Jurnal Natural Science Educational Research 8 (2) 2025
e-ISSN: 2654-4210

ANALISIS BUTIR SOAL EVALUASI PADA TOPIK ZAT DAN PERUBAHANNYABERBASIS PROGRAM ITEMAN

Riski Dwi Handayani[1], Juan Dwi Prayoga [2] dan Nadiya Munika Putri[3]

[1] Pendidikan IPA,FMIPA, Universitas Negeri Semarang Semarang, 50229, Indonesia
handayaniriski720@students.unnes.ac.id
[2] Pendidikan IPA,FMIPA, Universitas Negeri Semarang Semarang, 50229, Indonesia
juandwiprayoga0105@students.unnes.ac.id
[3] Pendidikan IPA,FMIPA, Universitas Negeri Semarang Semarang, 50229, Indonesia
nadiyaamp@students.unnes.ac.id

Abstrak Tujuan dari penelitian ini difokuskan pada evaluasi mutu setiap item soal dengan topik " Zat dan perubahannya" dalam Pembelajaran IPA di SMPN 2 Ampelgading tahun ajaran 2024/2025. Evaluasi dalam pembelajaran penting untuk menilai pencapaian tujuan Pendidikan secara menyeluruh, termasuk aspek kognitif, keterampilan, dan sikap ilmiah siswa, sesuai tuntutan literasi sains abad ke-21. Pendekatan kuantitatif dengan desain deskriptif kualittatif digunakan, melibatkan 30 siswa kelas VII D. Instrumen berupa 30 soal pilihan ganda divalidasi secara kualitatif oleh enam ahli dan diuji validitas isi menggunakan indeks Aiken's V. Analisis kuantitatif dilakukan dengan Program Iteman 4.3 berdasarkan teori tes klasik. Hasil validasi menunjukkan seluruh soal memiliki nilai Aiken's V ≥ 0,89, menandakan validitas isi sangat baik. Namun, analisis lebih lanjut menunjukkan beberapa soal bermasalah, terutama dalam daya pembeda dan tingkat kesukaran. Misalnya, soal nomor 2 memiliki daya beda negatif dan tingkat kesukaran sangat tinggi. Temuan ini menegaskan perlunya revisi redaksi dan penyesuaian tingkat kompleksitas soal agar sesuai dengan kemampuan siswa. Penelitian ini menunjukkan pentingnya pengembangan instrumen evaluasi yang tidak hanya valid secara konten, tetapi juga empiris efektif untuk meningkatkan kualitas pembelajaran IPA yang kontekstual dan mendorong keterampilan berpikir Tingkat tinggi.

Kata Kunci: Analisis, daya beda, tingkat kesukaran

Abstract *This research is intended to evaluate the degree of test items on the topic "Substances and Their Changes" in science learning at SMPN 2 Ampelgading during the 2024/2025 academic year. Learning evaluation is essential to assess the achievement of educational goals comprehensively, including cognitive aspects, skills, and students' scientific attitudes, in line with 21st-century science literacy demands. A quantitative approach with a descriptive qualitative design was employed, involving 30 students of class VII D. The instrument, consisting of 30 multiple-choice items, was qualitatively validated by six experts and content validity was tested using Aiken's V index. Quantitative analysis was conducted using the Iteman 4.3 program based on classical test theory. The validation results showed that all items had Aiken's V values ≥ 0.89, indicating very strong content validity. However, further analysis revealed several problematic items, particularly in terms of discrimination power and difficulty level. For example, item number 2 had a negative discrimination index and was classified as very difficult.*

These findings highlight the need for revision in item wording and adjustment of complexity levels to match students' abilities. This study emphasizes the importance of developing evaluation instruments that are not only content-valid but also empirically effective in enhancing contextual science learning and fostering higher-order thinking skills.

Keywords: Analysis, discrimination index, difficulty level

Pendahuluan

Evaluasi dalam proses pembelajaran merupakan elemen krusial yang berfungsi untuk mengetahui sejauh mana capaian pembelajaran telah berhasil diraih. Selain sebagai alat untuk memberikan informasi balik kepada guru, evaluasi juga membantu peserta didik mengenali perkembangan belajar mereka. Bila dirancang dengan baik, evaluasi memungkinkan guru untuk menilai efektivitas metode mengajar yang digunakan, mengenali kendala pemahaman siswa dalam belajar, serta menetapkan langkah-langkah pembelajaran lanjutan yang lebih tepat sasaran.

Di SMPN 2 Ampelgading, pelaksanaan pembelajaran Ilmu Pengetahuan Alam (IPA) tidak hanya terbatas pada penyampaian konten materi, tetapi juga mengedepankan penilaian yang menyeluruh dan relevan dengan kondisi nyata siswa. Evaluasi dilakukan tidak hanya dalam bentuk sumatif yang diberikan di akhir pembelajaran, tetapi juga dalam bentuk formatif yang berlangsung selama proses belajar, dengan tujuan memantau perkembangan siswa dan menyesuaikan pendekatan pengajaran secara real time. Evaluasi ini mencakup ranah pemahaman konsep, kemampuan praktik sains, serta sikap berpikir ilmiah yang menjadi bagian yang utama dalam penguatan literasi sains era digital saat ini.

Keterampilan sains mengharuskan siswa selain untuk memahami aspek-aspek ilmiah, melainkan juga mampu menerapkan dan menilai informasi ilmiah dalam kehidupan sehari-hari. Karena itu, evaluasi dalam pembelajaran IPA harus disusun agar mampu menguji keterampilan berpikir kritis dan analitis tingkat tinggi (Higher Order Thinking Skills/HOTS) serta tidak berhenti pada aspek penghafalan atau pemahaman semata. Lestari, Wibowo, dan Sugiyarto (2018) menekankan pentingnya penyusunan alat ukur yang dirancang untuk menilai kemampuan berpikir tingkat tinggi (HOTS) di tingkat SMP guna mendorong keterampilan berpikir secara rasional, evaluatif, dan imajinatif dalam tantangan yang dihadapi dalam kehidupan nyata.

Salah satu pendekatan yang mendukung penilaian berbasis HOTS adalah penerapan penilaian autentik. Dalam mata pelajaran IPA, penilaian autentik berarti mengaitkan materi sains dengan konteks kehidupan riil siswa. Di SMPN 2 Ampelgading yang memiliki karakteristik sosial dan geografis khas, seperti kedekatan dengan potensi sumber daya alam dan lingkungan pertanian, pendekatan ini menjadi sangat relevan. Kegiatan berbasis konteks lokal seperti eksperimen dengan bahan alam sekitar, pengamatan lingkungan hidup, atau kunjungan lapangan menjadi media yang efektif untuk membantu siswa mengalami sains secara langsung dan bermakna.

Menurut Handayani dan Trisnawati (2020), penilaian autentik mampu mendorong keterampilan berpikir kritis secara signifikan karena melibatkan siswa secara aktif dalam kegiatan yang kontekstual dan aplikatif. Aktivitas seperti penyusunan laporan praktikum, pelaksanaan proyek

konservasi, hingga analisis fenomena lingkungan lokal, menjadi wadah penerapan keterampilan berpikir analitis, evaluatif, dan sintesis yang merupakan inti dari HOTS. Di samping itu, kemampuan dalam bidang sains juga mencakup keterampilan untuk menafsirkan data, menarik kesimpulan berdasarkan bukti, serta membuat keputusan secara rasional dalam kehidupan sehari-hari. Evaluasi yang berfokus pada literasi ilmiah, sebagaimana diungkapkan oleh Nasution, Murniati, dan Hidayat (2021), mampu memberikan gambaran menyeluruh tentang kompetensi siswa dan menjadi dasar bagi guru untuk merancang pendekatan pembelajaran yang lebih adaptif terhadap kebutuhan mereka.

Penerapan penilaian yang mengacu pada konteks kehidupan siswa juga berkontribusi pada peningkatan motivasi belajar. Purnamasari (2019) menyebutkan bahwa keterkaitan antara materi dengan pengalaman nyata siswa menjadikan mereka lebih terlibat aktif dalam proses pembelajaran. Kondisi geografis dan potensi lokal yang masih terjaga di wilayah SMPN 2 Ampelgading memberikan peluang besar untuk mengembangkan penilaian kontekstual yang bermakna dan relevan. Namun demikian, keberhasilan pelaksanaan evaluasi sangat ditentukan oleh kemampuan guru dalam merancang dan mengimplementasikan penilaian sesuai kurikulum yang berlaku. Seperti dikemukakan oleh Sari dan Wulandari (2017), masih banyak guru yang menghadapi tantangan dalam merancang instrumen penilaian yang valid dan dapat diandalkan, serta terbatasnya pelatihan mengenai penilaian alternatif berbasis konteks, satu di antara bentuk assesmen yang kerap digunakan dalam aktivitas pembelajaran IPA adalah soal pilihan ganda. Meskipun efisien, soal jenis ini

memerlukan pengembangan yang matang agar mampu mencerminkan kemampuan siswa secara akurat. Soal yang tidak dirancang dengan memperhatikan validitas isi, daya pembeda, dan tingkat kesulitan yang tepat, dapat mengurangi kualitas hasil evaluasi yang diperoleh.

Berdasarkan keadaan tersebut, kajian ini dilakukan untuk menelaah kualitas item evaluasi pada konten pembelajaran "Zat dan Perubahannya" dalam pembelajaran IPA di SMPN 2 Ampelgading tahun ajaran 2024/2025. Studi ini dilakukan untuk menganalisis validitas konten, level kesulitan, dan efektivitas soal dalam membedakan tingkat kemampuan siswa, dalam rangka memastikan bahwa instrumen evaluasi yang digunakan benar-benar mampu mendukung pembelajaran yang bermakna, kontekstual, dan mengoptimalkan kemampuan siswa dalam berpikir kritis, analitis, dan kreatif.

Metode Penelitian

Studi ini menggunakan pendekatan kuantitatif yang dipadukan dengan metode deskriptif kualitatif dan desain evaluatif guna menilai mutu butir soal pada materi Zat dan Perubahannya. Partisipan dalam penelitian ini adalah 30 siswa kelas VII D di SMP Negeri 2 Ampelgading pada tahun ajaran 2024/2025. Sebagai alat evaluasi, digunakan 30 soal pilihan ganda yang telah disusun oleh pihak peneliti. Analisis validitas isi secara kualitatif dilakukan sebelum penyebaran soal melalui telaah ahli oleh enam validator, yang terdiri atas dua guru IPA SMP Negeri 38 Semarang dan empat mahasiswa Pendidikan IPA Universitas Negeri Semarang angkatan 2021 dan 2022. Setelah itu Soal di uji menggunakan rumus indeks (Aiken's V) dan yang dinyatakan Valid diujikan secara daring (*online*).

$$V = \frac{\sum s}{[n(c-1)]} \quad (1)$$

Keterangan:
V = Koefisien validitas isi
n = banyaknya validator
s = $r\text{-}l_o$
r = angka yang diberikan validator
l_o = angka penialaian validitas terendah
c = angka penilaian validitas yang tertinggi

Butir soal dinyatakan valid apabila nilai V yang didapat lebih dari sama dengan nilai V pada tabel.

Analisis kuantitatif dilakukan menggunakan Program Iteman versi 4.3 untuk mengevaluasi karakteristik butir soal berdasarkan empat parameter utama dalam kerangka Teori Tes Klasik. Aspek awal yang dianalisis adalah reliabilitas dari tes, yang diperoleh melalui penerapan rumus kuder-Richardson (KR-20) guna mengukur sejauh mana konsistensi antar item soal. Aspek selanjutnya adalah Tingkat kesulitan yang dihitung untuk menentukan presentase siswa yang mampu menjawab tiap item dengan benar, lalu dikategorikan ke dalam lima Tingkat kesulitan berdasarkan interval yang tercantum dalam tabel. Kriteria klasifikasi indeks kesulitan soal mengacu pada pendapat Iskandar dan Rizal (2017).

Tabel 1. Rentang nilai dan kategori kesukaran butir soal"

P	Kategori
P = 0,00	Sangat Sukar
0,00 < P ≤ 0,30	Sukar
0,30 < P ≤ 0,70	Sedang
0,70 < P < 1,00	Mudah
P = 1,00	Sangat Mudah

Satu di antara karakteristik penting dari instrumen tes yang berkualitas adalah kemampuan membedakan tingkat kemampuan peserta didik, yang disebut daya pembeda. Daya pembeda dianalisis guna mengukur kemampuan soal dalam membedakan siswa dengan tingkat kemampuan berbeda. Proses ini dilakukan dengan membandingkan kinerja dua grup, yaitu grup dengan capaian skor tertinggi dan grup dengan capaian skor terendah berdasarkan hasil tes secara keseluruhan. Nilai daya beda setiap butir soal kemudian diklasifikasikan ke dalam lima kategori sesuai dengan kriteria yang disajikan pada tabel berikut. Menurut :

Tabel 2. Indikator kategori daya pembeda soal

Nilai Dp	Interpretasi
$Dp \leq 0{,}00$	Sangat Buruk
$0{,}00 < Dp \leq 0{,}20$	Buruk
$0{,}20 < Dp \leq 0{,}40$	Cukup
$0{,}40 < Dp \leq 0{,}70$	Baik
$0{,}70 < Dp \leq 1{,}00$	Sangat Baik

Efektivitas pengecoh (distraktor) dievaluasi berdasarkan seberapa sering pilihan tersebut dipilih oleh peserta didik. Sebuah distraktor dinilai berfungsi secara efektif apabila dipilih oleh responden yang tidak mengetahui jawaban yang benar, namun tidak lebih menarik dibandingkan kunci jawaban. Semakin tinggi frekuensi pemilihan distraktor oleh peserta yang menjawab salah, maka semakin efektif distraktor tersebut dalam menjalankan fungsinya.

Hasil Penelitian dan Pembahasan

Pokok Materi	Aspek yang diungkap			Jumlah
	Pemahaman (50%)	Aplikasi (26,7%)	Analisis (23,3%)	
Wujud zat dan model partikel (30%)	5	1	3	9
Perubahan wujud zat (23,3%)	4	3	-	7
Perubahan fisika dan kimia (30%)	3	3	3	9
Kerapatan zat (16,7%)	3	1	1	5
Jumlah	15	10	15	40

Validitas

Tabel 3. Hasil uji validitas konten

Nomor Soal	Nilai V	Kategori	Nomor Soal	Nilai V	Kategori
1	1,00	Valid	16	0,89	Valid
2	0,94	Valid	17	0,94	Valid
3	0,89	Valid	18	1,00	Valid
4	1,00	Valid	19	1,00	Valid
5	1,00	Valid	20	1,00	Valid
6	0,89	Valid	21	0,94	Valid
7	0,89	Valid	22	1,00	Valid
8	1,00	Valid	23	0,94	Valid
9	1,00	Valid	24	1,00	Valid
10	0,89	Valid	25	0,89	Valid
11	1,00	Valid	26	0,94	Valid
12	1,00	Valid	27	0,94	Valid
13	0,89	Valid	28	0,94	Valid
14	1,00	Valid	29	1,00	Valid
15	0,89	Valid	30	0,94	Valid

Berdasarkan evaluasi validitas isi terhadap 30 item pertanyaan pada topik "Zat dan Perubahannya", semua item pertanyaan menunjukkan hasil yang memuaskan secara teoritis dengan Aiken's V ≥ 0,89, yang berarti bahwa konten soal telah sesuai dengan indikator pembelajaran dan telah divalidasi secara signifikan oleh para ahli (Ulfa, 2022; An Nabil et al., 2022). Tingginya validitas isi ini menunjukkan bahwa pertanyaan-pertanyaan tersebut, dari sisi konten, telah mencerminkan kompetensi yang ingin diukur dalam bidang Ilmu Pengetahuan Alam (IPA) untuk jenjang SMP. Walaupun begitu, tingginya validitas isi tidak menjamin bahwa sebuah soal dapat digunakan dalam evaluasi pembelajaran jika tidak didukung oleh validitas empiris yang kuat (Karno & Wibisono, 2021). Salah satu temuan signifikan dari analisis kuantitatif mengindikasikan bahwa terdapat butir soal yang secara teknis memiliki masalah, khususnya dari segi daya pembeda. Salah satu contoh yang paling signifikan adalah pertanyaan yang kedua. Meski pertanyaan ini telah lulus uji validitas konten dengan nilai Aiken's V 0,94 dan masuk dalam kategori sangat valid, analisis daya pembeda menunjukkan angka -0,146 yang dianggap sangat buruk. Nilai negatif tersebut menunjukkan bahwa siswa berkemampuan tinggi lebih sering memberikan jawaban yang salah dibandingkan siswa berkemampuan rendah, yang jelas menunjukkan adanya kelemahan dalam soal untuk membedakan tingkat penguasaan siswa terhadap materi.

Selain daya beda yang tidak baik, soal nomor 2 juga memperlihatkan tingkat kesulitan yang sangat tinggi, dengan P = 0,133, yang menunjukkan hanya sekitar 13% dari keseluruhan responden yang mampu menjawab soal tersebut secara tepat. Ini menunjukkan bahwa soal tersebut tidak hanya gagal membedakan kemampuan siswa, tetapi juga terlalu sulit dan tidak sesuai dengan kemampuan umum siswa SMP pada tingkat kelas yang diuji (Karno & Wibisono, 2021). Berdasarkan analisis pilihan jawaban, diketahui bahwa dua opsi yang menyesatkan dalam soal ini, yaitu A dan C, tidak berfungsi dengan baik. Distraktor tersebut tidak dapat menarik perhatian siswa yang tidak mengetahui jawaban yang tepat, atau bahkan membingungkan siswa yang sudah memahami materi. Kegagalan penyamar dalam menjalankan fungsinya ini juga berperan pada rendahnya kemampuan untuk membedakan dan tingginya tingkat kesulitan soal. Untuk meningkatkan mutu soal ini, beberapa

langkah perbaikan yang bisa dilakukan termasuk merampingkan redaksi soal agar lebih jelas, menyesuaikan tingkat kompleksitas kalimat dengan kemampuan berpikir siswa SMP, serta mengevaluasi setiap pilihan jawaban dengan mengedepankan logika konseptual yang dapat merangsang potensi berpikir siswa. Sebagai bentuk pelaksanaan perbaikan, misalnya jika pertanyaan awalnya adalah "Transformasi air menjadi uap ketika dipanaskan adalah perubahan…?", maka pernyataan ini dapat diperluas menjadi "Proses pemanasan menyebabkan air mengalami evaporasi dan berubah menjadi uap air". Perubahan ini disebut sebagai transformasi…" untuk memberikan siswa konteks yang lebih jelas dan familiar. Kunci jawaban "fisika karena tidak menghasilkan zat baru" dapat juga dilengkapi dengan opsi yang salah namun masuk akal, seperti "kimia karena adanya perubahan suhu" atau "kimia karena tidak bisa kembali ke bentuk asal", yang mencerminkan pemahaman yang salah dan mampu mengukur seberapa baik siswa memahami konsep perubahan fisika dan kimia dengan benar.

Reliabilitas

Reliabilitas merupakan indikator penting dalam menilai konsistensi suatu instrumen atau metode dalam menghasilkan data yang stabil dan dapat dipercaya. Instrumen dikatakan reliabel apabila mampu memberikan hasil pengukuran yang konsisten terhadap aspek yang sama dalam berbagai situasi yang setara (Himawan & Nurgiyantoro, 2022). Konsistensi ini juga mencerminkan ketepatan alat ukur dalam merepresentasikan atribut yang diukur. Dalam konteks evaluasi pembelajaran, pengujian reliabilitas biasanya dilakukan melalui pendekatan kuantitatif menggunakan bantuan perangkat lunak analisis butir soal. Salah satu software yang populer digunakan adalah Iteman, Dimana memungkinkan peneliti menganalisis ciri butir soal serta menghitung koefisien reliabilitas berdasarkan Teori Tes Klasik (Budiastuti & Bandur, 2018).

Tabel 4. Tabel hasil analisis reliabilitas

Karakteristik	Nilai
N of Items	30
N of Examinees	30
Alpha	0,697
SEM	2,431

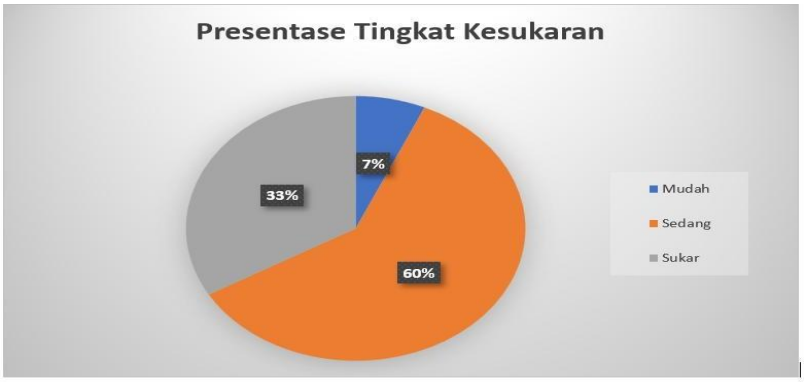
Nilai Alpha dari hasil perhitungan diperoleh sebesar 0,697, berdasarkan pada tabel diatas. Hal ini menunjukkan bahwa Soal Evaluasi pada Topik Zat dan Perubahannya tahun pelajaran 2024/2025 di SMP N 2 Ampelgading masuk kategori sedang. Ini selaras dengan ungkapan Nuryanti et al (2018), yang menunjukkan bahwa perolehan skor Alpha pada klasifikasi reliabilitas 0.00 - 0.20 (sangat rendah), 0.21- 0.40 (rendah), 0.41—0.70 (sedang), 0.71—0.90 (tinggi), dan 0.71—1.00 (sangat tinggi).

Analisis Tingkat Kesukaran

Tabel 5. Tabel hasil analisis tingkat kesukaran

Nomor Soal	Nilai P	Kategori	Nomor soal	Nilai P	Kategori
1	0,433	Sedang	16	0,533	Sedang
2	0,133	Sukar	17	0,367	Sedang
3	0,200	Sukar	18	0,300	Sukar
4	0,233	Sukar	19	0,600	Sedang
5	0,233	Sukar	20	0,267	Sukar
6	0,467	Sedang	21	0,600	Sedang
7	0,167	Sukar	22	0,733	Mudah
8	0,467	Sedang	23	0,600	Sedang
9	0,700	Sedang	24	0,533	Sedang
10	0,567	Sedang	25	0,233	Sukar
11	0,567	Sedang	26	0,500	Sedang
12	0,300	Sukar	27	0,367	Sedang
13	0,667	Sedang	28	0,467	Sedang
14	0,700	Sedang	29	0,833	Mudah
15	0,233	Sukar	30	0,333	Sedang

Indeks Tingkat Kesukaran (ITK) digunakan untuk mengetahui sejauh mana suatu butir soal tergolong mudah atau sulit bagi peserta didik. Menurut Iskandar dan Rizal (2017), kategori tingkat kesukaran soal dibedakan menjadi lima, yaitu Sangat Sukar ($P = 0{,}00$), Sukar ($0{,}00 < P \leq 0{,}30$), Sedang ($0{,}30 < P \leq 0{,}70$), Mudah ($0{,}70 < P < 1{,}00$), dan Sangat Mudah ($P = 1{,}00$). Untuk keperluan penelitian ini, analisis digunakan oleh penulis adalah nilai *proportion correct* (P) untuk menentukan kategori tingkat kesukaran. Dari data yang ditampilkan pada tabel tersebut, disimpulkan bahwa sebanyak 30 butir soal telah dianalisis, tidak terdapat soal dengan kategori sangat sukar maupun sangat mudah. Kategori sukar mencakup 10 butir soal, yaitu soal nomor 2, 3, 4, 5, 7, 12, 15, 18, 20, dan 25. Nilai P untuk butir soal ini berada pada rentang 0,133 hingga 0,300. Soal-soal dalam kategori sedang berjumlah 18 butir, yakni nomor 1, 6, 8, 9, 10, 11, 13, 14, 16, 17, 19, 21, 23, 24, 26, 27, 28, dan 30, dengan nilai P antara 0,333 hingga 0,700. Sementara itu, hanya 2 soal yang tergolong mudah, yaitu nomor 22 dan 29, dengan nilai P masing-masing sebesar 0,733 dan 0,833. Data tersebut dapat dilihat pada pie chart di bawah :



Gambar 1. Pie chart presentase Tingkat kesukaran

Daya Beda

Suatu soal dikatakan memiliki kemampuan membedakan apabila dapat memisahkan peserta didik yang menjawab benar dengan yang menjawab keliru disebut sebagai daya pembeda soal (Solichin, 2017). Kajian terhadap tujuan dari pengukuran daya pembeda adalah untuk mengetahui sejauh mana suatu soal mampu membedakan peserta didik dengan tingkat pencapaian akademik yang berbeda. Analisis ini dilakukan untuk mengevaluasi mutu butir soal secara faktual agar dapat dijadikan acuan dalam upaya perbaikan kualitas soal (Saputri & Larasati, 2023). Tingkat daya pembeda soal dikelompokkan ke dalam lima kategori, yaitu: sangat rendah, rendah, sedang, baik, dan sangat baik. Berdasarkan pendapat Mania et al. (2020), soal dengan nilai daya pembeda ≤ 0,00 tergolong sangat buruk; nilai antara 0,01 hingga 0,20 termasuk kategori buruk; rentang 0,21 sampai 0,40 disebut cukup; nilai antara 0,41 hingga 0,70 dianggap baik; dan nilai lebih dari 0,70 hingga 1,00 termasuk dalam kategori sangat baik. Pengukuran Daya beda pada soal ini penulis menggunakan program iteman pada menu poin biser. Tabel berikut menampilkan hasil analisis.

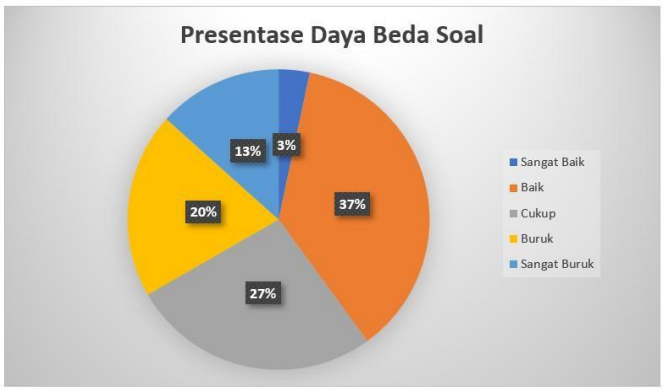
Tabel 6. Tabel hasil analisis daya beda

Nomor Soal	Poin Biser	Kategori	Nomor Soal	Poin Biser	Kategori
1	0,216	Cukup	16	0,480	Baik
2	-0,146	Sangat Buruk	17	0,232	Cukup
3	-0,020	Sangat Buruk	18	0,358	Cukup
4	-0,117	Sangat Buruk	19	0,255	Cukup
5	0,788	Sangat Baik	20	0,183	Buruk
6	0,600	Baik	21	0,191	Buruk
7	0,489	Baik	22	0,524	Baik
8	0,522	Baik	23	0,526	Baik
9	0,290	Cukup	24	0,198	Buruk
10	0,321	Cukup	25	0,179	Buruk
11	0,541	Baik	26	0,109	Buruk
12	0,188	Buruk	27	0,232	Cukup
13	0,437	Baik	28	0,506	Baik
14	0,631	Baik	29	0,370	Cukup
15	-0,283	Sangat Buruk	30	0,591	Baik

Ada beberapa simpulan yang dapat diambil dari tabel di atas. Dari 30 soal tersebut tercatat empat soal yang masuk dalam kelompok klasifikasi sangat buruk, yaitu pada nomor 2,3,4, dan 15 karena memiliki indeks daya beda pada angka ≤ 0,00. Soal yang dikelompokkan berdasarkan kategori buruk terdapat sebanyak 6 butir soal, yang terdiri atas soal nomor 12,20,21,24,25,26. Butir tersebut tergolong buruk karena memiliki daya beda pada angka 0,00 < Dp ≤ 0,20. Soal yang dikelompokkan

berdasarkan kategori cukup terdapat sebanyak 8 butir soal, yang terdiri atas soal nomor 1,9,10,17,18,19,27, dan 29. Soal yang dikelompokkan berdasarkan kategori cukup karena memiliki daya beda pada angka $0{,}20 < Dp \leq 0{,}40$. Soal yang dikelompokkan berdasarkan kategori baik terdapat sebanyak 11 butir soal, yang terdiri atas soal nomor 6,7,8,11,13,14,16,22,23,28, dan 30. Butir tersebut tergolong baik karena memiliki daya beda pada angka $0{,}40 < Dp \leq 0{,}70$. Butir soal dengan kategori sangat baik berjumlah 1 butir soal, yaitu butir soal nomor 5.

Hasil evaluasi kualitas soal berdasarkan nilai daya pembeda mengindikasikan bahwa beberapa soal masuk dalam kategori baik hingga sangat baik. Terdapat sebanyak 20 soal yang tergolong dalam kategori baik dan sangat baik, yaitu soal nomor 5, 6, 7, 8, 11, 13, 14, 16, 22, 23, 28, dan 30. Sementara itu, meskipun termasuk dalam kategori cukup, soal-soal tersebut berdasarkan daya pembeda tetap dapat membedakan siswa yang telah menguasai materi dari yang belum. Namun demikian, dalam soal-soal di kategori ini, terdapat peluang yang cukup tinggi bagi siswa yang masih mengalami kesulitan dalam memahami materi terkait menebak jawaban dengan benar. Dengan kata lain, butir soal berkategori cukup tetap dapat dimanfaatkan dalam pelaksanaan tes berikutnya, asalkan telah dilakukan revisi atau perbaikan. Adapun soal-soal yang dikategorikan buruk dan sangat buruk dicirikan oleh daya beda yang rendah hingga negatif, sehingga tidak dapat mengidentifikasi perbedaan antara siswa yang memahami materi dan yang tidak. Oleh sebab itu, penggunaan kembali soal dalam kategori tersebut tidak direkomendasikan, dan lebih tepat jika digantikan dengan soal baru yang memiliki kualitas lebih baik. Data tersebut dapat dilihat pada pie chart dibawah :



Gambar 2. Pie chart presentase daya beda

Analisis Efektivitas Pengecoh

Tabel 7. Hasil analisis efektivitas pengecoh

No	A			B			C			D			Keputusa n Pengecoh
	PE	PB	K	PE	PB	K	PE	PB	K	PE	PB	K	
1	0,067	-0,303	√	0,100	-0,313	√	0,433	0,216	*	0,400	0,128	√	OK
2	0,300	-0,068	×	0,133	-0,146	*	0,433	0,405	×	0,133	-0,352	√	A & C Di Ganti
3	0,500	0,250	×	0,267	-0,312	√	0,033	0,116	×	0,200	-0,020	*	A & C Di Ganti
4	0,233	-0,117	*	0,567	0,400	×	0,133	-0,330	√	0,067	-0,146	√	B Di Ganti
5	0,133	-0,261	√	0,333	-0,204	√	0,300	-0,324	√	0,233	0,788	*	OK
6	0,467	0,600	*	0,133	-0,238	√	0,367	-0,497	√	0,033	0,116	√	OK
7	0,167	0,489	*	0,267	-0,188	√	0,233	-0,228	√	0,333	-0,006	√	OK
8	0,067	-0,240	√	0,467	0,522	*	0,133	-0,398	√	0,333	-0,138	√	OK
9	0,167	-0,140	√	0,700	0,290	*	0,100	-0,208	√	0,033	-0,102	√	OK
10	0,567	0,321	*	0,067	-0,052	√	0,167	-0,496	√	0,200	0,098	√	OK
11	0,567	0,541	*	0,067	-0,272	√	0,133	-0,123	√	0,233	-0,376	√	OK

12	0,200	-0,059	√	0,267	0,059	√	0,233	-0,209	√	0,300	0,188	*	OK
13	0,067	-0,209	√	0,100	-0,313	√	0,667	0,437	*	0,167	-0,161	√	OK
14	0,033	-0,232	√	0,700	0,631	*	0,200	-0,449	√	0,067	-0,272	√	OK
15	0,467	0,350	√	0,200	-0,117	×	0,233	-0,283	*	0,100	-0,026	×	B dan D Diganti
16	0,200	-0,391	√	0,200	-0,176	√	0,067	-0,052	√	0,533	0,480	*	OK
17	0,367	0,232	*	0,300	-0,153	√	0,133	-0,307	√	0,200	0,156	√	OK
18	0,333	-0,055	√	0,300	0,358	*	0,267	-0,312	√	0,100	0,000	√	OK
19	0,100	-0,052	√	0,600	0,255	*	0,133	-0,490	√	0,167	0,154	√	OK
20	0,067	-0,178	√	0,167	-0,349	√	0,267	0,183	*	0,500	0,188	×	D Di Ganti
21	0,133	-0,146	√	0,600	0,191	*	0,233	-0,062	√	0,033	-0,102	√	OK
22	0,033	-0,276	√	0,100	-0,182	√	0,133	-0,375	√	0,733	0,524	*	OK
23	0,000	-	-	0,367	-0,514	√	0,600	0,526	*	0,033	-0,058	√	A Di Ganti
24	0,067	-0,178	√	0,267	0,130	√	0,533	0,198	*	0,133	-0,330	√	OK
25	0,233	0,216	×	0,100	-0,313	√	0,433	-0,147	√	0,233	0,179	*	A Di Ganti
26	0,100	-0,313	√	0,167	0,259	×	0,500	0,109	*	0,233	-0,135	√	B Di Ganti
27	0,400	0,064	√	0,367	0,232	*	0,167	-0,287	√	0,067	-0,146	√	OK
28	0,100	-0,287	√	0,333	-0,354	√	0,467	0,506	*	0,100	0,000	√	OK
29	0,000	-	-	0,167	-0,370	√	0,833	0,370	*	0	0,000	-	A dan D di Ganti
30	0,200	0,000	√	0,167	-0,349	√	0,300	-0,324	√	0,333	0,591	*	OK

Rotama et al, (2020) menyatakan bahwa tingkat keberfungsian pengecoh dapat ditinjau melalui nilai *proportional endorsing* dalam aplikasi Iteman. Distraktor atau pengecoh merupakan alternatif jawaban selain opsi jawaban yang dirancang untuk menjebak peserta yang belum menguasai materi ujian. Setiap pengecoh idealnya memiliki daya tarik yang cukup sehingga dapat dipilih sebagai jawaban oleh peserta. Efektivitas pengecoh diukur dari jumlah peserta tes yang memilihnya, yaitu minimal sebanyak 5% dari total responden (Forthmann et al., 2020). Berdasarkan tabel di atas analisis butir 30 soal evaluasi materi Zat dan perubahannya untuk kelas VII menunjukkan bahwa terdapat 28 butir soal yang termasuk kategori pengecoh yang berfungsi dengan baik yaitu pada nomor 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,24,25,26,27,28,30. Sedangkan 2 butir soal lainnya termasuk pengecoh yang tidak berfungsi dengan baik yaitu butir 23 dan 29.

Tersedia contoh soal yang dilengkapi dengan pengecoh yang semuanya bekerja secara efektif artinya semua opsi jawaban dipilih oleh peserta didik yaitu seperti pada butir soal nomor 1. Hal ini dapat dilihat dari *proportional endorsing* (PE), dimana pada opsi A = 0,067, opsi B = 0,100, opsi C = 0,433, opsi D = 0,400. kunci jawaban ada pada opsi C dan opsi tersebut paling banyak dipilih oleh peserta didik daripada opsi lain. Adapun contoh soal yang memiliki pengecoh kurang efektif dapat ditemukan pada nomor 29, dimana pada opsi A = 0,000, opsi B = 0,167. Opsi C = 0,833, opsi D = 0,000. Pada opsi A dan D bernilai 0,000 yang berarti jawaban tersebut tidak dipilih oleh peserta didik.

Rangkuman Analisis Butir soal

Tabel 8. Rangkuman analisis butir soal Berdasarkan analisis yang telah dilakukan, maka dapat dirangkum sebagai berikut :

No	Tingkat Kesukaran (0,3-0,8)	Daya Pembeda ≥0,30	Distribusi respon				Keputusan	Keterangan
			A	B	C	D		
1	√	×	√	√	*	√	Diterima dengan syarat	Revisi
2	×	×	×	*	×	√	Ditolak	Soal Tidak baik

3	×	×	×	√	×	*	Ditolak	Soal Tidak baik
4	×	×	*	×	√	√	Ditolak	Soal Tidak baik
5	×	√	√	√	√	*	Diterima dengan syarat	Revisi
6	√	√	*	√	√	√	Diterima	Sudah baik
7	×	√	*	√	√	√	Diterima dengan syarat	Revisi
8	√	√	√	*	√	√	Diterima	Sudah baik
9	√	×	√	*	√	√	Diterima dengan syarat	Revisi
10	√	√	*	√	√	√	Diterima	Sudah baik
11	√	√	*	√	√	√	Diterima	Sudah baik
12	√	×	√	√	√	*	Diterima dengan syarat	Revisi
13	√	√	√	√	*	√	Diterima	Sudah baik
14	√	√	√	*	√	√	Diterima	Sudah baik
15	×	×	√	×	*	×	Ditolak	Soal tidak baik
16	√	√	√	√	√	*	Diterima	Sudah baik
17	√	×	*	√	√	√	Diterima dengan syarat	Revisi
18	√	√	√	*	√	√	Diterima	Sudah baik
19	√	×	√	*	√	√	Diterima dengan syarat	Revisi
20	×	×	√	√	*	×	Ditolak	Soal tidak baik
21	√	×	√	*	√	√	Diterima dengan syarat	Revisi
22	√	√	√	√	√	*	Diterima	Sudah baik
23	√	√	-	√	*	√	Diterima	Sudah baik
24	√	×	√	√	*	√	Diterima dengan syarat	Revisi
25	×	×	×	√	√	*	Ditolak	Soal tidak baik
26	√	×	√	×	*	√	Diterima dengan syarat	Revisi
27	√	×	√	*	√	√	Diterima dengan syarat	Revisi
28	√	√	√	√	*	√	Diterima	Sudah baik
29	×	√	-	√	*	-	Diterima dengan syarat	Revisi
30	×	√	√	√	√	*	Diterima dengan syarat	Revisi

Salah satu saran yang telah diperbaharui dalam studi ini adalah pertanyaan kelima. Soal ini menunjukkan tingkat validitas isi yang sangat tinggi dengan nilai Aiken's V = 1,00, tetapi masih terdapat kekurangan teknis dalam hasil analisis kuantitatif. Tingkat kesulitan pertanyaan berada pada kategori sulit (p = 0,233), menunjukkan bahwa hanya sejumlah kecil siswa yang mampu memberikan jawaban yang tepat. Walaupun daya pembeda soal tergolong sangat baik (0,788), ketidakseimbangan efektivitas pengecoh menjadikan soal ini tidak optimal untuk menilai kemampuan siswa secara menyeluruh (Winta Pebrina, 2021; Karno & Wibisono, 2021). Evaluasi mengenai efektivitas distraktor menunjukkan bahwa siswa cenderung menghindari pilihan yang salah tanpa melakukan proses berpikir yang mendalam. Ini menunjukkan bahwa distraktor tidak cukup kuat atau tidak cukup menantang untuk mendorong pemikiran kritis siswa (Dewi & Prasetyo, 2023; Yamtinah et al., 2023). Revisi soal sebelumnya, "Transformasi air dikenal sebagai transformasi...", dengan kunci jawaban "fisika", dianggap terlalu mudah dan cenderung hanya untuk menghafal. Sebagai hasil, masalah ini tidak berhasil mendorong kemampuan berpikir ilmiah yang lebih kompleks. Untuk menyelesaikan masalah tersebut, redaksi soal diperbaiki menjadi: "Pada suhu ruangan, es batu yang ditinggalkan akan mencair menjadi air. "Perubahan ini melibatkan perubahan fisik akibat…". Revisi ini menciptakan konteks yang lebih relevan dan nyata, sehingga memudahkan siswa untuk

menghubungkan konsep ilmiah dengan pengalaman sehari-hari (Fajriah & Asiskawati, 2024). Pilihan jawaban juga diperbaharui menjadi: A. Membentuk materi baru, B. Tidak membentuk materi baru (kunci), C. Ada perubahan warna, D. Tidak dapat dikembalikan ke bentuk semula. Distraktor dirancang berdasarkan miskonsepsi umum yang dimiliki siswa, sehingga soal ini tidak hanya menilai penguasaan konsep, tetapi juga berperan sebagai alat untuk mendiagnosis pemahaman siswa (Dewi & Prasetyo, 2023; Fajriah & Asiskawati, 2024). Dengan cara ini, revisi ini memastikan bahwa soal tetap memiliki validitas isi yang tinggi sambil meningkatkan kualitas kognitif, efektivitas distraktor, dan tingkat tantangan yang seimbang bagi siswa. Pendekatan ini sejalan dengan prinsip penilaian yang berfokus pada literasi sains, yang menekankan nilai pemikiran kritis, pemahaman konseptual, dan penerapan ilmiah dalam kehidupan sehari-hari (Yamtinah et al., 2023; Fajriah & Asiskawati, 2024). Soal hasil perbaikan ini menunjukkan integrasi antara validitas substansial dan validitas empiris yang seimbang, yang merupakan karakteristik utama instrumen evaluasi berkualitas dalam pendidikan IPA tingkat SMP.

Salah satu butir soal yang sebaiknya ditolak atau dibuang adalah soal nomor 2. Berdasarkan hasil analisis, butir ini tidak memenuhi kriteria memiliki tingkat kesulitan yang sesuai, karena terletak di luar rentang 0,30 hingga 0,80. Hal ini menunjukkan bahwa soal ini memiliki tingkat kesulitan yang tidak sesuai, karena terlalu gampang atau terlalu menantang guna dikerjakan oleh sebagian besar peserta tes. Selain itu, daya pembeda soal ini juga berada di bawah standar minimal ($\geq 0{,}30$), yang berarti soal ini tidak berperan dalam membedakan tingkat pemahaman siswa terhadap materi yang telah dipelajari. Analisis terhadap distribusi respons menunjukkan bahwa hanya satu pilihan jawaban yang dipilih oleh sebagian besar peserta, sedangkan pengecoh lainnya tidak berfungsi secara efektif. Kondisi ini mengindikasikan bahwa daya tarik pengecoh sangat lemah. Dengan mempertimbangkan ketiga aspek tersebut tingkat kesukaran, daya beda, dan distribusi respons. Soal nomor 2 dikategorikan sebagai soal yang tidak baik dan tidak layak digunakan, Karena itu, direkomendasikan agar soal ini diganti dengan butir lain yang memenuhi syarat sebagai soal berkualitas.

Kesimpulan dan Saran

Data yang diperoleh dari hasil pengolahan dan penelaahan menunjukkan bahwa instrumen evaluasi untuk topik "Zat dan Perubahannya" memiliki validitas isi yang sangat tinggi (Aiken's $V \geq 0{,}89$), namun masih ditemukan kelemahan secara empiris, khususnya pada daya pembeda dan tingkat kesukaran beberapa soal. Misalnya, soal nomor 2 walaupun telah tervalidasi secara isi, tidak mampu membedakan kemampuan siswa dan tergolong sangat sulit. Reliabilitas instrumen tergolong sedang ($\alpha = 0{,}697$), dan meskipun sebagian besar pengecoh bekerja dengan baik, beberapa di antaranya kurang efektif. Temuan ini menunjukkan pentingnya peninjauan ulang kualitas butir soal secara menyeluruh dari aspek teoritis dan empiris agar instrumen benar-benar mencerminkan kemampuan siswa. Oleh karena itu, guru dan penyusun soal perlu mempertimbangkan keseimbangan antara validitas isi, daya pembeda, tingkat kesukaran, serta efektivitas pengecoh agar soal tidak hanya memenuhi standar kualitas, tetapi juga mampu merangsang kemampuan berpikir ilmiah siswa. Hal ini menegaskan pentingnya pelatihan guru dalam melakukan analisis butir soal serta mendorong pengembangan soal yang berbasis literasi sains untuk mendukung asesmen yang bermakna dan relevan dengan tuntutan kompetensi abad ke-21.

Daftar Pustaka

Budiastuti, D., & Bandur, A. (2018). Validitas dan Reliabilitas Penelitian. Dalam Binus. Citra Wacana Medika.

Forthmann,Boris,Natalie Förster, Birgit Schütze,Karin Hebbecker , Janis Flessner, Martin T.Peters and Elmar Souvignier. (2020). How Much g Is in the Distractor? Re-Thinking Item-Analysis of Multiple-Choice Items.J. Intell. 2020, 8, 11

Handayani, R., & Trisnawati, E. (2020). Authentic assessment dalam pembelajaran IPA untuk meningkatkan kemampuan berpikir kritis siswa. Jurnal Pendidikan Sains Indonesia, 8(2), 134– 142.

Himawan, R., & Nurgiyantoro, B. (2022). Analisis butir soal latihan penilaian akhir semester ganjil mata pelajaran bahasa Indonesia kelas VIII SMPN 1 Bambanglipuro Bantul menggunakan program ITEMAN. KEMBARA: Jurnal Keilmuan Bahasa, Sastra, Dan Pengajarannya, 8(1), 160-180.

Lestari, S. D., Wibowo, T., & Sugiyarto, K. (2018). Pengembangan instrumen penilaian HOTS pada materi IPA SMP. Jurnal Inovasi Pendidikan IPA, 4(1), 45–52.

Mania, S., Fitriani, F., Majid, AF, Ichiana, NN, & Abrar, AIP (2020). Analisis pertanyaan ujian akhir sekolah. Al-Asma: Jurnal Pendidikan Islam, 2 (2), 274-284.

Nasution, R. E., Murniati, A. R., & Hidayat, A. (2021). Evaluasi pembelajaran IPA melalui asesmen literasi sains di sekolah menengah pertama. Jurnal Evaluasi Pendidikan, 12(1), 67–74.

Nuryanti, S., Masykuri, M., & Susilowati, E. (2018). Analisis Iteman dan model Rasch pada pengembangan instrumen kemampuan berpikir kritis peserta didik sekolah menengah kejuruan. *Jurnal Inovasi Pendidikan IPA, 4*(2), 224-233.

Purnamasari, A. (2019). Penilaian kontekstual dalam pembelajaran IPA: Studi kasus di SMP berbasis lingkungan. Jurnal Pendidikan IPA Indonesia, 8(1), 65–72.

Saputri, H. A. S., & Larasati, N. J. (2023). ANALISIS INSTRUMEN ASSESMEN: VALIDITAS, RELIABILITAS, TINGKAT KESUKARAN DAN DAYA BEDA BUTIR SOAL. Didaktik:

Jurnal Ilmiah PGSD STKIP Subang, 9(5), 2986-2995.

Sari, D. K., & Wulandari, E. (2017). Evaluasi pembelajaran sains: Konsep dan implementasinya dalam kurikulum 2013. Jurnal Ilmu Pendidikan, 23(1), 12–20.

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.