

# Analisis Butir Soal Evaluasi Pembelajaran Bab Suhu Kalor dan Pemuaian Kelas VII SMP Menggunakan Aplikasi Iteman

Najwa Hamidah<sup>1</sup>, Dwi Tami<sup>2</sup>, Lu'luk Atul Jannah<sup>3</sup>

<sup>1</sup>Universitas Negeri Semarang, Semarang

<sup>1</sup>najwahamidah2@students.unnes.ac.id, <sup>2</sup>dwitami058@students.unnes.ac.id, <sup>3</sup>luluatujannahjekulo@students.unnes.ac.id

## ARTICLE INFO

### Article history:

Received 17 June 2025

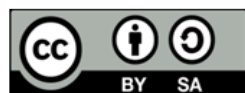
Revised 2 June 2026

Accepted 4 June 2026

Available online 30 June 2026

### Keywords:

analisis butir soal; evaluasi pembelajaran;  
teori tes klasik; iteman



This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.  
Copyright © 2026 by Author. Published by Universitas  
Sebelas Maret.

## ABSTRAK

Penelitian ini bertujuan untuk menganalisis kualitas butir soal evaluasi pembelajaran dalam topik Suhu, Kalor, dan Pemuaian di kelas VII SMP. Penelitian menggunakan pendekatan kuantitatif deskriptif berbasis Teori Tes Klasik (TTK) dan memanfaatkan perangkat lunak ITEMAN versi 4.3. Sebanyak 30 soal pilihan ganda yang sebelumnya telah divalidasi oleh tujuh ahli dianalisis dengan memperhatikan lima aspek utama, yaitu validitas isi, tingkat kesukaran, daya pembeda, efektivitas pengecoh, dan reliabilitas instrumen. Hasil kajian menunjukkan bahwa sebagian besar soal tergolong mudah dari sisi tingkat kesukaran. Namun begitu, daya pembeda dan reliabilitas instrumen masih rendah, menandakan soal-soal tersebut belum sepenuhnya mampu membedakan kemampuan peserta didik secara akurat. Validitas isi tergolong cukup baik dengan nilai rata-rata Aiken's V sebesar 0,898, meskipun terdapat dua soal yang belum memenuhi kriteria dan disarankan untuk diperbaiki. Kesimpulan dari temuan ini menekankan perlunya evaluasi kuantitatif untuk memastikan bahwa instrumen benar-benar berfungsi sebagaimana mestinya dalam

mengukur pemahaman peserta didik. Maka dari itu, penelitian ini merekomendasikan adanya perbaikan pada butir soal yang kurang baik serta pelatihan lanjutan bagi guru dalam hal penyusunan dan analisis soal berbasis data.

## ABSTRACT

*This study aimed to analyze the quality of test items used in learning evaluations for the topic of Heat, Temperature, and Expansion in seventh-grade junior high school science classes. It followed a descriptive quantitative approach based on Classical Test Theory (CTT) and utilized ITEMAN version 4.3 for item analysis. A total of 30 multiple-choice questions, previously validated by seven experts, were examined using five main indicators: content validity, item difficulty, discriminating power, distractor effectiveness, and instrument reliability. The findings revealed that most items fell into the "easy" category based on their difficulty index. However, both the discriminating power and the overall reliability of the test were still considered low. This suggests that the items were not yet effective in differentiating students' actual levels of understanding. In terms of content validity, the results were relatively strong, with an average Aiken's V score of 0.898. Even so, two items did not meet the criteria and should be revised. The study underlines the importance of quantitative analysis in ensuring that assessment instruments genuinely measure student comprehension. Based on the results, the research recommends revising weaker items and providing more intensive training for teachers in constructing and analyzing assessments using data-driven methods to improve the overall quality and fairness of student evaluations.*

## 1. PENDAHULUAN

Evaluasi pembelajaran menjadi bagian penting yang tidak bisa dipisahkan dari keseluruhan proses pendidikan. Perannya sangat penting karena membantu melihat seberapa jauh tujuan pembelajaran sudah tercapai. Selain itu, evaluasi juga berfungsi untuk menilai seberapa efektif proses belajar-mengajar (Fuad et al., 2023). Dari evaluasi butir soal ini, guru dan siswa bisa mendapat umpan balik yang membangun. Jika dilakukan secara sistematis dan tepat, hasil evaluasi mampu memberikan gambaran objektif tentang capaian kompetensi siswa. Hasil dari evaluasi ini pun bukan sekadar angka, tetapi jadi acuan penting dalam membuat keputusan yang berdasarkan data nyata. Dalam pembelajaran sains, nilai evaluasi menjadi hal yang penting. Karena sains tidak hanya menekankan pada pemahaman konsep, tetapi juga mengandalkan cara berpikir logis serta keterampilan dalam melakukan proses ilmiah (Sutrisna, 2021).

Salah satu materi penting dalam pembelajaran sains adalah Suhu, Kalor, dan Pemuaian. Materi ini bukan sekadar konsep teori, tapi juga memiliki hubungan langsung dengan peristiwa sehari-hari. Contohnya bisa dilihat dari bagaimana logam memuai saat terkena panas, cara kerja termometer, sampai dengan pemuaian pada rel kereta. Karena itu, pemahaman siswa terhadap topik ini harus dievaluasi secara menyeluruh agar tidak hanya sekadar hafalan, tetapi benar-benar memahami konsep dan aplikasinya. Tes pilihan ganda menjadi alat ukur yang sering digunakan karena dinilai praktis dan mampu mencakup banyak indikator sekaligus, karena proses pemeriksaannya cepat dan objektif (Kurniawan, 2022). Namun, kualitas soal dalam bentuk ini sering kali belum dievaluasi secara mendalam, sehingga berisiko tidak mencerminkan capaian belajar secara akurat.

Analisis butir soal menjadi langkah penting untuk menjamin kualitas instrumen evaluasi. Tujuannya agar setiap soal benar-benar mampu mengukur kompetensi siswa secara tepat. Salah satu cara yang bisa dipakai yaitu pendekatan konvensional dalam tes, yang menyediakan beberapa ukuran untuk menilai mutu soal (Darmansyah et al., 2025). Ukuran ini meliputi tingkat kesukaran, daya pembeda, dan efektivitas pengecoh. Tingkat kesukaran memberi tahu seberapa menantang soal itu bagi siswa. Daya pembeda menunjukkan apakah soal bisa membedakan siswa yang benar-benar menguasai materi dengan yang belum. Sedangkan efektivitas pengecoh berkaitan dengan seberapa “menarik” pilihan jawaban yang salah bagi siswa yang belum memahami materi secara keseluruhan. Ketiga ukuran ini memberi gambaran lengkap soal karakteristik dan kekuatan setiap butir soal (Muluki, 2020).

Untuk meningkatkan mutu evaluasi, penggunaan teknologi dalam analisis soal menjadi sangat penting. Salah satu alat yang bisa dimanfaatkan adalah program komputer bernama Iteman (*Item and Test Analysis*). Iteman bekerja dengan pendekatan konvensional dan dirancang khusus untuk menganalisis soal pilihan ganda secara kuantitatif. Program ini mampu menghasilkan data statistik yang menunjukkan kualitas suatu soal, seperti tingkat kesukaran, daya pembeda, dan efektivitas pengecoh. Lewat bantuan Iteman, guru atau pendidik bisa mengevaluasi soal dengan cara yang lebih cepat, efisien, dan hasilnya pun lebih akurat (Qadir et al., 2024). Tidak hanya itu, hasil analisis dari program ini bisa dijadikan dasar kuat untuk memperbaiki soal-soal yang belum optimal.

Walaupun teknologi seperti Iteman sudah tersedia, kenyataannya penggunaannya di sekolah masih jauh dari maksimal. Banyak guru belum terbiasa menggunakan program analisis soal, bahkan sebagian di antaranya belum memahami dasar-dasar pendekatan konvensional dalam tes (Rosardi, 2015). Akibatnya, banyak soal yang dipakai dalam evaluasi siswa tidak melewati proses analisis terlebih dulu. Sehingga, bisa menyebabkan kualitas soal tersebut diragukan dan hasil penilaiannya tidak benar-benar mencerminkan kemampuan siswa. Situasi ini menjadi tantangan besar dalam upaya memperbaiki mutu pembelajaran, apalagi dalam konteks sains yang sangat butuh evaluasi yang sah dan bisa dipercaya. Karena itu, pelatihan dan pendampingan untuk guru dalam hal analisis soal perlu diperkuat agar memiliki bekal yang cukup untuk menghasilkan instrumen evaluasi yang lebih baik dan berbasis data (Wijayati, 2025).

Melihat kondisi di atas, penting untuk melakukan kajian mendalam terhadap kualitas butir soal pada materi Suhu, Kalor, dan Pemuaian. Kajian ini menggunakan pendekatan angka melalui teori konvensional dalam tes, dengan bantuan program Iteman. Tujuan utamanya yaitu mengevaluasi sejauh mana soal-soal yang ada sudah memenuhi kriteria soal yang baik. Dengan kajian ini, dapat juga diidentifikasi soal mana yang perlu direvisi atau diperbaiki. Hasilnya diharapkan bisa menjadi panduan bagi guru dalam menyusun instrumen evaluasi yang lebih akurat dan tepat sasaran. Selain itu, penelitian ini juga bisa dijadikan bahan belajar bagi calon guru untuk memahami pentingnya menganalisis soal. Dengan mengetahui karakter soal secara lebih dalam, guru bisa mengambil keputusan pembelajaran yang lebih tepat dan berdampak langsung pada kualitas hasil belajar siswa.

## 2. METODE

Penelitian ini menggunakan pendekatan kuantitatif deskriptif dengan tujuan untuk menganalisis kualitas butir soal evaluasi pembelajaran pada bab Suhu, Kalor, dan Pemuaian kelas VII SMP. Pendekatan ini digunakan karena mampu memberikan gambaran yang jelas dan objektif mengenai kualitas instrumen berdasarkan data numerik yang diperoleh dari hasil tes siswa. Melalui pendekatan deskriptif, penelitian ini menggambarkan karakteristik soal secara sistematis, sementara pendekatan kuantitatif memungkinkan analisis statistik yang mendalam terhadap hasil jawaban siswa (Rahim, Jainuddin & Maheswari, 2025). Pengumpulan data dilakukan melalui tes pilihan ganda, dan hasilnya dianalisis untuk mengetahui sejauh mana soal memenuhi kriteria sebagai alat ukur yang baik. Model analisis yang digunakan dalam penelitian ini mengacu pada Teori Tes Klasik (Classical Test Theory), yang umum digunakan dalam evaluasi pendidikan untuk memprediksi hasil tes dengan menghubungkan skor yang diperoleh peserta dengan karakteristik soal secara umum. Dalam pendekatan ini, kualitas suatu butir soal ditentukan oleh sejumlah parameter seperti tingkat kesukaran, daya pembeda, efektivitas pengecoh, dan reliabilitas instrumen. Namun, karakteristik yang dihasilkan oleh teori tes klasik bersifat relatif terhadap kelompok peserta tes, sehingga hasil analisis bisa berubah tergantung pada kemampuan populasi yang diuji (Hardianti, Liliawati, & Tayubi, 2023).

Instrumen utama berupa 30 butir soal pilihan ganda yang kami sekelompok susun berdasarkan kisi-kisi yang kami buat dengan disesuaikan kompetensi dasar dan indikator capaian pembelajaran dalam kurikulum. Soal-soal tersebut mencakup berbagai level kognitif dalam Taksonomi Bloom revisi, mulai dari level C1 (mengingat) hingga

C6 (mencipta). Dalam proses penyusunan soal pilihan ganda, perlu dipertimbangkan tiga aspek penting, yaitu kesesuaian isi materi, ketepatan konstruksi soal, dan kejelasan penggunaan bahasa, agar setiap butir soal dapat memenuhi kaidah penulisan yang baik dan benar (Pramanda & Aisiah, 2023). Setelah penyusunan, soal-soal tersebut selanjutnya divalidasi oleh tujuh validator yang meliputi empat guru IPA SMP dan tiga mahasiswa Prodi Pendidikan IPA tingkat akhir yang telah memiliki pengalaman dalam penyusunan instrumen evaluasi. Validasi dilakukan dengan menelaah kisi-kisi, kartu soal, dan isi butir soal, serta mengisi lembar validasi yang telah kami sediakan. Proses ini dilakukan untuk mengetahui sejauh mana instrumen yang disusun sudah sesuai dengan kompetensi yang ingin diukur. Sesuai dengan pendapat Nabi et al. (2022), validitas isi digunakan untuk menilai sejauh mana butir-butir dalam suatu instrumen benar-benar mengukur keterampilan atau kompetensi yang seharusnya diukur sesuai dengan tujuan evaluasi.

Penilaian pada lembar validasi dilakukan menggunakan skala Likert dengan rentang nilai 1 sampai 4 dan disertai keterangan pada setiap butir soal. Menurut Busnawir et al. (2025), skala Likert merupakan alat ukur yang efektif untuk menilai sikap atau persepsi karena memberikan gradasi penilaian yang jelas dan mudah diinterpretasikan. Data yang diperoleh dari keseluruhan penilaian para validator kemudian dianalisis menggunakan indeks Aiken's V untuk menentukan tingkat validitas isi dari masing-masing soal. Salah satu keunggulan utama dari indeks Aiken terletak pada kemudahan perhitungannya dan kejelasan interpretasinya, karena indeks ini menyajikan nilai numerik yang secara langsung menggambarkan sejauh mana item tertentu dinilai relevan oleh para ahli (Utami, Ilahi, & Ratih, 2024). Menurut Aiken (1980), nilai Aiken's V dapat dikategorikan ke dalam tiga tingkat validitas, yaitu rendah jika bernilai kurang dari 0,4, sedang jika berada di antara 0,4 hingga 0,8, dan tinggi jika nilainya lebih dari 0,8. Soal-soal yang belum mencapai kriteria validitas minimum direvisi berdasarkan masukan dari para validator agar lebih sesuai dengan indikator pembelajaran yang dituju. Setelah direvisi, soal diujicobakan kepada 35 siswa SMP N 38 Semarang kelas VII menggunakan platform digital Google Form. Jumlah tersebut telah memenuhi syarat minimal untuk penelitian kuantitatif, di mana menurut Sugiyono (2016), jumlah sampel minimal dalam penelitian kuantitatif adalah 30 responden.

Data hasil uji coba dianalisis menggunakan aplikasi Iteman versi 4.3, yang merupakan perangkat lunak berbasis Teori Tes Klasik. Analisis dilakukan terhadap empat parameter atau kriteria utama sebagai berikut:

**Tabel 1.** Kriteria Reliabilitas

| Interval             | Kriteria      |
|----------------------|---------------|
| $0,80 < r \leq 1,00$ | Sangat tinggi |
| $0,60 < r \leq 0,80$ | Tinggi        |
| $0,40 < r \leq 0,60$ | Sedang        |
| $0,20 < r \leq 0,40$ | Rendah        |
| $0,00 < r \leq 0,20$ | Sangat Rendah |

Sumber: (Fietri et al., 2021)

**Tabel 2.** Kriteria Tingkat Kesukaran Butir Soal

| Koefisien   | Kategori |
|-------------|----------|
| 0,00 — 0,30 | Sukar    |
| 0,31 — 0,70 | Sedang   |
| 0,71 — 1,00 | Mudah    |

Sumber: (Dia & Fadhilah, 2022)

**Tabel 3.** Kriteria Efektivitas Pengecoh

| Koefisien   | Kategori    | Keputusan               |
|-------------|-------------|-------------------------|
| 0,40 — 1,00 | Baik        | digunakan tanpa revisi  |
| 0,30 — 0,39 | Cukup baik  | digunakan dengan revisi |
| 0,20 — 0,29 | Kurang baik | perlu revisi            |
| 0,00 — 0,19 | Tidak baik  | dibuang atau diganti    |

Sumber: (Radja et al., 2023)

**Tabel 4.** Kriteria Efektivitas Pengecoh

| Koefisien | Kaputusan         |
|-----------|-------------------|
| 0,05      | Pengecoh Diterima |
| < 0,05    | Pengecoh Direvisi |
| 0,00      | Pengecoh Ditolak  |

Sumber: (Arbiatin & Mulabbiyah, 2022)

Hasil dari keseluruhan analisis digunakan untuk mengevaluasi kualitas setiap butir soal serta memberikan rekomendasi perbaikan terhadap soal-soal yang belum memenuhi kriteria sebagai instrumen evaluasi yang valid dan reliabel. Dengan metode ini, diharapkan instrumen yang dikembangkan dapat menjadi alat ukur yang tepat dalam mengevaluasi pemahaman siswa pada materi Suhu, Kalor, dan Pemuain.

### 3. HASIL DAN PEMBAHASAN

Sebelum soal-soal diberikan kepada siswa kelas VII SMP, sebanyak 30 butir soal disusun terlebih dahulu dengan mengacu pada kisi-kisi yang sudah dirancang sesuai dengan indikator pembelajaran dalam bab Suhu, Kalor, dan Pemuain. Setiap soal dilengkapi dengan perangkat pendukung seperti kartu soal, lembar validasi, dan kisi-kisi soal. Proses validasi melibatkan tujuh orang validator dengan, empat guru IPA dan tiga mahasiswa tingkat akhir dari Program Studi Pendidikan IPA. Validasi isi ini bertujuan untuk menilai apakah soal-soal yang disusun sudah mengikuti prinsip-prinsip penyusunan soal pilihan ganda yang baik (Sukmasari & Rosana, 2017). Penilaian dilakukan menggunakan lembar validasi yang mencakup tiga aspek utama: isi materi, konstruksi soal, dan kejelasan bahasa.

Subjek dalam penelitian ini melibatkan tujuh validator ahli dan 35 siswa kelas VII SMP Negeri 38 Semarang yang menjadi responden dalam uji coba soal. Mengacu pada ketentuan dari Herawati (2022), jumlah ini sudah mencukupi untuk dianalisis menggunakan statistik deskriptif. Jawaban siswa kemudian diketik ulang dan dianalisis memakai program Itean versi 4.3 yang berbasis pada Teori Tes Klasik. Analisis difokuskan pada empat parameter utama: tingkat kesukaran, daya pembeda, efektivitas distraktor, dan reliabilitas tes (Farida & Musyarofah, 2021). Sebelum membahas keempat aspek tersebut lebih lanjut, penting untuk ditegaskan bahwa semua soal telah memenuhi standar validitas isi, seperti yang ditunjukkan oleh hasil telaah dari para ahli.

Uji reliabilitas dalam penelitian ini dilakukan menggunakan program Itean dengan rumus *Alpha Cronbach*. Hasilnya, nilai Alpha yang diperoleh sebesar 0,446, yang menurut kriteria dari Arisandi (2020) termasuk dalam kategori rendah karena berada di bawah 0,50. Berarti bahwa konsistensi antar butir soal masih belum kuat dan perlu ditingkatkan agar alat evaluasi yang digunakan menjadi lebih andal. Dari segi tingkat kesukaran, hasil analisis terhadap 30 soal menunjukkan bahwa sebagian besar, yaitu 28 soal, masuk dalam kategori mudah. Dua soal lainnya berada pada tingkat sedang, dan tidak ada soal yang tergolong sulit. Kategori ini ditentukan berdasarkan jumlah siswa yang berhasil menjawab benar masing-masing soal, sehingga memberikan gambaran sejauh mana soal-soal tersebut menantang bagi peserta didik.

Berdasarkan hasil analisis efektivitas distraktor, sebanyak 6 soal memiliki 1 distraktor yang harus diganti (butir 12, 14, 23, 28, 29, 30), 5 soal memiliki 2 distraktor yang harus diganti (butir 1, 3, 6, 7, 16), 7 soal memiliki 3 distraktor yang harus diganti (butir 5, 10, 13, 15, 19, 20, 25), dan 3 soal memiliki 4 distraktor yang harus diganti (butir 2, 4, 9). Sementara itu, 9 soal memiliki distraktor yang berfungsi dengan baik (butir 8, 11, 17, 18, 21, 22, 24, 26, 27).

Secara umum, meskipun soal-soal pada materi Suhu, Kalor, dan Pemuain sudah memenuhi aspek validitas isi dan menunjukkan variasi tingkat kesukaran, masih terdapat beberapa hal yang perlu dibenahi. Daya pembeda dan reliabilitas instrumen masih tergolong rendah, sehingga belum sepenuhnya mampu membedakan tingkat kemampuan siswa dengan tepat. Selain itu, beberapa distraktor atau opsi jawaban salah juga belum bekerja secara efektif dan masih perlu disesuaikan agar bisa menjalankan fungsinya secara maksimal. Temuan ini menjadi landasan penting untuk melakukan revisi lanjutan pada soal-soal tersebut, agar instrumen evaluasi yang dihasilkan bisa lebih tajam dan akurat dalam mengukur pemahaman siswa terhadap materi Suhu, Kalor, dan Pemuain.

### 3.1. Spesifikasi Butir Soal

**Tabel 5.** Spesifikasi Butir Soal

| Sub Bab      | <i>Behavioral objectives (Bloom's Taxonomy Cognitive Level)</i> |                                   |                           |                    |                     |                   | <b>Total (100%)</b> |
|--------------|-----------------------------------------------------------------|-----------------------------------|---------------------------|--------------------|---------------------|-------------------|---------------------|
|              | <i>Remember (%)</i>                                             | <i>Understand (%)</i>             | <i>Apply (%)</i>          | <i>Analyze (%)</i> | <i>Evaluate (%)</i> | <i>Create (%)</i> |                     |
| Suhu         |                                                                 | 9<br>(1, 2, 3, 4, 5, 6, 7, 9, 10) | 1<br>(8)                  |                    |                     |                   | 10                  |
| Kalor        |                                                                 |                                   | 5<br>(11, 12, 14, 15, 17) | 3<br>(13, 18, 20)  | 1<br>(16)           | 1<br>(19)         | 10                  |
| Pemuaian     | 2<br>(21, 24)                                                   | 4<br>(25, 27, 28, 30)             | 2<br>(23, 29)             | 2<br>(22, 26)      |                     |                   | 10                  |
| <b>Total</b> | <b>2</b>                                                        | <b>13</b>                         | <b>8</b>                  | <b>5</b>           | <b>1</b>            | <b>1</b>          | <b>30</b>           |

Berdasarkan Tabel Spesifikasi Butir Soal, terlihat bahwa sebagian besar soal berada pada level kognitif *Understand*, dengan jumlah 13 dari total 30 soal. Sub bab “Suhu” paling mendominasi pada level ini dengan 9 soal, yang menunjukkan fokus besar pada pemahaman dasar konsep suhu. Untuk level *Apply*, ada 8 soal yang tersebar merata di ketiga sub bab, meski lebih banyak muncul di bagian “Kalor.” Sementara itu, level *Analyze* mencakup 5 soal, sebagian besar berasal dari sub bab “Kalor” dan “Pemuaian,” mencerminkan pentingnya kemampuan siswa dalam menganalisis fenomena yang berkaitan dengan perubahan suhu, kalor, dan pemuaian. Hanya ada 2 soal di level *Remember*, semuanya dari sub bab “Pemuaian,” yang menunjukkan bahwa hafalan bukan prioritas dalam evaluasi ini. Level *Evaluate* dan *Create* masing-masing hanya diwakili oleh 1 soal, keduanya dari sub bab “Kalor,” menandakan bahwa pengujian terhadap kemampuan berpikir evaluatif dan kreatif masih sangat terbatas. Meskipun begitu, pembagian jumlah soal pada setiap sub bab cukup merata, masing-masing terdiri dari 10 soal. Total soal sebanyak 30 ini sudah mencakup enam level kognitif dari Taksonomi Bloom, meskipun distribusinya lebih menekankan pada pemahaman konsep dan penerapan dasar dibanding pengembangan keterampilan berpikir tingkat tinggi.

### 3.2. Validitas Isi

Validitas isi merupakan bagian penting dalam pengembangan instrumen evaluasi karena berkaitan langsung dengan sejauh mana materi soal mewakili kompetensi dan indikator yang ditetapkan dalam kurikulum. Dalam pembelajaran IPA, khususnya materi Suhu, Kalor, dan Pemuaian, validitas isi menjadi penentu apakah butir soal benar-benar sesuai dengan isi materi dan tingkat kognitif yang diharapkan (Zubaidah, 2018). Soal dengan validitas isi yang baik tidak hanya relevan tapi juga bermakna dalam mengukur pencapaian siswa. Dalam penelitian ini, validitas isi diuji oleh tujuh validator, gabungan dari guru SMP dan mahasiswa tingkat akhir dari Program Studi Pendidikan IPA. Validasi dilakukan terhadap 30 soal pilihan ganda yang sebelumnya disusun berdasarkan kisi-kisi yang mengacu pada kompetensi dasar serta indikator pembelajaran. Penilaian dilakukan melalui lembar validasi yang mencakup tiga aspek utama: kesesuaian materi dengan indikator, logika serta sistematika soal, dan kejelasan penggunaan bahasa. Skor diberikan dengan skala Likert 1 sampai 4, di mana skor paling tinggi mencerminkan kualitas terbaik (Widiyanto & Hayati, 2022). Dari hasil validasi tersebut, sebanyak 28 soal masuk dalam kategori tinggi, sedangkan dua lainnya dinilai sedang karena memperoleh skor CVI sebesar 0,762.

Soal nomor 1 dan 4 dinyatakan berada dalam kategori sedang untuk validitas isi. Nilai sedang ini muncul karena ada beberapa kelemahan penting pada kedua soal tersebut. Soal nomor 1, misalnya, belum benar-benar sesuai dengan indikator yang ingin diukur, sehingga arah pertanyaannya masih belum jelas. Sementara soal nomor 4 menggunakan redaksi yang kurang jelas dan strukturnya tidak rapi, bahkan ada pilihan jawaban yang maknanya bisa ditafsirkan ganda. Selain itu, soal ini juga belum berhasil menampilkan kedalaman berpikir yang sesuai dengan level kognitif dalam taksonomi Bloom yang sudah diperbarui. Karena itu, keduanya direkomendasikan untuk direvisi secara menyeluruh agar kualitas soal meningkat dan indikator pembelajaran bisa terwakili dengan lebih tepat.

Soal nomor 1 dinilai berada dalam kategori sedang karena tidak sepenuhnya mewakili indikator pembelajaran yang ingin dicapai, yaitu kemampuan siswa dalam mengidentifikasi alat ukur suhu secara umum. Soalnya hanya menyoroti pengukuran suhu tubuh manusia, sehingga ruang lingkupnya terlalu sempit dan tidak menggambarkan variasi alat ukur suhu yang lebih luas. Selain itu, distraktor atau pilihan jawaban yang salah terlalu mudah ditebak, karena tidak cukup kuat untuk mengecoh siswa yang belum paham. Hal ini membuat soal kurang efektif dalam menguji kemampuan berpikir kritis. Supaya lebih baik, soal ini perlu direvisi dengan memperluas konteksnya, misalnya dengan mencantumkan situasi yang melibatkan pengukuran suhu benda, zat cair, atau udara agar cakupan indikator lebih terpenuhi.

Walaupun ada dua soal yang masuk kategori sedang, secara keseluruhan validitas isi instrumen menunjukkan hasil yang cukup kuat. Rata-rata skor CVI mencapai angka 0,898, yang menggambarkan adanya tingkat kesepakatan tinggi di antara para validator. Sebagian besar soal memperoleh skor antara 0,857 sampai 1,000, artinya butir-butir soal tersebut sudah cukup menggambarkan indikator pembelajaran secara akurat, tersusun dengan urutan yang logis, dan menggunakan bahasa yang mudah dimengerti siswa. Proses validasinya pun dilakukan secara runtut, dimulai dari penyusunan kisi-kisi soal, perumusan butir soal, penilaian dari para ahli, pemberian masukan secara kualitatif, lalu ditutup dengan revisi soal berdasarkan masukan tersebut (Widiyanto & Hayati, 2022).

Validitas isi sangat bergantung pada keterlibatan para ahli dalam memastikan bahwa soal yang disusun benar-benar relevan dan layak secara materi. Zubaidah (2018) menjelaskan bahwa selain mencakup isi materi, validitas isi juga menilai seberapa baik soal tersebut bisa mengevaluasi kemampuan berpikir siswa sesuai dengan level kognitif yang ditargetkan. Dalam proses validasi, para validator juga memberikan masukan yang bersifat membangun, termasuk koreksi terhadap istilah ilmiah yang kurang tepat dan penyusunan kalimat agar lebih efektif dan mudah dipahami. Secara keseluruhan, proses validasi dalam penelitian ini dilakukan secara menyeluruh, mencakup dua pendekatan sekaligus: penilaian kuantitatif lewat skor CVI dan evaluasi kualitatif lewat saran langsung dari para ahli. Hasilnya, 28 soal berhasil masuk kategori tinggi, sementara dua lainnya tergolong sedang. Hal ini menunjukkan bahwa secara umum, instrumen yang dikembangkan sudah memiliki kualitas yang baik dan layak digunakan untuk mengevaluasi pembelajaran IPA, khususnya pada materi Suhu, Kalor, dan Pemuain.

**Tabel 6.** Hasil Perhitungan Koefisien AIKEN

| Butir  | Penilai |    |     |     |     |     |     | S  |    |    |    |    |    |    | $\Sigma s$ | V     | Keterangan |
|--------|---------|----|-----|-----|-----|-----|-----|----|----|----|----|----|----|----|------------|-------|------------|
|        | I       | II | III | IV  | V   | VI  | VII | s1 | s2 | s3 | s4 | s5 | s6 | s7 |            |       |            |
| (1-30) | 99      | 97 | 108 | 116 | 120 | 120 | 116 | 67 | 69 | 78 | 86 | 90 | 90 | 86 | 566        | 0,898 | Tinggi     |

Dari hasil validasi instrumen pembelajaran untuk materi Suhu, Kalor, dan Pemuain, diperoleh nilai koefisien Aiken's V sebesar 0,898, yang termasuk dalam kategori validitas tinggi. Angka ini didapat dari penilaian tujuh ahli terhadap 30 butir soal, dengan total skor  $\Sigma s$  mencapai 566. Nilai yang tinggi ini menunjukkan bahwa setiap soal dalam instrumen dianggap sangat sesuai, relevan, dan jelas dalam menyampaikan konsep oleh para validator. Artinya, instrumen ini sudah memenuhi syarat untuk digunakan dalam proses pembelajaran. Tingkat validitas yang tinggi juga menjadi penanda bahwa instrumen ini mampu mendukung pencapaian tujuan pembelajaran secara tepat serta efektif, karena dapat mengukur pemahaman siswa terhadap konsep-konsep sains yang diajarkan secara akurat.

**Tabel 7.** Ringkasan Statistik Butir Soal

| Score        | Items | Mean  | SD    | Min Score | Max Score | Mean P | Mean Rpbis |
|--------------|-------|-------|-------|-----------|-----------|--------|------------|
| Scored items | 0,446 | 2,102 | 0,273 | 0,073     | 0,533     | 0,429  | 0,136      |

Berdasarkan ringkasan statistik butir soal, nilai rata-rata (mean) sebesar 2,102 menunjukkan bahwa kemampuan siswa secara umum berada pada kategori sedang. Standar deviasi sebesar 0,273 menandakan bahwa nilai siswa cenderung tersebar secara merata, artinya tidak terlalu banyak perbedaan antara yang tinggi dan rendah. Nilai minimum yang tercatat adalah 0,073 dan maksimum 0,533, yang memperlihatkan adanya variasi skor di antara peserta. Sementara itu, nilai rata-rata tingkat kesukaran (Mean P) sebesar 0,429 juga menempatkan soal dalam kategori sedang, tidak terlalu mudah tapi juga tidak terlalu sulit. Namun, nilai rata-rata daya pembeda (Mean Rpbis) hanya mencapai 0,136, yang termasuk kategori rendah. Menandakan banyak soal yang belum bisa membedakan dengan baik antara siswa yang benar-benar menguasai materi dan yang belum. Proporsi soal yang dianalisis secara statistik berada di angka 0,446, atau hampir setengah dari jumlah total soal. Maka, diperlukan perbaikan khususnya pada soal-soal dengan daya pembeda rendah agar kualitas instrumen semakin meningkat. Meskipun tingkat kesukarannya sudah pas, pengembangan tetap diperlukan supaya instrumen benar-benar efektif dalam menilai pemahaman siswa.

### 3.3. Reliabilitas

Saat mengevaluasi konsistensi instrumen tes, reliabilitas merupakan indikator penting. Instrumen yang reliabel akan menghasilkan skor yang stabil dan konsisten saat digunakan dalam skenario dunia nyata. Reliabilitas, menurut Cohen, Manion, dan Morrison (2018), merupakan faktor penting dalam evaluasi pendidikan karena menggambarkan bagaimana alat ukur dipengaruhi oleh masalah pengukuran. Dalam penelitian ini, reliabilitas subjek dianalisis menggunakan beberapa pendekatan berbeda, seperti. Pendekatan tersebut meliputi Koreksi Spearman-Brown (S-B), Alpha Cronbach, dan Split-Half (acak, pertama-terakhir, ganjil-genap).

**Tabel 8.** Hasil Reliabilitas Butir Soal

| <i>Score</i>            | <i>Alpha</i> | <i>SEM</i> | <i>Split-Half<br/>(Random)</i> | <i>Split-Half<br/>(First-Last)</i> | <i>Split-Half<br/>(Odd-Even)</i> | <i>S-B<br/>Random</i> | <i>S-B<br/>First-Last</i> | <i>S-B<br/>Odd-Even</i> |
|-------------------------|--------------|------------|--------------------------------|------------------------------------|----------------------------------|-----------------------|---------------------------|-------------------------|
| <i>Scored<br/>items</i> | 0,446        | 2,102      | 0,273                          | 0,073                              | 0,533                            | 0,429                 | 0,136                     | 0,712                   |

Nilai alpha Cronbach sebesar 0,44 menunjukkan bahwa ambang batas reliabilitas tes berada dalam kategori rendah. Hal ini menunjukkan bahwa instrumen TES kurang konsisten dalam menilai kemampuan siswa. Seperti yang dinyatakan oleh Tavakol dan Dennick (2011), nilai alpha optimum minimal 0,70 agar dapat dianggap sebaik mungkin dalam konteks evaluasi pendidikan. Pengujian split-half menunjukkan variasi reliabilitas yang luas, dari 0,073 (nama depan-belakang) hingga 0,533 (ganjil-genap). Nilai split-half menunjukkan pembagian soal berdasarkan urutan awal-akhir tidak seimbang dari konstruksi soal sisi atau tingkat kesulitan, yang tersirat dari hasil reliabilitas yang sangat rendah. Sebaliknya, metode ganjil-genap menghasilkan hasil terbaik di antara metode split-half, mungkin menghasilkan nilai koreksi S-B tertinggi sebesar 0,712, yang telah mengurangi ambang batas reliabilitas minimum.

Dengan menganalisis semua metode, dapat disimpulkan bahwa instrumen tes ini belum sepenuhnya reliabel. Satu-satunya metode Spearman-Brown odd-even yang memberikan hasil yang bermakna. Berdasarkan analisis sebelumnya, hal ini juga dapat dikaitkan dengan ketidakseimbangan tingkat kesulitan antara butir soal, distributor yang tidak efektif, dan daya pembeda yang kurang baik. Hasil penelitian menunjukkan bahwa perlu dilakukan perbaikan pada aspek konstruksi butir soal, seperti revisi butir soal terkait hari pembeda rendah, distributor re-entry, dan distribusi tingkat kesulitan yang merata. Hal ini akan meningkatkan validitas dan konsistensi internal tes AIKEN (Ningsih et al., 2022).

### 3.3.1. Tingkat Kesukaran Butir Soal

**Tabel 9.** Rekapitulasi Tingkat Kesukaran Butir Soal

| Teoritis                                                                                                                  |                                         |                                        | Empiris                                                                   |                                                    |                                    |
|---------------------------------------------------------------------------------------------------------------------------|-----------------------------------------|----------------------------------------|---------------------------------------------------------------------------|----------------------------------------------------|------------------------------------|
| Butir soal<br>mudah<br>28 soal<br>(93%)                                                                                   | Butir soal<br>sedang<br>2 butir<br>(7%) | Butir soal<br>sukar<br>0 butir<br>(0%) | Butir soal<br>mudah<br>18 butir<br>(60%)                                  | Butir soal<br>sedang<br>12 butir<br>(40%)          | Butir soal suka<br>0 butir<br>(0%) |
| 2, 3, 5, 6, 7, 8,<br>9, 10, 11, 12,<br>13, 14, 15, 16,<br>17, 18, 19, 20,<br>21, 22, 23, 24,<br>25, 26, 27, 28,<br>29, 30 | 1 & 4                                   | -                                      | 1, 2, 3, 4, 5, 6,<br>9, 10, 12, 14,<br>15, 22, 23, 24,<br>25, 26, 28, 29, | 7, 8, 11, 13,<br>16, 17, 18, 19,<br>20, 21, 27, 30 | -                                  |

Analisis butir soal menunjukkan perbedaan mencolok antara perkiraan tingkat kesulitan soal oleh guru (teoritis) dan kenyataan saat dikerjakan siswa (empiris). Awalnya, sebagian besar soal (93% atau 28 butir) dianggap 'mudah sedang', namun secara empiris jumlahnya turun drastis menjadi 60% (18 butir), sementara soal 'sukar' melonjak dari 7% menjadi 40% (12 butir). Pergeseran ini, di mana 10 butir soal yang semula 'mudah sedang' menjadi 'sukar' bagi siswa, dapat dijelaskan oleh beberapa faktor. Pertama, kemungkinan siswa memiliki miskonsepsi atau pemahaman konsep yang belum mendalam, sejalan dengan penelitian Fauziana (2021) yang menunjukkan analisis butir soal sebagai alat diagnostik untuk kesulitan belajar. Kedua, kualitas teknis soal seperti ambiguitas bahasa atau pengecoh yang menjebak juga bisa membuat soal lebih sulit, seperti disoroti Yusuf (2020). Data ini mendesak perbaikan soal dan strategi pengajaran untuk meningkatkan kualitas asesmen dan pembelajaran sepenuhnya.

### 3.3.2. Distribusi Respons (kunci jawaban dan distraktor/pengecoh)

**Tabel 10.** Distribusi Respons Butir Soal

| No. | A     |        |   | B     |        |   | C     |        |   | D     |       |   | Keputusan Pengecoh |
|-----|-------|--------|---|-------|--------|---|-------|--------|---|-------|-------|---|--------------------|
|     | PE    | PB     | K | PE    | PB     | K | PE    | PB     | K | PE    | PB    | K |                    |
| 1.  | 0,943 | 0,080  | * | 0,057 | -0,080 | √ | 0,000 | 0,080  | - | 0,000 | -     | - | C & D diganti      |
| 2.  | 0,086 | 0,075  | - | 0,086 | 0,000  | - | 0,829 | -0,055 | * | 0,000 | -     | - | A, B, C, D diganti |
| 3.  | 0,000 | 0,054  | - | 0,029 | -0,054 | √ | 0,000 | 0,054  | - | 0,971 | 0,054 | * | A & C diganti      |
| 4.  | 0,971 | -0,195 | * | 0,000 | -0,195 | - | 0,029 | 0,195  | - | 0,000 | -     | - | A, B, C, D diganti |
| 5.  | 0,000 | 0,000  | - | 0,000 | 0,000  | - | 0,000 | 0,000  | - | 1,000 | 0,000 | * | A, B, C diganti    |
| 6.  | 0,000 | 0,245  | - | 0,971 | 0,245  | * | 0,029 | -0,245 | √ | 0,000 | -     | - | A & D diganti      |

| No. | A     |        |   | B     |        |       | C      |        |       | D     |        |   | Keputusan Pengecoh |
|-----|-------|--------|---|-------|--------|-------|--------|--------|-------|-------|--------|---|--------------------|
|     | PE    | PB     | K | PE    | PB     | K     | PE     | PB     | K     | PE    | PB     | K |                    |
| 7.  | 0,086 | -0,019 | √ | 0,029 | 0,179  | -     | 0,657  | -0,021 | *     | 0,229 | -0,034 | √ | B & C diganti      |
| 8.  | 0,286 | -0,102 | √ | 0,171 | -0,115 | √     | 0,514  | 0,275  | *     | 0,029 | -0,288 | √ | OK                 |
| 9.  | 0,029 | 0,192  | - | 0,057 | 0,008  | -     | 0,000  | -      | -     | 0,914 | -0,120 | * | A, B, C, D diganti |
| 10. | 0,143 | -0,124 | √ | 0,000 | -      | -     | 0,800  | -0,047 | *     | 0,057 | 0,268  | - | B, C, D diganti    |
| 11. | 0,171 | -0,333 | √ | 0,771 | 0,328  | *     | 0,029  | -0,070 | √     | 0,029 | -0,070 | √ | OK                 |
| 12. | 0,743 | 0,219  | * | 0,171 | -0,160 | √     | 0,000  | -      | -     | 0,086 | -0,127 | √ | C diganti          |
| 13. | 0,086 | -0,136 | √ | 0,286 | 0,234  | -     | 0,571  | -0,249 | *     | 0,057 | 0,239  | - | B, C, D diganti    |
| 14. | 0,114 | -0,136 | √ | 0,829 | 0,230  | *     | 0,000  | -      | -     | 0,057 | -0,187 | √ | C diganti          |
| 15. | 0,086 | 0,219  | - | 0,057 | -0,137 | √     | 0,057  | 0,042  | -     | 0,800 | -0,099 | * | A, C, D diganti    |
| 16. | 0,629 | -0,116 | * | 0,057 | -0,152 | √     | 0,114  | 0,365  | -     | 0,200 | -0,062 | √ | A & C diganti      |
| 17. | 0,200 | -0,114 | √ | 0,543 | 0,426  | *     | 0,114  | -0,041 | √     | 0,143 | -0,439 | √ | OK                 |
| 18. | 0,543 | 0,502  | * | 0,171 | -0,483 | √     | 0,171  | -0,022 | √     | 0,114 | -0,187 | √ | OK                 |
| 19. | 0,143 | 0,042  | - | 0,200 | -0,057 | √     | 0,514  | -0,158 | *     | 0,143 | 0,248  | - | A,C,D diganti      |
| 20. | 0,143 | -0,017 | √ | 0,200 | 0,047  | -     | 0,143  | 0,102  | -     | 0,514 | -0,097 | * | B, C, D diganti    |
| 21. | 0,143 | -0,224 | √ | 0,657 | 0,459  | *     | 0,057  | -0,362 | √     | 0,143 | -0,159 | √ | OK                 |
| 22. | 0,086 | -0,153 | √ | 0,029 | -0,129 | √     | 0,029  | -0,260 | √     | 0,857 | 0,307  | * | OK                 |
| 23. | 0,000 | -      | - | 0,114 | -0,237 | √     | 0,857  | 0,340  | *     | 0,029 | -0,261 | √ | A diganti          |
| 24. | 0,857 | 0,340  | * | 0,086 | -0,192 | √     | 0,029  | -0,261 | √     | 0,029 | -0,130 | √ | OK                 |
| 25. | 0,886 | -0,169 | * | 0,057 | 0,138  | -     | 0,029  | -0,058 | √     | 0,029 | 0,189  | - | A, B, D diganti    |
| 26. | 0,029 | -0,258 | √ | 0,829 | 0,171  | *     | 0,057  | 0,139  | √     | 0,086 | -0,192 | √ | OK                 |
| 27. | 0,686 | 0,153  | * | 0,171 | 0,033  | √     | 0,029  | -0,139 | √     | 0,114 | -0,190 | √ | OK                 |
| 28. | 0,000 | -      | - | 0,029 | -0,120 | √     | 0,029  | -0,056 | √     | 0,943 | 0,126  | * | A diganti          |
| 29. | 0,029 | -0,122 | √ | 0,104 | *      | 0,057 | -0,038 | √      | 0,000 | -     | -      | - | D diganti          |
|     |       |        |   | 0,914 |        |       |        |        |       |       |        |   |                    |
| 30. | 0,343 | -0,118 | √ | 0,314 | 0,124  | *     | 0,257  | -0,262 | √     | 0,086 | 0,403  | - | D diganti          |

Keterangan:

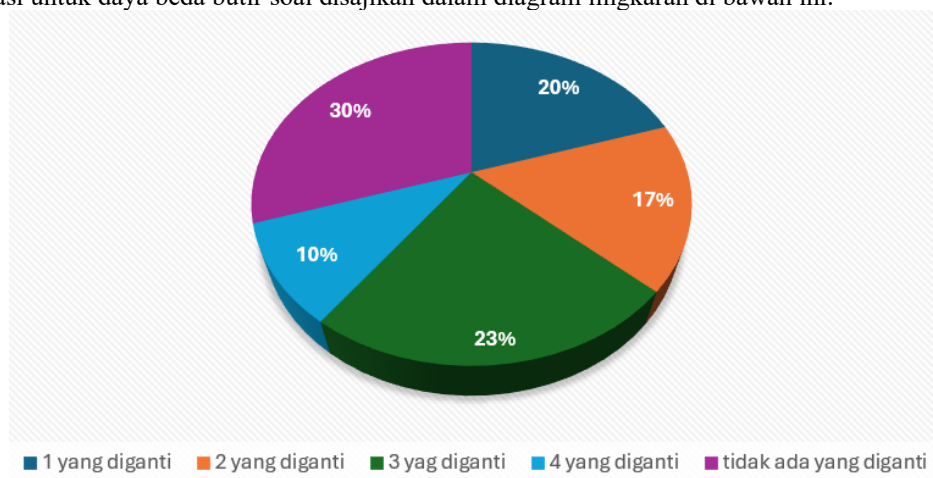
- \* : Kunci jawaban
- √ : Distraktor berfungsi dengan baik
- : Distraktor tidak berfungsi dengan baik (sebaiknya diganti)
- PE : Prop. Endorsing
- PB : Point Biser
- K : Keterangan

### 3.3.3. Daya Pembeda

**Tabel 11.** Daya Pembeda

| Butir diterima<br>(Total Rpbis $\geq 0,40$ ) | Butir diperbaiki<br>( $0,20 \leq \text{Total Rpbis} < 0,40$ ) | Butir ditolak<br>(Total Rpbis $< 0,20$ ) |
|----------------------------------------------|---------------------------------------------------------------|------------------------------------------|
| 11 butir                                     | 8 butir                                                       | 11 butir                                 |
| 6, 8, 11, 12, 14, 17, 18, 21, 22, 23, 24     | 1, 3, 5, 26, 27, 28, 29, 30                                   | 2, 4, 7, 9, 10, 13, 15, 16, 19, 20, 25   |

Rekapitulasi untuk daya beda butir soal disajikan dalam diagram lingkaran di bawah ini:



**Gambar 1.** Distribusi Respon Berdasarkan Distraktor

**Keterangan:**

- 1) Distraktor yang harus diganti 6 yaitu butir 12, 14, 23, 28, 29, 30
- 2) Distraktor yang harus diganti 5 yaitu butir 1, 3, 6, 7, 16
- 3) Distraktor yang harus diganti 7 yaitu butir 5, 10, 13, 15, 19, 20, 25
- 4) Distraktor yang harus diganti 3 yaitu butir 2, 4, 9
- 5) Distraktor tidak ada yang diganti 9 yaitu butir 8, 11, 17, 18, 21, 22, 24, 26, 27

Dalam upaya berkelanjutan untuk meningkatkan mutu pendidikan, evaluasi pembelajaran memegang peranan sentral. Salah satu aspek krusial dari evaluasi ini adalah analisis butir soal, sebuah proses yang tidak hanya mengukur capaian siswa, tetapi juga memberikan wawasan mendalam tentang kualitas instrumen penilaian dan efektivitas proses pengajaran (Sholihin et al., 2024). Analisis butir soal memungkinkan pendidik untuk mengidentifikasi kekuatan dan kelemahan setiap pertanyaan dalam mengukur pemahaman siswa, sekaligus mendeteksi area materi yang mungkin belum sepenuhnya dikuasai (Fauziana, 2021). Laporan ini bertujuan untuk menyajikan hasil analisis butir soal yang telah dilaksanakan di SMPN 38 Semarang, meninjau berbagai indikator psikometris seperti tingkat kesukaran, daya pembeda, dan fungsi *distractor* (pengecoh). Data yang diperoleh dari analisis ini diharapkan dapat menjadi landasan objektif untuk perbaikan instrumen evaluasi di masa mendatang dan perancangan strategi pembelajaran yang lebih efektif, sebagaimana kualitas butir soal mempengaruhi validitas hasil pengukuran (Yusuf, 2020).

Analisis butir soal di SMPN 38 Semarang menunjukkan adanya beberapa isu signifikan pada kualitas instrumen ujian. Meskipun secara teoritis sebagian besar soal (93%) diprediksi 'mudah sedang', hasil pengerjaan siswa (empiris) justru menunjukkan penurunan drastis pada kategori ini menjadi 60%, dan peningkatan tajam pada soal sukar menjadi 40%. Artinya, 10 butir soal yang dianggap mudah sedang ternyata sulit bagi siswa termasuk nomor 7, 8, 11, 13, 16, 17, 18, 19, 20, 21, 27, 30. Kesenjangan ini mengindikasikan kemungkinan miskonsepsi atau kurangnya pemahaman mendalam siswa (seperti dibahas Fauziana, 2021) dan adanya masalah teknis pada soal seperti bahasa yang membingungkan atau pengecoh yang tidak efektif (ditegaskan Yusuf, 2020). Selain itu, analisis daya pembeda mengungkap 11 butir soal tidak mampu membedakan siswa berkemampuan tinggi dan rendah misalnya nomor 2, 4, 7, 9, 10, 13, 15, 16, 19, 20, 25, dan 63.4% butir soal memiliki masalah pada pilihan pengecohnya. Temuan-temuan ini secara keseluruhan menyoroti perlunya revisi substansial pada soal ujian dan evaluasi ulang strategi pengajaran di SMPN 38 Semarang, demi mencapai keselarasan antara kurikulum, pengajaran, dan asesmen yang efektif seperti yang ditekankan (Sholihin et al., 2024).

### 3.3.4. Rangkuman Analisis Butir Soal

Analisis butir soal merupakan salah satu aspek terpenting dalam menilai kualitas instrumen yang digunakan untuk penelitian. Proses ini dilakukan untuk menentukan bagaimana setiap topik dapat digunakan untuk mendidik siswa berdasarkan kemampuannya. Dua parameter utama dalam analisis ini adalah indeks kesulitan soal (*difficulty index*) dan indeks daya pembeda soal (*daya pembeda*). Menurut Arifin (2017), butir soal memiliki tingkat kesulitan optimal antara 0,30 dan 0,80. Soal dengan indeks kesulitan di bawah 0,30 dianggap sulit, sedangkan soal dengan indeks 0,80 dianggap sangat mudah dipahami. Sebaliknya, butir soal yang baik harus memiliki nilai  $\geq 0,30$ , seperti yang dikemukakan oleh Zainul & Nasution (2020), sehingga siswa dapat dicocokkan sesuai dengan kemampuan tinggi dan rendahnya.

**Tabel 12.** Rangkuman Analisis Butir Soal

| No. | Tingkat Kesulitan | Daya Pembeda | Distribusi Respon |   |   |   | Keputusan | Keterangan |
|-----|-------------------|--------------|-------------------|---|---|---|-----------|------------|
|     | 0,3 - 0,8         |              | A                 | B | C | D |           |            |
| 1.  | 0,943<br>X        | 0,080<br>X   | *                 | √ | - | - | Ditolak   | Revisi     |
| 2.  | 0,829<br>X        | -0,055<br>X  | √                 | √ | * | - | Ditolak   | Revisi     |
| 3.  | 0,971<br>X        | 0,054<br>X   | -                 | √ | - | * | Ditolak   | Revisi     |
| 4.  | 0,971<br>X        | -0,195<br>X  | *                 | - | √ | - | Ditolak   | Revisi     |
| 5.  | 1,000<br>X        | 0,000<br>X   | -                 | - | - | * | Ditolak   | Revisi     |
| 6.  | 0,971<br>X        | 0,245<br>X   | -                 | * | √ | - | Ditolak   | Revisi     |

| No. | Tingkat Kesulitan | Daya Pembeda | Distribusi Respon |   |   |   | Keputusan              | Keterangan |
|-----|-------------------|--------------|-------------------|---|---|---|------------------------|------------|
|     | 0,3 - 0,8         | $\geq 0,30$  | A                 | B | C | D |                        |            |
| 7.  | 0,657<br>√        | -0,021<br>X  | √                 | √ | * | √ | Diterima dengan syarat | Revisi     |
| 8.  | 0,514<br>√        | 0,275<br>X   | √                 | √ | * | √ | Diterima               | Sudah Baik |
| 9.  | 0,914<br>X        | -0,120<br>X  | √                 | √ | - | * | Ditolak                | Revisi     |
| 10. | 0,800<br>√        | -0,047<br>X  | √                 | - | * | √ | Diterima dengan syarat | Revisi     |
| 11. | 0,328<br>√        | 0,456<br>√   | √                 | * | √ | √ | Diterima               | Sudah Baik |
| 12. | 0,743<br>√        | 0,21<br>X    | *                 | √ | - | √ | Diterima dengan syarat | Revisi     |
| 13. | 0,571<br>√        | -0,249<br>X  | √                 | √ | * | √ | Diterima dengan syarat | Revisi     |
| 14. | 0,829<br>X        | 0,230<br>X   | √                 | * | - | √ | Ditolak                | Revisi     |
| 15. | 0,800<br>√        | -0,099<br>X  | √                 | √ | √ | * | Diterima dengan syarat | Revisi     |
| 16. | 0,629<br>√        | -0,116<br>X  | *                 | √ | √ | √ | Diterima dengan syarat | Revisi     |
| 17. | 0,543<br>√        | 0,426<br>√   | √                 | * | √ | √ | Diterima               | Sudah Baik |
| 18. | 0,543<br>√        | 0,502<br>√   | *                 | √ | √ | √ | Diterima               | Sudah Baik |
| 19. | 0,514<br>√        | -0,158<br>X  | √                 | √ | * | √ | Diterima dengan syarat | Revisi     |
| 20. | 0,514<br>√        | -0,097<br>X  | √                 | √ | √ | * | Diterima dengan syarat | Revisi     |
| 21. | √<br>0,657        | √<br>0,459   | √                 | * | √ | √ | Diterima               | Sudah Baik |
| 22. | X<br>0,857        | √<br>0,307   | √                 | √ | √ | * | Diterima dengan syarat | Revisi     |
| 23. | X<br>0,857        | √<br>0,340   | -                 | √ | * | √ | Diterima dengan syarat | Revisi     |
| 24. | X<br>0,857        | √<br>0,340   | *                 | √ | √ | √ | Diterima dengan syarat | Revisi     |
| 25. | X<br>0,886        | X<br>-0,169  | *                 | √ | √ | √ | Ditolak                | Revisi     |
| 26. | X<br>0,829        | X<br>0,171   | √                 | * | √ | √ | Ditolak                | Revisi     |
| 27. | √<br>0,686        | X<br>0,153   | *                 | √ | √ | √ | Diterima dengan syarat | Revisi     |
| 28. | X<br>0,943        | X<br>0,126   | -                 | √ | √ | * | Ditolak                | Revisi     |
| 29. | X<br>0,914        | X<br>0,104   | √                 | * | √ | - | Ditolak                | Revisi     |
| 30. | X<br>0,314        | X<br>0,124   | √                 | * | √ | √ | Diterima dengan syarat | Revisi     |

Butir soal dengan daya pembeda negatif (misalnya, soal nomor 4 dan 13) mungkin perlu ditinjau secara menyeluruh karena mungkin mengandung konsep yang tidak efektif atau tidak efektif (Fitriyah & Setiawan, 2019). Lebih jauh, distribusi respons responden juga dapat berfungsi sebagai indikator sekunder. Misalnya, dalam Soal No. 6, meskipun tingkat kesulitannya cukup tinggi dan daya pembeda belum mencapai 0,30, pola distribusi jawaban menunjukkan tepat dominasi jawaban, sehingga masih dapat ditinjau oleh distributor. Hal ini sejalan dengan temuan penelitian Sari et al. (2021), yang menyatakan bahwa peningkatan kualitas subjek dapat dicapai melalui penggunaan analisis diskriminator.

### 3.4. 3.4. Soal Flag

Berdasarkan hasil analisis kuantitatif menggunakan program Iteman versi 4.3, soal nomor 2 terdeteksi mengalami permasalahan serius yang ditandai dengan adanya flag K,LR, yang mengindikasikan kelemahan pada aspek kunci jawaban serta daya pembeda yang rendah (*low discrimination*). Nilai tingkat kesukaran (P) sebesar 0,829 menunjukkan bahwa soal termasuk kategori mudah, namun nilai korelasi point-biserial (Rpbis) sebesar -0,055 justru tergolong sangat buruk. Korelasi negatif ini menunjukkan bahwa siswa dengan kemampuan tinggi justru lebih banyak menjawab salah dibanding siswa berkemampuan rendah. Menurut Qadir, Huda, dan Hermina (2024), daya pembeda negatif adalah indikator bahwa suatu soal tidak mampu membedakan antara peserta didik yang menguasai materi dengan yang tidak, sehingga harus direvisi atau bahkan dibuang.

Secara kualitatif, bentuk awal soal hanya menguji kemampuan siswa mengingat fakta tentang jenis alat ukur suhu tubuh, dengan redaksi sebagai berikut:

*“Alat yang digunakan untuk mengukur suhu tubuh manusia adalah...”*

- A. Thermometer
- B. Termometer alkohol
- C. Termometer klinis
- D. Barometer

Kunci jawaban yang benar adalah opsi C. Namun, distraktor yang digunakan terlalu lemah dan mudah dikenali sebagai pilihan yang salah, sehingga soal tidak menantang bagi siswa dan cenderung hanya mengukur aspek ingatan (C1) yang bersifat rendah dalam taksonomi kognitif. Hal ini sejalan dengan temuan Widiyanto dan Hayati (2022), yang menyatakan bahwa soal dengan dominasi hafalan dan distraktor tidak berfungsi seringkali memiliki daya pembeda rendah dan reliabilitas instrumen yang buruk secara keseluruhan.

Sebagai tindak lanjut, soal direvisi agar lebih kontekstual dan mengukur kemampuan berpikir tingkat tinggi (*higher-order thinking skills*). Berikut bentuk soal setelah direvisi:

*“Seorang teknisi akan mengukur suhu dalam ruang pembeku (freezer) yang suhunya mencapai  $-20^{\circ}\text{C}$ . Alat ukur suhu yang paling tepat digunakan adalah...”*

- A. Termometer raksa
- B. Termometer klinis
- C. Termometer digital
- D. Termometer alkohol

Kunci jawaban berubah menjadi opsi D, yaitu termometer alkohol. Revisi ini menekankan pada konteks aplikasi nyata dan penguasaan konsep karakteristik alat ukur suhu. Distraktor juga dirancang agar setara secara konseptual, sehingga lebih menguji pemahaman daripada sekadar hafalan. Menurut Immanuel et al. (2024), soal dengan konteks autentik dan distraktor yang seimbang cenderung memiliki daya pembeda yang lebih baik dan meningkatkan validitas instrumen. Dengan perubahan ini, diharapkan daya pembeda soal dapat meningkat dan fungsinya sebagai alat evaluasi yang adil serta akurat dapat tercapai. Langkah revisi ini juga sejalan dengan prinsip evaluasi formatif dalam Kurikulum Merdeka yang menekankan pengukuran kompetensi dan pemahaman siswa secara menyeluruh.

Berdasarkan laporan Iteman, soal nomor 5 menunjukkan karakteristik statistik yang sangat tidak ideal. Nilai tingkat kesukaran (P) sebesar 1,000 menandakan bahwa seluruh siswa (100%) menjawab soal ini dengan benar. Sementara itu, nilai daya pembeda (Rpbis) adalah 0,000, yang berarti soal tidak memiliki kemampuan membedakan antara siswa berkemampuan tinggi dan rendah. Dalam teori evaluasi, soal dengan karakteristik seperti ini tidak memiliki fungsi diagnostik dan sebaiknya dibuang, karena tidak memberikan kontribusi apapun terhadap pengukuran performa peserta didik. Temuan ini sejalan dengan pendapat Suyadi & Hartono (2017) bahwa soal yang dijawab benar oleh seluruh siswa dapat dianggap sebagai soal sangat mudah yang tidak lagi menantang, bahkan berpotensi merusak validitas instrumen. Selain itu, menurut Qadir et al. (2024), soal dengan daya pembeda 0 atau negatif tidak layak digunakan dalam instrumen penilaian karena tidak dapat mengidentifikasi perbedaan tingkat penguasaan kompetensi.

*“Benda yang mengalami pemuaian jika dipanaskan adalah...”*

- A. Karet
- B. Kayu

C. Logam

D. Batu

**Kunci: C (Logam)**

Kunci jawaban adalah opsi C, yaitu logam. Namun, keempat opsi tersebut tidak memberikan tantangan yang seimbang. Distraktor yang digunakan terlalu mudah dikenali sebagai jawaban yang salah oleh seluruh siswa. Bahkan bagi siswa dengan penguasaan materi yang rendah sekalipun, soal ini tetap dapat dijawab dengan benar karena bersifat intuitif dan telah dipelajari sejak jenjang pendidikan dasar. Hal ini sejalan dengan temuan Widiyanto dan Hayati (2022), yang menyatakan bahwa soal dengan karakteristik terlalu mudah dan tidak berfungsi membedakan kemampuan peserta didik akan berdampak negatif terhadap validitas dan reliabilitas tes secara keseluruhan.

Dari sisi revisi, soal ini dinilai tidak memungkinkan untuk diperbaiki secara substansial. Konsep yang diujikan terlalu dangkal, serta tidak membuka peluang pengembangan ke dalam bentuk soal dengan konteks penerapan atau penalaran. Upaya revisi terhadap redaksi atau distraktor tidak akan mengubah substansi evaluasi menjadi lebih bermakna, karena materi pokoknya sendiri terlalu sederhana untuk diuji dalam evaluasi tingkat SMP. Dengan mempertimbangkan data kuantitatif dan kajian kualitatif, maka soal ini sebaiknya dibuang dari bank soal, dan diganti dengan soal baru yang lebih menantang, kontekstual, serta mampu mengukur pemahaman mendalam terhadap konsep pemuatan.

#### 4. KESIMPULAN

Penelitian ini menunjukkan bahwa dari 30 butir soal pilihan ganda pada topik Suhu, Kalor, dan Pemuatan yang dianalisis menggunakan program Itean, sebagian besar tergolong mudah namun memiliki daya pembeda yang rendah dan reliabilitas yang belum memadai. Hanya sebagian kecil soal yang dinilai baik berdasarkan analisis statistik, sedangkan sisanya perlu direvisi atau bahkan ditolak karena tidak mampu membedakan kemampuan siswa dengan tepat. Selain itu, dua soal dinyatakan tidak valid berdasarkan hasil validasi isi oleh para ahli, sehingga perlu perbaikan mendalam agar sesuai dengan indikator pembelajaran dan taksonomi kognitif. Implikasi dari hasil penelitian ini cukup penting, baik secara praktis maupun teoritis. Bagi guru, temuan ini menjadi masukan bahwa penyusunan soal evaluasi perlu disertai analisis kualitas secara kuantitatif, agar instrumen yang digunakan benar-benar mencerminkan kemampuan siswa secara adil dan akurat. Dari sisi kebijakan pendidikan, hasil ini dapat mendorong sekolah dan dinas pendidikan untuk menyelenggarakan pelatihan penyusunan serta analisis soal berbasis program seperti Itean. Secara teoritis, penelitian ini mendukung pentingnya pengembangan instrumen evaluasi yang tidak hanya valid secara isi, tetapi juga terbukti andal secara empiris. Penelitian selanjutnya dapat memperluas cakupan materi, meningkatkan kualitas soal melalui konteks yang lebih variatif, dan melibatkan jumlah responden yang lebih besar untuk menghasilkan instrumen yang lebih representatif dan reliabel.

#### DAFTAR PUSTAKA

- Abdullah, G., Apriyanto, A., Patahuddin, A., Janah, R., Dia, E. E., Retnoningsih, R., ... & Setyaningrum, V. (2024). *Buku Ajar Evaluasi Pembelajaran*. PT. Sonpedia Publishing Indonesia.
- Aiken, L. R. (1980). Content Validity and Reliability of Single Items or Questionnaires. *Educational and Psychological Measurement*, 40(4), 955-959.
- Aiken, L. R. (2016). *Psychological Testing and Assessment*. Pearson Education.
- Arbiatin, E., & Mulabbiyah, M. (2020). Analisis Kelayakan Butir Soal Tes Penilaian Akhir Semester Mata Pelajaran Matematika Kelas Vi Di Sdn 19 Ampenan Tahun Pelajaran 2019/2020. *ElMidad*, 12(2), 146–171. <https://doi.org/10.20414/elmidad.v12i2.2627>
- Arikunto, S. (2012). *Dasar-dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara.
- Arifin, Z. (2017). *Evaluasi Pembelajaran*. Jakarta: Remaja Rosdakarya.
- Arisandi, W. F. (2020). Analisis Tingkat Kesukaran dan Daya Pembeda Soal Mata Pelajaran Sejarah Kelas VII SMP Negeri 1 Bandar Dua Kabupaten Pidie Jaya. *Riwayat: Educational Journal of History and Humanities*, 3(1). DOI:[10.31764/paedeutika.v12i2.4956](https://doi.org/10.31764/paedeutika.v12i2.4956)
- Busnawir, B., Judijanto, L., Abdullah, G., Abdurahman, A., Lumbu, A., Zamsir, Z., ... & Subhaktiyasa, P. G. (2025). *Evaluasi Pembelajaran: Prinsip, Teknik, & Aplikasi*. PT. Sonpedia Publishing Indonesia.
- Cohen, L., Manion, L., & Morrison, K. (2018). *Research Methods in Education* (8th ed.). Routledge.
- Darmansyah, S. N., Nurfazira, S., Azizah, N. N., Putri, H. E., & Nuraeni, F. (2025). Analisis Instrumen Tes Pemahaman Siswa Kelas VI di SD Labschool UPI Purwakarta Terhadap Materi Literasi Digital. *Jurnal*

*Kolaboratif Sains*, 8(1), 361-376.

- Dia, E. E., & Fadhilah, E. N. (2022). Tingkat Kesukaran Soal Ulangan Kelas Xi Di Ma As-Sulaimaniyah Mojoagung Jombang. *Prosiding Seminar Nasional Sastra, Lingua, Dan Pembelajarannya (Salinga)*, 2(1), 126–133. <https://doi.org/10.33503/salinga.v2i1.2157>
- Farida, A. M., & Musyarofah, A. (2021). Validitas dan Reliabilitas dalam Analisis Butir Soal. *Al-Mu'arrib: Jurnal Pendidikan Bahasa Arab*, 1. DOI: [10.32923/al-muarrib.v1i1.2100](https://doi.org/10.32923/al-muarrib.v1i1.2100)
- Fauziana. (2021). Analisis butir soal sebagai informasi diagnostik untuk mengidentifikasi miskonsepsi peserta didik. *Jurnal Luminous: Riset Ilmiah Pendidikan Fisika*,
- Fietri, W. A., Zulyusri, & Violita. (2021). Analisis Butir Soal Biologi Kelas XI Madrasah Aliyah Sakinah Kerinci Menggunakan Program Komputer Anates 4.0 For Windows. *Natural Science*, 7(1), 2477–6181. <https://doi.org/10.15548/nsc.v7i1.2329>
- Fitriyah, L., & Setiawan, D. (2019). Identifikasi Distraktor Tidak Berfungsi dalam Tes Pilihan Ganda. *Jurnal Pendidikan Indonesia*, 8(1), 45–53. DOI: [10.22373/jppm.v2i2.4502](https://doi.org/10.22373/jppm.v2i2.4502)
- Fuad, R., Iswantir, M., Akhyar, M., & Gusli, R. A. (2023). Strategi manajemen madrasah efektif dalam peningkatan kualitas belajar mengajar. *Idarah Tarbawiyah: Journal of Management in Islamic Education*, 4(2), 207-218. DOI: [10.32832/ijme.v4i2.15437](https://doi.org/10.32832/ijme.v4i2.15437)
- Herawati, I. (2022). Pengembangan Instrumen Penilaian HOTS Untuk Mengukur Dimensi Pengetahuan Fisika Siswa Kelas XI SMAN 14 Bandar Lampung. *Jurnal Pendidikan Taman Widya Humaniora*, 1(3), 299–323.
- Imania, K. A. N., & Bariah, S. H. (2019). Rancangan pengembangan instrumen penilaian pembelajaran berbasis daring. *Petik: Jurnal Pendidikan Teknologi Informasi dan Komunikasi*, 5(1), 31-47. <https://doi.org/10.33369/petik.v5i1.7923>
- Immanuel, C., Manik, S. D. P., Nababan, A. P., Sianturi, B. Y., & Hasnah, A. (2024). Analisis Butir Soal Pilihan Ganda Mata Pelajaran Matematika Materi Bangun Datar Berbasis Budaya Lokal di SDN 106163 Bandar Klippa. *Jurnal Ilmiah Multidisiplin Terpadu*, 8(6). DOI: <https://doi.org/10.35445/alishlah.v16i4.5671>
- Inanna, I., RAHMATULLAH, R., & Hasan, M. (2021). Evaluasi pembelajaran: teori dan praktek.
- Kurniawan, D. D. (2019). Analisis Butir Soal Ujian Akhir Semester Matematika dengan Teori Respon Butir. *Briliant: Jurnal Riset dan Konseptual*, 4(2). DOI: <https://doi.org/10.28926/briliant.v4i2.316>
- Kurniawan, H. (2022). *Pengantar praktis penyusunan instrumen penelitian*. Deepublish.
- Muluki, A. (2020). Analisis kualitas butir tes semester ganjil mata pelajaran IPA Kelas IV MI Radhiatul Adawiyah. *Jurnal Ilmiah Sekolah Dasar*, 4(1), 86-96.
- Nabil, N. R. A., Wulandari, I., Yamtinah, S., Ariani, S. R. D., & Ulfa, M. (2022). Analisis indeks Aiken untuk mengetahui validitas isi instrumen asesmen kompetensi minimum berbasis konteks sains kimia. *Jurnal Penelitian Pendidikan*, 25(2), 184-191. DOI: <https://doi.org/10.20961/paedagogia.v25i2.64566>
- Nofiana, M., & Karyanto, P. (2023). Pengembangan instrumen evaluasi two-tier multiple choice question untuk mengukur keterampilan berpikir tingkat tinggi pada materi kingdom plantae. *Inkuiri*, 67291. <https://doi.org/10.20961/inkuiri.v3i2.9694>
- Ningsih, D. S., Priyono, B., & Ratnasari, M. (2022). Analisis Reliabilitas Instrumen Evaluasi Pembelajaran IPA SMP. *Jurnal Evaluasi Pendidikan*, 13(1), 12–20. DOI: [10.19109/elidare.v9i2.20229](https://doi.org/10.19109/elidare.v9i2.20229)
- Permadani, W. A., Widyasari, N., Hakim, L., & Arif, A. (2024). Analisis Butir Soal Komputer Akuntansi Pilihan Ganda Menggunakan Aplikasi ANATES untuk Kelas XI. *WANARGI: Jurnal Manajemen Dan Akuntansi*, 2(1), 225-231. <https://doi.org/10.62017/wanargi.v2i2.3679>
- Pramanda, A., & Aisiah, A. (2023). Analisis Konstruksi Soal Tes Pilihan Ganda Mata Pelajaran Sejarah di SMA Negeri 6 Pariaman. *Jurnal Family Education*, 3(3), 335-339.
- Purwati, R., Wijayanti, L., & Sari, A. N. (2021). Analisis Validitas, Reliabilitas, Tingkat Kesukaran dan Daya Beda Soal Ujian. *Jurnal Evaluasi Pembelajaran*, 9(2), 145–155. <https://doi.org/10.52434/jp.v15i2.1287>
- Qadir, A., Huda, N., & Hermina, D. (2024). Analisis Butir Tes: Tingkat Kesukaran, Daya Pembeda Dan Efektivitas Pengecoh. *Al-Furqan: Jurnal Agama, Sosial, dan Budaya*, 3(3), 1450-1467. DOI: [10.31764/paedagogia.v12i2.4956](https://doi.org/10.31764/paedagogia.v12i2.4956)

- Rahim, A., Jainuddin, J., & Maheswari, K. A. M. (2025). Studi Deskriptif Kemampuan Pemahaman Aljabar Calon Guru Matematika. *MathSciEd of Journal*, 1(1), 01-07.
- Radja, S. P., Bano, V. O., & Ina, A. T. (2023). Analisis Kualitas Butir Soal Hasil Belajar Peserta Didik Berdasarkan Tingkat Kesukaran, Daya Beda Dan Efektivitas Pengecoh Di Sman 1 Pandawai. *Jurnal Edusavana*, 1(1), 30-41.
- Rosardi, R. G. (2015). Pilar Karakter Mandiri dan Peduli dalam Pembelajaran Social Studies; Penerapan Strategi Pemecahan Masalah dalam Pembelajaran. *Menyemai Generasi Emas Indonesia Menuju ASEAN Economic Community*.
- Safitri, I., Lestarani, D., Imtikhanah, R. D. N. W., Akbarini, N. R., Sari, M. W., Fitrah, M., & Hapsan, A. (2024). *Teori Pengukuran dan Evaluasi*. CV. Ruang Tentor.
- Sari, D. P., Hidayat, A., & Widodo, S. A. (2021). Analisis Kualitas Soal Ujian pada Mata Pelajaran IPA. *Jurnal Penilaian Pendidikan*, 12(1), 29-36. <https://doi.org/10.15294/jpbsi.v11i2.58091>
- Sugiyono, S. (2016). *Metode penelitian kuantitatif, kualitatif, R&D*. Bandung: Alfabeta, 1(11).
- Sukmasari, V. P., & Rosana, D. (2017). Pengembangan penilaian proyek pembelajaran IPA berbasis discovery learning untuk mengukur keterampilan pemecahan masalah. *Jurnal Inovasi Pendidikan IPA*, 3(1), 101-110. <https://doi.org/10.21831/jipi.v3i1.11072>
- Sutama, S. (2019). Pengembangan kurikulum matematika: Penilaian pembelajaran matematika berorientasi program for international student assessment di sekolah menengah pertama. In *Seminar Nasional Pendidikan Sultan Agung IV* (Vol. 1, No. 1).
- Sutrisna, N. (2021). Analisis kemampuan literasi sains peserta didik SMA di Kota Sungai Penuh. *Jurnal inovasi penelitian*, 1(12), 2683-2694. DOI: [10.47492/jip.v1i12.530](https://doi.org/10.47492/jip.v1i12.530)
- Suyadi, D., & Hartono, R. (2017). Analisis Kualitas Soal Ujian Akhir Semester Ganjil Berdasarkan Teori Tes Klasik. *Jurnal Ilmiah Pendidikan Fisika Al-Biruni*, 6(2), 175-189. <http://dx.doi.org/10.24127/ajpm.v12i3.7575>
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53-55. <https://doi.org/10.5116/ijme.4dfb.8dfd>
- Utami, L., Ilahi, D. P., & Ratih, A. (2024). Analisis Indeks Aiken untuk mengetahui Validitas Isi Instrumen Scinetific Habbits of Mind. *Journal of Research and Education Chemistry*, 6(1), 59-59. [https://doi.org/10.25299/jrec.2024.vol6\(1\).17430](https://doi.org/10.25299/jrec.2024.vol6(1).17430)
- Widiyanto, B., & Hayati, M. N. (2022). Pengembangan Modul Pengelolaan Laboratorium IPA Bab Keselamatan Dan Keamanan Kerja Untuk Meningkatkan Keterampilan Proses Sains Calon Guru IPA. *PSEJ (Pancasakti Science Education Journal)*, 7(2), 112-122. DOI: [10.24905/psej.v7i2.166](https://doi.org/10.24905/psej.v7i2.166)
- Wijayati, I. W. (2025). Urgensi Memahami Ketiganya Bagi Guru dan Mahasiswa Pendidikan. *Asesmen Pembelajaran: Teori dan Praktik*, 34.
- Yani, M. (2020). Efektivitas Distractor pada Tes Pilihan Ganda untuk Mendeteksi Kesalahan Siswa dalam Menyelesaikan Soal Matematika. *Al Khawarizmi: Jurnal Pendidikan dan Pembelajaran Matematika*, 4(1). DOI: [10.22373/jppm.v2i2.4502](https://doi.org/10.22373/jppm.v2i2.4502)
- Zubaidah, S. (2018, October). Mengenal 4C: Learning and innovation skills untuk menghadapi era revolusi industri 4.0. In *2nd Science Education National Conference* (Vol. 13, No. 2, pp. 1-18).
- Zainul, A., & Nasution, H. (2020). Analisis Butir Soal Ujian Sekolah Berbasis Komputer (USBK) SMP. *Jurnal Evaluasi Pendidikan*, 11(2), 123-132. <https://doi.org/10.21831/jep.v11i2.33910>