

REVISI INSTRUMEN EVALUASI BERBASIS ANATES UNTUK MENINGKATKAN KUALITAS PEMBELAJARAN K3LH DI SMK JURUSAN AKUNTANSI

Sri Indah Kurnia Sari^{1*}, Luqman Hakim², Vivi Pratiwi³, Cinta Nazwa Sofia⁴, Sonia Handayani⁵, Miftakhul Jannah⁶, Zulfa Khoiroturrosidah⁷, Zahrotun Nisa'i Sholikha⁸, Sherly Endah Kurniawati⁹

^{1,2,3,4,5,6,7,8,9} Universitas Negeri Surabaya, Indonesia

*e-mail korespondensi: 24080304056@mhs.unesa.ac.id

Abstract: The accuracy of measuring student competence in vocational education is highly dependent on the quality of the evaluation instruments used. This study aims to deeply evaluate the psychometric quality of test items in Health, Safety, and Environment (HSE) subjects through Anates software to diagnose instrument weaknesses. This evaluation serves as a strategic foundation for efforts to improve the quality of learning in Vocational High Schools (SMK) especially in accounting major. Using a quantitative descriptive method with 20 respondents, various parameters were tested, including validity, reliability, difficulty level, discriminating power, and distractor effectiveness. The research findings indicate that although the instrument has high reliability ($r = 0,92$), there are serious constraints regarding validity, with 43% of the test items found to be invalid. Additionally, 57% of the distractors proved to function ineffectively, causing the questions to fail in optimally distinguishing student ability levels. This empirical data is used as a basis for improving problematic test items. The study recommends systematic improvements in item construction to ensure more accurate and high-quality HSE instruments for vocational students.

Keywords: Item Analysis, HSE, Anates Software, Instrument Revision, Learning Quality

Abstrak: Akurasi pengukuran kompetensi siswa dalam pendidikan vokasi sangat bergantung pada kualitas instrumen evaluasi yang digunakan. Penelitian ini bertujuan mengevaluasi secara mendalam kualitas psikometrik butir soal mata pelajaran Keselamatan, Kesehatan Kerja, dan Lingkungan Hidup (K3LH) melalui aplikasi Anates untuk mendiagnosis kelemahan instrumen. Evaluasi ini menjadi landasan strategis dalam upaya meningkatkan kualitas pembelajaran di Sekolah Menengah Kejuruan (SMK) khususnya dalam jurusan akuntansi. Dengan metode deskriptif kuantitatif terhadap 20 responden, dilakukan pengujian parameter validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh. Temuan riset menunjukkan bahwa meski instrumen memiliki reliabilitas tinggi ($r = 0,92$), terdapat kendala serius pada aspek validitas dengan temuan 43% butir soal tidak valid. Selain itu, 57% pengecoh terbukti tidak berfungsi efektif, sehingga soal gagal membedakan tingkat kemampuan siswa secara optimal. Data empiris ini digunakan sebagai basis revisi terhadap butir soal yang bermasalah. Penelitian merekomendasikan perbaikan konstruksi soal secara sistematis demi menjamin instrumen K3LH yang lebih akurat dan berkualitas bagi siswa vokas.

Kata Kunci: Analisis Butir Soal, K3LH, Aplikasi Anates, Revisi Instrumen, Kualitas Pembelajaran.

Copyright (c) 2025 The Authors. This is an open-access article under the CC BY-SA 4.0 license (<https://creativecommons.org/licenses/by-sa/4.0/>)

PENDAHULUAN

Evaluasi pembelajaran adalah suatu rangkaian kegiatan penting untuk mengetahui ketercapaian tujuan pendidikan melalui instrumen evaluasi yang terukur dan sistematis (Abdillah & Hikmah, 2024). Tujuan evaluasi pendidikan adalah mengumpulkan informasi sebagai bukti perkembangan siswa setelah mengikuti pembelajaran serta menilai efektivitas pendekatan pengajaran yang digunakan dalam periode tertentu (Tampubolon et al., 2023). Evaluasi dapat digunakan pendidik sebagai landasan untuk peningkatan mutu serta penguatan layanan pendidikan dalam penetapan nilai akhir (Maulana et al., 2024). Proses evaluasi tersebut dapat berupa tes atau ujian yang dilaksanakan di awal, tengah atau akhir proses pembelajaran di dalam suatu semester (Phafiandita et al., 2022). Agar proses evaluasi ini memberikan data yang valid, diperlukan alat ukur atau instrumen yang memiliki standar kualitas yang tinggi. Terdapat dua prinsip dasar yang harus diperhatikan dalam pembuatan soal evaluasi, yaitu harus dapat mengukur apa yang semestinya diukur dan akurat dalam membuat keputusan tentang siswa (W. Ramadhan et al., 2023).

Salah satu bentuk instrumen yang paling sering digunakan dalam dunia pendidikan adalah tes objektif pilihan ganda (Putri et al., 2022). Suatu butir soal dikatakan bermutu apabila dapat mengidentifikasi perbedaan kemampuan antar siswa, serta memenuhi kriteria kevalidan dan keandalan (Augustia et al., 2025). Kualitas butir soal menjadi faktor utama dalam menghasilkan penilaian yang objektif dan akurat melalui analisis butir soal. Analisis butir soal adalah proses pengevaluasian dan pengukuran karakteristik suatu tes atau ujian (Pudjiati & Madani, 2023). Tanpa analisis butir soal yang memadai, instrumen evaluasi berpotensi tidak menginterpretasikan kemampuan peserta didik secara sebenarnya. Hal ini menjadi krusial ketika instrumen tersebut digunakan untuk mengukur kompetensi yang berkaitan dengan keselamatan jiwa di lingkungan kerja.

Dalam industri modern, penerapan Keselamatan dan Kesehatan Kerja serta Lingkungan Hidup (K3LH) menjadi komponen utama untuk memastikan lingkungan kerja yang aman dan sehat (Yudhistiro & Rahayu, 2025). Sejalan dengan itu, pada pembelajaran K3LH bidang vokasi akuntansi, keberadaan instrumen evaluasi yang berkualitas sangat penting karena berhubungan langsung dengan kesiapan peserta didik dalam menerapkan prosedur keselamatan. Pembelajaran K3LH tidak hanya menekankan aspek teoretis, tetapi juga aspek praktis yang menuntut peserta didik memahami potensi bahaya dan tindakan preventif (Rahmawati et al., 2024b). Meskipun akuntansi dianggap bidang administratif, risiko seperti cedera ergonomi dan gangguan kelistrikan tetap memerlukan penilaian kompetensi yang akurat (Rahmawati et al., 2024a). Kompleksitas materi ini menuntut adanya alat bantu analisis yang lebih canggih guna memastikan kualitas soal yang disusun guru.

Kemajuan teknologi saat ini telah memfasilitasi penilaian otomatis yang mempermudah guru dari tugas administratif melalui perangkat lunak seperti Anates (Hardiansyah et al., 2025). Penggunaan perangkat lunak Anates memberikan kemudahan bagi pendidik karena mampu

menganalisis validitas, reliabilitas, indeks kesukaran, dan daya pembeda secara otomatis (Iriyanti et al., 2024). Anates juga menyediakan laporan hasil analisis yang mudah dipahami sehingga pendidik dapat segera mengambil keputusan mengenai perbaikan butir soal. Hasil pengujian menunjukkan bahwa Anates mampu mengidentifikasi kualitas soal secara lebih akurat dibandingkan perhitungan manual (Purwanti et al., 2021). Dengan kemudahan yang ditawarkan, seharusnya kualitas instrumen K3LH di SMK sudah teruji dengan baik, namun kenyataannya masih ditemukan celah dalam praktiknya.

Meskipun penelitian tentang analisis butir soal K3LH sudah pernah dilakukan, fokus utama sebelumnya sering kali terbatas pada tahap pemetaan kualitas tanpa disertai tindak lanjut revisi yang mendalam. Perbedaan signifikan pada penelitian ini terletak pada integrasi antara diagnosa empiris dengan langkah perbaikan konkret, khususnya pada materi K3LH bagi siswa SMK jurusan akuntansi. Celah ini perlu diperhatikan mengingat instrumen evaluasi yang tidak direvisi secara berkala dapat menyebabkan bias dalam pengukuran kompetensi keselamatan kerja siswa. Melalui pendekatan ini, penelitian tidak hanya berhenti pada tahap identifikasi masalah, melainkan berfokus pada pemberian solusi perbaikan instrumen secara terstruktur. Hal tersebut penting untuk memastikan bahwa setiap butir soal mampu berfungsi sebagai alat ukur yang valid dalam mendukung standar kompetensi di lingkungan pendidikan vokasi.

Penelitian ini berupaya memperdalam kajian yang telah ada dengan melakukan analisis komprehensif serta langkah revisi butir soal K3LH melalui bantuan aplikasi Anates. Analisis yang dilakukan mencakup parameter validitas, reliabilitas, tingkat kesukaran, daya pembeda dan distraktor untuk menghasilkan instrumen yang benar-benar akurat. Hasil analisis yang diperoleh kemudian digunakan sebagai landasan untuk mengidentifikasi faktor penyebab ketidaklayakan soal serta merumuskan solusi perbaikan yang bersifat solutif. Selain itu, hasil studi ini menawarkan model pengembangan evaluasi bagi pendidik vokasi dalam menyusun instrumen evaluasi yang berkualitas dan dapat dipertanggungjawabkan. Sebagai kesimpulan, revisi instrumen berbasis Anates diharapkan mampu meningkatkan kualitas pembelajaran K3LH di SMK secara signifikan.

METODE

Penelitian ini merupakan penelitian deskriptif kuantitatif untuk mencari informasi dan data yang dapat digunakan untuk mendeskripsikan kualitas soal elemen Keselamatan, Kesehatan Kerja, dan Lingkungan Hidup (K3LH) (Fitriani, 2021). Fokus utama penelitian adalah menghasilkan data empiris mengenai karakteristik butir soal. Butir soal dianalisis secara kuantitatif untuk informasi tentang mengetahui daya pembeda, validitas, reliabilitas, tingkat kesukaran, dan efektifitas distraktor (Akhmadi, 2021). Sedangkan, penggunaan metode deskriptif digunakan untuk memetakan kualitas instrumen secara detail guna menentukan tindakan revisi yang diperlukan sebagai upaya menjamin akurasi evaluasi hasil belajar.

Data penelitian dikumpulkan dengan memanfaatkan kuesioner daring berbasis Google Form yang didistribusikan kepada responden yang telah ditentukan (Syafiriya et al., 2024). Responden penelitian terdiri dari 20 peserta yang mencakup siswa kelas XI SMK jurusan akuntansi dan mahasiswa baru dari rumpun ilmu akuntansi dan administrasi. Pemilihan responden yang mencakup mahasiswa baru dimaksudkan untuk menciptakan heterogenitas kemampuan yang ekstrem guna menguji indeks daya pembeda soal secara akurat. Penggabungan mahasiswa baru dianggap sah dan relevan karena kurikulum pendidikan tinggi di bidang tersebut memuat materi serupa pada mata kuliah K3LH di semester awal. Perbedaan strata pendidikan ini memastikan bahwa butir soal benar-benar mampu membedakan responden yang memiliki penguasaan konsep mendalam dengan responden yang baru pada tahap pengenalan awal, sehingga stabilitas instrumen dapat teruji saat dihadapkan pada variasi kognitif yang luas.

Data jawaban responden diinput ke dalam aplikasi Anates untuk dievaluasi berdasarkan parameter validitas, reliabilitas, tingkat kesukaran, dan fungsi distraktor secara otomatis. Instrumen yang diuji terdiri dari 63 butir pilihan ganda yang dirancang untuk mengukur kompetensi siswa pada elemen K3LH melalui analisis dan evaluasi. Hasil analisis dari Anates dikonversi menjadi persentase untuk mengidentifikasi butir soal yang masuk dalam kategori layak, perlu direvisi, atau harus diganti (Cahyaningrum et al., 2023). Prosedur penelitian sendiri meliputi tiga tahap teknis yaitu pelaksanaan tes kepada responden, pengolahan data digital menggunakan Anates, dan interpretasi hasil analisis. Hasil analisis tersebut selanjutnya dijadikan sebagai dasar rekomendasi perbaikan struktur kalimat, pengecoh, maupun substansi soal agar sesuai dengan tujuan pembelajaran K3LH dalam meningkatkan kualitas evaluasi hasil belajar di SMK.

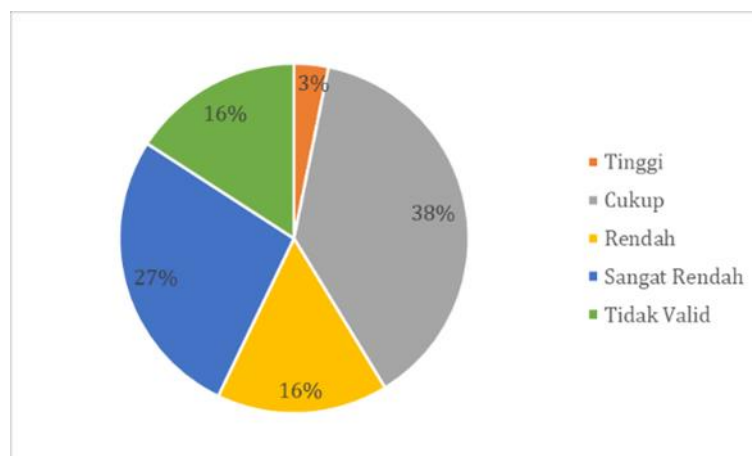
HASIL DAN PEMBAHASAN

Penelitian ini menghasilkan analisis dan pemaknaan kualitas butir soal pada materi K3LH di jenjang SMK terkhusus jurusan akuntansi dengan menggunakan perangkat lunak Anates. Evaluasi kualitas instrumen dilakukan melalui pengujian karakteristik butir soal, meliputi tingkat kesukaran, daya pembeda, reliabilitas, serta efektivitas pengecoh. Uraian berikut menyajikan hasil analisis yang diperoleh pada setiap tahap yang dilakukan dalam penelitian ini.

Hasil

Uji Validitas

Pengujian validitas dilakukan guna memastikan instrumen yang diujikan valid atau tidak. Analisis validitas sebuah instrumen dapat diketahui melalui aplikasi Anates. Dari data yang dianalisis didapatkan hasil berikut.

**Gambar 1.** Grafik Persentase Validitas

Sumber: data diolah (2025)

Hasil pengujian melalui aplikasi Anates menunjukkan bahwa sebagian besar instrumen K3LH belum mencapai standar kelayakan psikometrik yang diharapkan. Grafik menunjukkan bahwa 43% atau sebanyak 27 butir soal masuk dalam klasifikasi tingkat validitas yang rendah hingga tidak valid. Sebaliknya, proporsi butir soal yang memiliki daya ukur akurat atau berkategori tinggi ditemukan sangat sedikit. Secara keseluruhan, instrumen ini didominasi oleh butir soal yang memiliki korelasi lemah hingga negatif terhadap skor total. Kondisi tersebut mengindikasikan bahwa instrumen tersebut belum dapat berfungsi secara optimal sebagai alat ukur kompetensi siswa dan memerlukan perbaikan sebelum diterapkan.

Tabel 1. Koefisien Korelasi dan Kategori Validitas

Koefisien Korelasi	Kategori Validitas	Keterangan
0,800 - 1,00	Sangat Tinggi	Memiliki daya ukur sangat baik.
0,600 - 0,800	Tinggi	Memiliki daya ukur yang baik.
0,400 - 0,600	Cukup	Masih dapat dipertahankan.
0,200 - 0,400	Rendah	Harus direvisi atau dibuang.
0,00 - 0,200	Sangat Rendah	Dibuang dan diperiksa kembali.

Uji Reliabilitas

Aspek reliabilitas diuji untuk mengukur sejauh mana instrumen ini dapat memberikan hasil yang stabil dan konsisten. Melalui pengolahan data menggunakan aplikasi Anates, tingkat kepercayaan perangkat tes ini dapat dipetakan berdasarkan respons siswa secara empiris. Adapun rangkuman hasil analisisnya adalah sebagai berikut.

RELIABILITAS TES

Rata2= 42,10
Simpang Baku= 7,30
KorelasiXY= 0,85
Reliabilitas Tes= 0,92
Nama berkas: C:\USERS\ASUS\DOWNLOADS\DATA K3LH 63 SOAL.ANA

No.Urut	No. Subyek	Kode>Nama Subyek	Skor Ganjil	Skor Genap	Skor Total
1	1	Melati Karuni...	25	25	50
2	2	Aurellia Pram...	28	26	54
3	3	Nabilah Naafi...	24	23	47
4	4	Natasya Mutia...	28	27	55
5	5	Nazwa Salsa S...	21	21	42
6	6	Lidya Novita ...	19	18	37
7	7	Maulita Erina...	20	22	42
8	8	Vanesha Nuran...	27	21	48
9	9	Neza Alexandra	24	24	48
10	10	Julia Andita ...	19	22	41
11	11	Widia Isnaini...	22	20	42
12	12	Fitri Tri Zul...	23	20	43
13	13	Wulan Dwi Kar...	22	18	40
14	14	Siska Meilani	18	18	36
15	15	Muhammad Isma...	16	18	34
16	16	Firda Ariani	25	22	47
17	17	Lutfiah Hani	21	18	39
18	18	Zulfa Afina P...	20	20	40
19	19	Aulia Gita	14	13	27
20	20	M. Mahdy Ariel	15	15	30

Gambar 2. Hasil Reliabilitas Tes

Sumber: data diolah (2025)

Berdasarkan hasil analisis menggunakan perangkat lunak Anates, instrumen K3LH yang diujikan menunjukkan koefisien yang sangat tinggi, yaitu sebesar 0,92. Artinya, Seluruh butir soal dalam perangkat tes ini memiliki stabilitas yang sangat baik, sehingga mampu menghasilkan skor yang konsisten meskipun digunakan dalam waktu yang berbeda. Tingkat kepercayaan yang dihasilkan menempatkan instrumen ini pada kategori reliabel yang sangat kuat untuk mengukur kompetensi siswa di elemen K3LH. Dengan demikian, secara statistik instrumen ini telah memenuhi syarat keandalan untuk digunakan dalam penilaian siswa di tingkat SMK. Namun, konsistensi hasil ini tetap perlu diselaraskan dengan kualitas parameter butir soal lainnya agar evaluasi yang dihasilkan benar-benar akurat.

Tabel 2. Kategori Reliabilitas Soal

No.	Nilai Koefisien	Kategori Reliabilitas Soal
1	0,000 sampai 0,199	Sangat Rendah
2	0,200 sampai 0,399	Rendah
3	0,400 sampai 0,599	Cukup
4	0,600 sampai 0,7999	Tinggi
5	0,800 sampai 1,00	Sangat Tinggi

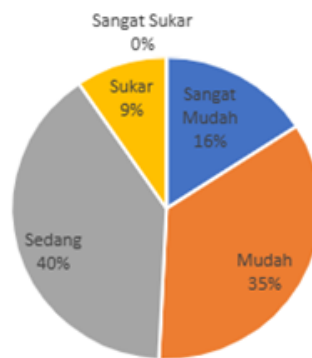
Tingkat Kesukaran

Kualitas setiap butir soal dalam penelitian ini diukur melalui indeks kesukaran yang direpresentasikan oleh nilai P (Proporsi) (Qadir et al., 2024). Nilai P menunjukkan rasio antara jumlah responden yang menjawab benar dengan total seluruh peserta tes, di mana semakin mendekati angka 1,00 atau 100%, maka soal tersebut dianggap semakin mudah karena mampu dijawab benar oleh mayoritas peserta. Sebaliknya, nilai P yang semakin mendekati 0,00 menandakan soal tersebut memiliki tingkat kesulitan yang tinggi.

Tabel 3. Kriteria tingkat kesukaran butir soal

No.	Besarnya P	Interpretasi
1	0% sampai 15%	Sangat Sukar
2	16% sampai 30%	Sukar
3	31% sampai 70%	Sedang
4	71% sampai 85%	Mudah
5	86% sampai 100%	Sangat Mudah

Berdasarkan data yang diolah melalui aplikasi Anates, distribusi tingkat kesukaran pada instrumen K3LH ini menunjukkan pola yang cukup bervariasi. Sebagian besar butir soal berada pada klasifikasi sedang, namun proporsi soal yang masuk dalam kategori mudah dan sangat mudah masih ditemukan dalam jumlah yang cukup dominan. Di sisi lain, kelompok soal dengan kategori sukar memiliki jumlah yang paling sedikit, sementara untuk kategori sangat sukar sama sekali tidak ditemukan dalam tes ini. Dari hasil tersebut dapat disimpulkan bahwa secara keseluruhan, instrumen cenderung lebih mudah diselesaikan oleh siswa dan memerlukan penyesuaian komposisi agar dapat memetakan kemampuan kognitif siswa secara lebih mendalam.



Gambar 3. Grafik hasil tingkat kesukaran

Sumber: data diolah (2025)

Daya Pembeda

Analisis daya pembeda dilakukan guna mengidentifikasi kapasitas diskriminasi setiap butir soal dalam memisahkan responden berdasarkan tingkat kompetensinya. Berdasarkan hasil olah data menggunakan aplikasi Anates, kualitas daya pembeda pada instrumen K3LH ini menunjukkan distribusi yang tidak ideal.

Tabel 5. Presentase daya pembeda

Kategori	Jumlah	Persentase	Nomor Butir Soal
Buruk Sekali	6	9%	20, 38, 41, 42, 43, 50
Buruk	26	41%	2, 3, 7, 10, 12, 15, 18, 19, 21, 22, 24, 25, 29, 37, 40, 43, 45, 46, 52, 54, 55, 57, 58, 59, 60, 61
Cukup	15	24%	4, 13, 14, 16, 23, 27, 31, 34, 35, 39, 47, 49, 51, 62, 63
Baik	12	19%	1, 5, 6, 8, 11, 17, 26, 32, 33, 48, 53, 56,
Baik Sekali	4	7%	9, 28, 30, 36

Temuan empiris menunjukkan bahwa instrumen didominasi oleh butir soal dengan indeks diskriminasi yang buruk hingga buruk sekali. Sebaliknya, butir soal yang memiliki kategori baik hingga baik sekali hanya ditemukan pada sebagian kecil instrumen tes. Kondisi ini mengindikasikan bahwa secara statistik, mayoritas butir soal tidak memiliki sensitivitas yang cukup untuk mengklasifikasikan kelompok peserta tes ke dalam strata kemampuan yang berbeda. Hasil ini menjadi dasar objektif bahwa instrumen memerlukan rekonstruksi struktural agar fungsinya sebagai alat ukur yang diskriminatif dapat terpenuhi..

Kualitas Pengecoh

Analisis efektivitas pengecoh dilakukan untuk memverifikasi kualitas opsi jawaban salah dalam struktur tes pilihan ganda. Berdasarkan pola sebaran jawaban peserta, didapatkan hasil analisis data mengenai proporsi distraktor yang efektif (fungsional) dan yang tidak efektif (non-fungsional) pada tabel di bawah ini.

Tabel 5. Persentase kualitas pengecoh

Kategori	Jumlah	Persentase	Nomor Butir Soal
Berfungsi	27	43%	1, 2, 3, 4, 6, 7, 11, 15, 17, 18, 19, 20, 21, 27, 28, 29, 35, 36, 45, 47, 48, 49, 51, 52, 53, 54, 63
Tidak Berfungsi	36	57%	5, 8, 9, 10, 12, 13, 14, 16, 22, 23, 24, 25, 26, 30, 31, 32, 33, 34, 37, 38, 39, 40, 41, 42, 43, 44, 46, 50, 55, 56, 57, 58, 59, 60, 61, 62

Dari hasil analisis melalui aplikasi Anates, pola penyebaran kualitas pengecoh pada instrumen K3LH ini menunjukkan kinerja yang belum optimal. Hasil pemetaan memperlihatkan bahwa proporsi butir soal dengan distraktor yang tidak berfungsi (non-fungsional) jauh lebih dominan dibandingkan dengan butir soal yang memiliki distraktor berfungsi efektif. Kondisi ini menunjukkan adanya kelemahan struktural pada pilihan jawaban, di mana mayoritas opsi pengecoh gagal untuk menarik perhatian peserta tes yang tidak menguasai materi.

Pembahasan

Validitas

Validitas atau *validity* merupakan derajat ketepatan dan signifikansi suatu instrumen pengukur dalam melakukan fungsinya sesuai tujuan yang diinginkan (M. F. Ramadhan et al., 2024). Validitas erat kaitannya terhadap sejauh mana instrumen penilaian benar-benar mengukur konsep yang dimaksud. Sebagai syarat mutlak, sebuah butir soal harus memiliki validitas tinggi agar mampu mengukur kompetensi secara akurat sesuai dengan tujuan evaluasi (Rahma et al., 2025). Sebaliknya butir soal yang memiliki validitas rendah harus direvisi kembali agar dapat menjadi alat ukur yang akurat. Rendahnya tingkat validitas pada instrumen ini kemungkinan besar disebabkan oleh konstruksi kalimat soal yang kurang spesifik atau ketidakselarasan antara indikator kompetensi dengan stimulus soal. Selain itu, faktor penggunaan kata dalam soal yang kurang tepat dapat memicu misinterpretasi di tingkat responden.

Dampak dari rendahnya validitas instrumen ini sangat krusial karena validitas tes harus dipertautkan dengan arah tujuan penilaian maupun keputusan yang hendak ditetapkan (M. F. Ramadhan et al., 2024). Jika validitas tes tidak dikaitkan dengan arah tujuan penilaian, maka keputusan atau kesimpulan mengenai kompetensi siswa akan menjadi bias dan tidak akuntabel. Penggunaan instrumen yang tidak valid akan merugikan siswa karena hasil yang diperoleh tidak mencerminkan kemampuan kognitif mereka yang sesungguhnya. Tanpa adanya pengujian validitas, pendidik tidak memiliki dasar yang kuat untuk menentukan apakah seorang siswa benar-benar telah menguasai materi K3LH atau tidak.

Urgensi perbaikan melalui validasi butir soal bertujuan untuk memberikan landasan teknis dalam melakukan revisi menyeluruh terhadap struktur instrumen. Validasi butir soal memberikan manfaat besar dalam menilai mutu tes sebelum diterapkan serta memudahkan perbaikan jika ditemukan ketidaksesuaian (Dantes et al., 2021). Langkah revisi yang harus ditempuh adalah menyinkronkan kembali butir soal dengan tujuan pembelajaran K3LH serta memperbaiki diksi bahasa agar lebih lugas dan tidak ambigu. Melalui skema perbaikan berbasis data ini, pendidik tidak hanya memperbaiki kualitas alat evaluasi tetapi juga meningkatkan profesionalisme dalam menyusun butir soal yang baik di masa depan. Dengan demikian, proses revisi instrumen merupakan penting dari siklus peningkatan kualitas pembelajaran di lingkungan SMK.

Reliabilitas

Reliabilitas atau *reliability* merupakan karakteristik instrumen yang menunjukkan sejauh mana suatu tes dapat dipercaya dan memberikan hasil yang konsisten, selama aspek yang diukur pada diri subjek tidak mengalami perubahan (Farida & Musyarofah, 2021). Dalam penelitian ini, angka reliabilitas yang didapatkan masuk dalam kategori sangat tinggi. Hal ini kemungkinan besar dipengaruhi oleh jumlah butir soal yang cukup banyak, yaitu 63 soal. Secara teori, semakin banyak

jumlah soal dan peserta tes, maka tingkat reliabilitasnya akan cenderung meningkat karena risiko kesalahan pengukuran menjadi lebih kecil (Sabela et al., 2025). Tingginya reliabilitas ini menunjukkan bahwa pola jawaban siswa dalam materi K3LH sudah bersifat tetap atau tidak berubah-ubah.

Namun demikian, reliabilitas yang tinggi tidak dapat menutupi kelemahan validitas. Instrumen tetap tidak layak digunakan apabila butir-butirnya tidak mengukur kompetensi yang tepat, sehingga revisi tetap diperlukan pada sebagian besar soal. Secara empiris, reliabilitas ditunjukkan oleh nilai koefisien reliabilitas dengan rentang nilai dari 0 sampai 1, di mana semakin tinggi nilai yang diperoleh, semakin tinggi juga konsistensi hasil pengukurannya (Farida & Musyarofah, 2021). Inilah alasan mengapa sebuah instrumen bisa memiliki skor reliabilitas yang sangat bagus, tetapi parameter lainnya seperti validitas atau daya pembeda justru rendah. Jika soalnya kurang relevan dengan materi pelajaran tetapi jawaban siswa tetap konsisten, maka nilai reliabilitasnya akan tetap terlihat tinggi.

Kondisi ini berdampak pada munculnya risiko salah ukur yang konsisten. Jika guru menggunakan soal yang reliabel tetapi tidak valid, guru akan terus-menerus mendapatkan gambaran yang salah tentang kemampuan siswa. Data yang dihasilkan bisa terlihat meyakinkan karena stabil, padahal tidak mencerminkan penguasaan materi K3LH yang sebenarnya. Oleh karena itu, hasil reliabilitas yang tinggi ini bukan menjadi jaminan untuk kualitas soal. Perlu dilakukan revisi pada butir-butir soal yang bermasalah agar instrumen tidak hanya stabil secara statistik, tetapi juga benar-benar tepat dalam mengukur kompetensi siswa sesuai dengan tujuan pembelajaran.

Tingkat Kesukaran

Kualitas suatu butir soal dapat ditinjau dari tingkat kesulitan yang dimiliki oleh setiap butir soal (Pradita et al., 2023). Secara umum, soal yang berkualitas baik adalah soal yang memiliki tingkat kesukaran seimbang, relevan, dan representatif (Dia & Syah, 2022). Meskipun hasil analisis menunjukkan soal kategori sedang sudah cukup banyak, keberadaan soal berkategori sangat mudah berisiko menurunkan efektivitas tes dalam memetakan penguasaan materi siswa secara mendalam. Soal yang terlalu mudah dianggap tidak layak karena cenderung hanya menilai kemampuan mengingat tingkat rendah, sehingga tidak mampu memberikan stimulus bagi kemampuan peserta didik dalam menyelesaikan masalah (Nurhalimah et al., 2022).

Dampak dari ketidakseimbangan tingkat kesukaran ini adalah hasil penilaian yang kurang mampu menggambarkan kompetensi siswa yang sebenarnya. Jika jumlah siswa yang menjawab tepat terlalu tinggi (soal terlalu mudah), maka fungsi instrumen sebagai alat pemetaan kemampuan menjadi tidak optimal (Setiyawan & Wijayanti, 2020). Kualitas butir soal sangat mempengaruhi cara siswa menyelesaikannya, siswa mungkin tidak tertantang jika soal terlalu mudah, namun akan putus asa jika soal terlalu sulit (Nurhalimah et al., 2022). Ketidadaan soal kategori sangat sukar dalam

instrumen ini menandakan bahwa perangkat tes tersebut belum mampu mengukur batas maksimal kemampuan analisis dan evaluasi siswa pada materi K3LH.

Langkah revisinya adalah dengan melakukan kalibrasi ulang pada butir soal yang memiliki nilai P terlalu tinggi agar beban kognitifnya dapat ditingkatkan ke ranah HOTS. Soal tes yang berkualitas haruslah seimbang agar setiap keputusan pendidikan yang diambil memiliki landasan yang akurat (Dia & Syah, 2022). Perbaikan redaksi soal dan penambahan stimulus analisis diperlukan agar instrumen ini tidak sekadar menguji hafalan, melainkan benar-benar mengukur kesiapan siswa SMK dalam menghadapi standar kompetensi industri yang menuntut kemampuan analisis tingkat tinggi.

Daya Pembeda

Daya pembeda merupakan indikator psikometrik yang menunjukkan kemampuan sebuah butir soal dalam membedakan peserta didik yang berkemampuan tinggi (kelompok atas) dan rendah (kelompok bawah) (Pradita et al., 2023). Umumnya, peserta dengan kemampuan tinggi akan berhasil menjawab lebih banyak soal dengan benar, sedangkan peserta dengan kemampuan rendah cenderung tidak mampu memberikan jawaban yang tepat (Rahma et al., 2025). Namun, data penelitian ini menunjukkan daya pembeda yang buruk, di mana lebih dari separuh butir soal gagal membedakan kedua kelompok tersebut. Secara ilmiah, hal ini merupakan indikasi bahwa butir soal kehilangan daya ukurnya akibat ambiguitas konstruksi atau kurangnya kedalaman stimulus.

Tingginya persentase butir dalam kategori jelek berkaitan erat dengan homogenitas jawaban responden. Hal ini terjadi karena soal yang disusun cenderung memiliki tingkat kesukaran yang sangat rendah, sehingga tidak menuntut proses kognitif kompleks (HOTS) yang dapat menjadi pemisah antara siswa yang benar-benar kompeten dan siswa yang hanya menjawab berdasarkan tebakan. Selain itu, daya pembeda yang lemah sering kali berakar pada malfungsi distraktor, yaitu ketika pilihan jawaban salah tidak bekerja secara efektif. Akibatnya, skor yang dihasilkan tidak mencerminkan perbedaan kompetensi yang sebenarnya, melainkan tertutup oleh faktor kebetulan.

Urgensi evaluasi pada aspek daya pembeda ini krusial untuk menjamin validitas prediktif instrumen. Jika daya pembeda tetap rendah, maka hasil penilaian akan menghasilkan interpretasi yang bias, di mana guru tidak dapat melakukan diagnosis yang akurat mengenai siapa yang membutuhkan pengayaan atau remedial. Langkah revisi harus difokuskan pada peningkatan kualitas pengecoh dan penguatan konteks pertanyaan agar lebih spesifik dan menantang. Perbaikan ini sangat diperlukan agar instrumen K3LH tidak hanya berfungsi sebagai alat administratif, tetapi menjadi instrumen evaluasi yang mampu menunjukkan gradasi pemahaman siswa SMK secara nyata sesuai tuntutan kurikulum.

Kualitas Pengecoh

Efektivitas distraktor merupakan parameter vital yang menjamin bahwa pilihan jawaban

yang salah mampu mengalihkan perhatian dan menciptakan keraguan yang konstruktif bagi peserta tes (Augustia et al., 2025). Sebuah distraktor dikatakan berfungsi secara psikometrik apabila dipilih oleh minimal 5% dari total peserta tes (Radja et al., 2023). Batasan kriteria 5% ini merupakan standar literatur internasional untuk menentukan apakah sebuah opsi tergolong functioning atau non-functional distractor (Rezigalla et al., 2024). Namun, temuan dalam penelitian ini mengungkap fakta bahwa mayoritas butir soal memiliki distraktor yang non-fungsional, yang berarti opsi tersebut diabaikan sepenuhnya oleh peserta didik.

Dominasi distraktor non-fungsional ini merefleksikan adanya cacat konstruksi dalam penyusunan opsi jawaban. Secara ilmiah, hal ini terjadi karena pilihan jawaban salah dibuat terlalu mencolok perbedaannya dengan kunci jawaban, tidak logis, atau tidak homogen dari segi panjang kalimat dan substansi materi. Akibatnya, peserta didik dapat dengan mudah melakukan eliminasi opsi yang tidak relevan tanpa perlu memahami konsep intinya. Kondisi ini menyebabkan inflasi skor semu, di mana keberhasilan siswa menjawab benar bukan disebabkan oleh pemahaman materi, melainkan karena lemahnya kualitas pengecoh yang disajikan.

Implikasi dari kegagalan fungsi distraktor ini berkontribusi langsung terhadap rendahnya daya pembeda butir soal yang dibahas sebelumnya. Ketika distraktor tidak bekerja, peluang menebak benar (*guessing probability*) meningkat drastis, sehingga instrumen kehilangan sensitivitasnya dalam mendiagnosis kemampuan siswa yang sesungguhnya. Oleh karena itu, perbaikan pada aspek ini mutlak diperlukan. Perbaikan harus diarahkan pada peningkatan derajat homogenitas antara kunci jawaban dan pengecoh, baik dari segi tata bahasa maupun konteks materi, agar instrumen K3LH ini benar-benar valid secara konten dan reliabel secara pengukuran..

SIMPULAN

Berdasarkan hasil analisis psikometrik, dapat disimpulkan bahwa instrumen evaluasi K3LH ini menunjukkan kelemahan validitas yang signifikan. Meskipun memiliki koefisien reliabilitas yang sangat tinggi (0,92), instrumen ini tidak memiliki ketepatan ukur yang memadai karena sebanyak 27 butir soal berada pada kategori tidak valid dan berkualitas sangat rendah. Rendahnya kualitas ini disebabkan oleh dominasi soal kategori mudah serta kegagalan fungsi pengecoh pada 36 butir soal. Akibatnya, instrumen kehilangan daya pembedanya dan tidak mampu memberikan gambaran objektif mengenai perbedaan kompetensi siswa yang sesungguhnya.

Sebagai langkah perbaikan, restrukturisasi total terhadap 27 butir soal yang tidak valid agar lebih selaras dengan indikator kompetensi. Diperlukan pula peningkatan beban kognitif soal ke taraf HOTS (Higher Order Thinking Skills) agar instrumen tidak sekadar menguji hafalan, melainkan kemampuan analisis prosedur keselamatan kerja. Selain itu, revisi pada 36 butir soal dengan pengecoh non-fungsional mutlak diperlukan agar pilihan jawaban lebih homogen dan masuk akal, sehingga mampu meminimalkan faktor kebetulan (*guessing factor*) dan menjamin akuntabilitas

penilaian di SMK.

DAFTAR RUJUKAN

- Abdillah, Y. H., & Hikmah, K. (2024). Analisis Butir Soal Bahasa Arab Siswa Kelas XI MA Islamiyah Sidoarjo Berbasis Software Anates Versi 4.0.9. *Pendas: Jurnal Ilmiah Pendidikan Dasar*, 9(4), 375–391. <https://doi.org/https://doi.org/10.23969/jp.v9i04.21488>
- Akhmadi, M. N. (2021). Analisis Butir Soal Evaluasi Tema 1 Kelas 4 SDN Plumbungan Menggunakan Program Anates. *Ed-Humanistics: Jurnal Ilmu Pendidikan*, 6(1), 799–806. <https://doi.org/https://doi.org/10.33752/ed-humanistics.v6i1.1464>
- Augustia, A. D., Agustia, C. N., Azzahra, D., Hakim, L., & Pratiwi, V. (2025). Analisis Validitas dan Reliabilitas Soal Pilihan Ganda dengan Menggunakan Software Anates pada Mata Pelajaran Perpajakan. *AKADEMIK: Jurnal Mahasiswa Ekonomi Dan Bisnis*, 5(1), 250–265. <https://doi.org/https://doi.org/10.37481/jmbev.v5i1.1165>
- Cahyaningrum, I. Y., Fuady, A., & Sunismi. (2023). Analisis Butir Soal Sumatif Akhir Semester Ganjil Mata Pelajaran Matematika Kelas VII dengan Berbantuan Aplikasi Software Anates. *Mathema: Jurnal Pendidikan Matematika*, 5(2), 67–81. <https://doi.org/https://doi.org/10.37481/jmbev.v5i1.1165>
- Dantes, N., Handayani, N. N. L., & Muliastri, N. K. E. (2021). Pelatihan Validasi Tes Kognitif Berbantuan Aplikasi SPSS 17.0 dan Aplikasi Anates Pada Guru SMAN 1 Singaraja. *Patria: Jurnal Pengabdian Kepada Masyarakat*, 3(2), 116–123. <https://doi.org/https://doi.org/10.24167/patria.v3i2.3243>
- Dia, E. E., & Syah, M. A. S. (2022). Tingkat Kesukaran Soal Penilaian Tengah Semester Mata Pelajaran Bahasa Indonesia di Madrasah. *Fourth Conference on Research and Community Services*, 3(2), 14–23. <https://ejournal.stkipjb.ac.id/index.php/CORCYS/article/view/2609>
- Farida, & Musyarofah, A. (2021). Validitas dan Reliabilitas dalam Analisis Butir Soal. *AL-Mu'arrib: Jurnal Pendidikan Bahasa Arab*, 1(1), 34–44. <https://doi.org/10.32923/al-muarrib.v1i1.2100>
- Fitriani, N. (2021). Analisis Tingkat Kesukaran, Daya Pembeda, dan Efektivitas Pengecoh Soal Pelatihan Kewaspadaan Kegawatdaruratan Maternal dan Neonatal. *Paedagogia: Jurnal Kajian, Penelitian Dan Pengembangan Kependidikan*, 12(2), 199–205. <https://doi.org/https://doi.org/10.31764/paedagogia.v12i2.4956>
- Hardiansyah, A., Harahap, R., & Vandika, A. Y. (2025). Kecerdasan Buatan Sebagai Mitra dalam Penilaian dan Evaluasi Pendidikan. *JIPKL: Jurnal Ilmu Pendidikan Dan Kearifan Lokal*, 2(5), 535–544. <https://lawinsight.net/index.php/JIPEKEL/article/view/317>
- Iriyanti, R. D., Agustina, N. D., Prinita, D. A., Hakim, L., & Pratiwi, V. (2024). Analisis Butir Soal Pilihan Ganda Mata Pelajaran Ekonomi Bisnis dan Administrasi Umum dengan Menggunakan Aplikasi Software Anates di SMK NU Gresik. *Jurnal Padma: Jurnal Pengabdian Kepada Masyarakat*, 4(2), 595–613. <https://doi.org/https://doi.org/10.56689/padma.v4i2.1573>
- Maulana, Y., Nurhayati, Y., Maryam, Yosepty, R., Ratnawulan, T., & Adzani, N. H. (2024). Implementasi Asesmen Nasional Data Pokok Pendidikan (Dapodik) Dalam Peningkatan Kualitas Pendidikan Di Sma Muhammadiyah 2 Majalaya. *EDUSAINTEK: Jurnal Pendidikan, Sains Dan Teknologi*, 11(2), 797–814. <https://doi.org/https://doi.org/10.47668/edusaintek.v11i2.1168>
- Nurhalimah, S., Hidayati, Y., Rosidi, I., & Hadi, W. P. (2022). Hubungan Antara Validitas Item dengan Daya Pembeda dan Tingkat Kesukaran Soal Pilihan Ganda PAS. *Journal Natural*

- Science Education Research, 4(3), 249–257.
<https://doi.org/https://doi.org/10.21107/nser.v4i3.8682>
- Phafiandita, A. N., Permadani, A., Pradani, A. S., & Wahyudi, M. I. (2022). Urgensi Evaluasi Pembelajaran di Kelas. *JIRA: Jurnal Inovasi Dan Riset Akademik*, 3(2), 111–121.
<https://doi.org/https://doi.org/10.47387/jira.v3i2.262>
- Pradita, E., Megawanti, P., & Yulianingsih. (2023). Analisis Tingkat Kesukaran, Daya Pembeda, dan Fungsi Distraktor PTS Matematika SMPN Jakarta. *Himpunan: Jurnal Ilmiah Mahasiswa Pendidikan Matematika*, 3(1), 109–118.
<https://jim.unindra.ac.id/index.php/himpunan/article/view/8645>
- Pudjiati, I., & Madani, F. (2023). Asesmen Autentik Analisis Butir Soal Dengan Rasch Model Di Sekolah Dasar : Literature Review. *Mutiara: Jurnal Penelitian Dan Karya Ilmiah*, 1(4), 1–12. <https://doi.org/https://doi.org/10.59059/mutiara.v1i4.325>
- Purwanti, L. M., Arianty, R., Syakilah, D. M., Ridlo, S., & Susilaningsih, E. (2021). Analisis Soal Tes Pilihan Ganda Berbasis Higher Order Thinking Skill Menggunakan Aplikasi Anates Versi 4.0.9. *Jurnal Pendidikan Universitas Garut*, 15(2), 460–473.
<https://doi.org/https://doi.org/10.52434/jp.v15i2.1287>
- Putri, H., Susiani, D., Wandani, N. S., & Putri, F. A. (2022). Instrumen Penilaian Hasil Pembelajaran Kognitif pada Tes Uraian dan Tes Objektif. *Jurnal Papeda: Jurnal Publikasi Dan Pendidikan Dasar*, 4(2), 139–148. [ttps://ejournal.unimudasorong.ac.id/index.php/jurnalpendidikandasar/article/view/1821](https://ejournal.unimudasorong.ac.id/index.php/jurnalpendidikandasar/article/view/1821)
- Qadir, A., Huda, N., & Hermana, D. (2024). Analisis Butir Tes: Tingkat Kesukaran, Daya Pembeda dan Efektivitas Pengecoh. *Al Furqan: Jurnal Agama, Sosial Dan Budaya*, 3(3), 1450–1467.
<https://doi.org/https://doi.org/10.35931/alfurqan.v3i3>
- Radja, S. P., Bano, V. O., & Ina, A. T. (2023). Analisis Kualitas Butir Soal Hasil Belajar Peserta Didik Berdasarkan Tingkat Kesukaran, Daya Beda Dan Efektivitas Pengecoh di SMAN 1 Pandawai. *Jurnal Edusavana*, 1(1), 30–41.
<https://ojs.unkriswina.ac.id/index.php/Edusavana/article/view/671>
- Rahma, A., Meivia, A., Hakim, L., & Pratiwi, V. (2025). Analisis Butir Soal K3LH pada Tingkat SMK Menggunakan Aplikasi Anates. *Journal of Learning and Teaching*, 2(1), 1–8.
<https://doi.org/https://doi.org/10.70692/hcenev429>
- Rahmawati, N., Qoiri, F. N. P., Cahyo, F. T., Hakim, L., & Pratiwi, V. (2024a). Analisis Butir Soal HOTS Elemen Keselamatan Kerja dan Lingkungan Hidup untuk Menguji Pemahaman Siswa SMK Jurusan Akuntansi Keuangan dan Lembaga Menggunakan Aplikasi Anates. *Hikamatzu Journal of Multidisiplin*, 2(1), 418–427.
<https://yasyahikamatzu.com/index.php/hjm/article/view/160>
- Rahmawati, N., Qoiri, F. N. P., Cahyo, F. T., Hakim, L., & Pratiwi, V. (2024b). Penerapan Modul K3LH (Kesehatan Keselamatan Kerja dan Lingkungan Hidup) Bagi Siswa Kelas X SMKN 4 Surabaya Jurusan Akuntansi Keuangan dan Lembaga. *Hikamatzu Journal of Multidisiplin*, 1(1), 408–414. <https://yasyahikamatzu.com/index.php/hjm/article/view/159>
- Ramadhan, M. F., Siroj, R. A., & Afgani, M. W. (2024). Validitas and Reliabilitas. *Journal on Education*, 6(2), 10967–10975. <https://doi.org/https://doi.org/10.31004/joe.v6i2.4885>
- Ramadhan, W., Malahati, F., Romadhon, K., & Ramadhan, S. (2023). Analisis Butir Soal Tipe Multiple Choice Questions pada Penilaian Harian Sekolah Dasar. *Tarbiyah Wa Ta'lim: Jurnal Penelitian Pendidikan Dan Pembelajaran*, 10(2), 93–105.
<https://doi.org/https://doi.org/10.21093/twt.v10i2.6155>
- Rezigalla, A. A., Eleragi, A. M., Elhussein, A. B., Alfaifi, J., Alghamdi, M. A., Al Ameer, A. Y.,

- Mohammed, O. A., Yahia, O., Mohammed, O. A., & Adam, M. I. (2024). Item analysis : the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*, 24(445), 1–7. <https://doi.org/doi.org/10.1186/s12909-024-05433-y>
- Sabela, O. R., Krisdayanty, D., Taqqiyah, A. Z., Hakim, L., Pratiwi, V., & Surabaya, U. N. (2025). Analisis Butir Soal HOTS Elemen Dokumen Berbasis Digital (FASE E) Menggunakan Program Anates. *Education Achievment: Journal of Science and Research*, 6(1), 251–262. <https://doi.org/https://doi.org/10.51178/jsr.v6i1.2328>
- Setiyawan, R. A., & Wijayanti, P. S. (2020). Analisis Kualitas Instrumen untuk Mengukur Kemampuan Pemecahan Masalah Siswa Selama Pembelajaran Daring di Masa Pandemi. *Jurnal Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika Dan Statistika*, 1(2), 130–139. <https://doi.org/https://doi.org/10.46306/lb.v1i2.26>
- Syafiriya, D. M., Rahmawati, P. A., Qolbi, Z., Pratiwi, V., & Hakim, L. (2024). Analisis Butir Soal HOTS untuk Menguji Pemahaman Siswa SMK Jurusan Akuntansi Menggunakan Software Anates. *Jurnal Nirta: Studi Inovasi*, 4(1), 149–171. <https://doi.org/https://doi.org/10.61412/jnsi.v4i1.107>
- Tampubolon, P. D. S. R., Rantung, D. A., & Naibaho, L. (2023). Kajian Pentingnya Belajar Mengevaluasi Pelaksanaan Pembelajaran Menggunakan Model Evaluasi Cipp. *EDUSAINTEK: Jurnal Pendidikan, Sains Dan Teknologi*, 10(2), 582–599. <https://doi.org/https://doi.org/10.47668/edusaintek.v10i2.757>
- Yudhistiro, A. A., & Rahayu, N. W. (2025). *Pengembangan Aplikasi Mobile Health, Safety, & Environment Menggunakan Ionic Framework*. 12(3), 1350–1369. <https://doi.org/https://doi.org/10.47668/edusaintek.v12i3.1754>