difficult) remains skewed toward easy items; future item revisions should aim for the recommended 30% easy, 50% medium, and 20% difficult proportion (Fatimah & Alfath, 2019) to more precisely differentiate students across all ability levels.

The distractor effectiveness rate of 83.33% (Table 10) falls within the good category (Ipin, 2018). The two items with non-functional distractors (questions 1 and 2) are also the items with the lowest discriminating power (Table 11), suggesting that their overly straightforward phrasing allowed most students to identify the correct answer without needing to consider the incorrect options — indicating that distractor quality and discriminating power are closely linked and should be revised together, not treated as separate issues.

This study also highlights several limitations. First, the trial was conducted on a limited sample of 38 students from a single study program, which may limit the generalizability of the findings. Second, the instrument consists of only 10 multiple-choice items, which may not comprehensively cover all indicators of clutch system competency. Third, the item analysis was performed using Classical Test Theory (CTT) rather than Item Response Theory (IRT). Fourth, the content validity assessment relied on the judgment of two expert validators, and the involvement of additional validators could enhance the robustness of the validity evidence.

Theoretically, this study provides empirical evidence that a systematic, expert-informed instrument development process can be demonstrated in practice within a specific technical-vocational competency, complementing prior general reviews of vocational assessment approaches (Fitriyani et al., 2021). In practice, this instrument offers a directly usable evaluation tool for lecturers teaching the clutch system material. At the same time, the development process itself — from grid construction through content and empirical validation — can serve as a replicable model for developing assessment instruments for other automotive competencies within the study program. For lecturers, the results suggest that revising the two low-discriminating items (questions 1 and 2) before further use would improve the instrument's diagnostic value. For the institution, adopting a similar validation procedure for other course assessments could help standardize assessment quality across the Automotive Engineering Education Study Program.

**CONCLUSION**

In conclusion, the clutch system test instrument developed in this study met the required criteria of a high-quality instrument based on Classical Test Theory: content validity was confirmed by expert validators, all 10 items were empirically valid, the KR-20 reliability coefficient reached the very high category, the difficulty level distribution covered the full range of student ability, 25 of 30 distractors (83.33%) functioned effectively, and all items showed sufficient to very good discriminating power. Beyond confirming instrument quality, these findings provide empirical evidence that integrating structured expert validation with comprehensive empirical item analysis can yield a robust assessment tool for a specific technical-vocational competency. This approach has been limited in prior automotive assessment research. Given the sample size and single-institution context of this trial, the instrument is suitable for classroom use in the Automotive Engineering Education Study Program. However, broader adoption would benefit from further validation across diverse student cohorts. In practice, this instrument can be used directly by lecturers to evaluate students' mastery of clutch system competency, and its development procedure can guide the study program in producing similarly validated instruments for other automotive courses.

It is important to note that these conclusions are limited by the relatively small, single-program sample (N = 38) and the exclusive use of Classical Test Theory; therefore, the findings should be interpreted within this scope. For future research, it is recommended to increase the trial sample size. For future research, it is recommended to expand the trial sample to include students from multiple institutions or programs to enhance the instrument's generalizability. Future studies may also consider increasing the number of test items to cover all competency indicators of the clutch system more comprehensively. The application of Item Response Theory (IRT) is suggested for a more precise analysis of item parameters, particularly item difficulty, discrimination, and guessing parameters. Additionally, longitudinal studies examining the consistency of instrument performance across different cohorts would further strengthen the evidence of reliability. Collaborative instrument development involving practitioners from the automotive industry is also encouraged to ensure the content relevance of the test items to real-world competency demands.

**REFERENCES**

Abdul, H. (2015). *Evaluasi Pembelajaran Berbasis Kompetensi*. Jakarta: Yrama Widya.

Addzikri, J. A., & Iskandar, R. (2026). Pengembangan Media Pembelajaran Online Berbasis Google Sites Untuk Meningkatkan Hasil Belajar Sistem Rem. *ARMADA : Jurnal Penelitian Multidisiplin*, *4*(6), 2107–2116. https://ejournal.45mataram.ac.id/index.php/armada/article/view/3044/2373

Anshari, M. I., Nasution, R., Irsyad, M., Alifa, A. Z., & Zuhriyah, I. A. (2024). Analisis Validitas dan Reliabilitas Butir Soal Sumatif Akhir Semester Ganjil Mata Pelajaran PAI. *Edukatif: Jurnal Ilmu Pendidikan*, 6(1), 964–975. https://doi.org/10.31004/edukatif.v6i1.5931

Arikunto, S. (2018*). Dasar-Dasar Evaluasi Pendidikan Edisi 3*. Semarang: PT Bumi Aksara.

Arikunto, S. (2021). *Prosedur Penelitian: Suatu Pendekatan Praktik*. Jakarta: PT Rineka Cipta.

Dewi, S. S., Hariastuti, R. M., & Utami, A. U. (2019). Analisis Tingkat Kesukaran dan Daya Pembeda Soal Olimpiade Matematika (OMI) Tingkat SMP Tahun 2018. *Transformasi Jurnal Pendidikan Matematika dan Matematika, 3*(1), 15–26. https://doi.org/10.36526/tr.v3i1.388

Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). Analisis Kualitas Soal Kemampuan Membedakan Rangkaian Seri Dan Paralel Melalui Teori Tes Klasik Dan Model Rasch. *Indonesian Journal of Educational Research and Review, 3*(1), 11–19.

Farida, N. (2017). Analisis Kualitas Soal pada Tingkat SMK. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *6*(2), 201–207.

Fatimah, L. L. U., & Alfath, K. E. (2019). Analisis Kesukaran Soal, Daya Pembeda dan Fungsi Distraktor. *Jurnal Komunikasi dan Pendidikan Islam*, *8*(2), 37–64.

Fitriyani, Y., Nurtanto, M., & Ramdani, S. D. (2021). Pendekatan Evaluasi Pendidikan Vokasi: Sebuah Tinjauan Literatur Sistematis. *Jurnal Pendidikan Vokasi Otomotif*, 4(1), 21–30.

Fuadiy, M. R. (2021). Evaluasi Pembelajaran Sebagai Sebuah Studi Literatur. *DIMAR: Jurnal Pendidikan Islam*, *3*(1), 173–197.

Gierl, M. J., Bulut, O., & Zhang, X. (2023). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research*, *87*(6), 10821116.

Handayani, R., & Iba, Z. (2020). Analisis Karakteristik Butir Soal: Tingkat Kesukaran dan Daya Beda. *Jurnal Pendidikan Sains, 8*(1), 35–43.

Hadiyanto, R. (2023). Pengembangan Instrumen Asesmen Uji Kompetensi Mahasiswa SMK untuk Jabatan Teknisi di Dunia Kerja Bidang Otomotif. *Jurnal Inovasi Pembelajaran Karakter*, 8(3), 38–47.

Holik, A., Sunyoto, S., & Ulfa, S. (2021). Pengembangan Instrumen Penilaian Unjuk Kerja pada Kompetensi Servis Berkala Sepeda Motor Injeksi. *Jurnal Pendidikan Teknik Mesin*, 24(1), 1–8.

Huda, K., Iskandar, R., Darsono, F. B., Hidayat, H., & Ahda, Z. Z. (2024). Pengembangan Buku Ajar untuk Mata Kuliah Praktikum Pemesinan Berbasis Project-Based Learning. *Jurnal Ilmiah Wahana Pendidikan*, *10*(4), 965–970. https://jurnal.peneliti.net/index.php/JIWP/article/view/9003

Ipin, A. (2018). Pengembangan Soal-Soal Pilihan Ganda untuk Mengukur Kemampuan Berpikir Kritis Siswa dalam Konsep Sistem Regulasi Manusia pada Jenjang SMA. *Mangifera Edu, 3*(1), 26–39.

Iskandar, R. (2019). *Pedoman Penilaian Hasil Belajar Peserta Didik SMK Kompetensi Keahlian Teknik Kendaraan Ringan pada Mata Pelajaran Pemeliharaan Sasis dan Pemindah Tenaga Kendaraan Ringan*. CV Jejak (Jejak Publisher).

Iskandar, R. (2024). Assessing the Digital Literacy Profile of Productive Automotive Engineering Teacher Candidate Students. *Journal of Education and Teaching (JET)*, *5*(1), 60–69. https://jet.or.id/index.php/jet/article/view/331

Iskandar, R., & Nashiroh, P. K. (2026). Validity and Reliability of a Research Instrument on the Learning Outcome of Automotive Engineering. *International Journal of Educational Practice and Policy*, *4*(2), 278–294. https://journal.globresco.com/index.php/IJEPP/article/view/890

Iskandar, R., Nashiroh, P. K., & Khumaedi, M. (2026). Qualitative and Quantitative Item Analysis of Essay Test Instruments in a Learning Planning and Microteaching Course. *Journal of Mathematics Instruction, Social Research and Opinion*, *5*(2), 1979–1988. https://journal-gehu.com/index.php/misro/article/view/1508/1843

Khoiruddin, M. A., & Iskandar, R. (2024). Pengembangan Gamifikasi untuk Meningkatkan Motivasi Belajar dan Hasil Belajar pada Materi Sistem AC. *Jurnal Pendidikan Vokasi Otomotif*, *7*(1), 194–214. Retrieved from https://journal.uny.ac.id/index.php/jpvo/article/view/78689

Malik, M. B., Iskandar, R., & Naryanto, R. F. (2024). Development of Android-based mobile learning media to increase learning results in vocational high schools. *Journal of Research in Instructional*, *4*(2), 425–438. https://jurnal.unipa.ac.id/index.php/jri/article/view/462

Maulana, R., Iskandar, R., & Naryanto, R. F. (2025). Pengembangan E-Modul Interaktif Sistem Penerangan Berbasis Pembelajaran Berdiferensiasi. *Panthera : Jurnal Ilmiah Pendidikan Sains Dan Terapan*, *5*(3), 219–231.

Mardapi, D. (2017). *Pengukuran Penilaian & Evaluasi Pendidikan*. Yogyakarta: Parama Publishing.

Nahdliyah, U. (2025). Analisis Reliabilitas Instrumen Tes pada Mata Pelajaran Vokasi. *Jurnal Pendidikan Teknik, 10*(1), 45–58.

Okta, Y. R., Budiman, F. A., Wahyudi, & Iskandar, R. (2025). Efektivitas media pembelajaran interaktif berbasis aplikasi Flash cs6 terhadap kemampuan berpikir tingkat tinggi pada materi sistem kopling. *SIBATIK JOURNAL: Jurnal Ilmiah Bidang Sosial, Ekonomi, Budaya, Teknologi, dan Pendidikan*, *4*(8), 2389–2406. https://publish.ojs-indonesia.com/index.php/SIBATIK/article/view/3264/1352

Putra, R. A., Wibowo, Y., & Susilowati, S. M. E. (2025). Pengembangan Instrumen Evaluasi Berbasis Model ADDIE. *Jurnal Inovasi Pendidikan, 7*(2), 112–125.

Ramadhani, F. (2023). Pengembangan E-Modul Sejarah Materi Masuknya Agama dan Kebudayaan Hindu-Buddha di Nusantara Untuk Meningkatkan Hasil Belajar Siswa Kelas X TKJ 1 SMK Al-Huda Jatimulyo. *Estoria: Journal of Social Science and Humanities*, *3*(2), 476–492. https://doi.org/https://doi.org/10.30998/je.v3i2.1806

Retnawati, H. (2016). *Analisis Kuantitatif Instrumen Penelitian (Panduan Peneliti, Mahasiswa, dan Psikometrian*). Yogyakarta: Parama Publishing.

Saputri, D. (2023). Lembar Validasi: Instrumen yang Digunakan untuk Menilai Produk yang Dikembangkan dalam Penelitian Pengembangan di Bidang Pendidikan**.** *Biology and Education Journal*, *3*(2), 133–151.

Siregar, S. U., Hasibuan, R., Julyanti, E., Siregar, M. (2021). Manajemen peningkatan kualitas pembelajaran matematika di SMA Labuhanbatu. *Journal of Education and Development, 9*(2), 285–290.

Republik Indonesia. (2003). *Undang-Undang Republik Indonesia Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional*. Jakarta: Departemen Pendidikan Nasional.

Solichin, M. (2017). Analisis Daya Beda Soal, Taraf Kesukaran, Validitas Butir Tes, Interpretasi Hasil Tes dan Validitas Ramalan dalam Evaluasi Pendidikan. *Dirasat: Jurnal Manajemen dan Pendidikan Islam, 2*(2), 192–213.

Sugeng, S. (2013). *Pemeliharaan sasis dan pemindah tenaga ringan untuk SMK/MAK kelas XI semester 1.* Jakarta: Direktorat Pembinaan Sekolah Menengah Kejuruan Direktorat Jenderal Manajemen Pendidikan Dasar dan Menengah.

Sugiyono. (2019). *Metode Penelitian Kuantitatif, Kualitatif, dan R&D (Edisi ke-2).* Bandung: Alfabeta.

Sukoco, M. G. A. M. L., Iskandar, R., & Naryanto, R. F. (2025). Diterimanya Hasil Belajar Evaluasi Pembelajaran Kejuruan: Peran Mediasi Motivasi Belajar. *Panthera : Jurnal Ilmiah Pendidikan Sains Dan Terapan*, *5*(4), 1050–1068. https://doi.org/10.36312/panthera.v5i4.698

Syahroni, M. I. (2022). Prosedur Penelitian Kuantitatif. *Al-Musthafa Journal of STIT Al-Aziziyah West Lombok, 2*(3), 211–213.

Sulistyorini, Y., Sunuyeko, N., Rochsun, N., S., & Mufidah, K (2023). Pengembangan evaluasi pembelajaran matematika berbasis game menggunakan platform Wordwall. *Prismatika :Jurnal Pendidikan dan Riset Matematika, 5*(2), 162–175.

Susanto, H., Rinaldi, A., & Novalia, N. (2015). Analisis Validitas, Reliabilitas, Tingkat Kesukaran, dan Daya Beda pada Butir Soal Ujian Akhir Semester Ganjil Mata Pelajaran Matematika. *Al-Jabar: Jurnal Pendidikan Matematika, 6*(2), 203–218.

Syahid, A. A., Juanda, A., & Supiandi, M. I. (2024). Penerapan Model ADDIE dalam Pengembangan Instrumen Penilaian Berbasis Kompetensi. *Jurnal Penelitian Pendidikan, 24*(1), 88–101.

Ulva, D. Y. & Amalia Fitri, A. (2022). Analisis Kebutuhan Modul Matematika untuk Meningkatkan Kemampuan Pemecahan Masalah Siswa SMP N 4 Batang. *Journal of Nusantara Education*, *2*(1), 11–21. https://doi.org/10.57176/jn.v2i1.38

Widianingrum, I. D. (2025). Implementation of the 2013 Curriculum on Islamic Religious Education Learning. *Quality Journal of Education Arabic and Islamic Studies,* 3(3), 214–221.

Widoyoko, E. P. (2012). *Evaluasi Program Pembelajaran: Panduan Praktis bagi Pendidik dan Calon Pendidik*. Yogyakarta: Pustaka Pelajar.

Yani, A., Asri, A. F., & Burhan, A. (2014). Analisis Tingkat Kesukaran, Daya Pembeda, dan Fungsi Distraktor Soal Ujian Semester Ganjil Mata Pelajaran Produktif di Smk Negeri 1 Indralaya Utara Tahun Pelajaran 2012/2013. *Jurnal Pendidikan Teknik Mesin, 1*(2), 98–115.

Zainal, A. (2017). Kriteria Instrumen dalam suatu Penelitian. *Journal of THEOREMS (The Original Research of Mathematics), 2*(1), 28–36.

Zulpan, S., Yusron, A., Hanida, R. S., & Marfu'ah, S. (2024). Validitas Tes Objektif Bentuk Pilihan Ganda. *Jurnal Pendidikan Matematika*, 1(3), 10. https://doi.org/10.47134/ppm.v1i3.812