



Analisi Butir Soal Pilihan Ganda Pembelajaran Kimia Kelas X SMA Srijaya Negara

Nabilah Muazarah, Muhammad Rafi Farhan, Efendi Nawawi

Program Studi Pendidikan Kimia, Fakultas Keguruan dan Ilmu Pendidikan Universitas Sriwijaya, Kampus Indralaya KM 32, Indralaya 30662, Sumatra Selatan, Indonesia

Informasi Artikel	Abstrak
<i>Sejarah Artikel:</i> Diterima: 15-05-2025 Disetujui: 11-07-2025 Dipublikasikan: 17-07-2025 Kata Kunci; Analisis butir soal, validitas, reliabilitas, tingkat kesulitan, daya diskriminasi, evaluasi pembelajaran <i>Keywords:</i> <i>Item Analysis, Validity, Reliability, Difficulty Level, Discrimination Power, Learning Evaluation</i>	Penelitian ini bertujuan untuk menganalisis kualitas butir soal pilihan ganda mata pelajaran kimia kelas X SMA Srijaya Negara berdasarkan validitas, reliabilitas, tingkat kesukaran, dan daya pembeda. Metode penelitian yang digunakan adalah deskriptif kuantitatif. Subjek penelitian adalah 10 soal pilihan ganda yang diujicobakan kepada 26 siswa. Data dianalisis menggunakan bantuan Microsoft Excel untuk menghitung koefisien validitas, reliabilitas dengan metode Kuder-Richardson 20 (KR-20), tingkat kesukaran, dan daya pembeda. Hasil penelitian menunjukkan bahwa 60% butir soal valid, namun reliabilitas instrumen tergolong rendah (0,502). Tingkat kesukaran soal didominasi oleh kategori sulit hingga sangat sulit (60%), dan daya pembeda soal bervariasi. Disimpulkan bahwa instrumen evaluasi ini memerlukan perbaikan signifikan untuk meningkatkan kualitasnya sebagai alat ukur hasil belajar siswa. Abstract <i>This research aimed to analyze the quality of multiple-choice items in chemistry for tenth-grade students at SMA Srijaya Negara based on validity, reliability, difficulty level, and discrimination power. The research method employed was descriptive quantitative. The subjects of the research were 10 multiple-choice questions tested on 26 students. Data were analyzed using Microsoft Excel to calculate validity coefficients, reliability using the Kuder-Richardson 20 (KR-20) method, difficulty level, and discrimination power. The results indicated that 60% of the items were valid; however, the instrument's reliability was low (0.502). The difficulty level of the questions was predominantly in the difficult to very difficult categories (60%), and the discrimination power of the items varied. It was concluded that this evaluation instrument requires significant improvement to enhance its quality as a measurement tool for student learning outcomes.</i>

© 2025 Universitas Riau

*Alamat korespondensi:
e-mail: muazarahnabilah@gmail.com
No. Telf: +62895620827983

1. Pendahuluan

Penilaian terhadap hasil belajar adalah suatu aktivitas yang sangat krusial dalam pendidikan. Proses dalam lembaga pendidikan formal pada akhirnya akan berujung pada hasil belajar yang

dinyatakan dalam bentuk nilai. Namun, mengevaluasi hasil belajar siswa tidak selalu mudah, karena tujuan pembelajaran mencakup aspek kognitif, afektif, dan psikomotor. Area pengetahuan dan sikap relatif sulit untuk diobservasi, meskipun masih dapat diukur. Oleh karena itu, langkah pertama dalam penilaian hasil belajar adalah merumuskan tujuan pembelajaran yang dapat diamati dan diukur. Setelah itu, instrumen disusun untuk mengamati serta mengukur hasil pembelajaran. Dengan menggunakan instrumen tersebut, data yang menggambarkan pencapaian pembelajaran siswa dapat diperoleh. Dari data ini, penilaian dapat membuat keputusan mengenai posisi atau status seorang siswa (Supriyadi, 2021). Salah satu metode untuk menilai hasil belajar siswa dan mengetahui seberapa baik tujuan pembelajaran tercapai adalah dengan melakukan evaluasi. Evaluasi adalah sebuah langkah untuk menguji apakah sebuah kegiatan telah dilaksanakan sesuai dengan tujuan yang sudah ditetapkan sebelumnya dan juga untuk membuat keputusan mengenai objek yang dievaluasi (Purniasari, et al. 2021).

Evaluasi adalah salah satu bagian penting dalam sistem pendidikan yang perlu dilakukan dengan cara yang teratur dan direncanakan untuk mengukur seberapa sukses atau target yang ingin dicapai selama proses belajar mengajar. Pada dasarnya, evaluasi memberikan penilaian atau nilai berdasarkan standar tertentu (Sudjana, 2002). Evaluasi dalam proses pembelajaran merupakan salah satu bagian yang sangat penting dan merupakan langkah yang perlu diambil untuk mengevaluasi seberapa efektif pembelajaran tersebut. Hasil yang didapat bisa berfungsi sebagai umpan balik untuk memperbaiki dan memperbaiki program serta kegiatan belajar. Evaluasi menjadi tahap akhir dari sebuah proses pendidikan, yang memungkinkan kita untuk memahami seberapa berhasil proses belajar itu sesuai dengan tujuan yang telah ditetapkan. Oleh sebab itu, evaluasi tidak kalah pentingnya dalam proses belajar. Evaluasi mencakup semua aspek dari pembelajaran, termasuk kemampuan berpikir (*kognitif*), perasaan, sikap, dan perilaku (*afektif*), serta kemampuan praktis (*psikomotor*) (Ismail. 2020).

Ulangan adalah proses untuk mengevaluasi sejauh mana siswa telah menguasai kompetensi dalam bidang pembelajaran tertentu. Ulangan ini juga dikenal sebagai evaluasi formatif, karena guru ingin memahami tingkat penguasaan materi dari setiap bab yang telah diajarkan. Salah satu alat evaluasi yang digunakan untuk memberi ulangan kepada siswa adalah soal pilihan ganda (objektif). Soal pilihan ganda yang akan diberikan kepada siswa perlu memiliki standar kualitas yang tinggi (Santosa & Badawi, 2022).

Sementara itu, soal yang baik adalah soal yang mampu mencerminkan keadaan siswa berdasarkan kemampuan dan keterampilan tertentu. Instrumen yang berkualitas sangat penting karena membantu guru dalam memilih soal yang paling sesuai, yang termasuk dalam kategori valid dan reliabel. Sebuah soal dianggap valid jika dapat memberikan informasi yang relevan dan sesuai dengan apa yang diajarkan dalam pembelajaran, dan dianggap reliabel jika memberikan hasil yang konsisten ketika diukur berulang kali. Setiap soal dievaluasi baik secara keseluruhan maupun per butir. Analisis pada butir soal mencakup tingkat kesulitan, daya pembeda, dan tingkat pengecoh. Analisis ini bertujuan untuk mengetahui apakah soal yang dikerjakan siswa terlalu mudah atau sulit dan untuk menilai kemampuan setiap butir soal dalam membedakan siswa yang sudah memahami materi dengan yang belum. Oleh karena itu, penting untuk melakukan analisis soal melalui pengujian validitas, reliabilitas, daya beda, tingkat kesulitan, dan fungsi pengecoh (Tarmizi, et al. 2020).

Beberapa kajian terdahulu yang telah mengeksplorasi tentang analisis butir soal dari berbagai mata pelajaran. Purniasari et al (2021) telah melakukan analisis butir soal ujian sekolah pada mata pelajaran kimia menggunakan model Iteman dan Rasch. Mulyani et al (2022) telah menganalisis butir soal dan kecukupan hots soal ujian akhir semester pada mata pelajaran kimia di SMK kelas X. Simamora et al (2021) telah melaporkan tentang analisis kualitas butir soal buatan guru kimia pada tes ujian tengah semester di kelas XII MIPA. Mira et al (2021) telah menganalisis butir soal pilihan ganda pada materi pokok larutan elektrolit dan nonelektrolit kelas X IPA yang jalankan di SMA N 1

Lamba Leda Manggarai. Keumala et al (2020) telah melaporkan tentang analisis butir soal ujian semester genap pada mata pelajaran kimia menggunakan program Proanaltes dilakukan di kelas X SMK Banda Aceh. Sakinah dan Ritonga (2017) juga telah melaporkan analisis butir soal ujian semester mata pelajaran kimia kelas x di Madrasah Aliyah di Kecamatan Pasir Penyau. Ritonga, (2017) juga telah melaporkan tentang analisis butir soal ujian semester ganjil pada mata pelajaran kimia kelas X di SMA/MA se-kecamatan Pekaitan.

Berdasarkan urian pada latar belakang, maka kajian ini bertujuan untuk menganalisis butir soal pilihan ganda pada mata pelajaran kimia kelas X SMA Srijaya Negara berdasarkan validitas, reliabilitas, tingkat kesukaran, dan daya pembeda.

2. Metode Penelitian

Jenis penelitian yang digunakan dalam penelitian ini adalah deskriptif kuantitatif. Analisis deskriptif merupakan metode yang digunakan untuk memahami data dengan cara menerangkan atau menggambarkan informasi yang telah dikumpulkan (Tilar & Hasriyanti, 2019).). Subjek dalam penelitian ini adalah tes objektif berupa pilihan ganda. Teknik pengumpulan soal ini yaitu pembuatan soal lalu dilakukan tes validitas, setelah itu dilakukan uji coba kepada anak-anak kelas X SMA Srijaya Negara. Setelah mendapatkan data lalu melakukan analisis butir soal yang ditinjau dari validitas, reliabilitas, tingkat kesukaran, daya pembeda dan efektivitas pengecoh dengan bantuan excel (Purniasari, et al. 2021).

3. Hasil dan Pembahasan

Setelah menganalisis butir soal mata pembelajaran kimia pada siswa SMA Srijaya Negara kelas X dari segi validitas butir soal, maka didapatkan hasil sebagai berikut:

a). Validitas

Berdasarkan temuan dari uji validitas pada instrumen penelitian, analisis menunjukkan bahwa 6 pertanyaan atau 60% dari keseluruhan alat ukur dinyatakan valid. Pertanyaan yang valid adalah nomor 2, 5, 6, 7, 9, dan 10. Pertanyaan nomor 2 memiliki koefisien korelasi 0,477809, yang menunjukkan hubungan sedang dengan skor total. Selanjutnya, pertanyaan nomor 5 memiliki koefisien korelasi 0,448421, yang juga mencerminkan tingkat validitas yang baik. Sementara itu, pertanyaan nomor 6 memiliki validitas yang lebih tinggi dengan koefisien korelasi 0,555650. Pertanyaan nomor 7 mendapatkan koefisien korelasi 0,429263, yang masih dalam kategori validitas yang bisa diterima. Pertanyaan nomor 9 mencatat koefisien korelasi tertinggi di antara semua pertanyaan, yaitu 0,692903, yang menunjukkan bahwa pertanyaan ini memiliki kemampuan diskriminasi yang sangat baik dan berperan penting dalam pengukuran konstruk yang dimaksud. Terakhir, pertanyaan nomor 10 dengan koefisien korelasi 0,444394 juga menunjukkan validitas yang memadai.

Di sisi lain, analisis menemukan 3 pertanyaan atau 30% dari keseluruhan instrumen tidak valid, yakni pertanyaan nomor 3, 4, dan 8. Pertanyaan nomor 3 memiliki koefisien korelasi 0,351117, yang berada di bawah nilai kritis untuk validitas. Begitu pula, pertanyaan nomor 4 memperoleh koefisien korelasi 0,39870, dan pertanyaan nomor 8 dengan koefisien korelasi 0,360318. Ketiga pertanyaan ini tidak menunjukkan hubungan yang cukup kuat dengan skor total, yang menunjukkan bahwa mereka mungkin mengukur konstruk yang berbeda atau memiliki kelemahan dalam rumusan soal. Masalah ini perlu diatasi karena pertanyaan-pertanyaan yang tidak valid dapat menurunkan akurasi pengukuran secara keseluruhan.

Masalah khusus terdeteksi pada pertanyaan nomor 1, yang tidak bisa dihitung validitasnya, terjadi karena munculnya error “DIV/0!” dalam perhitungan koefisien korelasi. Situasi ini sering muncul ketika tidak ada variasi pada jawaban responden, yang berarti semua responden memberikan jawaban yang sama pada pertanyaan tersebut. Dalam analisis validitas, pertanyaan dengan variasi

nol tidak dapat dievaluasi keabsahannya karena kurangnya kemampuan diskriminasi. Ini menyebabkan 10% dari instrumen tidak dapat dianalisis validitasnya dengan tepat, sehingga diperlukan peninjauan ulang terhadap penyusunan pertanyaan tersebut.

b). Reliabilitas

Reliabilitas menunjukkan seberapa konsisten atau stabil hasil pengukuran saat alat yang sama digunakan berkali-kali dalam kondisi yang serupa. Dalam dunia pendidikan, alat tes yang mampu diandalkan akan memberikan hasil yang cenderung sama jika diujikan pada waktu yang berbeda atau pada kelompok yang setara.

Dari data yang ada, nilai koefisien reliabilitas (R_{11}) yang ditemukan adalah 0,502096361. Berdasarkan klasifikasi umum untuk koefisien reliabilitas dalam penelitian pendidikan, nilai ini ada di kategori "Sedang," tetapi di bagian bawah tabel tertera keterangan "Tidak Reliabel." Ini menunjukkan bahwa walaupun alat tersebut mempunyai tingkat konsistensi internal yang sedang, nilai tersebut belum mencapai batas minimum yang diperlukan untuk dianggap reliabel dalam pengukuran pendidikan. Secara umum, sebuah alat dianggap memiliki reliabilitas yang baik jika koefisiennya mencapai setidaknya 0,70. Dengan nilai koefisien reliabilitas sedikit lebih dari 0,50, alat ini menunjukkan beberapa masalah dalam hal konsistensi pengukuran. Ini berarti bahwa jika alat ini digunakan untuk mengukur hal yang sama pada waktu yang berbeda atau pada responden yang setara, ada kemungkinan besar hasilnya bisa bervariasi. Variasi ini dapat mempengaruhi keakuratan interpretasi dan keputusan yang diambil berdasarkan hasil tes.

Dengan mengamati komponen perhitungan dalam data (p , q , $p \times q$, varians total), dapat diketahui bahwa rumus yang dipakai untuk menghitung reliabilitas adalah Kuder-Richardson 20 (KR-20), yang ditujukan untuk mengukur konsistensi internal alat dengan item-item pertanyaan dikotomi (benar-salah, 1-0). Dengan memasukkan data: jumlah item soal $k = 10$, jumlah hasil kali p dan q (Σpq) = 1,718934911, serta varians total (σ^2) = 3,136094675, didapatkan hasil $r_{11} = (10/9) \times (1 - 1,718934911/3,136094675) = (10/9) \times (1 - 0,548112) = (10/9) \times 0,451888 = 0,502096361$. Nilai ini menunjukkan bahwa sekitar 50,2% varians dari skor total adalah varians skor bersih, sedangkan 49,8% lainnya adalah varians dari kesalahan pengukuran. Dengan kata lain, hampir setengah dari variasi skor yang didapat oleh responden mungkin dipengaruhi oleh faktor-faktor yang tidak berkaitan dengan konstruk yang diukur, seperti keberuntungan, kelelahan, atau kondisi pengukuran lainnya.

Analisis proporsi jawaban benar (p) dari setiap soal memberikan pemahaman menyeluruh mengenai tingkat kesulitan dan peran masing-masing soal dalam reliabilitas instrumen. Untuk soal No. 1, yang memiliki $p = 0$, menunjukkan bahwa tidak ada responden yang menjawab dengan benar. Hal ini menjadikannya sebagai soal yang sangat sulit dan tidak memberikan kontribusi sama sekali terhadap reliabilitas ($p \times q = 0$). Soal No. 2, dengan $p = 0,5$ (50% atau 13 dari 26 responden menjawab benar), memiliki tingkat kesulitan yang ideal dan berkontribusi secara maksimal terhadap reliabilitas ($p \times q = 0,25$). Soal No. 3, dengan $p = 0,3076923077$ (30,77% atau 8 responden menjawab benar), dikategorikan sebagai cukup sulit dan memberikan kontribusi yang baik ($p \times q = 0,2133017815$). Dalam kategori sangat sulit, terdapat soal No. 14 dengan $p = 0,1153846154$ (11,54% atau 3 responden menjawab benar) yang berkontribusi relatif kecil ($p \times q = 0,1020710059$). Soal No. 15, dengan $p = 0,2692307692$ (26,92% atau 7 responden benar), tergolong sulit dan memiliki kontribusi yang memadai ($p \times q = 0,1967455621$). Di sisi lain, soal No. 16, yang memiliki $p = 0,1538461538$ (15,38% atau 4 responden menjawab benar), dinyatakan sangat sulit dengan kontribusi yang lebih kecil ($p \times q = 0,1301775148$). Soal No. 17 tercatat dengan $p = 0,5384615385$ (53,85% atau 14 responden menjawab benar), menunjukkan tingkat kesulitan menengah dan berkontribusi secara optimal ($p \times q = 0,2485207101$). Soal No. 18, yang memiliki $p = 0,1923076923$ (19,23% atau 5 responden menjawab benar), termasuk sangat sulit dengan kontribusi yang relatif kecil ($p \times q =$

0,1553254438). Soal No. 19, dengan $p = 0,2692307692$ (26,92% atau 7 responden menjawab benar), dikategorikan sebagai sulit dan berkontribusi dengan baik ($p \times q = 0,1967455621$). Terakhir, soal No. 20 dengan $p = 0,3461538462$ (34,62% atau 9 responden menjawab benar) berada dalam kategori cukup sulit dengan kontribusi yang baik ($p \times q = 0,2263313609$). Secara keseluruhan, analisis menunjukkan bahwa tidak ada distribusi yang merata pada tingkat kesulitan setiap soal. Ada 10% soal dengan tingkat kesulitan ekstrem, 20% sedang, 20% cukup sulit, 20% sulit, dan 30% sangat sulit. Ini dapat menjadi faktor penting yang memengaruhi rendahnya reliabilitas instrumen.

Nilai $p \times q$ untuk setiap soal terlihat memberikan variasi dalam jawaban dan seberapa besar kontribusinya terhadap reliabilitas instrumen. Dalam teori pengukuran, nilai $p \times q$ yang terbaik adalah 0,25, yaitu ketika $p = q = 0,5$. Dari sepuluh soal, hanya dua soal (20%) yang menunjukkan kontribusi optimal terhadap reliabilitas ($p \times q$ mendekati 0,25), yaitu soal No. 2 ($p \times q = 0,25$) dan soal No. 17 ($p \times q = 0,2485207101$). Soal No. 3 ($p \times q = 0,2133017815$) dan soal No. 20 ($p \times q = 0,2263313609$) menunjukkan kontribusi yang baik. Diikuti oleh soal No. 15 ($p \times q = 0,1967455621$) dan soal No. 19 ($p \times q = 0,1967455621$) yang memberikan kontribusi yang cukup. Sedangkan soal No. 14 ($p \times q = 0,1020710059$), soal No. 16 ($p \times q = 0,1301775148$), dan soal No. 18 ($p \times q = 0,1553254438$) menunjukkan kontribusi yang relatif kecil. Yang paling mengkhawatirkan adalah soal No. 1 yang memiliki $p \times q = 0$, yang tidak memberikan kontribusi sama sekali terhadap reliabilitas instrumen. Distribusi kontribusi yang tidak optimal ini menjadi penyebab utama rendahnya reliabilitas instrumen.

Total varians dari instrumen terukur sebesar 3,136094675, yang merefleksikan variasi skor total yang diperoleh dari semua responden. Nilai varians keseluruhan ini dipengaruhi oleh berbagai faktor, seperti heterogenitas kemampuan responden, jumlah item soal, dan distribusi dari tingkat kesulitan soal. Dengan total varians 3,136094675 berdasarkan sepuluh soal, bisa dikatakan bahwa keragaman skor total di tingkat yang moderat. Namun, karena jumlah soal yang terbatas, nilai varians ini belum cukup tinggi untuk menjamin reliabilitas yang kuat. Untuk meningkatkan total varians, penting untuk menambah jumlah soal berkualitas dan menyeimbangkan distribusi tingkat kesulitan.

Jumlah jawaban yang benar untuk setiap soal menunjukkan bahwa dari total 260 kemungkinan jawaban, terdapat 70 jawaban yang benar ($26 \text{ responden} \times 10 \text{ soal}$), yang berarti sekitar 26,92%. Persentase yang diperoleh ini menunjukkan bahwa instrumen secara umum sulit bagi responden. Ini bisa jadi salah satu alasan mengapa tingkat reliabilitasnya rendah. Analisis tentang skor total dari responden menunjukkan adanya variasi, dengan skor berkisar antara 2 hingga 7 poin. Hal ini menandakan bahwa instrumen mampu membedakan responden dengan kemampuan yang berbeda, meskipun reliabilitasnya masih di bawah standar yang diharapkan. Beberapa penyebab utama yang dapat dikenali yang menyebabkan rendahnya kehandalan alat ukur mencakup keberagaman item soal dengan variasi tingkat kesulitan yang tidak terkontrol, adanya item soal yang sangat sulit seperti Soal No. 1, serta jumlah item soal yang terbatas (hanya 10 soal). Selain itu, juga ada kemungkinan bahwa konstruk yang diukur oleh item soal tidak cukup homogen, distribusi tingkat kesulitan yang tidak ideal dengan sebagian besar soal tergolong sulit hingga sangat sulit, dan kontribusi $p \times q$ yang masih rendah karena hanya 20% dari item soal yang memberikan kontribusi yang baik terhadap kehandalan.

Kehandalan yang rendah (0,502096361) mengarah pada beberapa implikasi penting saat alat ini digunakan. Pertama, konsistensi dalam pengukuran menjadi kurang baik, yang berarti hasil yang diperoleh mungkin akan bervariasi jika alat ini digunakan pada waktu yang berbeda atau dengan sampel yang serupa. Kedua, terdapat batasan saat menafsirkan skor, sehingga skor yang diperoleh harus ditafsirkan dengan hati-hati karena terdapat kemungkinan kesalahan pengukuran yang signifikan. Ketiga, alat ini mungkin tidak cukup efektif dalam membedakan responden dengan kemampuan yang hampir sama secara konsisten. Keempat, rendahnya kehandalan dapat berdampak

negatif pada validitas alat, karena validitas maksimal yang bisa dicapai hanya sekitar 0,708 (akar kuadrat dari 0,502).

Untuk memperbaiki kehandalan alat ukur, beberapa saran yang bisa diterapkan antara lain adalah melakukan revisi atau mengubah item soal yang tingkat kesulitannya ekstrem, khususnya Soal No. 1, menambah jumlah item soal berkualitas menjadi minimal 19-20 item sesuai dengan rumus Spearman-Brown, menyeimbangkan distribusi tingkat kesulitan dengan komposisi ideal yaitu 30% soal mudah, 40% soal sedang, dan 30% soal sulit, meningkatkan homogenitas konstruk dengan memastikan setiap item soal mengukur aspek yang tepat, meningkatkan kualitas item soal melalui analisis daya beda dan analisis distraktor (untuk soal pilihan ganda), melakukan uji coba alat ukur secara berulang dengan sampel yang representatif, mempertimbangkan penggunaan rumus kehandalan alternatif seperti test-retest atau parallel form, serta memberikan pelatihan kepada pengembang soal mengenai prinsip-prinsip pembuatan item soal yang baik.

Sebagai kesimpulan, alat ukur ini memiliki koefisien kehandalan 0,502096361 yang termasuk dalam kategori "Sedang" namun dinyatakan "Tidak Reliabel" karena belum memenuhi standar minimal yang diharapkan dalam pengukuran pendidikan. Untuk meningkatkan kehandalan alat, perlu dilakukan rangkaian perbaikan komprehensif seperti yang telah disarankan. Dengan melaksanakan saran tersebut, diharapkan alat ukur dapat mencapai tingkat kehandalan yang lebih baik, sehingga dapat memberikan pengukuran yang lebih konsisten dan berarti dalam konteks penelitian atau evaluasi pendidikan.

c). Tingkat kesukaran soal

Berdasarkan analisis mengenai tingkat kesulitan dari sepuluh soal yang diberikan kepada dua puluh enam responden, ditemukan proporsi jawaban benar (P) untuk setiap soal yang menunjukkan variasi dalam tingkat kesulitan. Soal nomor satu memiliki nilai P sebesar 0, yang menunjukkan bahwa tidak ada siswa yang menjawab dengan benar, sehingga soal ini dikategorikan sebagai sangat sulit dan tidak memberikan data yang berguna untuk pengukuran. Soal nomor tiga, lima belas, enam belas, dan delapan belas juga dikategorikan sebagai sulit, dengan nilai P masing-masing 0,307692; 0,115384; 0,153846; dan 0,192307. Hal ini menunjukkan bahwa banyak siswa mengalami kesulitan saat menjawab soal-soal tersebut, yang mungkin karena soal-soal ini terlalu kompleks atau kurang jelas. Soal yang termasuk dalam kategori sedang adalah nomor dua ($P = 0,5$), nomor empat belas (0,269230), nomor tujuh belas (0,538461), nomor sembilan belas (0,269230), dan nomor dua puluh (0,346153). Umumnya, soal-soal dengan tingkat kesulitan sedang dianggap baik karena mampu membedakan siswa yang memahami materi dengan yang tidak. Namun, dari sepuluh soal yang dianalisis, enam soal atau 60% termasuk dalam kategori sulit hingga sangat sulit, sedangkan hanya empat soal (40%) berada dalam kategori sedang. Ketidakseimbangan dalam distribusi tingkat kesulitan ini dapat memengaruhi kualitas keseluruhan tes, terutama dalam hal kemampuan membedakan dan keandalan instrumen. Oleh karena itu, disarankan untuk merevisi atau mengganti soal-soal yang terlalu sulit seperti nomor satu dan lima belas, serta menyusun ulang soal dengan tingkat kesulitan sedang agar lebih seimbang, untuk meningkatkan efektivitas dan akurasi instrumen dalam mengukur kemampuan siswa secara lebih menyeluruh.

d). Daya pembeda

Berdasarkan analisis yang terdapat pada tabel, evaluasi terhadap soal nomor 16 sampai 20 menunjukkan daya pembeda dan tingkat kesulitan masing-masing. Soal nomor 16 memiliki tingkat kesulitan 0,308, yang berarti cukup sulit, dengan hanya sekitar 30,8% peserta yang menjawab dengan benar. Daya pembedanya termasuk dalam kategori baik. Untuk soal nomor 17, angka kesulitannya adalah 0,769, menandakan bahwa soal ini cukup mudah karena sekitar 76,9% peserta mampu menjawab dengan tepat, dan daya pembedanya sangat baik (0,308), sehingga efektif dalam

membedakan kemampuan peserta. Dalam soal nomor 18, tingkat kesulitannya adalah 0,308, dengan daya pembedanya 0,077 yang tergolong rendah, tetapi masih dianggap baik. Soal nomor 19 menunjukkan tingkat kesulitan 0,462, yang dianggap sedang dengan daya pembedanya 0,077, dan termasuk dalam kategori baik. Di sisi lain, soal nomor 20 mencatat indeks kesulitan 0,538 yang juga menunjukkan tingkat kesulitan sedang, sementara daya pembedanya adalah 0,154 dan termasuk baik. Dari kelima soal ini, nomor 17 memiliki kualitas terbaik karena daya pembedanya sangat baik, sedangkan soal lain perlu direvisi, terutama nomor 18 dan 19 yang menunjukkan daya pembedaan relatif rendah meskipun masih dalam kategori baik. Selain itu, analisis jawaban peserta di bagian atas tabel menunjukkan konsistensi sesuai dengan nilai indeks yang diperoleh, di mana terlihat kelompok atas umumnya mendapatkan skor lebih tinggi dibandingkan kelompok bawah.

4. Kesimpulan

Berdasarkan analisis yang telah dilakukan, dapat disimpulkan bahwa evaluasi hasil belajar sangat penting dalam sistem pendidikan, dan harus dilaksanakan secara sistematis dan terukur. Evaluasi ini memberikan gambaran menyeluruh tentang pencapaian siswa dalam aspek kognitif, afektif, dan psikomotor. Dalam penelitian ini, instrumen evaluasi yang digunakan adalah tes objektif tipe pilihan ganda yang dianalisis dengan metode kuantitatif. Hasil dari uji validitas menunjukkan bahwa mayoritas soal adalah valid, meskipun beberapa di antaranya perlu direvisi atau dihapus. Uji reliabilitas menggunakan metode KR-20 menunjukkan bahwa instrumen tersebut belum cukup reliabel, dengan koefisien reliabilitas mencapai 0,502, tergolong sedang dan di bawah standar minimum. Hal ini disebabkan oleh distribusi kesulitan soal yang tidak merata, dominasi dari soal yang sulit, dan kontribusi yang rendah dari beberapa soal terhadap keandalan instrumen. Selain itu, analisis mengenai tingkat kesulitan menunjukkan bahwa 60% soal ada dalam kategori sulit hingga sangat sulit, yang dapat mempengaruhi efektivitas pengukuran. Daya pembeda dari beberapa soal bervariasi, di mana sebagian dapat membedakan peserta didik dengan baik, sementara lainnya memerlukan perbaikan. Oleh karena itu, perlu dilakukan perbaikan terhadap item soal, menambah jumlah soal dengan proporsi yang tepat, serta meningkatkan kualitas soal agar instrumen dapat mencapai reliabilitas dan validitas yang optimal sebagai alat evaluasi dalam proses pembelajaran.

Daftar Pustaka

- Ismail, M. I. (2020). *Evaluasi Pembelajaran: Konsep Dasar, Prinsip, Teknik, dan Prosedur*. Depok: PT Raja Grafindo Persada.
- Keumala, M., Khaldun, I., Hanum, L. (2020). Analisis Butir Soal Ujian Semester Genap Pelajaran Kimia Kelas X SMK SMTI Banda Aceh Tahun Ajaran 2018/2019 Menggunakan Program Proanaltes. *Jurnal Ilmiah Mahasiswa Jurusan Pendidikan Kimia*. 4(2): 27-32
- Mira, K., Hayon, V. H., & Tinenti, Y. R. (2021). Analisis butir soal pilihan ganda pada materi pokok larutan elektrolit dan nonelektrolit kelas X IPA SMA Negeri 1 Lamba Leda Manggarai Timur. *Wahana: Tridarma Perguruan Tinggi*, 73(2): 209-220.
- Mulyani, S., Krismonita, M., & Yamtinah, S. (2022). Analisis butir soal dan kecukupan hots soal ujian akhir semester mata pelajaran kimia SMK Kelas X. *Jurnal Penelitian Pendidikan*, 25(2): 162-172
- Purniasari, L., Masykuri, M., dan Ariani, S. R. D. (2021). Analisis Butir Soal Ujian Sekolah Mata Pelajaran 2019/2020 Menggunakan Model Iteman dan Rasch. *Jurnal Pendidikan Kimia*, 10(2): 205-214.
- Ritonga, P. S. (2017). Analisis Butir Soal Ujian Semester Ganjil Pelajaran Kimia Kelas X di SMA/MA Sekecamatan Pekaitan. *Konfigurasi: Jurnal Pendidikan Kimia dan Terapan*, 1(1): 25-32.
- Sakinah, P., & Ritonga, P. S. (2017). Analisis Butir Soal Ujian Semester Mata Pelajaran Kimia Kelas X Madrasah Aliyah di Kecamatan Pasir Peny. *Konfigurasi: Jurnal Pendidikan Kimia dan Terapan*, 1(1): 129-137.

- Santosa, S., & Badawi, J. A. (2022). Analisis Bugtir Soal Pilihan Ganda Tema Pertumbuhan dan Perkembangan Makhluk Hidup Kelas III Madrasah Ibtidaiyah. *Jurnal Basicedu*, 8(2):1678-1686.
- Simamora, H., Hartono, H., & Effendi, E. (2021). Analisis Kualitas Butir Soal Buatan Guru Kimia Pada Tes Ujian Tengah Semester Ganjil Kelas XII MIPA. *Hydrogen: Jurnal Kependidikan Kimia*, 9(1): 8-18.
- Sudjana, N. (2002). *Dasar-dasar Proses Belajar Mengajar*. Bandung: Sinar Baru.
- Supriyadi. (2021). *Evaluasi Pendidikan*. Pekalongan: PT. Nasya Expanding Management.
- Tarmizi, P., Setiono, P., Amaliyah, Y., & Agrian, A. (2020). Analisis butir soal pilihan ganda tema sehat itu penting kelas V SD Negeri 04 Kota Bengkulu. *ELSE (Elementary School Education Journal): Jurnal Pendidikan dan Pembelajaran Sekolah Dasar*, 4(2): 124-132.