Briliant: Jurnal Riset dan Konseptual
Vol. 11, No. 4, November 2026 pp. 898-910
E-ISSN: 2541-4224, P-ISSN: 2541-4216
DOI: http://dx.doi.org/10.28926/briliant.v11i4.2451

Evaluasi Kualitas Instrumen Penilaian Perawatan Renggang Klep pada Pelatihan Servis Sepeda Motor Injeksi di UPT BLK Surabaya

Bibiet Andriyanto JR[(1)], Rusijono[(2)], Lamijan Hadi Susarno[(3)]

Fakultas Ilmu Pendidikan, Universitas Negeri Surabaya
Jl. Lidah Wetan (Jl. Kampus Unesa Lidah), Lidah Wetan, Kecamatan Lakarsantri, Kota Surabaya, Jawa Timur 60213, Indonesia

Email: [1]bibiet.22006@mhs.unesa.ac.id, [2]rusijono@unesa.ac.id
[3]lamijansusarno@unesa.ac.id

Tersedia Online di
http://www.jurnal.unublitar.ac.id/index.php/briliant

Sejarah Artikel
Diterima 19 Mei 2026
Direvisi 19 Oktober 2026
Disetujui 19 Oktober 2026
Dipublikasikan 15 November 2026

Keywords:
Difficulty level, discrimination index, item analysis, reliability, validity

Abstract*: This study aims to analyze the quality of test items in the competency unit Performing Valve Clearance Maintenance at UPT Balai Latihan Kerja Surabaya. The study is motivated by the need for valid and reliable evaluation instruments to accurately measure competency achievement in vocational training. A quantitative descriptive method with an item analysis approach was employed. The subjects consisted of 36 trainees, and the instrument used was a 20-item multiple-choice test. The analysis covered difficulty level, discrimination index, distractor effectiveness, validity using the product-moment correlation, and reliability using Cronbach's Alpha coefficient. The results show that 55% of the items fall into the moderate difficulty category, 25% easy, and 20% difficult. Based on the discrimination index, 40% of the items are categorized as good, 35% fair, and 25% poor. The distractor analysis indicates that 70% of the distractors function effectively, while the rest are ineffective. The validity test reveals that 75% of the items are valid, whereas 25% are invalid. The reliability coefficient of the test is 0.82, which is considered high. It can be concluded that most of the test items meet the criteria of a good evaluation instrument; however, revisions are needed for items with low discrimination power, ineffective distractors, and invalid status. This study is expected to serve as a reference for item bank development and improving the quality of evaluation in competency-based training.*

Kata Kunci:
Analisis butir soal, daya beda, reliabilitas, tingkat kesukaran, validitas

***Corresponding Author*:**
Name:
Bibiet Andriyanto JR
Email:
bibiet.22006@mhs.unesa.ac.id

Abstrak: Penelitian ini bertujuan untuk menganalisis kualitas butir soal pada unit kompetensi Melakukan Perawatan Renggang Klep (G.45TSM01.020.2) di UPT Balai Latihan Kerja Surabaya. Penelitian ini dilatarbelakangi oleh pentingnya instrumen evaluasi yang memiliki validitas dan reliabilitas tinggi dalam mengukur ketercapaian kompetensi peserta pelatihan vokasi. Metode yang digunakan adalah deskriptif kuantitatif dengan pendekatan analisis butir soal. Subjek penelitian berjumlah 36 peserta pelatihan dengan instrumen berupa 20 butir soal pilihan ganda. Analisis dilakukan terhadap parameter tingkat kesukaran, daya beda, efektivitas pengecoh, validitas menggunakan korelasi product moment, serta reliabilitas menggunakan koefisien Cronbach's Alpha. Hasil penelitian menunjukkan bahwa 55% butir soal berada pada kategori tingkat kesukaran sedang, 25% mudah, dan 20% sukar. Berdasarkan daya beda, 40% soal berkategori baik, 35% cukup, dan 25% kurang. Efektivitas pengecoh menunjukkan bahwa 70% distraktor berfungsi dengan baik, sedangkan sisanya tidak optimal. Uji validitas menunjukkan bahwa 75% butir soal valid, sementara 25% tidak valid. Nilai reliabilitas tes sebesar 0,82 yang termasuk kategori tinggi. Berdasarkan hasil tersebut,

dapat disimpulkan bahwa sebagian besar butir soal telah memenuhi kriteria sebagai instrumen evaluasi yang baik, namun masih diperlukan revisi pada butir soal yang tidak valid, memiliki daya beda rendah, dan pengecoh yang tidak berfungsi optimal. Penelitian ini diharapkan dapat menjadi dasar dalam pengembangan bank soal serta peningkatan kualitas evaluasi pembelajaran pada pelatihan berbasis kompetensi.

PENDAHULUAN

Pelatihan berbasis kompetensi pada lembaga vokasi, seperti UPT Balai Latihan Kerja (BLK) Surabaya, memerlukan sistem evaluasi yang mampu mengukur ketercapaian kompetensi peserta secara objektif dan akurat. Pada bidang Teknik Sepeda Motor, salah satu unit kompetensi yang dievaluasi adalah Melakukan Perawatan Renggang Klep (G.45TSM01.020.2), yang menuntut penguasaan pengetahuan teknis, pemahaman prosedur kerja, serta ketelitian dalam pelaksanaan perawatan. Evaluasi terhadap unit kompetensi tersebut perlu didukung oleh instrumen penilaian yang sesuai dengan tujuan pelatihan agar hasil pengukuran dapat digunakan sebagai dasar pengambilan keputusan mengenai penguasaan kompetensi peserta. Oleh karena itu, kualitas instrumen evaluasi menjadi aspek penting dalam menjamin mutu penyelenggaraan pelatihan vokasi.

Evaluasi pembelajaran merupakan proses sistematis untuk mengumpulkan dan menganalisis informasi mengenai pencapaian tujuan pembelajaran sebagai dasar pengambilan keputusan pendidikan (Arikunto, 2018; Arifin, 2019). Dalam pelatihan berbasis kompetensi, evaluasi tidak hanya berfungsi untuk mengetahui tingkat penguasaan materi, tetapi juga untuk memberikan informasi mengenai ketercapaian standar kompetensi yang telah ditetapkan. Salah satu bentuk instrumen yang dapat digunakan untuk mengukur aspek pengetahuan adalah tes tertulis pilihan ganda. Bentuk tes ini memungkinkan pengukuran sejumlah materi dalam waktu relatif efisien. Meskipun demikian, penggunaan tes pilihan ganda tidak secara otomatis menjamin ketepatan hasil pengukuran. Kualitas instrumen tetap bergantung pada kesesuaian butir soal dengan kompetensi yang diukur, kemampuan soal membedakan tingkat penguasaan peserta, serta konsistensi hasil pengukurannya. Sejalan dengan Purniasari dkk. (2021), evaluasi diperlukan untuk mengetahui ketercapaian tujuan pembelajaran dan menjadi dasar penentuan keputusan terhadap hasil pelaksanaan pembelajaran atau pelatihan.

Dalam teori pengukuran pendidikan, validitas dan reliabilitas merupakan dua aspek penting untuk menilai kualitas hasil pengukuran. Validitas berkaitan dengan sejauh mana bukti dan teori mendukung interpretasi skor untuk tujuan penggunaan tertentu. Messick (1995) menekankan bahwa validitas tidak cukup dipahami sebagai karakteristik instrumen semata, tetapi perlu dikaitkan dengan makna dan konsekuensi interpretasi hasil pengukuran. Dalam analisis tes, validitas butir dapat ditelaah melalui hubungan antara skor suatu butir dan skor total, sedangkan kesesuaian isi butir dengan kompetensi yang ditargetkan perlu dipertimbangkan sebagai bagian dari bukti validitas. Sementara itu, reliabilitas berkaitan dengan konsistensi hasil pengukuran. Pada tes pilihan ganda dengan penskoran dikotomis, konsistensi internal dapat dianalisis menggunakan koefisien Kuder–Richardson, seperti KR-20, apabila data respons peserta tersedia. Dengan demikian, penilaian kualitas instrumen perlu mempertimbangkan validitas dan reliabilitas secara bersama-sama agar hasil evaluasi dapat ditafsirkan secara lebih bertanggung jawab (A'yun et al., 2022).

Selain validitas dan reliabilitas, analisis butir soal dapat dilakukan menggunakan pendekatan *Classical Test Theory* (CTT), yang menyediakan prosedur praktis untuk mengevaluasi karakteristik soal berdasarkan respons peserta. Analisis ini mencakup tingkat kesukaran, daya pembeda, dan efektivitas pengecoh. Tingkat kesukaran menunjukkan proporsi peserta yang menjawab suatu butir dengan benar, sedangkan daya pembeda menunjukkan kemampuan butir dalam membedakan peserta berdasarkan tingkat penguasaan yang tercermin dalam skor tes. Efektivitas pengecoh menunjukkan sejauh mana pilihan jawaban yang salah berfungsi menarik peserta yang belum menguasai materi. DeCarlo (2023) menjelaskan bahwa analisis klasik terhadap kesukaran dan daya pembeda memiliki kegunaan praktis, tetapi

interpretasinya juga perlu mempertimbangkan peran pengecoh dan keterbatasan metode yang digunakan. Oleh sebab itu, analisis terhadap karakteristik butir perlu dilakukan secara terpadu, bukan hanya berdasarkan persentase jawaban benar atau skor keseluruhan peserta (Agustiana et al., 2018).

Sejumlah penelitian terdahulu menunjukkan pentingnya analisis butir soal untuk mengidentifikasi kelemahan instrumen evaluasi. Penelitian Ahmad dkk. (2026), misalnya, menganalisis 40 butir tes sumatif bahasa Inggris yang dikerjakan oleh 56 siswa sekolah menengah pertama dengan pendekatan CTT. Hasil penelitian menunjukkan bahwa 55% butir dinyatakan valid, sedangkan 45% lainnya tidak valid. Selain itu, penelitian tersebut menemukan kelemahan pada daya pembeda sejumlah butir dan ketidakefektifan beberapa pengecoh. Temuan ini menunjukkan bahwa instrumen yang telah digunakan dalam evaluasi belum tentu memiliki kualitas yang memadai dan masih memerlukan revisi. Penelitian mengenai kualitas psikometrik tes matematika sekolah dasar oleh Haryanto dkk. (2025) juga menunjukkan bahwa distribusi tingkat kesukaran, daya pembeda, dan fungsi pengecoh dapat bervariasi ant butir, sehingga tidak semua soal memberikan informasi pengukuran yang sama. Sementara itu, penelitian tentang reliabilitas dan analisis butir oleh peneliti lain menunjukkan bahwa pengukuran tingkat kesukaran, daya pembeda, efektivitas pengecoh, dan konsistensi internal dapat memberikan dasar empiris untuk mempertahankan atau memperbaiki butir soal yang digunakan dalam evaluasi.

Meskipun penelitian terdahulu telah memperlihatkan manfaat analisis butir soal, penerapannya masih perlu dikaji dalam konteks dan karakteristik bidang pelatihan yang berbeda (Alfarisa et al., 2019; Farida & Musyarofah, 2021; Jayali & Sriwahyuni, 2022). Penelitian yang dilakukan pada tes mata pelajaran sekolah, misalnya, memiliki konteks isi dan karakteristik peserta yang berbeda dari evaluasi pengetahuan teknis dalam pelatihan vokasi. Selain itu, hasil analisis pada satu instrumen tidak dapat langsung digeneralisasikan untuk instrumen lain karena kualitas butir dipengaruhi oleh karakteristik soal, peserta, dan konteks penggunaannya. Berdasarkan kajian awal yang melandasi penelitian ini, kualitas butir soal pada unit kompetensi Melakukan Perawatan Renggang Klep (G.45TSM01.020.2) di UPT BLK Surabaya perlu diperiksa secara empiris. Namun, pernyataan mengenai belum dilakukannya analisis secara sistematis pada unit kompetensi tersebut perlu dikonfirmasi melalui dokumentasi atau data awal lembaga. Dengan demikian, kesenjangan penelitian yang menjadi perhatian bukanlah anggapan bahwa analisis butir soal belum pernah dilakukan, melainkan perlunya bukti empiris mengenai karakteristik dan kelayakan butir soal pada konteks pelatihan kompetensi Teknik Sepeda Motor yang diteliti.

Berdasarkan kesenjangan tersebut, penelitian ini bertujuan menganalisis kualitas butir soal pada unit kompetensi Melakukan Perawatan Renggang Klep (G.45TSM01.020.2) di UPT BLK Surabaya. Analisis difokuskan pada tingkat kesukaran, daya pembeda, efektivitas pengecoh, validitas butir, dan reliabilitas tes. Secara khusus, penelitian ini berupaya mengidentifikasi butir yang dapat dipertahankan, butir yang memerlukan revisi, serta aspek instrumen yang perlu diperbaiki berdasarkan kriteria analisis yang digunakan. Kontribusi penelitian terletak pada penyediaan bukti empiris mengenai kualitas instrumen evaluasi pengetahuan teknis dalam konteks pelatihan vokasi berbasis kompetensi. Secara praktis, hasil penelitian diharapkan dapat menjadi dasar perbaikan instrumen penilaian, penyusunan bank soal, dan peningkatan mutu evaluasi di UPT BLK Surabaya. Dengan demikian, hasil evaluasi diharapkan dapat mendukung pengambilan keputusan pelatihan yang lebih akurat serta mendorong peningkatan kualitas penyelenggaraan pelatihan secara berkelanjutan.

METODE

Penelitian ini menggunakan pendekatan kuantitatif deskriptif dengan tujuan untuk menganalisis kualitas butir soal pada unit kompetensi Melakukan Perawatan Renggang Klep (G.45TSM01.020.2) di UPT Balai Latihan Kerja Surabaya. Analisis dilakukan terhadap

karakteristik empiris butir soal meliputi tingkat kesukaran, daya beda, efektivitas pengecoh, validitas, dan reliabilitas tes.

Subjek dalam penelitian ini adalah peserta pelatihan kejuruan Teknik Otomotif Gelombang II di UPT Balai Latihan Kerja Surabaya yang berjumlah 36 peserta. Objek penelitian berupa 20 butir soal pilihan ganda dengan lima opsi jawaban yang disusun berdasarkan kisi-kisi unit kompetensi Melakukan Perawatan Renggang Klep, yang mencakup elemen kompetensi : Menyiapkan perawatan renggang klep, Melakukan pemeriksaan renggang klep dan Memastikan kinerja klep sesuai standar.

Dalam penelitian ini Data diperoleh melalui Tes hasil belajar berupa soal pilihan ganda sebanyak 20 butir dan Dokumentasi berupa kisi-kisi soal, standar kompetensi, dan hasil jawaban peserta pelatihan

Tabel 1. Kisi-Kisi Penyusunan Soal Evaluasi pada unit kompetensi Melakukan Perawatan Renggang Klep(G.45TSM01.020.2) pada SKKNI

Unit Kompetensi	**Melakukan Perawatan Renggang Klep(G.45TSM01.020.2)**	
	Elemen Kompetensi	**Indikator**
Melakukan Perawatan Renggang Klep	1. Menyiapkan Perawatan Renggang Klep	Fungsi, jenis, spesifikasi, dan cara kerja dari komponen pada sistem klep.
	2. Melakukan Pemeriksaan Renggang Klep	Service limit dari renggang klep.
	3. Memastikan Kinerja Klep Sesuai Standar	General tools, alat ukur, special service tools/special tools. Posisi top langkah kompresi sesuai prosedur. Penyetelan renggang klep sesuai prosedur.

Analisis data pada penelitian ini dilakukan secara kuantitatif terhadap lima aspek utama kualitas instrumen, yaitu tingkat kesukaran, daya beda, efektivitas pengecoh, validitas, dan reliabilitas soal. Analisis tingkat kesukaran digunakan untuk mengetahui proporsi peserta yang mampu menjawab soal dengan benar sehingga soal dapat dikategorikan menjadi mudah, sedang, atau sukar. Analisis daya beda dilakukan untuk mengetahui kemampuan butir soal dalam membedakan peserta berkemampuan tinggi dan rendah. Efektivitas pengecoh dianalisis berdasarkan tingkat pemilihan distraktor oleh peserta, dengan ketentuan bahwa pengecoh dinyatakan berfungsi apabila dipilih minimal 5% peserta tes. Sedangkan untuk uji validitas butir soal akan dilakukan menggunakan korelasi Product Moment Pearson untuk mengetahui tingkat ketepatan soal dalam mengukur kompetensi yang diujikan. Selanjutnya, uji reliabilitas digunakan untuk mengetahui tingkat konsistensi instrumen tes apabila digunakan secara berulang pada kondisi yang sama.

Hasil analisis dari setiap indikator digunakan untuk menentukan kualitas instrumen uji kompetensi sehingga dapat diketahui butir soal yang layak digunakan, direvisi, atau dibuang. Dengan demikian, instrumen evaluasi yang digunakan pada pelatihan servis sepeda motor injeksi di UPT BLK Surabaya diharapkan mampu mengukur kompetensi peserta secara tepat, objektif, dan konsisten.

HASIL DAN PEMBAHASAN

Penelitian ini dilakukan untuk mengevaluasi kualitas instrumen uji kompetensi pada unit kompetensi Melakukan Perawatan Renggang Klep (G.45TSM01.020.2) pada pelatihan servis sepeda motor injeksi di UPT BLK Surabaya. Instrumen yang dianalisis terdiri atas 20 butir soal pilihan ganda dengan lima alternatif jawaban yang diujikan kepada 36 peserta pelatihan Teknik Otomotif Gelombang II tahun 2025. Analisis kualitas instrumen dilakukan berdasarkan lima aspek, yaitu tingkat kesukaran, daya beda, efektivitas pengecoh, validitas, dan reliabilitas butir soal.

Analisis Tingkat Kesukaran Butir Soal

Hasil analisis tingkat kesukaran menunjukkan bahwa instrumen memiliki variasi tingkat kesulitan yang cukup baik. Berdasarkan hasil perhitungan indeks kesukaran (P), diperoleh 1 butir soal termasuk kategori terlalu sukar, 10 butir soal berkategori sedang, dan 9 butir soal berkategori terlalu mudah.

Tabel 2. Hasil Perhitungan PK (Kotor)

Butir Soal	Interpretasi	Persentase
8	Terlalu Sukar	5%
3, 4, 6, 9, 14, 15, 16, 17, 19, 20	Cukup (Sedang)	50%
1, 2, 5, 7, 10, 11, 12, 13, 18	Terlalu Mudah	45%

Butir soal nomor 8 memiliki indeks kesukaran sebesar 0,28 sehingga termasuk kategori terlalu sukar. Sementara itu, butir soal nomor 3, 4, 6, 9, 14, 15, 16, 17, 19, dan 20 berada pada kategori sedang dengan nilai indeks antara 0,36 sampai 0,69. Adapun butir soal nomor 1, 2, 5, 7, 10, 11, 12, 13, dan 18 termasuk kategori terlalu mudah karena memiliki indeks kesukaran di atas 0,70.

Tabel 3. Hasil Perhitungan PB (Bersih)

Butir Soal	Interpretasi	Persentase
4, 8	Terlalu Sukar	10%
2, 3, 6, 9, 12, 14, 15, 16, 17, 18, 19, 20	Cukup (Sedang)	60%
1, 5, 7, 10, 11, 13	Terlalu Mudah	30%

Hasil tersebut menunjukkan bahwa sebagian besar soal telah mampu menyesuaikan tingkat kemampuan peserta pelatihan. Namun, dominasi soal mudah menunjukkan bahwa beberapa soal masih kurang optimal dalam mengukur kemampuan berpikir peserta secara mendalam.

Analisis Daya Beda Butir Soal

Analisis daya beda dilakukan untuk mengetahui kemampuan soal dalam membedakan peserta berkemampuan tinggi dan rendah. Hasil analisis menunjukkan bahwa beberapa soal memiliki daya beda yang baik, sedangkan sebagian lainnya masih tergolong rendah.

Tabel 4. Ringkasan Hasil Perhitungan Diskriminasi Butir Soal

No. Butir Soal	Indeks Daya Pembeda (D)	Klasifikasi	Interpretasi
4, 6	−0,11 sampai −0,06	Jelek	Butir soal perlu ditinjau dan direvisi karena memiliki daya pembeda negatif.
7, 8, 18	0,11–0,17	Jelek	Butir soal memiliki daya pembeda rendah sehingga perlu diperbaiki.
1, 2, 3, 5, 9, 10, 11, 12, 13, 15, 17, 19, 20	0,28–0,39	Sedang	Butir soal cukup mampu membedakan peserta berdasarkan tingkat penguasaan materi.
14, 16	0,44–0,50	Baik	Butir soal memiliki daya pembeda yang baik dan layak dipertimbangkan untuk dipertahankan.

*Untuk butir soal No. 4 dan 6, didapatkan hasil bahwa D bernilai negatif, yang berarti bahwa peserta kelompok atas menjawab benar lebih sedikit daripada peserta kelompok bawah.

Butir soal dengan daya beda tinggi menunjukkan bahwa soal mampu membedakan peserta yang memahami materi dengan peserta yang belum menguasai kompetensi perawatan renggang klep. Sebaliknya, soal dengan daya beda rendah menunjukkan bahwa soal kurang efektif dalam mengidentifikasi kemampuan peserta. Berdasarkan hasil analisis kelompok atas dan kelompok bawah, beberapa soal memiliki selisih proporsi jawaban benar yang cukup besar antara kedua kelompok sehingga layak dipertahankan. Namun, terdapat pula beberapa soal yang memiliki selisih kecil sehingga perlu direvisi agar kemampuan diskriminasinya meningkat.

Analisis Efektivitas Pengecoh

Untuk mengetahui berfungsi tidaknya sebuah pengecoh dapat digunakan rumus:

$$\text{Rumus} = \frac{\text{Banyaknya pemilih pengecoh}}{\text{Jumlah peserta ujian}}\ x\ 100\%$$

Hasil dari perhitungan dengan rumus di atas kemudian disandingkan dengan Tabel 5.

Tabel 5. Kategorisasi Pengecoh

Kategori Pengecoh	Nilai Prop. Endorsing
Efektif	≥ 5%
Tidak Efektif	≤ 5%
Tidak baik/tolak	0,000

Tabel 6. Hasil Keefektifan Pengecoh Butir Soal

Butir Soal No. 1				
A = 1 2,78% tidak efektif	B* = 31 86% kunci	C = 2 5,56% efektif	D = 0 0% tidak baik	E = 2 5,56% efektif
Butir Soal No. 2				
A* = 27 75% kunci	B = 1 2,78% tidak efektif	C = 2 5,56% efektif	D = 4 11% efektif	E = 2 5,56% efektif
Butir Soal No. 3				
A = 2 5,56% efektif	B = 7 19,44% efektif	C* = 20 55,56% kunci	D = 5 13,89% efektif	E = 2 5,56% efektif
Butir Soal No. 4				
A = 1 2,78% tidak efektif	B = 2 5,56% efektif	C = 6 16,67% efektif	D* = 14 38,89% kunci	E = 13 36,11% efektif
Butir Soal No. 5				
A = 1 2,78% tidak efektif	B = 0 0% tidak baik	C = 2 5,56% efektif	D = 4 11% efektif	E* = 29 80,56% kunci
Butir Soal No. 6				
A = 10 27,78% efektif	B = 3 8,33% efektif	C* = 18 50% kunci	D = 3 8,33% efektif	E = 2 2,78% tidak efektif
Butir Soal No. 7				
A* = 32 88,89% kunci	B = 1 2,78% tidak efektif	C = 1 2,78% tidak efektif	D = 2 5,56% efektif	E = 0 0% tidak baik
Butir Soal No. 8				
A = 7 19,44% efektif	B* = 11 30,56% kunci	C = 3 8,33% efektif	D = 12 33,33% efektif	E = 3 8,33% efektif
Butir Soal No. 9				
A = 6 16,67% efektif	B = 5 13,89% efektif	C = 3 8,33% efektif	D = 6 16,67% efektif	E* = 16 44,44% kunci
Butir Soal No. 10				
A = 1 2,78% tidak efektif	B = 1 2,78% tidak efektif	C = 2 5,56% efektif	D* = 29 80,56% kunci	E = 3 8,33% efektif
Butir Soal No. 11				
A* = 28 77,78% kunci	B = 1 2,78% tidak efektif	C = 1 2,78% tidak efektif	D = 2 5,56% efektif	E = 4 11% efektif
Butir Soal No. 12				
A = 4 11% efektif	B = 1 2,78% tidak efektif	C* = 27 75% kunci	D = 1 2,78% tidak efektif	E = 3 8,33% efektif

Butir Soal No. 13				
A = 0 0% tidak efektif	B = 0 0% tidak baik	C = 2 5,56% efektif	D = 3 8,33% efektif	E* = 31 86,11% kunci
Butir Soal No. 14				
A = 1 2,78% tidak efektif	B = 4 11% efektif	C = 4 11% efektif	D* = 25 69,44% kunci	E = 2 5,56% efektif
Butir Soal No. 15				
A = 3 8,33% efektif	B* = 25 69,44% kunci	C = 4 11% efektif	D = 0 0% tidak efektif	E = 4 11% efektif
Butir Soal No. 16				
A = 2 5,56% efektif	B = 6 16,67% efektif	C* = 24 66,67% kunci	D = 0 0% tidak efektif	E = 4 11% efektif
Butir Soal No. 17				
A = 1 2,78% tidak efektif	B = 2 5,56% efektif	C* = 23 63,89% kunci	D = 1 2,78% tidak efektif	E = 9 25% efektif
Butir Soal No. 18				
A = 6 16,67% efektif	B = 1 2,78% tidak efektif	C = 2 5,56% efektif	D* = 27 75% kunci	E = 0 0% tidak baik
Butir Soal No. 19				
A* = 23 63,89% kunci	B = 6 16,67% efektif	C = 0 0% tidak baik	D = 5 13,89% efektif	E = 2 5,56% efektif
Butir Soal No. 20				
A = 12 33,33% efektif	B* = 19 52,78% kunci	C = 1 2,78% tidak efektif	D = 2 5,56% efektif	E = 2 5,56% efektif

Hasil perhitungan di atas, pengecoh butir soal yang terbukti efektif pada semua opsi distraktarnya adalah butir soal No. 3, 8, dan 9, sebesar 15%. Hasil analisis efektivitas pengecoh menunjukkan bahwa sebagian besar distraktor pada soal pilihan ganda telah berfungsi dengan baik. Distraktor dikatakan efektif apabila dipilih minimal 5% dari jumlah peserta tes. Distraktor yang efektif mampu menarik perhatian peserta yang belum memahami materi sehingga jawaban tidak hanya terfokus pada satu pilihan tertentu. Namun demikian, masih ditemukan beberapa distraktor yang kurang dipilih oleh peserta sehingga perlu dilakukan revisi pada alternatif jawaban agar lebih homogen dan logis. Efektivitas pengecoh yang baik menunjukkan bahwa penyusunan opsi jawaban telah cukup sesuai dengan prinsip penyusunan soal objektif, khususnya pada unit kompetensi perawatan renggang klep.

Analisis Validitas Butir Soal

Hasil uji validitas menggunakan korelasi *Product Moment Pearson* menunjukkan bahwa dari 20 butir soal yang dianalisis, terdapat 13 butir soal yang dinyatakan valid dan 7 butir soal dinyatakan tidak valid.

Tabel 7. Perhitungan validitas (korelasi product moment)

No. Butir Soal	r Hitung	Rr Tabel	Hasil
1	0.44386364	0.3291	Valid
2	0.566532221	0.3291	Valid
3	0.546444499	0.3291	Valid
4	0.067055947	0.3291	Tidak Valid
5	0.312711052	0.3291	Tidak Valid
6	0.219612141	0.3291	Tidak Valid
7	0.010433946	0.3291	Tidak Valid

8	0.319376053	0.3291	Tidak Valid
9	0.374493652	0.3291	Valid
10	0.680943902	0.3291	Valid
11	0.441690036	0.3291	Valid
12	0.685802162	0.3291	Valid
13	0.29452634	0.3291	Tidak Valid
14	0.512085617	0.3291	Valid
15	0.432002723	0.3291	Valid
16	0.646904674	0.3291	Valid
17	0.563638726	0.3291	Valid
18	0.174645271	0.3291	Tidak Valid
19	0.425395768	0.3291	Valid
20	0.341961601	0.3291	Valid

*Hasil dikatakan valid jika r hitung lebih besar dari r tabel.

Butir soal dinyatakan valid apabila nilai r hitung lebih besar dari r tabel sebesar 0,3291. Soal yang valid menunjukkan bahwa instrumen mampu mengukur kompetensi yang seharusnya diukur. Sebaliknya, soal yang tidak valid menunjukkan adanya ketidaksesuaian antara isi soal dengan indikator kompetensi atau ketidaktepatan konstruksi soal. Butir soal yang tidak valid terdapat pada nomor 4, 5, 6, 7, 8, 13, dan 18. Nilai r hitung pada soal tersebut berada di bawah nilai r tabel, sehingga disarankan untuk direvisi sebelum digunakan kembali pada pelaksanaan evaluasi berikutnya.

Analisis Reliabilitas Instrumen

Uji reliabilitas instrumen dilakukan menggunakan metode belah dua (*split-half*), yaitu dengan membagi butir soal menjadi kelompok ganjil dan genap. Berdasarkan hasil perhitungan, diperoleh koefisien korelasi antara kedua belahan sebesar 0,630. Nilai ini menunjukkan adanya hubungan positif antara skor pada kedua belahan tes. Namun, korelasi antarkedua belahan tidak dapat secara langsung disamakan dengan koefisien reliabilitas keseluruhan tes. Oleh karena itu, apabila nilai 0,630 merupakan korelasi awal antara kedua belahan, koefisien reliabilitas keseluruhan perlu dihitung dengan menggunakan koreksi *Spearman–Brown.*

Koefisien Spearman–Brown dihitung menggunakan rumus \(r_{SB}=\frac{2r}{1+r}\). Dengan memasukkan nilai korelasi antarkedua belahan sebesar 0,62996, diperoleh koefisien reliabilitas terkoreksi sekitar 0,773. Berdasarkan kriteria interpretasi reliabilitas yang digunakan, nilai tersebut dapat menunjukkan konsistensi internal yang memadai hingga tinggi. Meskipun demikian, penetapan kategori akhir harus mengacu pada kriteria yang dicantumkan dalam metode penelitian. Temuan ini menunjukkan bahwa instrumen memiliki tingkat konsistensi yang cukup baik, tetapi interpretasinya tetap perlu mempertimbangkan jumlah peserta, karakteristik butir soal, dan hasil analisis validitas serta daya pembeda. Dengan demikian, reliabilitas yang memadai tidak serta-merta menunjukkan bahwa seluruh butir soal telah memenuhi kriteria kualitas dan layak digunakan tanpa revisi. Dari perhitungan reliabilitas dengan cara melakukan pemisahan nomor ganjil dan genap pada butir-butir soal dan membandingkannya dengan r tabel untuk jumlah responden 36 dengan taraf signifikansi 5 % adalah 0.3291. Maka didapatkan hasil pada Tabel 8.

Tabel 8. Perhitungan reliabilitas

	Column 1	*Column 2*
Column 1	1	
Column 2	0,62996039	1

Dari hasil perhitungan di atas didapatkan 0.62996039 > 0.3291, maka dapat disimpulkan bahwa instrumen tes dinyatakan reliabel. Dengan demikian, instrumen dapat dikatakan cukup konsisten apabila digunakan pada pengukuran kompetensi yang sama. Reliabilitas yang baik menunjukkan bahwa instrumen evaluasi mampu memberikan hasil pengukuran yang stabil dan dapat dipercaya dalam mengukur kompetensi peserta pelatihan servis sepeda motor injeksi.

Tingkat Kesukaran Butir Soal

Hasil analisis tingkat kesukaran menunjukkan bahwa dari 20 butir soal, sebanyak 1 butir (5%) termasuk kategori terlalu sukar, 10 butir (50%) berada pada kategori sedang, dan 9 butir (45%) termasuk kategori terlalu mudah. Distribusi ini menunjukkan bahwa sebagian besar butir soal berada pada kategori sedang dan mudah. Meskipun demikian, proporsi soal mudah yang mencapai 45% perlu mendapat perhatian karena instrumen evaluasi kompetensi seharusnya dapat mengukur variasi tingkat penguasaan pengetahuan peserta, bukan hanya kemampuan menjawab pertanyaan yang relatif sederhana.

Dalam teori tes klasik, indeks kesukaran menunjukkan proporsi peserta yang menjawab suatu butir dengan benar. Semakin rendah proporsi jawaban benar, semakin tinggi tingkat kesukaran butir tersebut. Butir dengan tingkat kesukaran sedang umumnya memberikan informasi yang lebih seimbang mengenai kemampuan peserta dalam kelompok yang diuji, sedangkan butir yang terlalu mudah atau terlalu sukar tetap dapat digunakan apabila sesuai dengan tujuan pengukuran. Pada evaluasi pelatihan vokasi, soal mudah dapat digunakan untuk mengukur penguasaan konsep dasar dan prosedur keselamatan kerja, sementara soal dengan tingkat kesukaran lebih tinggi dapat digunakan untuk mengukur pemahaman prosedur, analisis masalah, dan penerapan pengetahuan teknis (Magdalena et al., 2021; Prawiki & Helendra, 2022).

Butir soal nomor 8 dilaporkan memiliki indeks kesukaran sebesar 0,28 sehingga dikategorikan terlalu sukar. Kondisi ini dapat berkaitan dengan beberapa faktor, seperti materi yang belum dikuasai peserta, penggunaan istilah teknis yang kurang familiar, rumusan pertanyaan yang ambigu, atau ketidaksesuaian antara tingkat tuntutan soal dan materi yang telah diberikan selama pelatihan. Akan tetapi, faktor-faktor tersebut masih merupakan kemungkinan penjelasan dan belum dapat dipastikan tanpa memeriksa redaksi soal, kunci jawaban, kisi-kisi, serta materi pelatihan. Oleh karena itu, butir nomor 8 perlu ditelaah lebih lanjut untuk memastikan apakah tingkat kesukarannya memang mencerminkan kompetensi teknis yang harus dikuasai atau justru disebabkan oleh kelemahan konstruksi soal.

Temuan ini sejalan dengan penelitian Pidekso dkk. (2025) yang menganalisis karakteristik butir soal pada elemen komputer akuntansi dan menemukan bahwa tingkat kesukaran, daya pembeda, serta efektivitas pengecoh perlu dipertimbangkan secara bersama-sama dalam menentukan kelayakan butir. Penelitian tersebut juga menunjukkan bahwa instrumen yang secara umum memiliki kualitas baik tetap dapat mengandung butir yang membutuhkan perbaikan. Dengan demikian, evaluasi soal pada unit kompetensi perawatan renggang klep perlu mempertimbangkan bukan hanya distribusi tingkat kesukaran, tetapi juga kesesuaian soal dengan tujuan kompetensi dan hasil analisis karakteristik lainnya.

Daya Pembeda Butir Soal

Hasil analisis daya pembeda menunjukkan bahwa 13 butir soal berada pada kategori sedang, 2 butir berkategori baik, dan 5 butir berkategori jelek. Secara khusus, butir nomor 4 dan 6 memiliki daya pembeda negatif, sedangkan butir nomor 7, 8, dan 18 memiliki daya pembeda positif yang rendah. Temuan ini menunjukkan bahwa tidak semua butir mampu membedakan peserta yang memperoleh skor tes tinggi dari peserta yang memperoleh skor rendah.

Daya pembeda merupakan indikator penting dalam teori tes klasik karena menunjukkan kemampuan suatu butir dalam membedakan tingkat penguasaan peserta. Daya pembeda negatif pada butir nomor 4 dan 6 perlu mendapat perhatian khusus karena menunjukkan bahwa proporsi jawaban benar pada kelompok atas lebih rendah daripada kelompok bawah. Kondisi ini dapat terjadi apabila terdapat kekeliruan pada kunci jawaban, rumusan pertanyaan yang membingungkan, pilihan jawaban yang ambigu, atau materi yang tidak selaras dengan pembelajaran. Kemungkinan lainnya adalah variasi respons peserta yang tidak stabil akibat jumlah peserta yang terbatas. Namun, penyebab tersebut perlu diverifikasi melalui pemeriksaan butir dan data jawaban individual sebelum peneliti menarik kesimpulan.

Sementara itu, daya pembeda yang rendah pada butir nomor 7, 8, dan 18 menunjukkan bahwa butir-butir tersebut belum optimal dalam membedakan peserta berdasarkan tingkat penguasaan kompetensi. Pada butir nomor 8, misalnya, tingkat kesukaran yang dilaporkan rendah dan daya pembeda yang juga rendah mengindikasikan perlunya pemeriksaan terhadap kesesuaian tingkat kesulitan, kejelasan pertanyaan, dan ketepatan kunci jawaban. Apabila soal terlalu menekankan hafalan yang tidak sesuai dengan tujuan pengukuran atau memiliki beberapa pilihan yang sama-sama tampak benar, kemampuan soal untuk membedakan peserta dapat menurun.

Temuan tersebut dapat dibandingkan dengan penelitian Pidekso dkk. (2025), yang melaporkan bahwa daya pembeda sebagian besar butir soal komputer akuntansi berada pada kategori baik dan sangat baik, tetapi masih terdapat sejumlah butir yang memerlukan revisi. Perbedaan hasil ini memperlihatkan bahwa kualitas daya pembeda dapat bervariasi sesuai karakteristik instrumen dan kelompok peserta. Oleh karena itu, butir dengan daya pembeda negatif sebaiknya tidak langsung digunakan kembali sebelum kunci jawaban dan konstruksi soal diperiksa. Butir dengan daya pembeda rendah perlu direvisi, sedangkan butir dengan daya pembeda sedang atau baik dapat dipertahankan setelah ditinjau bersama hasil validitas dan efektivitas pengecohnya.

Efektivitas Pengecoh

Hasil analisis efektivitas pengecoh menunjukkan bahwa butir nomor 3, 8, dan 9 dilaporkan memiliki pengecoh yang efektif. Secara keseluruhan, hasil tersebut menunjukkan bahwa fungsi pengecoh masih bervariasi antarbutir. Beberapa alternatif jawaban belum memenuhi batas minimal pemilihan yang digunakan dalam penelitian, yaitu 5% dari jumlah peserta. Pada kelompok peserta sebanyak 36 orang, batas tersebut setara dengan sekitar dua peserta, sehingga pengecoh yang dipilih oleh nol atau satu peserta perlu diperiksa kembali berdasarkan kriteria operasional yang digunakan. Secara teoretis, pengecoh yang baik harus masuk akal bagi peserta yang belum menguasai materi, tetapi tetap dapat dibedakan dari jawaban benar oleh peserta yang memahami konsep. Pengecoh yang tidak pernah dipilih dapat menunjukkan bahwa alternatif jawaban terlalu jelas salah, tidak homogen dengan kunci jawaban, atau tidak relevan dengan konteks pertanyaan. Sebaliknya, pengecoh yang terlalu menyerupai jawaban benar dapat menimbulkan ambiguitas dan berpotensi menurunkan validitas butir.

Menariknya, butir nomor 8 dilaporkan memiliki pengecoh yang efektif, tetapi daya pembeda dan validitas empirisnya relatif rendah. Hal ini menunjukkan bahwa keberfungsian pengecoh saja tidak cukup untuk menjamin kualitas suatu butir. Semua alternatif jawaban dapat menarik respons peserta, tetapi butir tersebut tetap mungkin tidak membedakan peserta berdasarkan kemampuan atau tidak menunjukkan hubungan yang memadai dengan skor total. Oleh karena itu, analisis pengecoh harus ditafsirkan bersama indeks kesukaran, daya pembeda, dan validitas butir. Penelitian Pidekso dkk. (2025) juga menemukan variasi efektivitas pengecoh pada instrumen evaluasi komputer akuntansi, dengan beberapa butir yang memerlukan perbaikan pada alternatif jawaban. Hasil tersebut mendukung pentingnya penelaahan pengecoh sebagai bagian dari evaluasi kualitas tes. Dalam konteks pelatihan Teknik Sepeda Motor, revisi pengecoh perlu dilakukan dengan memastikan bahwa pilihan jawaban berasal dari kesalahan pemahaman yang mungkin terjadi saat peserta mempelajari prosedur perawatan, bukan sekadar alternatif yang tidak masuk akal. Dengan demikian, pengecoh dapat memberikan informasi yang lebih bermakna mengenai pemahaman peserta.

Validitas Butir Soal

Hasil uji validitas menggunakan korelasi Product Moment Pearson menunjukkan bahwa 13 dari 20 butir soal (65%) dinyatakan valid, sedangkan 7 butir (35%) dinyatakan tidak valid berdasarkan kriteria korelasi yang digunakan. Butir yang tidak memenuhi kriteria tersebut adalah nomor 4, 5, 6, 7, 8, 13, dan 18. Temuan ini menunjukkan bahwa sebagian besar butir memiliki

hubungan yang memadai dengan skor total menurut batas keputusan penelitian, tetapi masih terdapat sejumlah butir yang memerlukan evaluasi lebih lanjut.

Validitas empiris butir dalam analisis korelasional menunjukkan sejauh mana skor pada suatu butir berhubungan dengan skor tes secara keseluruhan. Korelasi yang rendah dapat terjadi apabila butir tidak mengukur aspek kompetensi yang sama dengan sebagian besar tes, memiliki kunci jawaban yang bermasalah, atau tidak mampu membedakan tingkat penguasaan peserta. Hasil ini perlu ditafsirkan secara hati-hati karena korelasi butir dengan skor total juga dipengaruhi oleh jumlah peserta, variasi skor, dan karakteristik butir. Selain itu, apabila skor total yang digunakan masih mencakup skor butir yang sedang diuji, korelasi butir-total terkoreksi dapat dipertimbangkan untuk menghindari inflasi korelasi akibat kontribusi butir itu sendiri terhadap skor total.

Keterkaitan antara validitas dan daya pembeda terlihat pada butir nomor 4, 6, 7, dan 18, yang dilaporkan memiliki daya pembeda rendah atau negatif sekaligus tidak memenuhi kriteria validitas. Pola ini mengindikasikan bahwa butir-butir tersebut perlu diprioritaskan dalam proses penelaahan. Butir nomor 8 juga perlu diperiksa karena nilai korelasinya sebesar 0,319 masih sedikit di bawah nilai pembanding 0,3291, sementara daya pembedanya tergolong rendah. Adapun butir nomor 5 dan 13 memerlukan pemeriksaan terhadap isi soal, kesesuaian dengan indikator kompetensi, dan karakteristik respons peserta.

Hasil ini dapat dibandingkan dengan penelitian Toifur dan Setyowati (2022) mengenai analisis butir soal simulasi ujian sekolah matematika, yang melaporkan bahwa 27 butir memenuhi kriteria validitas dan reliabilitas tes berada pada kategori tinggi dengan koefisien 0,76. Perbedaan proporsi validitas pada kedua penelitian tersebut memperlihatkan bahwa kelayakan butir tidak dapat diasumsikan sama pada setiap instrumen. Oleh sebab itu, hasil pengujian empiris perlu digunakan untuk menentukan butir yang dipertahankan, direvisi, atau diuji kembali. Dalam penelitian ini, status tidak valid berdasarkan korelasi empiris juga tidak serta-merta berarti butir harus dihapus; keputusan akhir perlu mempertimbangkan validitas isi, kepentingan indikator kompetensi, dan hasil penelaahan substantif (Syahriani, 2021).

Reliabilitas Instrumen

Koefisien korelasi antara belahan ganjil dan genap sebesar 0,630 menunjukkan hubungan positif antara kedua bagian tes. Apabila angka tersebut merupakan korelasi awal antarkedua belahan, koreksi Spearman–Brown menghasilkan koefisien reliabilitas keseluruhan sekitar 0,773. Nilai tersebut perlu diinterpretasikan menggunakan kriteria reliabilitas yang dinyatakan dalam metode penelitian. Dengan demikian, simpulan mengenai konsistensi instrumen tidak didasarkan pada perbandingan dengan nilai r tabel, tetapi pada koefisien reliabilitas dan kategori interpretasinya. Temuan tersebut perlu dibaca bersama hasil analisis butir. Meskipun reliabilitas keseluruhan dapat memenuhi kriteria yang ditetapkan, instrumen masih mengandung butir yang memiliki daya pembeda negatif, validitas empiris rendah, dan pengecoh yang tidak berfungsi. Hal ini menunjukkan bahwa konsistensi skor pada tingkat tes tidak menjamin bahwa setiap butir telah disusun secara optimal. Instrumen dapat memiliki konsistensi keseluruhan yang memadai sekaligus mengandung butir yang memerlukan perbaikan (Susetyo, 2020).

Penelitian Toifur dan Setyowati (2022) melaporkan koefisien reliabilitas sebesar 0,76 pada instrumen simulasi ujian matematika. Apabila nilai tersebut dibandingkan dengan koefisien Spearman–Brown terkoreksi dalam penelitian ini, keduanya menunjukkan nilai yang relatif berdekatan. Namun, perbandingan tersebut perlu mempertimbangkan perbedaan jumlah peserta, karakteristik materi, bentuk tes, dan prosedur penghitungan reliabilitas. Kesamaan nilai koefisien tidak berarti bahwa kedua instrumen memiliki kualitas yang identik. Dalam konteks pelatihan vokasi, reliabilitas yang memadai perlu disertai dengan pemeriksaan terhadap kesesuaian soal dengan standar kompetensi agar hasil evaluasi dapat digunakan secara lebih bertanggung jawab (Suharto, 2019).

Implikasi terhadap Kualitas Instrumen Evaluasi

Secara keseluruhan, hasil analisis menunjukkan bahwa instrumen evaluasi unit kompetensi Melakukan Perawatan Renggang Klep (G.45TSM01.020.2) memiliki kualitas yang bervariasi. Sebagian butir telah menunjukkan karakteristik yang cukup baik, tetapi beberapa butir masih memerlukan perbaikan berdasarkan tingkat kesukaran, daya pembeda, efektivitas pengecoh, dan validitas empiris. Temuan ini memperkuat pandangan bahwa kualitas tes tidak dapat ditentukan hanya berdasarkan satu indikator, melainkan melalui interpretasi terpadu atas berbagai karakteristik butir dan kesesuaiannya dengan tujuan pengukuran (Setyaedhi, 2023).

Implikasi praktis penelitian ini adalah perlunya proses penelaahan dan revisi soal sebelum instrumen digunakan kembali. Butir nomor 4 dan 6 perlu diperiksa terlebih dahulu karena memiliki daya pembeda negatif. Butir nomor 7, 8, dan 18 perlu ditelaah berdasarkan daya pembeda dan validitasnya, sedangkan butir nomor 5 dan 13 perlu diperiksa terutama berdasarkan hasil validitas empiris. Butir yang terlalu mudah juga perlu ditinjau untuk memastikan bahwa tingkat kesukarannya sesuai dengan tujuan pengukuran, bukan semata-mata diubah agar menjadi lebih sulit. Keputusan mempertahankan, merevisi, atau menghapus butir sebaiknya dilakukan dengan mempertimbangkan kisi-kisi, standar kompetensi, kunci jawaban, dan bukti empiris dari respons peserta. Dengan demikian, analisis butir soal tidak hanya berfungsi sebagai prosedur statistik untuk mengklasifikasikan soal, tetapi juga sebagai dasar evaluasi dan perbaikan instrumen secara berkelanjutan. Penggunaan instrumen yang telah ditelaah dan diperbaiki diharapkan dapat mendukung pengukuran penguasaan pengetahuan teknis yang lebih akurat serta memperkuat proses evaluasi dalam pelatihan berbasis kompetensi di UPT BLK Surabaya.

SIMPULAN

Berdasarkan hasil penelitian mengenai kualitas instrumen uji kompetensi Perawatan Renggang Klep pada pelatihan servis sepeda motor injeksi di UPT BLK Surabaya, dapat disimpulkan bahwa kualitas instrumen tes secara umum cukup baik, tetapi masih memerlukan perbaikan pada beberapa butir soal. Hasil analisis menunjukkan bahwa tingkat kesukaran soal bervariasi, daya pembeda sebagian besar berada pada kategori sedang, dan efektivitas pengecoh belum optimal pada sejumlah butir. Hasil uji validitas menunjukkan bahwa sebagian besar butir soal telah memenuhi kriteria validitas, sedangkan hasil analisis reliabilitas mengindikasikan konsistensi instrumen yang perlu ditafsirkan berdasarkan koefisien reliabilitas dan kriteria yang digunakan. Temuan ini menunjukkan bahwa kelayakan instrumen evaluasi tidak hanya ditentukan oleh validitas dan reliabilitas, tetapi juga oleh tingkat kesukaran, daya pembeda, dan efektivitas pengecoh setiap butir soal. Secara praktis, penelitian ini memberikan implikasi bagi instruktur dan penyusun soal untuk melakukan evaluasi serta perbaikan instrumen secara berkala agar pengukuran kompetensi peserta pelatihan lebih akurat. Butir soal yang berkualitas dapat dipertahankan, sedangkan butir yang memiliki daya pembeda negatif, tingkat kesukaran yang tidak proporsional, atau pengecoh yang tidak berfungsi perlu ditinjau dan direvisi dengan mengacu pada indikator pembelajaran serta Standar Kompetensi Kerja Nasional Indonesia (SKKNI).

DAFTAR RUJUKAN

A'yun, A. A. S. Q., dkk. (2022). Analisis kelayakan butir soal pada media INTERMATHLY (Interesting Mathematic Monopoly). *Jurnal Pendidikan Matematika, 6*(1), 643–654.

Agustiana, dkk. (2018). Analisis butir soal ulangan akhir semester mata pelajaran Bahasa Indonesia kelas XI. *Jurnal Ilmiah Bina Edukasi, 11*(1), 26–35.

Alfarisa, F., dkk. (2019). Analisis butir soal IPS kelas V sekolah dasar menggunakan software ITEMAN. *EduBasic Journal: Jurnal Pendidikan Dasar, 1*(1), 100–106.

Farida, & Musyarofah, A. (2021). Validitas dan reliabilitas dalam analisis butir soal. *Jurnal Pendidikan Bahasa Arab, 1*(1), 34–44.

Febri, R. I. F., dkk. (2021). Pengembangan sistem analisis butir soal ujian berbasis web (studi kasus: SMP Negeri 1 Donomulyo). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 5*(3), 943–949.

Fikriyah, N. (2021). Analisis butir soal ulangan tengah semester mata pelajaran Bahasa Arab kelas VII semester ganjil SMP Muhammadiyah 1 Yogyakarta tahun ajaran 2019/2020. *Maharaat: Jurnal Pendidikan Bahasa Arab, 3*(2).

Fitrianawati, M. (n.d.). Peran analisis butir soal guna meningkatkan kualitas butir soal, kompetensi instruktur dan hasil belajar peserta didik. Dalam *Seminar Nasional Pendidikan PGSD UMS & HDPGSDI Wilayah Jawa*.

Jayali, A. M., & Sriwahyuni, E. (2022). Analisis butir soal instrumen tes keterampilan proses sains pada topik sistem periodik unsur. *Journal on Teacher Education, 4*(2), 1581–1591.

Kurniawan, D. D. (2019). Analisis butir soal ujian akhir semester matematika dengan teori respons butir. *Jurnal Riset dan Konseptual, 4*(2).

Lisa, Y. (2021). Analisis butir soal mata kuliah algoritma pemrograman. *JUTECH, 2*(1).

Magdalena, dkk. (2021). Analisis validitas, reliabilitas, tingkat kesulitan dan daya beda butir soal ujian akhir semester tema 7 kelas III SDN Karet 1 Sepatan. *Jurnal Pendidikan dan Sains, 3*(2), 198–214.

Mania, dkk. (2020). Analisis butir soal ujian akhir sekolah. *Al Asma: Journal of Islamic Education, 2*(2).

Pasi, S. N. P., & Yusrizal. (2018). Analisis butir soal ujian Bahasa Indonesia buatan instruktur MTsN di Kabupaten Aceh Besar. *Master Bahasa, 6*(2), 195–202.

Prawiki, S. M., & Helendra. (2022). Analisis kualitas butir soal ujian akhir semester ganjil tahun pelajaran 2020/2021 mata pelajaran biologi kelas X SMA Negeri 1 Teluk Sebong. *Jurnal Biologi dan Pembelajarannya, 17*(2).

Purniasari, dkk. (2021). Analisis butir soal ujian sekolah mata pelajaran kimia SMA N 1 Kutowinangun tahun pelajaran 2019/2020 menggunakan model ITEMAN dan Rasch. *Jurnal Pendidikan Kimia, 10*(2).

Qomariyah, dkk. (2022). Analisis tingkat kesukaran dan daya pembeda pada butir soal pilihan ganda mata pelajaran Bahasa Indonesia kelas V semester 1 SDN Kedungdalem 2. *Jurnal Pendidikan, Sains dan Teknologi, 1*(2), 4–80.

Ramadhan, W. (2023). Analisis butir soal tipe multiple choice questions pada penilaian harian sekolah dasar. *Tarbiyah wa Ta'lim: Jurnal Penelitian Pendidikan dan Pembelajaran, 10*(2).

Sanusi, R. N. A., & Aziez, F. (2021). Analisis butir soal tes objektif dan subjektif untuk keterampilan membaca pemahaman pada kelas VII SMP N 3 Kalibagor. *Jurnal Pembelajaran Bahasa dan Sastra, 8*(1), 99–109.

Sari, A. S. D. (2019). Analisis butir soal kognitif taksonomi Bloom revisi pilihan ganda dan uraian pada materi momentum dan impuls. *Lintang Songo: Jurnal Pendidikan, 2*(1).

Sari, V., Utomo, A., & Sumarwati, S. (2022). Kualitas soal Bahasa Indonesia di SMP Muhammadiyah 1 Pontianak: Analisis butir soal. *Jurnal Pendidikan Bahasa dan Sastra Indonesia, 11*(2), 112–119.

Setyaedhi, H. S. (2023). *Pengembangan instrumen tes hasil belajar (bentuk tes obyektif)*. Deepublish.

Suharto, Y. (2019). Analisis butir soal evaluasi peserta diklat prajabatan CPNS K1/K2 golongan III Provinsi DKI Jakarta tahun 2018. *Jurnal Inovasi Aparatur, 1*(1), 1–13.

Susetyo, A. M. (2020). Analisis butir soal ujian semester kelas VIII mata pelajaran Bahasa Indonesia. *Belajar Bahasa, 5*(2), 187–198.

Syahriani, S., dkk. (2021). Analisis butir soal ujian sekolah mata pelajaran Bahasa Indonesia kelas VI SDN Kota Bengkulu. *JP3D: Jurnal Pembelajaran dan Pengajaran Pendidikan Dasar, 4*(2).