RISET :
Jurnal Ilmiah Multidisiplin Ilmu
Published by PT. Ahlal Publisher Nusantara, Indonesia
Volume 1 No 4 Tahun 2026 Issue E-ISSN : 3123-8475
Journal Homepage: https://jurnalriset.com/index.php/riset/index

Validitas dan Reliabilitas Instrumen Penilaian sebagai Dasar Analisis Hasil Tes serta Implementasi Umpan Balik, Program Perbaikan, dan Pengayaan dalam Pembelajaran

Yusril Alkarim Harahap[1], Zulhimma[2]

[1,2] Universitas Islam Negeri Syehk Ali Hasan Ahmad Addary, Padangsidimpuan, Indonesia

Article Info

Corresponding Author:
Penulis Korespondensi
yusrilhrp200@gmail.com
zulhimma@uinsyahada.ac.id

History:
Submitted: 30-05-2026
Revised: 30-05-2026
Accepted: 29-06-2026

Keyword:
educational evaluation, test instruments, validity, reliability, difficulty level, discriminative power.

Kata Kunci:
evaluasi pendidikan, instrumen tes, validitas, reliabilitas, tingkat kesukaran, daya pembeda.

https://doi.org/10.66914/riset

Abstract

The quality of test instruments is one of the key factors determining the accuracy of evaluation results in the educational process. Instruments that do not meet scientific requirements have the potential to produce inaccurate data, which can lead to errors in educational decision-making. This article aims to comprehensively examine the concepts of validity, reliability, difficulty level, and discriminative power as key indicators of test instrument quality in learning evaluation. The study employed a qualitative approach using library research, specifically through a review of relevant literature, including books on psychometrics, educational evaluation, educational measurement standards, as well as national and international scientific articles. Data collection was conducted through documentary analysis, while data analysis utilized content analysis, involving the stages of data reduction, data presentation, and drawing conclusions. The results of the study indicate that validity ensures that the instrument is capable of measuring the intended construct, while reliability indicates the consistency of measurement results. Additionally, difficulty level analysis is used to determine the appropriate difficulty level of test items relative to students' abilities, whereas discriminant power identifies each item's ability to distinguish between high- and low-ability students. These four aspects complement one another in producing an assessment instrument that meets scientific standards. Therefore, testing for validity, reliability, difficulty level, and discriminative power must be conducted systematically before the instrument is used to ensure that learning assessment results are more objective, accurate, and accountable.

Abstrak

Kualitas instrumen tes merupakan salah satu faktor utama yang menentukan ketepatan hasil evaluasi dalam proses pendidikan. Instrumen yang tidak memenuhi persyaratan ilmiah berpotensi menghasilkan data yang tidak akurat sehingga berdampak pada kesalahan dalam pengambilan keputusan pendidikan. Artikel ini bertujuan untuk mengkaji secara komprehensif konsep validitas, reliabilitas, tingkat kesukaran, dan daya pembeda sebagai indikator utama kualitas instrumen tes dalam evaluasi pembelajaran. Penelitian menggunakan pendekatan kualitatif dengan jenis library research, yaitu melalui kajian terhadap berbagai literatur yang relevan berupa buku psikometri, evaluasi pendidikan, standar pengukuran pendidikan, serta artikel ilmiah nasional dan internasional. Pengumpulan data dilakukan melalui studi dokumentasi, sedangkan analisis data menggunakan content analysis dengan tahapan reduksi data, penyajian data, dan penarikan kesimpulan. Hasil kajian menunjukkan bahwa validitas berfungsi memastikan instrumen mampu mengukur konstruk yang seharusnya diukur, sedangkan reliabilitas menunjukkan konsistensi hasil pengukuran. Selain itu, analisis tingkat kesukaran digunakan untuk menentukan proporsi kesulitan butir soal agar sesuai dengan kemampuan peserta didik, sementara daya pembeda berfungsi mengidentifikasi kemampuan setiap butir soal dalam membedakan peserta didik berkemampuan tinggi dan rendah. Keempat aspek tersebut saling melengkapi dalam menghasilkan instrumen evaluasi yang memenuhi standar ilmiah. Oleh karena itu, pengujian validitas, reliabilitas, tingkat kesukaran, dan daya pembeda perlu dilakukan secara sistematis sebelum instrumen digunakan agar hasil evaluasi pembelajaran lebih objektif, akurat, dan dapat dipertanggungjawabkan.

PENDAHULUAN

Evaluasi merupakan salah satu komponen yang tidak dapat dipisahkan dari proses pendidikan karena berfungsi sebagai dasar dalam mengetahui tingkat ketercapaian tujuan pembelajaran. Melalui evaluasi, pendidik dapat memperoleh informasi mengenai keberhasilan proses pembelajaran, perkembangan kompetensi peserta didik, serta efektivitas strategi pembelajaran yang telah diterapkan. Informasi tersebut menjadi dasar bagi pengambilan keputusan dalam memperbaiki proses pembelajaran, menyusun program tindak lanjut, maupun menetapkan kebijakan pendidikan. Oleh sebab itu, kualitas hasil evaluasi sangat bergantung pada kualitas instrumen yang digunakan sebagai alat pengumpul data.(Anas Sudijono 2018)

Instrumen tes merupakan salah satu alat evaluasi yang paling banyak digunakan untuk mengukur hasil belajar peserta didik, terutama pada ranah kognitif. Namun demikian, sebuah tes tidak cukup hanya mampu menghasilkan skor, melainkan juga harus memenuhi persyaratan ilmiah agar hasil pengukuran dapat dipercaya. Instrumen yang disusun tanpa melalui proses pengujian kualitas berpotensi menghasilkan data yang bias sehingga dapat menimbulkan kesalahan dalam penilaian maupun pengambilan keputusan. Oleh karena itu, sebelum digunakan dalam kegiatan evaluasi, setiap instrumen perlu dianalisis secara komprehensif melalui pengujian validitas, reliabilitas, tingkat kesukaran, dan daya pembeda butir soal.(Hanipah 2023)

Validitas menjadi aspek pertama yang harus dipenuhi karena berkaitan dengan sejauh mana instrumen mampu mengukur apa yang seharusnya diukur sesuai dengan tujuan pengukuran. Di samping itu, reliabilitas menunjukkan tingkat konsistensi hasil pengukuran apabila instrumen digunakan berulang kali dalam kondisi yang relatif sama. Sebuah instrumen yang valid tetapi tidak reliabel tetap belum layak digunakan karena hasil pengukurannya tidak stabil. Sebaliknya, instrumen yang reliabel tetapi tidak valid juga tidak mampu memberikan informasi yang benar mengenai konstruk yang diukur. Oleh karena itu, validitas dan reliabilitas merupakan dua karakteristik utama yang saling melengkapi dalam menentukan kualitas suatu instrumen evaluasi pendidikan.(Efendi, Tatang Muhtar, and Yusuf Tri Herlambang 2023)

Selain validitas dan reliabilitas, kualitas instrumen juga ditentukan oleh

karakteristik setiap butir soal. Analisis tingkat kesukaran diperlukan untuk mengetahui apakah suatu butir soal tergolong terlalu mudah, terlalu sukar, atau berada pada tingkat kesukaran yang ideal. Sementara itu, analisis daya pembeda bertujuan mengetahui kemampuan suatu butir soal dalam membedakan peserta didik yang memiliki kemampuan tinggi dengan peserta didik yang memiliki kemampuan rendah. Kedua analisis tersebut sangat penting dalam proses pengembangan instrumen karena dapat menjadi dasar dalam mempertahankan, merevisi, atau mengganti butir soal sehingga instrumen yang dihasilkan memiliki kualitas yang lebih baik.(Yanti, Y. 2024)

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kualitatif dengan jenis penelitian kepustakaan (library research). Penelitian kepustakaan merupakan metode penelitian yang memanfaatkan berbagai sumber literatur sebagai objek utama kajian untuk memperoleh pemahaman yang komprehensif mengenai suatu konsep, teori, maupun hasil penelitian terdahulu. Dalam penelitian ini, fokus kajian diarahkan pada konsep validitas, reliabilitas, tingkat kesukaran, dan daya pembeda instrumen tes sebagai komponen utama dalam pengembangan instrumen evaluasi pendidikan yang berkualitas.

Sumber data penelitian terdiri atas data primer dan data sekunder. Data primer diperoleh dari buku-buku psikometri, evaluasi, dan pengukuran pendidikan yang relevan, sedangkan data sekunder berasal dari artikel jurnal ilmiah, prosiding, serta dokumen pendukung lainnya yang berkaitan dengan topik penelitian. Teknik pengumpulan data dilakukan melalui studi dokumentasi dengan menelaah dan menyeleksi berbagai sumber pustaka berdasarkan kredibilitas, relevansi, dan validitas akademiknya. Analisis data menggunakan content analysis melalui tahapan reduksi data, penyajian data, dan penarikan kesimpulan. Data yang telah diperoleh diklasifikasikan, dianalisis, dan disintesis berdasarkan tema pembahasan untuk menghasilkan pemahaman yang komprehensif mengenai validitas, reliabilitas, tingkat kesukaran, dan daya pembeda instrumen tes.(John W. Creswell dan J. David Creswell 2018)

PEMBAHASAN

A. Hakikat Pengukuran, Penilaian, dan Evaluasi dalam Pendidikan

Dalam khazanah ilmu pendidikan, sering kali terjadi tumpang tindih penggunaan istilah pengukuran (measurement), penilaian (assessment), dan evaluasi (evaluation). Pengukuran adalah proses kuantifikasi atribut atau karakteristik dari suatu objek, orang, atau kejadian berdasarkan aturan-aturan tertentu. Hasil dari pengukuran selalu berbentuk angka atau skor numerik. Penilaian adalah kegiatan menafsirkan, mendeskripsikan, dan mengorganisasikan data hasil pengukuran untuk membuat gambaran mengenai capaian belajar siswa. Sedangkan evaluasi adalah proses sistematis yang mengumpulkan, menganalisis, dan menginterpretasikan informasi untuk menentukan sejauh mana tujuan pembelajaran telah dicapai, guna mengambil keputusan atau kebijakan tertentu.(Sudijono 2011)

Untuk melakukan evaluasi yang adil dan akurat, alat ukur yang digunakan dalam proses pengukuran harus memiliki kualitas yang terstandardisasi. Kualitas alat ukur inilah yang dijaga melalui analisis validitas, reliabilitas, dan karakteristik butir soal. Jika alat ukurnya cacat, maka angka pengukuran akan keliru, penilaian menjadi bias, dan keputusan evaluasi yang diambil pun akan salah sasaran.

Konsep Mutakhir Validitas Instrumen Tes Validitas berasal dari kata validity, yang merujuk pada sejauh mana suatu alat ukur secara tepat dan cermat melakukan fungsi ukurnya. Berdasarkan pandangan kontemporer dari Standards for Educational and Psychological Testing, validitas bukanlah karakteristik yang melekat secara absolut pada instrumen itu sendiri, melainkan mengacu pada derajat kesahihan interpretasi dan penggunaan skor hasil tes. Validitas dikelompokkan menjadi tiga bukti utama:

Validitas Isi (Content Validity)

Validitas isi menunjukkan sejauh mana butir-butir soal dalam instrumen mencerminkan keseluruhan domain materi yang hendak diukur secara representatif. Validitas ini sangat krusial pada tes hasil belajar kognitif (achievement test). Pengujian validitas isi tidak menggunakan analisis statistik berbasis skor respon siswa, melainkan melalui analisis rasional-logis oleh para pakar atau penelaah sejawat (expert judgment).

Langkah utama dalam

membangun validitas isi adalah menyusun kisi-kisi tes (test blueprint) yang memetakan hubungan antara kompetensi dasar, materi pelajaran, indikator soal, dan level kognitif (misalnya taksonomi Bloom revisi). Para pakar kemudian menilai kesesuaian antara butir soal dengan indikator yang ingin dicapai. Untuk mengkuantifikasi validitas isi dari penilaian para pakar, sering digunakan indeks Aiken's V atau Lawshe's CVR (Content Validity Ratio). Formula Indeks V Aiken dirumuskan sebagai berikut:

$$V = \Sigma s / [n \times (c - 1)]$$

Di mana s = r - lo (skor pilihan pakar dikurangi skor terendah), n adalah jumlah pakar, dan c adalah skor tertinggi yang dapat dipilih. Nilai V berkisar antara 0 hingga 1, di mana nilai yang mendekati 1 menunjukkan validitas isi yang sangat kuat.(Susan M. Brookhart and Anthony J. Nitko 2019)

Validitas Konkuren (Concurrent Validity): Kriteria eksternal diambil pada waktu yang bersamaan atau dalam rentang waktu yang sangat dekat. Contohnya adalah mengorelasikan tes kemampuan matematika baru yang dibuat oleh guru dengan tes matematika standar yang sudah baku (seperti prasat ujian nasional atau olimpiade).

Validitas Konstruk (Construct Validity) Validitas konstruk berkaitan dengan sejauh mana instrumen mampu mengukur konstruk teoretis atau variabel laten psikologis yang bersifat abstrak, seperti motivasi belajar, kecemasan akademis, efikasi diri, atau berpikir kritis. Konstruk ini tidak dapat diamati secara langsung melainkan melalui indikator-indikator keperilakua. Pembuktian validitas konstruk memerlukan penelaahan empiris yang mendalam menggunakan analisis statistika multivariat, yaitu Analisis Faktor (Faktor Analisis Eksploratori maupun Konfirmatori). Melalui analisis faktor, peneliti dapat menguji apakah butir-butir soal yang dikelompokkan ke dalam dimensi tertentu benar-benar mengelompok dan memiliki muatan faktor (factor loading) yang signifikan (biasanya > 0,40) pada konstruk teoretis yang dihipotesiskan.(C. H. Lawshe 1975)

B. Validitas Kriteria (Criterion-Related Validity)

Validitas kriteria (criterion-related validity) adalah jenis validitas empiris yang didapat dengan membandingkan skor instrumen yang dibuat dengan skor instrumen lain yang sudah diakui sebagai kriteria atau standar perbandingan (benchmark). Tujuan dari pengujian ini adalah untuk mengetahui seberapa besar hasil pengukuran suatu

alat berhubungan dengan ukuran eksternal yang relevan. Secara statistik, kriteria validitas umumnya dianalisis dengan menggunakan koefisien korelasi Pearson Product Moment (rxy). Semakin besar koefisien korelasi yang diperoleh, semakin tinggi pula validitas instrumen terhadap kriteria yang digunakan. Oleh sebab itu, validitas kriteria biasanya diterapkan untuk menunjukkan bahwa suatu instrumen dapat mengukur konstruk yang sama atau meramalkan hasil yang berhubungan dengan variabel tertentu.

Berdasarkan dimensi waktu pengumpulan data kriteria, validitas kriteria dibagi menjadi dua jenis, yaitu sebagai berikut.

. Validitas Prediktif (Kemampuan Meramalkan) Validitas prediktif menggambarkan kemampuan suatu alat untuk meramalkan keadaan, tindakan, atau kinerja yang akan muncul di masa depan. Pada tipe validitas ini, data kriteria didapatkan setelah pelaksanaan tes atau pengukuran awal sehingga hasil instrumen dapat dibandingkan dengan pencapaian responden di waktu yang akan datang. Semakin kuat keterkaitan antara skor instrumen dengan hasil yang diprediksi, semakin baik validitas prediktif dari instrumen itu. Salah satu contoh umum adalah pemanfaatan skor Tes Potensi Akademik (TPA) atau Ujian Tulis Berbasis Komputer (UTBK) untuk meramalkan Indeks Prestasi Kumulatif (IPK) mahasiswa di akhir masa pendidikan. Jika hubungan antara skor UTBK dan IPK menunjukkan nilai yang tinggi serta signifikan, maka instrumen tersebut memiliki validitas prediktif yang baik.

Validitas Konkuren (Concurrent Validity) Validitas konkuren adalah validitas yang didapat dengan membandingkan skor instrumen yang sedang dibuat dengan skor instrumen lain yang sudah terstandarisasi dan diukur pada saat yang sama atau dalam periode yang sangat mendekati. Tujuan dari penelitian ini adalah untuk mengukur seberapa cocok hasil dari kedua alat tersebut. Jika korelasi yang didapatkan tinggi, maka alat yang dibuat dapat dianggap memiliki validitas konkuren yang baik. Contohnya, seorang pengajar membuat ujian kemampuan matematika lalu membandingkan hasil ujian itu dengan skor ujian matematika standar yang sudah terverifikasi dan diberikan kepada siswa pada waktu yang bersamaan.(Anthony J. Nitko dan Susan M. Brookhart 2019)

C. Validitas Konstruk (*Construct Validity*)

Validitas konstruk (*construct*

validity) merupakan tingkat ketepatan suatu instrumen dalam mengukur konstruk teoretis atau variabel laten yang bersifat abstrak dan tidak dapat diamati secara langsung, seperti motivasi belajar, efikasi diri (*self-efficacy*), kecemasan akademik, sikap, minat belajar, kreativitas, maupun kemampuan berpikir kritis. Konstruk tersebut hanya dapat diukur melalui sejumlah indikator atau manifestasi perilaku yang disusun berdasarkan landasan teori yang kuat. Oleh karena itu, penyusunan instrumen harus diawali dengan identifikasi dimensi konstruk, penentuan indikator, dan penyusunan butir-butir pernyataan yang merepresentasikan setiap indikator secara proporsional.(Robert F. DeVellis dan Thor Widaman 2016)

Pembuktian validitas konstruk dilakukan melalui dua tahap, yaitu validasi teoretis dan validasi empiris. Pada tahap validasi teoretis, para ahli (*expert judgment*) menilai kesesuaian butir instrumen dengan definisi konseptual dan indikator variabel. Selanjutnya, validasi empiris dilakukan menggunakan teknik analisis statistik multivariat, terutama *Analisis Faktor Eksploratori (Exploratory Factor Analysis/EFA) dan Analisis Faktor Konfirmatori (Confirmatory Factor Analysis/CFA).* EFA digunakan untuk mengeksplorasi struktur faktor yang terbentuk dari data empiris, sedangkan CFA bertujuan menguji kesesuaian struktur faktor dengan model teoritis yang telah dirumuskan sebelumnya.(Joseph F. Hair Jr., William C. Black, Barry J. Babin, dan Rolph E. Anderson 2019)

D. Konsep dan Estimasi Reliabilitas Instrumen Tes

Reliabilitas merepresentasikan derajat konsistensi, keajegan, kestabilan, atau keterpercayaan dari hasil pengukuran instrumen. Sebuah tes dikatakan reliabel jika skor yang diperoleh peserta didik cenderung stabil atau mirip ketika tes tersebut diujikan kembali pada waktu berbeda, dengan penguji yang berbeda, atau dengan set soal yang setara. Faktor yang memengaruhi reliabilitas antara lain panjang tes, heterogenitas kelompok, tingkat kesulitan tes, dan kondisi emosional peserta didik saat ujian.(Robert L. Linn and Norman E. Gronlund 2000)

Metode Belah Dua (Split-Half Method)

Salah satu kelemahan metode tes-ulang (test-retest) adalah adanya efek bawaan (practice effect) di mana siswa mengingat soal pada ujian pertama saat

mengerjakan ujian kedua. Untuk mengatasi hal tersebut, digunakan metode konsistensi internal, salah satunya adalah Metode Belah Dua (Split-Half Method). Keunggulan utama metode ini adalah efisiensi waktu, karena tes hanya perlu diujikan satu kali saja kepada sekelompok responden.

Setelah tes diujikan, seluruh butir soal dibelah menjadi dua bagian yang setara. Terdapat beberapa teknik pembelahan yang umum digunakan dalam evaluasi pendidikan:

Pembelahan Ganjil-Genap (Odd-Even Split): Mengelompokkan butir soal bernomor ganjil (1, 3, 5, dst.) ke dalam belahan pertama, dan butir soal bernomor genap (2, 4, 6, dst.) ke dalam belahan kedua. Teknik ini paling direkomendasikan karena dapat mengontrol efek kelelahan siswa dan tingkat kesulitan soal yang biasanya meningkat di akhir tes.

Pembelahan Awal-Akhir (First-Last Split): Membagi tes secara linier, misalnya jika ada 40 soal, soal nomor 1-20 menjadi belahan pertama, dan nomor 21-40 menjadi belahan kedua. Teknik ini kurang cocok jika soal di akhir memiliki bobot kesulitan yang jauh lebih tinggi atau jika siswa kehabisan waktu di akhir tes.(Suharsimi Arikunto 2018)

Formula Spearman-Brown dan Pembuktian Matematis

Pembelahan instrumen berimplikasi pada penurunan jumlah butir soal menjadi setengah dari panjang tes semula. Karena reliabilitas sangat dipengaruhi oleh panjang tes (makin banyak soal, cenderung makin reliabel), maka koefisien korelasi yang dihitung antara belahan pertama dan belahan kedua baru mencerminkan reliabilitas dari setengah tes tersebut. Untuk mengoreksi koefisien tersebut agar mencerminkan reliabilitas instrumen secara utuh, digunakan Formula Spearman-Brown Prophecy.

Rumus matematis Spearman-Brown adalah sebagai berikut:

$$r_{11} = 2r_{½½} / (1 + r_{½½})$$

Di mana:

r11 = Koefisien reliabilitas instrumen secara keseluruhan (tes penuh).

r½½ = Koefisien korelasi Pearson Product Moment antara skor belahan pertama dan skor belahan kedua.

Kriteria interpretasi koefisien reliabilitas secara umum merujuk pada ketentuan Guilford, di mana jika r11 ≥ 0,70, instrumen dinyatakan memiliki reliabilitas yang memadai untuk tes hasil belajar kognitif sekolah

Contoh Kasus Perhitungan Reliabilitas

Belah Dua :

Berikut disajikan simulasi data skor dari 5 orang siswa yang mengerjakan tes pilihan ganda yang telah dibelah menggunakan teknik ganjil-genap. Skor berkisar antara 0 (salah) dan 5 (benar) untuk masing-masing belahan.

Nama Siswa	Skor Ganjil (X)	Skor Genap (Y)	X^2	Y^2	XY
Siswa A	4	3	16	9	12
Siswa B	5	5	25	25	25
Siswa C	2	3	4	9	6
Siswa D	3	2	9	4	6
Siswa E	1	2	1	4	2
TOTAL (Σ)	**15**	**15**	**55**	**51**	**51**

Langkah 1: Menghitung Korelasi Belahan ($r_{½½}$) menggunakan rumus korelasi Pearson:

Langkah 1: Menghitung Korelasi Belahan ($r_{½½}$) menggunakan ru

$$r\_half = [N\Sigma XY - (\Sigma X)(\Sigma Y)] / \sqrt{[\ \{N\Sigma X^2 - (\Sigma X)^2\}\{N\Sigma Y^2 - (\Sigma Y}$$

$$r\_half = [5(51) - (15)(15)] / \sqrt{[\ \{5(55) - (15)^2\}\{5(51) - (15)^2\}\ ]}$$

$$r\_half = [255 - 225] / \sqrt{[\ \{275 - 225\}\{255 - 225\}\ ]}$$

$$r\_half = 30 / \sqrt{[\ 50 \times 30\ ]} = 30 / \sqrt{1500} = 30 / 38{,}73 = 0{,}775$$

Langkah 2: Memasukkan nilai korelasi belahan ke dalam Formula Spearman-Brown untuk mendapatkan reliabilitas total (r11):

$$r\_11 = (2 \times 0{,}775) / (1 + 0{,}775)$$

$$r\_11 = 1{,}55 / 1{,}775 = 0{,}873$$

Berdasarkan hasil perhitungan di atas, diperoleh koefisien reliabilitas keseluruhan tes sebesar 0,873. Nilai ini berada pada rentang yang sangat tinggi, sehingga instrumen tes tersebut dikategorikan sebagai instrumen yang memiliki konsistensi internal sangat kuat dan reliabel untuk digunakan.(Jum C. Nunnally and Ira H. Bernstein 1994)

E. Metode Perhitungan Indeks Daya Pembeda

Prosedur standar untuk menghitung indeks daya pembeda pada sampel berskala kecil hingga sedang di sekolah adalah sebagai berikut:

1. Seluruh lembar jawaban siswa diurutkan berdasarkan total skor pasca-ujian, mulai dari skor tertinggi hingga skor terendah.
2. Kelompok peserta didik dibelah menjadi dua bagian utama: kelompok atas (upper group) dan kelompok bawah (lower group). Jika jumlah sampel besar ($N \geq 30$), porsi ideal yang diambil menurut para ahli psikometri adalah 27% kelompok teratas dan 27% kelompok terbawah. Untuk kelas kecil, porsi dapat disesuaikan menjadi 50% atas dan 50% bawah.
3. Dihitung jumlah siswa yang menjawab benar pada kelompok atas (BA) dan jumlah

siswa yang menjawab benar pada kelompok bawah (BB).

Formula matematis Indeks Daya Pembeda (D) dinyatakan sebagai:

$$D = (B_A / J_A) - (B_B / J_B) = P_A - P_B$$

Di mana:

JA = Jumlah total sampel dalam kelompok atas.

JB = Jumlah total sampel dalam kelompok bawah.

PA = Proporsi kelompok atas yang menjawab benar soal tersebut.

PB = Proporsi kelompok bawah yang menjawab benar soal tersebut.

Klasifikasi interpretasi nilai Indeks Daya Pembeda (D) mengacu pada standar umum evaluasi sebagai berikut:

Rentang Nilai Indeks D	Kategori Kualitas Soal	Rekomendasi Tindak Lanjut
0,40 – 1,00	Sangat Baik (Excellent)	Diterima langsung dan disimpan di bank soal.
0,30 – 0,39	Baik (Good)	Diterima dengan revisi minor pada opsi pengecoh jika perlu.
0,20 – 0,29	Cukup/Marginal (Fair)	Harus direvisi secara substantif sebelum digunakan lagi.
< 0,20 atau Negatif	Jelek/Buruk (Poor)	Dibuang atau ditolak total dari sistem bank soal.

Simulasi Komparatif Analisis Daya Pembeda

Sebagai contoh, guru menguji daya pembeda untuk dua butir soal (Soal No. 1 dan Soal No. 2) di kelas dengan jumlah kelompok atas (JA) = 10 siswa dan kelompok bawah (JB) = 10 siswa.

Kasus Soal No. 1: Di kelompok atas, ada 9 siswa menjawab benar (BA=9). Di kelompok bawah, ada 3 siswa menjawab benar (BB=3).

Perhitungan: D = (9/10) - (3/10) = 0,90 - 0,30 = 0,60 (Kategori: Sangat Baik).

Kasus Soal No. 2: Di kelompok atas, hanya 2 siswa menjawab benar (BA=2). Di kelompok bawah, ada 7 siswa menjawab benar (BB=7).

Perhitungan: D = (2/10) - (7/10) = 0,20 - 0,70 = -0,50 (Kategori: Jelek/Negatif).

Soal No. 2 memiliki daya pembeda negatif, artinya soal tersebut menyesatkan. Siswa yang bodoh justru lebih banyak menjawab benar dibandingkan siswa yang pintar. Fenomena ini.

KESIMPULAN

Berdasarkan hasil kajian, dapat disimpulkan bahwa kualitas instrumen tes merupakan faktor fundamental dalam menjamin ketepatan hasil evaluasi pendidikan. Instrumen yang baik harus memenuhi aspek validitas, sehingga mampu mengukur konstruk atau kompetensi yang menjadi tujuan pengukuran, serta reliabilitas, sehingga menghasilkan skor yang konsisten dan dapat dipercaya. Selain itu, analisis

tingkat kesukaran diperlukan untuk memastikan butir soal memiliki tingkat kesulitan yang proporsional sesuai dengan kemampuan peserta didik, sedangkan daya pembeda berfungsi mengidentifikasi kemampuan setiap butir soal dalam membedakan peserta didik yang berkemampuan tinggi dan rendah. Keempat komponen tersebut saling berkaitan dan menjadi indikator utama dalam menentukan mutu suatu instrumen evaluasi. Oleh karena itu, pengembangan instrumen tes hendaknya dilakukan secara sistematis melalui penyusunan kisi-kisi, pengujian validitas dan reliabilitas, serta analisis karakteristik butir soal sebelum digunakan dalam proses pembelajaran. Dengan demikian, instrumen yang dihasilkan mampu memberikan data yang objektif, akurat, konsisten, dan dapat dipertanggungjawabkan sebagai dasar pengambilan keputusan untuk meningkatkan kualitas pembelajaran dan hasil belajar peserta didik.

DAFTAR PUSTAKA

Anas Sudijono. 2018. *Pengantar Evaluasi Pendidikan*. jakarta: Rajawali Pers.

Anthony J. Nitko dan Susan M. Brookhart. 2019. *Educational Assessment of Students*. Hoboken: NJ: Pearson.

C. H. Lawshe. 1975. "A Quantitative Approach to Content Validity,"." *Personnel Psychology* 28(4):567. https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1744-6570.1975.tb01393.x.

Efendi, Pitri Maharani, Tatang Muhtar, and Yusuf Tri Herlambang. 2023. "Relevansi Kurikulum Merdeka Dengan Konsepsi Ki Hadjar Dewantara: Studi Kritis Dalam Perspektif Filosofis-Pedagogis." *Jurnal Elementaria Edukasia* 6(2):548–61. doi:10.31949/jee.v6i2.5487.

Hanipah, Sri. 2023. "Analisis Kurikulum Merdeka Belajar Dalam Memfasilitasi Pembelajaran Abad Ke-21." *Jurnal Bintang Pendidikan Indonesia (JUBPI)* 1(2):264–75.

John W. Creswell dan J. David Creswell. 2018. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches,*. Thousand Oaks: CA: SAGE Publications,.

Joseph F. Hair Jr., William C. Black, Barry J. Babin, dan Rolph E. Anderson. 2019. *Multivariate Data Analysi*. Andover: Hampshire: Cengage Learning,.

Jum C. Nunnally and Ira H. Bernstein. 1994. *Psychometric Theory,*. New York: McGraw-Hill,.

Robert F. DeVellis dan Thor Widaman. 2016. *Scale Development: Theory and Applications,*. Thousand Oaks: Oaks, CA: SAGE Publications,.

Robert L. Linn and Norman E. Gronlund. 2000. *Measurement and Assessment in Teaching*. Upper Saddle River,: NJ: Pearson Education.

Sudijono, Anas. 2011. *Pengantar Evaluasi Pendidikan*. jakarta: Rajawali Pers.

Suharsimi Arikunto. 2018. *Dasar-Dasar Evaluasi Pendidikan*. jakarta: Bumi Aksara.

Susan M. Brookhart and Anthony J. Nitko. 2019. *Educational Assessment of Students*. New York: Pearson,.

Yanti, Y., M. 2024. "Implementation Of The Independent Learning Curriculum For Students." *PPSDP International Journal of Education* 3(2):413–24.