J-TP MIPA: Jurnal Transformasi Pendidikan MIPA
Vol. 1, No. 1, 31 Maret 2026, pp. 37–50

Pengembangan Instrumen Evaluasi Berbasis Keterampilan Berpikir Tingkat Tinggi (HOTS) pada Materi Klasifikasi Makhluk Hidup

Fitri Hayati Kurnia[1], Asep Dikdik[2], Arini Nur Fariha[3], Salma Nurillah[4]

[1]1 MTs Al-Jawahir, Kabupaten Bandung, Jawa Barat, Indonesia
[2]SD Negeri Cisarua, Kabupaten Bandung Barat, Jawa Barat, Indonesia
[3]STAI Inovatif Daarul Ihsan, Kota Cimahi, Jawa Barat, Indonesia
[4]PAUD Annur Faridatul Irsyad, Kabupaten Garut, Jawa Barat, Indonesia
[1]fitrihayatikurnia8@gmail.com, [2]asep.dikdik@upi.edu, [3]arininurfariha19@upi.edu, [4]salma.nrllh13@upi.edu

*Penulis Korespondensi: fitrihayatikurnia8@gmail.com

(Diterima 5 Maret 2026, Disetujui 14 Maret 2026, Tersedia Online 31 Maret 2026)

Abstract: The enhancement of Higher-Order Thinking Skills (HOTS) in 21st-century science education necessitates the development of valid and reliable assessment instruments.This study developed and validated a Higher-Order Thinking Skills (HOTS)-based evaluation instrument covering cognitive levels of analyzing (C4) and evaluating (C5) for the topic of living organism classification at the Madrasah Tsanawiyah (MTs) level. The 3-D Model (Define, Design, Develop) served as the development framework. The instrument comprised 20 multiple-choice items built from stimulus-based scenarios aligned with the Revised Bloom's Taxonomy. Content validity was assessed by one science education expert through a structured validation sheet interpreted using Guilford's feasibility criteria, yielding 70% feasibility (sufficiently feasible); targeted revisions were applied. Reliability was analyzed through the split-half method using ANATES, producing a Spearman-Brown coefficient of 0.82 (very high). A small-scale trial involved 14 Grade VII students at MTs Inovatif Daarul Ihsan Bandung. Item analysis revealed 70% of items in the easy category, 25% moderate, and 5% difficult. Discrimination indices showed four items excellent ($D \geq 0.70$), six good ($0.40 \leq D < 0.70$), five adequate ($0.20 \leq D < 0.40$), and five poor ($D < 0.20$); six items required revision. The instrument demonstrates adequate psychometric properties for measuring students' higher-order thinking abilities in living organism classification.

Keywords: 3-D Development Model; Evaluation Instrument; Guilford Feasibility Criteria; HOTS; Living Organism Classification

Abstrak: Penguatan Higher-Order Thinking Skills (HOTS) dalam pendidikan sains abad ke-21 menuntut pengembangan instrumen evaluasi yang valid dan reliabel**.**Penelitian ini mengembangkan dan memvalidasi instrumen evaluasi berbasis Higher-Order Thinking Skills (HOTS) yang mencakup level kognitif menganalisis (C4) dan mengevaluasi (C5) pada materi klasifikasi makhluk hidup di tingkat Madrasah Tsanawiyah (MTs). Model 3-D (Define, Design, Develop) digunakan sebagai kerangka pengembangan. Instrumen terdiri dari 20 soal pilihan ganda berbasis stimulus yang selaras dengan Taksonomi Bloom Revisi. Validitas isi dinilai oleh satu ahli pendidikan IPA melalui lembar validasi terstruktur yang diinterpretasikan menggunakan kriteria kelayakan Guilford, menghasilkan persentase kelayakan 70% (cukup layak); revisi tertarget dilakukan. Reliabilitas dianalisis menggunakan teknik split-half melalui aplikasi ANATES dengan koefisien Spearman-Brown 0,82 (sangat tinggi). Uji coba terbatas melibatkan 14 peserta didik kelas VII MTs Inovatif Daarul Ihsan Bandung. Analisis butir menunjukkan 70% butir berkategori mudah, 25% sedang, dan 5% sulit. Daya pembeda menghasilkan empat butir sangat baik ($D \geq 0{,}70$), enam butir baik ($0{,}40 \leq D < 0{,}70$), lima butir cukup ($0{,}20 \leq D < 0{,}40$), dan lima butir kurang ($D < 0{,}20$); enam butir direvisi. Instrumen memiliki properti psikometri yang memadai untuk mengukur kemampuan berpikir tingkat tinggi peserta didik pada materi klasifikasi makhluk hidup.

Kata Kunci: Model Pengembangan 3-D; Instrumen Evaluasi; Kriteria Kelayakan Guilford; HOTS; Klasisifikasi Makhluk Hidup

1. Pendahuluan

Pendidikan sains di abad ke-21 menuntut perubahan fundamental dalam cara guru merancang proses pembelajaran dan instrumen evaluasi. Kurikulum Merdeka secara eksplisit mengarahkan penguatan kemampuan kognitif tingkat tinggi yang dikenal sebagai *Higher-Order Thinking Skills* (HOTS) sebagai kompetensi inti bagi seluruh peserta didik [3]. Berdasarkan Taksonomi Bloom Revisi [2], level menganalisis (C4) dan mengevaluasi (C5) dipandang

sebagai fondasi intelektual yang memungkinkan peserta didik menavigasi kompleksitas masalah dunia nyata. Tanpa instrumen evaluasi yang dirancang secara khusus untuk mengukur kedua level tersebut, klaim mengenai implementasi HOTS dalam pembelajaran tetap tidak dapat diverifikasi secara empiris.

Urgensi HOTS melampaui mandat kurikulum dan mencerminkan kebutuhan kompetensi global yang lebih luas. Thornhill-Miller et al. [3] menegaskan bahwa kreativitas, berpikir kritis, komunikasi, dan kolaborasi (keterampilan 4C) merupakan prasyarat fundamental keberhasilan individu di era modern. Santos-Meneses dan Drugova [4] menunjukkan bahwa lemahnya integrasi HOTS dalam sistem pendidikan secara langsung menurunkan kapasitas pemecahan masalah peserta didik. Di Indonesia, kinerja peserta didik dalam *Programme for International Student Assessment* (PISA) dan *Trends in International Mathematics and Science Study* (TIMSS) yang konsisten berada di bawah rata-rata regional mengindikasikan bahwa kualitas instrumen evaluasi layak mendapat perhatian serius [5].

Materi klasifikasi makhluk hidup di tingkat MTs/SMP memiliki posisi strategis untuk pengembangan HOTS. Materi ini secara inheren mensyaratkan peserta didik menganalisis karakteristik morfologis dan filogenetik organisme, serta mengevaluasi kriteria pengelompokan berdasarkan bukti ilmiah; sehingga secara alami mengoperasionalkan C4 dan C5 [19]. Meski demikian, Darmawan et al. [6] mengidentifikasi bahwa kapasitas guru dalam merancang asesmen sains yang menuntut berpikir tingkat tinggi masih sangat terbatas. Delgado dan Luna [7] menyimpulkan bahwa mayoritas instrumen evaluasi sains masih terkonsentrasi pada domain kognitif rendah (C1–C3), sedangkan Muhayimana et al. [8] menunjukkan dominasi *Lower-Order Thinking Skills* (LOTS) dalam ujian formal. Sepriyanti et al. [5] membuktikan bahwa peserta didik yang terpapar instrumen HOTS tervalidasi menunjukkan peningkatan kemampuan berpikir tingkat tinggi yang lebih signifikan dibandingkan dengan instrumen konvensional.

Berbagai penelitian pengembangan telah memberikan kontribusi bermakna. Mohamad et al. [9], Lamhatin et al. [10], Kurnia et al. [11], dan Ayubi et al. [12] memvalidasi pendekatan pengembangan instrumen HOTS melalui berbagai konteks sains. Nurjanah et al. [13] dan Ahmad et al. [14] menetapkan prosedur validasi isi sebagai tahap kritis dalam pengembangan instrumen. Meski demikian, tiga kesenjangan penelitian masih belum terselesaikan: pertama, instrumen berbasis HOTS yang dirancang khusus untuk materi klasifikasi makhluk hidup belum tersedia, hampir seluruh pengembangan berfokus pada fisika, kimia, dan matematika [7–16]; kedua, instrumen yang mengintegrasikan C4 dan C5 secara proporsional untuk materi ini masih sangat langka dalam literatur terindeks [9–10]; ketiga, potensi analitis materi klasifikasi makhluk hidup belum dieksploitasi secara optimal melalui instrumen yang tepat [6–8]. Kebaruan penelitian ini terletak pada pengembangan dan validasi psikometri instrumen evaluasi HOTS yang *pertama kali* menargetkan secara khusus materi klasifikasi makhluk hidup di tingkat MTs. Penelitian ini bertujuan: (1) mengembangkan instrumen evaluasi berbasis HOTS (C4 dan C5) untuk materi klasifikasi makhluk hidup; (2) menganalisis kualitas butir soal; dan (3) menentukan kelayakan konten instrumen berdasarkan penilaian ahli.

2. Metode dan Eksperimen

Penelitian ini menggunakan pendekatan *Research and Development* (R&D) dengan Model 3-D (*Define, Design, Develop*) [1]. Richey dan Klein [17] menegaskan bahwa model ini tepat untuk penelitian desain instruksional karena setiap tahap memasukkan mekanisme evaluasi independen yang menjamin kualitas produk akhir. Subjek penelitian adalah 14 peserta didik kelas VII Fase D di MTs Inovatif Daarul Ihsan Kabupaten Bandung, dipilih melalui *purposive sampling* dengan memastikan keterwakilan kemampuan tinggi, sedang, dan rendah berdasarkan nilai rapor semester sebelumnya. Uji coba dilaksanakan pada tanggal 9 Maret 2026 dengan durasi 40 menit.

Pada tahap *Define*, dilakukan: (1) analisis kurikulum berdasarkan Capaian Pembelajaran IPA Fase D [18] dan indikator C4–C5 Taksonomi Bloom Revisi [2](2) analisis kebutuhan melalui wawancara terstruktur dengan guru IPA; dan (3) pemetaan konsep materi klasifikasi makhluk hidup. Tahap *Design* menghasilkan cetak biru 20 butir soal pilihan ganda dengan distribusi 60% C4 (butir 1–8 dan 13–16) dan 40% C5 (butir 9–12 dan 17–20), dilengkapi dengan stimulus berupa teks, tabel, dan gambar. Tahap *Develop* mencakup validasi ahli, revisi, dan uji coba terbatas kepada 14 peserta didik.

Validitas isi instrumen dinilai melalui *expert judgment* oleh satu ahli pendidikan IPA menggunakan lembar validasi terstruktur. Roebianto et al. [25] menegaskan bahwa untuk instrumen dengan konteks dan dampak terbatas, seperti instrumen kelas di jenjang sekolah menengah, penilaian oleh satu ahli secara metodologis dapat diterima. Persentase kelayakan dihitung sebagai rasio indikator "sesuai" terhadap total indikator dan diinterpretasikan menggunakan kriteria Guilford [24]. Perlu ditegaskan bahwa persentase kelayakan ini merupakan prosedur validitas isi yang berbeda dari penghitungan indeks Aiken's V; keduanya tidak dapat dipertukarkan [25]. Empat prosedur analisis data diterapkan sebagai berikut.

Pertama, persentase kelayakan ahli diinterpretasikan menggunakan kriteria Guilford [24] yang disajikan pada Tabel 1.

Tabel 1. Kriteria Interpretasi Kelayakan Instrumen

Persentase (%)	Kategori	Simpulan
90–100	Sangat Layak	Sudah bisa digunakan tanpa adanya perbaikan
80–89	Layak	Sudah bisa digunakan dengan anjuran melakukan sedikit perbaikan; tidak melakukan perbaikan juga tidak masalah
70–79	Cukup Layak	Dapat digunakan dengan adanya perbaikan
60–69	Tidak Layak	Belum bisa digunakan dan harus melakukan revisi

Sumber: Guilford (1985) [24]

Kedua, reliabilitas dihitung menggunakan formula Spearman-Brown dari teknik *split-half* melalui aplikasi ANATES, teknik ini dipilih karena sesuai untuk tes pilihan ganda dengan distribusi butir ganjil-genap yang setara pada sampel kecil [20]. Formula yang mendasarinya adalah:

$$r_{11} = (2 \times r_{½½}) / (1 + r_{½½})$$

Keterangan: r_{11} = koefisien reliabilitas keseluruhan tes; $r_{½½}$ = korelasi Pearson antara skor butir ganjil dan genap.

Ketiga, indeks kesukaran butir dihitung dan dikategorikan menggunakan ANATES [15]:

$$P = B / N$$

Keterangan: P = indeks kesukaran; B = jumlah peserta yang menjawab benar; N = jumlah seluruh peserta. Kriteria: mudah ($P > 0{,}70$), sedang ($0{,}30 \leq P \leq 0{,}70$), sulit ($P < 0{,}30$) [15].

Keempat, daya pembeda dihitung menggunakan kelompok 27% atas dan bawah melalui ANATES. Formula yang mendasarinya adalah [15–22]:

$$D = (BA / NA) - (BB / NB)$$

Keterangan: D = indeks daya pembeda; BA/BB = banyak peserta kelompok atas/bawah yang menjawab benar; NA/NB = jumlah peserta masing-masing kelompok. Kriteria: sangat baik ($D \geq 0{,}70$), baik ($0{,}40 \leq D < 0{,}70$), cukup ($0{,}20 \leq D < 0{,}40$), kurang ($D < 0{,}20$) [22–15]. Butir direvisi jika $D < 0{,}20$ (kurang) atau $P < 0{,}30$ (sulit); butir dengan D cukup ($0{,}20 \leq D < 0{,}40$) dipertahankan dengan catatan karena masih memberikan diskriminasi minimal yang dapat diterima untuk instrumen tahap awal [22].

3. Hasil dan Pembahasan

Temuan dari setiap tahap Model 3-D disajikan dan didiskusikan secara terpadu, dengan setiap tahap menghasilkan produk yang menjadi masukan bagi tahap berikutnya.

A. Tahap *Define*

1) Analisis Kurikulum

Kurikulum Merdeka mengarahkan pembelajaran agar berpusat pada peserta didik, mendorong kemampuan berpikir kritis dan analitis melalui kegiatan belajar yang bermakna dan kontekstual [18]. Berdasarkan Capaian Pembelajaran (CP) IPA Fase D, peserta didik diharapkan mampu memahami karakteristik makhluk hidup serta mengelompokkan organisme berdasarkan persamaan dan perbedaan ciri. Dimensi bernalar kritis Profil Pelajar Pancasila secara eksplisit mendorong peserta didik untuk menganalisis informasi dan mengevaluasi berbagai kemungkinan berdasarkan data [18]. Temuan ini menetapkan justifikasi regulatif untuk mengembangkan instrumen

berbasis C4 dan C5, sekaligus memastikan kesesuaian instrumen dengan tuntutan kurikulum yang berlaku. Informasi ini selanjutnya menjadi acuan utama dalam merancang indikator butir soal pada tahap *Design.*

2) Analisis Kebutuhan Evaluasi Pembelajaran

Wawancara terstruktur dengan guru IPA di salah satu MTs di Kabupaten Bandung mengungkap bahwa instrumen evaluasi yang digunakan masih didominasi oleh soal yang berorientasi pada kemampuan mengingat (C1) dan memahami (C2). Guru menyampaikan bahwa penyusunan soal berbasis HOTS masih menjadi tantangan utama, dengan keterbatasan referensi contoh soal dan waktu perancangan sebagai faktor penghambat. Temuan ini konsisten dengan Muhayimana et al. [8] dan Delgado dan Luna [7] yang mendokumentasikan dominasi LOTS dalam instrumen evaluasi formal. Kondisi ini secara empiris mengonfirmasi urgensi pengembangan instrumen berbasis HOTS dan memberikan justifikasi kebutuhan lapangan yang memperkuat relevansi penelitian ini. Hasil analisis kebutuhan ini selanjutnya menjadi dasar dalam menentukan spesifikasi butir soal pada tahap *Design.*

3) Analisis Konsep Materi Klasifikasi Makhluk Hidup

Analisis konsep mengidentifikasi lima konsep utama dalam materi klasifikasi makhluk hidup yang menjadi basis pengembangan butir HOTS: (1) *ciri-ciri makhluk hidup*, yang menuntut peserta didik menganalisis dan membedakan karakteristik biologis dari data yang disajikan (C4); (2) *dasar klasifikasi makhluk hidup*, yang mensyaratkan kemampuan mengklasifikasikan organisme berdasarkan persamaan dan perbedaan ciri (C4); (3) *sistem klasifikasi*, yang menuntut evaluasi ketepatan pengelompokan organisme berdasarkan sistem tertentu (C5); (4) *tingkatan takson*, yang memerlukan analisis hubungan kekerabatan organisme berdasarkan hierarki taksonomi (C4); dan (5) *sistem lima kingdom*, yang menuntut evaluasi ketepatan penempatan organisme ke dalam kingdom berdasarkan karakteristiknya (C5) [19]. Hubungan hierarkis antara kelima konsep ini memungkinkan konstruksi butir yang menuntut integrasi pengetahuan. Peserta didik harus memahami konsep dasar sebelum dapat menganalisis atau mengevaluasi kasus yang disajikan. Pemetaan indikator HOTS untuk setiap level kognitif disajikan pada Tabel 2, yang selanjutnya menjadi landasan perancangan cetak biru pada tahap *Design.*

Tabel 2. Indikator Level Kognitif HOTS untuk Materi Klasifikasi Makhluk Hidup

Level	Proses Kognitif	Definisi Operasional
C4	Menganalisis	(1) Peserta didik mampu menganalisis ciri-ciri makhluk hidup dan membedakannya dengan benda tak hidup berdasarkan pengamatan atau data yang disajikan. (2) Peserta didik mampu menganalisis hubungan kekerabatan antarmakhluk hidup berdasarkan tingkatan takson. (3) Peserta didik mampu menganalisis ciri-ciri organisme dan menentukan kingdom yang sesuai berdasarkan sistem lima kingdom.
C5	Mengevaluasi	Peserta didik mampu menjelaskan konsep sistem klasifikasi makhluk hidup dan mengevaluasi ketepatan pengelompokan organisme berdasarkan sistem klasifikasi tertentu.

B. Tahap *Design*

Berdasarkan temuan tahap *Define*, yang mengidentifikasi dominasi LOTS di kelas dan memetakan lima konsep dengan potensi HOTS tinggi, tahap *Design* menghasilkan cetak biru 20 butir soal pilihan ganda dengan distribusi 60% C4 (12 butir) dan 40% C5 (8 butir). Setiap butir dikonstruksi menggunakan stimulus kontekstual berupa teks deskripsi organisme, tabel data morfologis, atau gambar filogeni, sehingga peserta didik harus menganalisis informasi yang disajikan sebelum menentukan jawaban. Lamhatin et al. [10] mengonfirmasi bahwa butir berbasis stimulus autentik lebih valid dalam mengukur berpikir tingkat tinggi daripada format konvensional. Distribusi proporsional butir antartopik dan level kognitif disajikan pada Tabel 3.

Tabel 3. Rancangan Cetak Biru Instrumen Evaluasi Berbasis HOTS

Capaian Pembelajaran	Topik	Indikator HOTS	Level Kognitif	No. Soal	Bentuk Soal
Peserta didik mampu mengidentifikasi dan mengelompokkan makhluk hidup berdasarkan persamaan dan perbedaan ciri.	Ciri-ciri makhluk hidup	Menganalisis perbedaan ciri organisme berdasarkan data yang disajikan	C4	1–4	Pilihan Ganda
Peserta didik mampu menjelaskan dasar klasifikasi makhluk hidup.	Dasar klasifikasi	Mengklasifikasikan organisme berdasarkan persamaan dan perbedaan ciri	C4	5–8	Pilihan Ganda
Peserta didik mampu menjelaskan sistem klasifikasi makhluk hidup.	Sistem klasifikasi	Mengevaluasi ketepatan pengelompokan organisme dalam sistem klasifikasi	C5	9–12	Pilihan Ganda
Peserta didik mampu menganalisis hubungan antar organisme berdasarkan tingkatan takson.	Tingkatan takson	Menganalisis hubungan antar organisme berdasarkan tingkatan takson	C4	13–16	Pilihan Ganda
Peserta didik mampu menentukan pengelompokan organisme berdasarkan sistem lima kingdom.	Sistem lima kingdom	Mengevaluasi ketepatan pengelompokan organisme ke dalam kingdom	C5	17–20	Pilihan Ganda

Proporsi 60:40 antara C4 dan C5 dipilih karena materi klasifikasi makhluk hidup secara struktural lebih kaya pada tuntutan analisis, peserta didik perlu membandingkan dan mengorganisasikan karakteristik organisme, sementara tuntutan evaluasi (menilai ketepatan sistem klasifikasi) menjadi level lanjutan yang memerlukan penguasaan C4 terlebih dahulu. Mohamad et al. [9] menyatakan bahwa instrumen HOTS yang efektif mempertahankan gradasi kognitif yang terstruktur antara level analisis dan evaluasi. Instrumen yang telah dirancang selanjutnya memasuki tahap *Develop.*

C. Tahap Develop

1) Validasi Ahli

Berdasarkan instrumen yang dirancang pada tahap *Design*, validasi isi dilakukan melalui *expert judgment* oleh satu ahli pendidikan IPA menggunakan lembar validasi terstruktur. Roebianto et al. [25] menegaskan bahwa untuk instrumen dengan konteks dan dampak terbatas, penilaian oleh satu ahli secara metodologis dapat diterima. Hasil

penilaian diinterpretasikan sebagai persentase kelayakan, bukan indeks Aiken's V, menggunakan kriteria Guilford [24]. Rekapitulasi hasil disajikan pada Tabel 4.

Tabel 4. Rekapitulasi Hasil Penilaian Ahli (*Expert Judgment*)

Aspek Penilaian	Jml. Indikator	Sesuai	Perlu Revisi	Kelayakan (%)
Kesesuaian materi	2	2	0	100%
Konstruksi soal	3	3	0	100%
Aspek HOTS	2	0	2	0%
Penggunaan bahasa	2	2	0	100%
Kesesuaian instrumen	1	0	1	0%
Total	10	7	3	70%

Sumber: data primer (2026); interpretasi mengacu pada Guilford (1985) [24]

Persentase kelayakan keseluruhan 70% masuk dalam kategori "Cukup Layak" berdasarkan kriteria Guilford [24], yang berarti instrumen dapat digunakan dengan adanya perbaikan, menjawab tujuan penelitian ketiga mengenai kelayakan konten instrumen pada tahap validasi. Aspek kesesuaian materi (100%), konstruksi soal (100%), dan penggunaan bahasa (100%) telah memenuhi kriteria sepenuhnya. Sebaliknya, aspek kemampuan HOTS (0%) dan kesesuaian instrumen (0%) memerlukan revisi substantif, yang berarti kata kerja operasional pada butir soal belum sepenuhnya mencerminkan tuntutan kognitif C4–C5 dan pilihan jawaban pengecoh belum mendiskriminasi secara optimal.

Revisi dilakukan melalui dua tindakan: (a) mengganti kata kerja operasional setiap butir agar selaras dengan deskriptor C4 (menganalisis, membedakan, mengorganisasikan) dan C5 (mengevaluasi, mengkritisi, memutuskan) [2]; dan (b) merekonstruksi pilihan jawaban pengecoh agar memerlukan pemahaman konseptual mendalam sebelum dapat dieliminasi, bukan sekadar intuisi. Temuan ini konsisten dengan Mohamad et al. [9] yang menetapkan keselarasan antara tuntutan kognitif butir dan indikator sebagai prasyarat keberhasilan instrumen HOTS. Perlu dicatat bahwa keterlibatan satu validator merupakan keterbatasan metodologis yang berpengaruh pada tingkat objektivitas penilaian [25]; perluasan ke panel minimal tiga validator direkomendasikan pada penelitian lanjutan.

2) Reliabilitas

Setelah revisi berdasarkan masukan ahli, instrumen diujicobakan kepada 14 peserta didik. Reliabilitas dianalisis menggunakan teknik *split-half* melalui aplikasi ANATES dengan formula Spearman-Brown. Hasil analisis disajikan pada Tabel 5.

Tabel 5. Hasil Uji Reliabilitas

Komponen	Nilai
Rata-rata skor	14,93
Simpangan baku	3,45
Korelasi XY (r½½)	0,69
Koefisien Reliabilitas (r_{11})	0,82
Kategori	Sangat Tinggi

Koefisien reliabilitas r_{11} = 0,82 masuk dalam kategori sangat tinggi, melampaui ambang batas minimum yang diterima ($r_{11} \geq 0{,}70$) [20]. Korelasi ganjil-genap $r_{½½}$ = 0,69 menunjukkan konsistensi internal yang kuat antarbutir soal. Nilai ini mengonfirmasi bahwa instrumen menghasilkan pengukuran yang stabil, selaras dengan Ayubi et al. [12]. Meskipun demikian, pada sampel kecil (n = 14), estimasi reliabilitas memiliki rentang kepercayaan yang lebar sehingga nilai r_{11} = 0,82 bersifat indikatif dan perlu dikonfirmasi pada sampel yang lebih besar.

3) Analisis Kualitas Butir Soal

Analisis butir soal dilakukan menggunakan aplikasi ANATES terhadap 20 butir yang diujicobakan. Kriteria yang digunakan: tingkat kesukaran mudah (P > 0,70), sedang (0,30 ≤ P ≤ 0,70), sulit (P < 0,30); daya pembeda sangat baik (D ≥ 0,70), baik (0,40 ≤ D < 0,70), cukup (0,20 ≤ D < 0,40), kurang (D < 0,20). Butir direvisi jika D kurang (D < 0,20) atau P sulit (P < 0,30). Hasil selengkapnya disajikan pada Tabel 6.

Tabel 6. Hasil Analisis Butir Soal

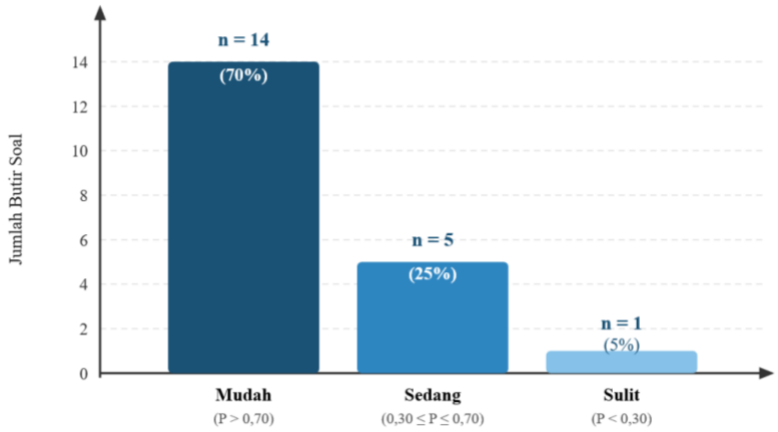
a) Tingkat Kesukaran

No.	Tk. Kesukaran (P)	Kategori Kesukaran	Daya Pembeda (D)	Kategori Daya Pembeda	Keputusan
1	0,93	Mudah	0,25	Cukup	Digunakan
2	0,79	Mudah	0,00	Kurang	Direvisi
3	0,79	Mudah	0,00	Kurang	Direvisi
4	0,86	Mudah	0,50	Baik	Digunakan
5	1,00	Mudah	0,00	Kurang	Direvisi
6	0,79	Mudah	0,25	Cukup	Digunakan
7	0,86	Mudah	0,50	Baik	Digunakan
8	0,86	Mudah	0,00	Kurang	Direvisi
9	0,86	Mudah	0,50	Baik	Digunakan
10	0,57	Sedang	0,25	Cukup	Digunakan
11	0,93	Mudah	0,25	Cukup	Digunakan
12	0,86	Mudah	0,50	Baik	Digunakan
13	0,86	Mudah	0,50	Baik	Digunakan
14	0,43	Sedang	0,00	Kurang	Direvisi
15	0,86	Mudah	0,50	Baik	Digunakan
16	0,14	Sulit	0,25	Cukup	Direvisi
17	0,64	Sedang	0,75	Sangat Baik	Digunakan
18	0,79	Mudah	0,75	Sangat Baik	Digunakan
19	0,50	Sedang	1,00	Sangat Baik	Digunakan
20	0,64	Sedang	0,75	Sangat Baik	Digunakan

Hasil analisis menunjukkan 14 butir (70%) berkategori mudah (P > 0,70), lima butir (25%) sedang (butir 10, 14, 17, 19, dan 20), dan satu butir (5%) sulit (butir 16, P = 0,14). Dominasi butir mudah. Paling ekstrem pada butir 5 dengan P = 1,00 yang dijawab benar oleh seluruh 14 peserta mencerminkan bahwa konsep dasar materi telah dikuasai oleh sebagian besar peserta didik pascapembelajaran. Butir dengan P = 1,00 tidak menghasilkan informasi psikometri yang bermakna karena tidak mampu membedakan kemampuan peserta didik sama sekali [15]; butir ini masuk dalam daftar revisi. Variasi tingkat kesukaran yang lebih seimbang diperlukan agar instrumen mampu mengukur rentang

kemampuan peserta didik secara komprehensif. Kim et al. [15] menegaskan bahwa pada sampel kecil (n < 30), indeks kesukaran sangat dipengaruhi oleh karakteristik sampel sehingga distribusi ini bersifat indikatif.

Dominasi butir soal berkategori mudah kemungkinan dipengaruhi oleh ukuran sampel yang relatif kecil pada tahap uji coba terbatas. Pada sampel yang terbatas, distribusi tingkat kesukaran dapat dipengaruhi oleh karakteristik kemampuan peserta didik sehingga hasil analisis masih bersifat indikatif dan perlu diuji kembali pada sampel yang lebih besar. Distribusi tingkat kesukaran disajikan pada Gambar 1.



Gambar 1. Distribusi Tingkat Kesukaran Butir Soal

b) Daya Pembeda

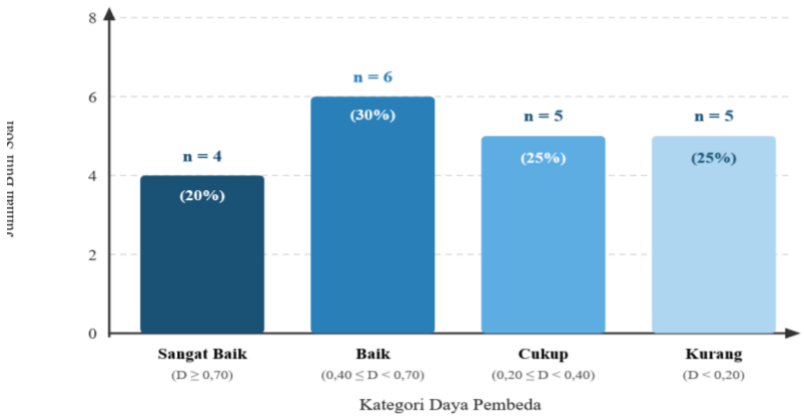
Hasil analisis daya pembeda menunjukkan empat butir sangat baik (D ≥ 0,70; butir 17, 18, 19, 20 dengan D = 0,75–1,00), enam butir baik (0,40 ≤ D < 0,70; butir 4, 7, 9, 12, 13, 15 dengan D = 0,50), lima butir cukup (0,20 ≤ D < 0,40; butir 1, 6, 10, 11, 16 dengan D = 0,25), dan lima butir kurang (D < 0,20; butir 2, 3, 5, 8, 14 dengan D = 0,00). Total enam butir direvisi: lima butir berkategori kurang (D = 0,00) dan satu butir berkategori sulit meskipun daya pembedanya cukup (butir 16, P = 0,14).

Lima butir dengan D = 0,00 memerlukan analisis mendalam karena mencerminkan dua mekanisme yang berbeda. Butir 2, 3, 5, dan 8 semuanya berkategori mudah (P ≥ 0,79) dan memiliki D = 0,00. Pola ini mencerminkan hubungan negatif antara tingkat kesukaran dan daya pembeda yang sudah terdokumentasi dengan baik dalam psikometri [23]: ketika hampir semua peserta menjawab benar, kelompok atas dan bawah memiliki proporsi jawaban benar yang identik sehingga D mendekati nol. Stimulus pada keempat butir ini belum cukup menuntut proses analisis mendalam, sehingga peserta dari semua tingkat kemampuan dapat menjawab dengan cara yang sama. Butir 14 (P = 0,43, sedang) dengan D = 0,00 menunjukkan kegagalan konstruksi yang berbeda: stimulusnya kemungkinan ambigu sehingga peserta berkemampuan tinggi tidak dapat mengidentifikasi jawaban yang lebih baik daripada peserta berkemampuan rendah.

Butir 16 direvisi bukan karena D-nya kurang (D = 0,25, masuk kategori cukup), melainkan karena tingkat kesukarannya sangat tinggi (P = 0,14, sulit); dalam konteks uji coba terbatas dengan n = 14, butir yang sangat sulit tidak memberikan data item yang dapat diandalkan.

Nilai daya pembeda yang sangat rendah pada beberapa butir menunjukkan bahwa butir tersebut belum mampu membedakan peserta didik dengan kemampuan tinggi dan rendah secara optimal. Hal ini kemungkinan disebabkan oleh tingkat kesukaran yang terlalu mudah atau kurang efektifnya pengecoh (distractor) pada pilihan jawaban, sehingga butir-butir tersebut perlu direvisi agar dapat memberikan diskriminasi yang lebih baik pada pengujian selanjutnya.

Lima butir dengan D = 0,25 (butir 1, 6, 10, 11) dipertahankan karena kategori cukup ($0{,}20 \leq D < 0{,}40$) masih memberikan diskriminasi minimal yang dapat diterima untuk instrumen tahap awal [22]; keputusan ini konsisten dengan Rezigalla et al. [23] yang menyatakan bahwa butir cukup tetap berkontribusi pada estimasi reliabilitas total. Butir-butir ini direkomendasikan untuk diuji ulang pada sampel lebih besar sebelum penggunaan formal. Butir berkategori sangat baik (D = 0,75–1,00) terkonsentrasi pada topik sistem lima kingdom (butir 17–20), karena butir-butir ini menuntut integrasi informasi dari beberapa kingdom secara bersamaan, yaitu sebuah tuntutan kognitif yang lebih tinggi, sehingga menghasilkan daya diskriminasi terbesar. Hal ini mengonfirmasi bahwa stimulus evaluatif yang menuntut sintesis multidimensi menghasilkan daya pembeda lebih tinggi secara konsisten [23]. Distribusi daya pembeda disajikan pada gambar. 2.



Gambar 2. Distribusi Daya Pembeda Butir Soal

Secara keseluruhan, instrumen mencapai profil psikometri yang memadai: reliabilitas sangat tinggi (r_{11} = 0,82), 14 dari 20 butir (70%) siap digunakan langsung, dan empat butir (20%) menunjukkan daya pembeda sangat baik. Profil ini memenuhi tiga tujuan penelitian yang ditetapkan: instrumen telah dikembangkan (tujuan 1), kualitas butir telah dianalisis (tujuan 2), dan kelayakan konten telah ditentukan melalui penilaian ahli dengan hasil 70%, cukup layak dengan perbaikan (tujuan 3).

4. Kesimpulan

Hasil penelitian menunjukkan bahwa instrumen yang dikembangkan terdiri dari 20 butir soal pilihan ganda yang secara proporsional menargetkan kemampuan menganalisis (C4) sebesar 60% dan mengevaluasi (C5) sebesar 40%. Berdasarkan hasil validasi ahli, instrumen memperoleh tingkat kelayakan konten sebesar 70% yang termasuk kategori cukup layak menurut kriteria Guilford, dengan perbaikan terfokus pada peningkatan tuntutan kognitif HOTS dan kesesuaian konstruk soal. Uji reliabilitas menggunakan metode split-half (Spearman-Brown) menghasilkan koefisien sebesar 0,82 yang menunjukkan konsistensi internal sangat tinggi. Analisis butir soal mengindikasikan variasi kualitas, dengan sebagian besar butir berada pada kategori baik hingga sangat baik, meskipun masih terdapat butir dengan daya pembeda rendah yang memerlukan revisi. Secara keseluruhan, temuan ini menegaskan bahwa instrumen yang dikembangkan telah memenuhi kriteria dasar sebagai alat evaluasi HOTS yang layak digunakan dengan beberapa penyempurnaan. Lebih lanjut, penelitian ini membuka peluang pengembangan lanjutan baik dari segi penyempurnaan kualitas butir maupun perluasan implementasi. Instrumen ini berpotensi diaplikasikan sebagai alat evaluasi pembelajaran berbasis HOTS di tingkat MTs, khususnya pada materi klasifikasi makhluk hidup, serta dapat menjadi referensi bagi pengembangan instrumen serupa pada materi lain. Oleh karena itu, penelitian selanjutnya disarankan untuk melanjutkan ke tahap disseminate dengan melibatkan guru sebagai pengguna utama, memperluas jumlah sampel agar parameter psikometri lebih stabil, menambah jumlah validator untuk memperkuat validitas isi, serta menguji kembali butir yang telah direvisi sebelum digunakan secara lebih luas. Dengan demikian, hasil penelitian ini tidak hanya menjawab tujuan penelitian, tetapi juga memberikan kontribusi praktis dan arah pengembangan ke depan dalam evaluasi pembelajaran berbasis HOTS.

Ucapan Terima Kasih

Penulis menyampaikan terima kasih kepada kepala sekolah, guru IPA, dan peserta didik MTs Al-Jawahir dan MTs Inovatif Daarul Ihsan Kabupaten Bandung atas izin, dukungan, dan partisipasi aktif selama proses penelitian, validasi, dan uji coba instrumen berlangsung.

Daftar Pustaka

[1] S. Thiagarajan, D. S. Semmel, and M. I. Semmel, Instructional development for training teachers of exceptional children: A sourcebook. Indiana: Indiana University, 1974.

[2] L. W. Anderson and D. R. Krathwohl, Eds., A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. New York: Longman, 2001.

[3] B. Thornhill-Miller et al., "Creativity, critical thinking, communication, and collaboration: Assessment, certification, and promotion of 21st century skills for the future of work and education," Journal of Intelligence, vol. 11, no. 3, p. 54, 2023. doi: 10.3390/jintelligence11030054.

[4] L. F. Santos-Meneses and E. A. Drugova, "Trends in critical thinking instruction in 21st-century research and practice," Thinking Skills and Creativity, vol. 49, p. 101383, 2023. doi: 10.1016/j.tsc.2023.101383.

[5] N. Sepriyanti, S. Nelwati, M. Kustati, and J. Afriadi, "The effect of 21st-century learning on higher-order thinking skills (HOTS) and numerical literacy of science students in Indonesia based on gender," Jurnal Pendidikan IPA Indonesia, vol. 11, no. 2, pp. 314–321, 2022. doi: 10.15294/jpii.v11i2.36384.

[6] D. Darmawan, D. Yatimah, K. Sasmita, and R. Syah, "Analysis of non-formal education tutor capabilities in exploring assessment for science learning," Jurnal Pendidikan IPA Indonesia, vol. 9, no. 2, pp. 267–275, 2020. doi: 10.15294/jpii.v9i2.24025.

[7] F. J. Delgado and M. J. Luna, "Designing assessments for higher-order thinking in science education," International Journal of Science Education, vol. 44, no. 9, pp. 1389–1406, 2022. doi: 10.1080/09500693.2022.2051201.

[8] T. Muhayimana, L. Kwizera, and M. R. Nyirahabimana, "Using Bloom's taxonomy to evaluate the cognitive levels of Primary Leaving English Exam questions in Rwandan schools," The Language Learning Journal, vol. 50, no. 1, pp. 51–63, 2022. doi: 10.1007/s41297-021-00156-2.

[9] S. N. A. Mohamad, N. A. Shukor, and Z. Tasir, "Systematic literature review on the elements of metacognition-based higher order thinking skills (HOTS) teaching and learning modules," Sustainability, vol. 14, no. 2, p. 813, 2022. doi: 10.3390/su14020813.

[10] F. Lamhatin, D. A. Fajarianingtyas, and A. Anekawati, "Pengembangan instrumen penilaian HOTS memuat keterampilan 4C menuju pembelajaran abad 21," EKSAKTA: Jurnal Penelitian dan Pembelajaran MIPA, vol. 7, no. 1, pp. 30–38, 2022. doi: 10.31604/eksakta.v7i1.30-38.

[11] L. D. Kurnia, S. Haryati, and R. Linda, "Pengembangan instrumen evaluasi Higher Order Thinking Skills menggunakan Quizizz pada materi termokimia," Jurnal Pendidikan Sains Indonesia, vol. 10, no. 1, pp. 176–190, 2022. doi: 10.24815/jpsi.v10i1.21727.

[12] M. Ayubi, J. Ikhsan, V. D. Arthamena, M. A. Chaniago, and E. Pradesta, "Validity and reliability of HOTS instrument for voltaic cell subject using the classical method," Unibulletin, vol. 13, no. 1, pp. 53–64, 2024. doi: 10.22521/unibulletin.2024.131.4.

[13] S. Nurjanah, E. Istiyono, W. Widihastuti, M. Iqbal, and S. Kamal, "The application of Aiken's V method for evaluating the content validity of instruments that measure the implementation of formative assessments," Journal of Research and Educational Research Evaluation, vol. 12, no. 2, pp. 125–133, 2023. doi: 10.15294/jere.v12i2.76451.

[14] B. H. Ahmad, M. H. Abd Razak, and M. K. Muhamad Fuad, "Learning management system instrument development based on Aiken's V technique," International Journal of Evaluation and Research in Education, vol. 13, no. 5, pp. 2924–2931, 2024. doi: 10.11591/ijere.v13i5.28925.

[15] Y. H. Kim, B. H. Kim, J. Kim, B. Jung, and S. Bae, "Item difficulty index, discrimination index, and reliability of the 26 health professions licensing examinations in 2022, Korea: A psychometric study," Journal of Educational Evaluation for Health Professions, vol. 20, p. 31, 2023. doi: 10.3352/jeehp.2023.20.31.

[16] S. Fuadiyah, G. H. Selaras, D. Melta, and D. Rahmi, "The urgency of higher order thinking skills (HOTS) based on development of biology assessment instruments for class XI students," in Proc. 3rd Int. Conf. Biology, Science and Education (IcoBioSE 2021), Atlantis Press, 2023, pp. 3–8. doi: 10.2991/978-94-6463-166-1_2.

[17] R. C. Richey and J. D. Klein, Design and development research: Methods, strategies, and issues. Routledge, 2021. doi: 10.4324/9781003090861.

[18] Kemendikbudristek, Keputusan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi tentang Capaian Pembelajaran pada Pendidikan Anak Usia Dini, Jenjang Pendidikan Dasar, dan Jenjang Pendidikan Menengah. Jakarta: Kemendikbudristek, 2022.

[19] V. Inabuy, A. Setiawan, L. Rusyati, and S. Feranie, "The effect of STEM-based learning on students' conceptual understanding of living things classification," Journal of Physics: Conference Series, vol. 1806, p. 012049, 2021. doi: 10.1088/1742-6596/1806/1/012049.

[20] L. V. Aiken, "Three coefficients for analyzing the reliability and validity of ratings," Educational and Psychological Measurement, vol. 45, no. 1, pp. 131–142, 1985. doi: 10.1177/0013164485451012.

[21] A. Latifah, F. Husaini, and A. Khoirotun Nisa, "Pengembangan instrumen penilaian berbasis HOTS," Didaktik: Jurnal Ilmiah PGSD STKIP Subang, vol. 9, no. 2, pp. 4486–4496, 2023. doi: 10.36989/didaktik.v9i2.1057.

[22] C. N. P. Olipas and R. G. Luciano, "Analyzing test performance of BSIT students and question quality: A study on item difficulty index and item discrimination index for test question improvement," International Journal of Information Technology and Computer Science (IJITCS), vol. 16, no. 3, pp. 1–11, 2024. doi: 10.5815/ijitcs.2024.03.01.

[23] A. A. Rezigalla, A. M. S. A. Eleragi, A. B. Elhussein, J. Alfaidi, M. A. Alghamdi, and A. Y. Al Ameer, "Item analysis: The impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items," BMC Medical Education, vol. 24, p. 445, 2024. doi: 10.1186/s12909-024-05433-y.

[24] J. P. Guilford, Fundamental Statistics in Psychology and Education, 5th ed. New York: McGraw-Hill, 1985.

[25] A. Roebianto, S. I. Savitri, I. Aulia, A. Suciyana, and L. Mubarokah, "Content validity: Definition and procedure of content validation in psychological research," Testing, Psychometrics, Methodology in Applied Psychology (TPM), vol. 30, no. 1, pp. 5–18, 2023. doi: 10.4473/TPM30.1.1.