Jurnal Pendidikan Ilmu Pengetahuan Alam (JP-IPA) **Vol. 07 No. 01. Februari 2026**
https://jurnal.habi.ac.id/index.php/JP-IPA *p-ISSN:* 2774-5945. *e-ISSN:* 2774-5937
Doi: https://doi.org/10.56842/jp-ipa Terakreditasi SINTA 5

EVALUASI PSIKOMETRIK INSTRUMEN KEMAMPUAN BERPIKIR TINGKAT TINGGI PADA MATERI SISTEM ORGAN MANUSIA

Mely Fatmawati[1*], Ida Farida[2], & Ade Yeti Nuryantini[3]
[1-3]UIN Sunan Gunung Djati Bandung, Indonesia
**Email: 2225945011@student.uinsgd.ac.id*

Diterima: 14 Januari 2026 *Direvisi: 08 Februari 2026* *Publikasi: 10 Februari 2026*

Abstract

This study aims to evaluate the psychometric quality of a higher order thinking skills instrument in the topic of the human organ system for junior high school science learning. A descriptive quantitative approach was applied to analyze the validity, reliability, difficulty index, discrimination power, and distractor effectiveness. The instrument consisted of ten multiple-choice items developed based on the revised Bloom's taxonomy at levels C4-C6. Content validity was examined by two experts using Aiken's V, while empirical validity was tested with the Pearson Product-Moment correlation. The reliability coefficient was calculated using KR-20. The results show that the content validity values ranged from 0.83 to 0.94 (very valid), and the reliability coefficient reached 0.78 (high category). Eight items demonstrated moderate difficulty, seven items had good discrimination power, and 70 percent of distractors functioned effectively. These findings indicate that the developed HOTS instrument is psychometrically sound and feasible for strengthening 21st-century science assessment, although several items require minor revision to improve balance in difficulty and distractor performance.

Keywords: *Psychometric Analysis; HOTS; Science Assessment; Human Organ System*

Abstrak

Penelitian ini bertujuan untuk mengevaluasi kualitas psikometrik instrumen kemampuan berpikir tingkat tinggi pada materi sistem organ manusia dalam pembelajaran IPA di tingkat SMP. Pendekatan yang digunakan adalah deskriptif kuantitatif dengan fokus pada analisis validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas distraktor. Instrumen berupa sepuluh butir soal pilihan ganda dikembangkan berdasarkan Taksonomi Bloom Revisi pada level C4–C6. Validitas isi dinilai oleh dua ahli menggunakan rumus Aiken's V, sedangkan validitas empiris diuji dengan korelasi Product Moment. Reliabilitas dihitung menggunakan rumus KR-20. Hasil penelitian menunjukkan nilai validitas isi berada pada rentang 0,83–0,94 (kategori sangat valid) dan reliabilitas sebesar 0,78 (kategori tinggi). Delapan butir berada pada tingkat kesukaran sedang, tujuh butir memiliki daya pembeda baik, dan 70 persen distraktor berfungsi efektif. Secara keseluruhan, instrumen kemampuan berpikir tingkat tinggi yang dikembangkan memenuhi kelayakan psikometrik dan dapat digunakan untuk memperkuat asesmen sains abad ke-21, meskipun beberapa butir masih memerlukan revisi kecil agar keseimbangan tingkat kesukaran dan efektivitas distraktor lebih optimal.

Kata kunci: *Analisis Psikometrik; Berpikir Tingkat Tinggi; Asesmen Sains; Sistem Organ Manusia*

PENDAHULUAN

Perubahan paradigma pendidikan sains dalam dekade terakhir menuntut asesmen yang tidak hanya mengukur kemampuan kognitif dasar, tetapi juga kemampuan berpikir tingkat tinggi yang menjadi esensi pembelajaran abad ke-21. Sejalan dengan kebijakan Kurikulum Merdeka, asesmen sains di sekolah diharapkan mampu menilai kemampuan analisis, evaluasi, dan kreasi peserta didik terhadap fenomena ilmiah yang kontekstual (Brookhart, 2010; Kim et al., 2021). Namun, berbagai kajian menunjukkan bahwa instrumen penilaian kemampuan berpikir tingkat tinggi yang

dikembangkan guru masih belum memenuhi karakteristik psikometrik yang memadai (Retnawati et al., 2018; Yustiqvar & Fauzi, 2022).

Penelitian terdahulu oleh Azizah dan Retnawati (2019) menyoroti kesulitan guru dalam menyusun soal berbasis kemampuan berpikir tingkat tinggi yang valid dan reliabel. Susanti dan Kurniawati (2021) menegaskan bahwa sebagian besar instrumen yang dikembangkan guru hanya menuntut kemampuan C1–C3 dalam taksonomi Bloom revisi. Sementara itu, penelitian oleh Zulkardi dan Putri (2021) menggunakan pendekatan model Rasch menunjukkan bahwa banyak butir soal sains belum mencapai unidimensionalitas dan memiliki distribusi kesukaran yang tidak seimbang. Studi Malau dan Sari (2023) menambahkan bahwa konteks autentik seperti isu kesehatan masyarakat dapat meningkatkan daya pembeda butir karena menuntut penalaran kritis siswa.

Pemilihan materi sistem organ manusia dalam penelitian ini didasarkan pada karakteristiknya yang kompleks, integratif, dan kontekstual. Materi ini menuntut peserta didik untuk tidak hanya memahami fungsi organ secara terpisah, tetapi juga menganalisis keterkaitan antarsistem organ dalam menjaga homeostasis serta mengevaluasi berbagai gangguan kesehatan yang sering dijumpai dalam kehidupan sehari-hari. Secara teoretis, topik sistem organ manusia memiliki potensi tinggi untuk mengukur kemampuan berpikir tingkat tinggi karena melibatkan analisis sebab-akibat, penalaran sistemik, serta evaluasi berbasis data biologis (Sari & Prasetyo, 2023).

Selain itu, isu kesehatan seperti gangguan pernapasan, pencernaan, dan ekskresi merupakan konteks autentik yang relevan dengan pengalaman peserta didik, sehingga efektif digunakan sebagai stimulus soal kemampuan berpikir tingkat tinggi (Rosdiana et al., 2020).

Selain itu, aspek validitas isi dan empiris sering diabaikan dalam pengembangan instrumen kemampuan berpikir tingkat tinggi. Johanson dan Brooks (2010) menekankan pentingnya uji coba instrumen skala kecil untuk memeriksa performa butir sebelum digunakan secara luas. Mardapi (2017) dan Azwar (2021) menjelaskan bahwa validitas empiris, reliabilitas KR-20, serta efektivitas distraktor merupakan indikator penting dalam menjamin keandalan alat ukur pendidikan. Sayangnya, hasil telaah Retnawati et al. (2018) memperlihatkan bahwa sebagian besar soal kemampuan berpikir tingkat tinggi guru IPA tidak lolos uji validitas empiris dan memiliki reliabilitas rendah.

Beberapa penelitian terbaru mencoba memperbaiki hal tersebut dengan pendekatan berbasis konteks kehidupan nyata. Sari dan Prasetyo (2023) serta Rosdiana, Putri, dan Rahmawati (2020) menunjukkan bahwa konteks sistem organ manusia memberikan peluang besar untuk mengukur kemampuan analisis sebab-akibat dan pemahaman konseptual. Namun, kajian mendalam tentang kualitas psikometrik instrumen kemampuan berpikir tingkat tinggi khusus pada topik sistem organ manusia masih terbatas, terutama pada jenjang SMP di Indonesia. Hal ini menunjukkan adanya gap penelitian dalam konteks pengembangan dan evaluasi instrumen kemampuan berpikir tingkat tinggi berbasis IPA yang valid, reliabel, serta efektif untuk mengukur kemampuan berpikir tingkat tinggi.

Urgensi penelitian ini terletak pada kebutuhan mendesak akan instrumen asesmen sains yang tidak hanya valid secara teoretis, tetapi juga empiris, sehingga dapat memperkuat praktik asesmen formatif maupun sumatif dalam pembelajaran IPA. Pengembangan instrumen yang memiliki kualitas psikometrik

tinggi akan membantu guru memperoleh gambaran kemampuan siswa secara objektif dan akurat (Crocker & Algina, 2008; Ebel & Frisbie, 1991).

Berdasarkan celah penelitian tersebut, penelitian ini bertujuan untuk mengevaluasi kualitas psikometrik instrumen kemampuan berpikir tingkat tinggi pada materi sistem organ manusia yang mencakup analisis validitas isi, validitas empiris, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas distraktor. Hasil penelitian ini diharapkan dapat menjadi rujukan empiris dalam pengembangan asesmen kemampuan berpikir tingkat tinggi berbasis IPA yang kontekstual, sahih, dan sesuai dengan tuntutan pembelajaran abad ke-21.

METODE PENELITIAN

Desain Penelitian

Penelitian ini menggunakan pendekatan deskriptif kuantitatif untuk mengevaluasi kualitas psikometrik instrumen kemampuan berpikir tingkat tinggi pada materi sistem organ manusia. Pendekatan ini dipilih karena sesuai dengan tujuan penelitian, yaitu menggambarkan secara empiris karakteristik validitas, reliabilitas, tingkat kesukaran, daya pembeda, serta efektivitas distraktor dari suatu instrumen asesmen. Penelitian deskriptif kuantitatif memberikan gambaran faktual berbasis data numerik tanpa memanipulasi variabel penelitian, sehingga hasilnya dapat merepresentasikan performa instrumen secara objektif.

Objek dan Partisipan Penelitian

Objek penelitian adalah instrumen tes kemampuan berpikir tingkat tinggi pada mata pelajaran Ilmu Pengetahuan Alam (IPA) yang dikembangkan berdasarkan Taksonomi Bloom Revisi pada level kognitif C4 (analisis), C5 (evaluasi), dan C6 (kreasi). Uji coba dilakukan terhadap 30 peserta didik kelas VIII salah satu SMP di Kabupaten Bandung yang dipilih menggunakan teknik *purposive sampling*, dengan pertimbangan kesesuaian materi ajar dan homogenitas kemampuan akademik. Jumlah partisipan memenuhi kriteria minimal pengujian instrumen skala kecil sebagaimana direkomendasikan oleh Johanson dan Brooks (2010), yaitu 25–30 responden untuk instrumen dengan jumlah butir di bawah 20.

Instrumen Penelitian

Instrumen penelitian berupa sepuluh butir soal pilihan ganda berbasis kemampuan berpikir tingkat tinggi yang dikembangkan oleh peneliti dengan mengacu pada kisi-kisi instrumen yang memuat kompetensi dasar, indikator, level kognitif, dan konteks stimulus. Setiap butir soal dilengkapi dengan stimulus kontekstual seperti grafik fungsi organ, ilustrasi sistem tubuh, dan kasus gangguan kesehatan. Instrumen ini dirancang untuk menilai kemampuan siswa dalam menganalisis hubungan antarsistem organ, mengevaluasi situasi kesehatan, dan merancang solusi berbasis bukti ilmiah. Kisi-kisi instrumen disusun dengan mempertimbangkan kesesuaian konten, konstruksi soal, dan bahasa yang komunikatif.

Validitas isi instrumen diuji oleh dua validator ahli, yaitu dosen bidang pendidikan IPA dan ahli evaluasi pembelajaran. Penilaian dilakukan dengan skala Likert 1–5 dan dianalisis menggunakan Aiken's V untuk menentukan tingkat kesepakatan antarvalidator. Nilai $V \geq 0{,}80$ dikategorikan sangat valid.

Teknik Pengumpulan Data

Pengumpulan data dilakukan melalui tes tertulis dengan durasi 30 menit pada kondisi kelas yang terkontrol. Sebelum pelaksanaan, peserta diberi penjelasan mengenai tujuan penelitian dan jaminan kerahasiaan data untuk menjaga etika penelitian. Setiap jawaban peserta didik dikodekan secara anonim untuk memastikan objektivitas analisis. Data hasil

jawaban dikonversi menjadi skor 1 untuk jawaban benar dan 0 untuk jawaban salah.

Teknik Analisis Data

Analisis data dilakukan melalui lima tahap utama:

1. Validitas Isi (*Content Validity*)
   Validitas isi instrumen dianalisis menggunakan rumus Aiken's V untuk mengukur tingkat kesepakatan antarvalidator terhadap aspek konten, konstruksi, dan bahasa butir soal. Rumus Aiken's V dituliskan sebagai berikut:
   $$V = \Sigma s / [n (c - 1)]$$
   Keterangan:
   s = r – lo
   r = skor yang diberikan oleh validator
   lo = skor terendah pada skala penilaian
   n = jumlah validator
   c = jumlah kategori penilaian

   Nilai $V \geq 0{,}80$ menunjukkan bahwa butir soal berada pada kategori sangat valid.

2. Validitas Empiris (*Item Validity*)
   Validitas empiris butir soal dihitung menggunakan korelasi Product Moment Pearson antara skor setiap butir dengan skor total tes. Rumus korelasi Product Moment adalah sebagai berikut:
   $$\mathbf{r} = [N\Sigma XY - (\Sigma X)(\Sigma Y)] / \sqrt{\{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]\}}$$
   Keterangan:
   r = koefisien korelasi butir
   X = skor butir soal
   Y = skor total tes
   N = jumlah peserta

   Butir soal dinyatakan valid apabila r-hitung > r-tabel pada taraf signifikansi 5%.

3. Reliabilitas Instrumen
   Reliabilitas instrumen dihitung menggunakan rumus *Kuder Richardson* 20 (KR-20) yang sesuai untuk tes pilihan ganda dengan skor dikotomi. Rumus KR-20 adalah sebagai berikut:
   $$\mathbf{r_{11}} = [k / (k - 1)] [1 - (\Sigma pq / \sigma^2)]$$
   keterangan:
   r_{11} = koefisien reliabilitas tes
   k = jumlah butir soal
   p = proporsi peserta yang menjawab benar pada setiap butir
   $q = 1 - p$
   σ^2 = varians total skor tes

   Instrumen dikatakan memiliki reliabilitas tinggi apabila $r_{11} \geq 0{,}70$.

4. Tingkat Kesukaran (Difficulty Index)
   Tingkat kesukaran butir soal dihitung menggunakan rumus:
   $$\mathbf{P} = B / N$$
   Keterangan:
   P = indeks kesukaran
   B = jumlah peserta yang menjawab benar
   N = jumlah seluruh peserta

   Kriteria tingkat kesukaran butir soal adalah sebagai berikut: $P > 0{,}70$: mudah, $0{,}30 \leq P \leq 0{,}70$: sedang, $P < 0{,}30$: sukar.

5. Daya Pembeda dan Efektivitas Distraktor
   Daya pembeda dihitung menggunakan perbedaan proporsi antara kelompok atas dan bawah (27%) berdasarkan metode Kelley (Ebel & Frisbie, 1991). Distraktor dianggap efektif jika dipilih minimal oleh 5% peserta (Zulkardi & Putri, 2021).

Semua hasil analisis disajikan dalam bentuk tabel dan diinterpretasikan berdasarkan kriteria psikometrik standar untuk menentukan kelayakan instrumen.

HASIL DAN PEMBAHASAN

Hasil analisis menunjukkan bahwa instrumen kemampuan berpikir tingkat tinggi pada materi sistem organ manusia memiliki kualitas psikometrik yang baik pada sebagian besar butir. Nilai validitas isi yang diperoleh dari dua validator ahli berkisar antara 0,83–

0,94, dengan rata-rata 0,89, menunjukkan kategori *sangat valid*. Nilai ini mengindikasikan bahwa isi butir soal telah sesuai dengan indikator pembelajaran, konteks sains, dan level kognitif kemampuan berpikir tingkat tinggi yang diharapkan.

Uji validitas empiris menggunakan korelasi *Product Moment* menunjukkan delapan butir valid dan dua butir perlu revisi minor. Butir yang belum optimal secara empiris tetap dipertahankan untuk dianalisis lebih lanjut karena secara konseptual telah sesuai dengan indikator kemampuan berpikir tingkat tinggi. Reliabilitas instrumen berdasarkan koefisien KR-20 sebesar 0,78, termasuk dalam kategori *tinggi*, yang berarti konsistensi internal antarbutir sangat baik.

Analisis tingkat kesukaran menunjukkan delapan butir berkategori sedang, satu butir mudah, dan satu butir sukar, sehingga distribusi butir relatif seimbang. Pada daya pembeda, tujuh butir berkategori baik ($D \geq 0{,}40$), sedangkan tiga lainnya cukup ($0{,}30 \leq D < 0{,}40$). Selain itu, 70% distraktor berfungsi efektif, menandakan bahwa pengecoh mampu menarik perhatian peserta dari berbagai tingkat kemampuan.

Tabel 1. Ringkasan Analisis Validitas, Reliabilitas, Tingkat Kesukaran, Daya Pembeda, dan Efektivitas Distraktor

No	Validitas (r)	Kategori	P	Kesukaran	D	Daya Pembeda	Distraktor Efektif	Kategori Akhir
1	0.45	Cukup	0.52	Sedang	0.38	Cukup	2/3	Revisi minor
2	0.52	Cukup	0.48	Sedang	0.44	Baik	3/3	Sangat baik
3	0.61	Tinggi	0.44	Sedang	0.50	Baik	3/3	Sangat baik
4	0.49	Cukup	0.38	Sukar	0.31	Cukup	1/3	Revisi mayor
5	0.58	Cukup	0.72	Mudah	0.25	Kurang	1/3	Revisi mayor
6	0.54	Cukup	0.56	Sedang	0.56	Baik	3/3	Sangat baik
7	0.47	Cukup	0.41	Sedang	0.31	Cukup	2/3	Revisi minor
8	0.63	Tinggi	0.59	Sedang	0.50	Baik	3/3	Sangat baik
9	0.59	Cukup	0.65	Sedang	0.44	Baik	3/3	Sangat baik
10	0.56	Cukup	0.47	Sedang	0.38	Cukup	2/3	Revisi minor

Temuan penelitian ini memperlihatkan bahwa sebagian besar butir instrumen kemampuan berpikir tingkat tinggi yang dikembangkan telah memenuhi kriteria psikometrik yang baik. Hasil validitas isi yang tinggi menunjukkan bahwa konstruk soal selaras dengan indikator kemampuan analisis, evaluasi, dan kreasi dalam Taksonomi Bloom Revisi (Anderson & Krathwohl, 2001).

Temuan Brookhart (2010) dan Kim et al. (2021) bahwa keterpaduan antara stimulus autentik dan konteks kehidupan nyata meningkatkan kesesuaian isi terhadap domain kognitif yang diukur. Koefisien reliabilitas sebesar 0,78 menandakan konsistensi internal yang kuat. Nilai ini sejalan dengan kriteria yang ditetapkan oleh Crocker dan Algina (2008), yang menyatakan bahwa reliabilitas ≥0,70 tergolong tinggi. Konsistensi ini menunjukkan bahwa seluruh butir mengukur konstruk yang sama, yaitu kemampuan berpikir tingkat tinggi dalam konteks sistem organ manusia.

Sebagian besar butir memiliki tingkat kesukaran sedang, sesuai prinsip konstruksi tes yang baik (Arikunto, 2010). Butir dengan kesukaran ekstrem (terlalu mudah atau sukar) dapat mengurangi daya diskriminatif instrumen. Temuan ini selaras dengan hasil penelitian Susanti dan Kurniawati (2021) yang menunjukkan bahwa distribusi butir sedang memberikan hasil evaluasi yang lebih

representatif terhadap variasi kemampuan peserta.

Dari sisi daya pembeda, tujuh butir berkategori baik menunjukkan kemampuan instrumen dalam membedakan peserta berkemampuan tinggi dan rendah. Nilai D tertinggi (0,56) terdapat pada butir nomor 6, yang memuat konteks isu kesehatan masyarakat tentang paparan asap rokok. Temuan ini mendukung penelitian Malau dan Sari (2023), bahwa konteks sosial yang relevan dapat menstimulasi penalaran kritis siswa sehingga meningkatkan daya pembeda butir.

Analisis distraktor menunjukkan bahwa sebagian besar pengecoh berfungsi efektif. Distraktor yang berfungsi dengan baik menandakan bahwa butir tidak bias dan mampu mengukur kemampuan konseptual secara adil (Zulkardi & Putri, 2021). Butir yang memiliki distraktor tidak efektif (seperti nomor 4 dan 5) perlu diperbaiki agar pilihan jawaban salah tetap logis bagi peserta didik berkemampuan rendah, sebagaimana disarankan oleh Ebel dan Frisbie (1991).

Hasil ini menegaskan bahwa penggunaan stimulus autentik dan kontekstual seperti grafik fungsi organ atau kasus gangguan sistem tubuh berdampak positif terhadap validitas dan daya pembeda butir. Sari dan Prasetyo (2023) serta Rosdiana et al. (2020) juga menemukan bahwa penggunaan konteks yang dekat dengan pengalaman sehari-hari siswa dapat meningkatkan kemampuan berpikir analitis dan evaluatif.

Secara praktis, temuan ini menunjukkan bahwa instrumen kemampuan berpikir tingkat tinggi yang dikembangkan dapat menjadi alat ukur yang layak untuk asesmen sains abad ke-21, terutama dalam mengukur kemampuan berpikir kritis dan pemecahan masalah peserta didik. Secara teoretis, hasil penelitian ini memperkuat kerangka pengembangan asesmen berbasis berpikir tingkat tinggi dalam konteks IPA dan memberikan contoh konkret bagaimana karakteristik psikometrik dapat digunakan sebagai dasar revisi instrumen.

Ke depan, instrumen serupa dapat diperluas pada topik lain seperti sistem ekosistem atau interaksi makhluk hidup dengan lingkungan untuk memperkuat literasi sains siswa. Penelitian lanjutan juga disarankan untuk melibatkan jumlah sampel lebih besar atau pendekatan *Rasch Model* guna memperoleh estimasi parameter butir yang lebih presisi (Zulkardi & Putri, 2021).

KESIMPULAN

Penelitian ini menyimpulkan bahwa instrumen kemampuan berpikir tingkat tinggi pada materi sistem organ manusia memiliki kualitas psikometrik yang baik dan layak digunakan dalam pembelajaran IPA tingkat SMP. Validitas isi dan empiris menunjukkan kesesuaian konstruk dengan indikator kemampuan berpikir tingkat tinggi, reliabilitas berada pada kategori tinggi, serta distribusi tingkat kesukaran dan daya pembeda memenuhi proporsi ideal untuk mengukur variasi kemampuan siswa. Efektivitas distraktor yang cukup tinggi juga memperkuat keandalan instrumen dalam membedakan peserta berkemampuan berbeda. Dengan demikian, instrumen ini berpotensi menjadi model asesmen sains yang mampu mendukung pengukuran kemampuan berpikir tingkat tinggi secara autentik dan kontekstual dalam kerangka pembelajaran abad ke-21.

UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih kepada UIN Sunan Gunung Djati Bandung atas dukungan fasilitas dan pendampingan akademik selama pelaksanaan penelitian ini. Penghargaan juga diberikan kepada para validator ahli yang telah memberikan masukan konstruktif terhadap instrumen yang dikembangkan, serta kepada

guru dan peserta didik SMP mitra penelitian yang berpartisipasi dalam uji coba instrumen. Penelitian ini tidak didanai oleh hibah eksternal, sehingga seluruh proses pelaksanaan didukung oleh sumber daya mandiri dari institusi dan peneliti.

DAFTAR PUSTAKA

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Arikunto, S. (2010). *Prosedur penelitian: Suatu pendekatan praktik*. Rineka Cipta.

Azizah, N., & Retnawati, H. (2019). The challenges of teachers in developing higher order thinking skills (HOTS) assessment. *Journal of Physics: Conference Series, 1200*(1), 012045. https://doi.org/10.1088/1742-6596/1200/1/012045

Azwar, S. (2021). *Reliabilitas dan validitas* (5th ed.). Pustaka Pelajar.

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice-Hall.

Johanson, G. A., & Brooks, G. P. (2010). Initial scale development: Sample size for pilot studies. *Educational and Psychological Measurement, 70*(3), 394–400. https://doi.org/10.1177/0013164409355692

Kim, K. J., Park, S. H., & Lee, Y. (2021). Development and validation of higher-order thinking assessment rubric for science. *International Journal of Science Education, 43*(3), 345–366. https://doi.org/10.1080/09500693.2020.1865582

Malau, H., & Sari, M. (2023). The effectiveness of contextual issues in improving students' critical thinking skills. *Jurnal Pendidikan IPA Indonesia, 12*(1), 45–57. https://doi.org/10.15294/jpii.v12i1.38901

Mardapi, D. (2017). *Pengukuran, penilaian, dan evaluasi pendidikan*. Nuha Medika.

Retnawati, H., Arifin, Z., & Chen, S. (2018). Teachers' difficulties in developing HOTS assessment: A case study. *Problems of Education in the 21st Century, 76*(4), 520–532.

Rosdiana, D., Putri, N. L., & Rahmawati, R. (2020). The role of context familiarity on students' difficulty level in answering HOTS questions. *Jurnal Pendidikan Sains, 8*(2), 103–112. https://doi.org/10.17977/um033v8i22020p103

Sari, P., & Prasetyo, Z. (2023). Students' conceptual reasoning on human organ systems through problem-based contexts. *Journal of Biological Education, 57*(2), 150–162. https://doi.org/10.1080/00219266.2020.1862803

Susanti, R., & Kurniawati, A. (2021). HOTS-based assessment in science learning: Instrument development and validation. *Journal of Physics: Conference Series, 1806*, 012121. https://doi.org/10.1088/1742-6596/1806/1/012121

Yustiqvar, M., & Fauzi, A. (2022). Psychometric evaluation of HOTS items for junior high school science assessments. *Journal of Educational Research and Evaluation, 11*(2), 258–270. https://doi.org/10.23887/jere.v11i2.43157

Zulkardi, Z., & Putri, R. I. I. (2021). Designing valid and reliable performance assessments in science: A Rasch-model approach. *International Electronic Journal of Mathematics Education, 16*(2), em0645. https://doi.org/10.29333/iejme/10988