KAJIAN LITERATUR PENGEMBANGAN INSTRUMEN TES HOTS DALAM EVALUASI PEMBELAJARAN

Elsa Puspita[1], Wahyu Na'ilah AL Farafishah[2], Aulia Syifa Prabowo[3], Zafira Irsyadah[4], Tasya Qurrota A'yun[5], Nabillah Suci Melati[6], Fachri Rahmat Tipano[7], Sri Sumarni[8], Minarsi[9]
[1-9]Universitaas Sriwijaya
[1]epuspita644@gmail.com,[2]wahyunailahalf@gmail.com,
[3]auliasyifaa0606@gmail.com, [4]zafirairsyadah2006@gmail.com,
[5]tasyaqurrata591@gmail.com, [6]melatinabillah@gmail.com,
[7]fachrirahmat36@gmail.com

ABSTRACT

The development of Higher Order Thinking Skills (HOTS) test instruments is important in learning evaluation because assessment must measure analyzing, evaluating, and creating, not only recalling. This study reviews trends in development procedures, quality aspects, and testing methods of HOTS test instruments. A descriptive qualitative literature review was conducted on 30 articles published in 2021–2026 and retrieved from SINTA and Google Scholar. Borg & Gall or Sugiyono models were used most often, followed by 4D and ADDIE. Most articles reported validity and reliability, whereas item difficulty and discrimination were reported unevenly, and several articles contained inconsistent reporting. HOTS instrument development therefore needs empirical testing and transparent reporting.

Keywords: *higher order thinking skills, test instrument, instrument quality*

ABSTRAK

Pengembangan instrumen tes HOTS penting dalam evaluasi pembelajaran karena penilaian perlu mengukur kemampuan menganalisis, mengevaluasi, dan mencipta, bukan hanya mengingat. Kajian ini bertujuan menelaah kecenderungan prosedur pengembangan, aspek kualitas, dan metode pengujian instrumen tes HOTS. Metode yang digunakan adalah kajian literatur dengan pendekatan deskriptif kualitatif terhadap 30 artikel tahun 2021-2026 dari SINTA dan Google Scholar. Hasilnya, model Borg & Gall atau Sugiyono paling banyak digunakan, diikuti 4D dan ADDIE. Validitas dan reliabilitas dilaporkan pada sebagian besar artikel, sedangkan tingkat kesukaran dan daya beda belum merata, dan pelaporan beberapa artikel tidak konsisten. Pengembangan instrumen HOTS perlu disertai pengujian empiris dan pelaporan yang transparan.

Kata Kunci : keterampilan berpikir tingkat tinggi, instrument tes, kualitas instrumen

A. Pendahuluan

Higher Order Thinking Skills (HOTS) mencakup kemampuan menganalisis (C4), mengevaluasi (C5), dan mencipta (C6), sehingga memerlukan instrumen penilaian yang

mampu mengukur proses berpikir peserta didik (Avina & Winarsih, 2020). Namun, soal yang tampak menantang belum tentu mengukur HOTS, sebab kesesuaian dengan indikator kognitif dan kualitas butirnya harus dibuktikan secara empiris. Miladanta et al., (2024) menguji kualitas soal HOTS melalui validitas, reliabilitas, tingkat kesukaran, dan daya pembeda; Supriyati (2023) mengembangkan tujuh soal penalaran HOTS yang valid dan reliabel.

Penelitian terdahulu memperlihatkan beragam model pengembangan, yaitu ADDIE pada materi laju reaksi Novyandini et al., (2021), dan 4D pada instrumen berbasis literasi numerasi Wardani & Siregar, (2023) Arriah & Romba (2023) menghasilkan 15 butir soal literasi numerasi berbasis HOTS yang valid dan reliabel (koefisien 0,91), sedangkan Merta et al. (2022) mengembangkan 30 butir instrumen HOTS muatan IPA kelas V SD dengan model Djemari Mardapi, dengan seluruh butir valid dan reliabilitas 0,918. Miladanta et al. (2024) menghasilkan 24 soal esai HOTS yang memenuhi aspek validitas, reliabilitas, daya pembeda, dan tingkat kesukaran. Rahman et al. (2026) menemukan dominasi ADDIE dan Borg & Gall, tetapi integrasi dimensi kognitif dan pengetahuan serta penggunaan model Rasch masih terbatas. Variasi tersebut menunjukkan perlunya pemetaan tahapan pengembangan dan pengujian kualitas instrumen HOTS secara terstruktur.

Karena model, bentuk soal, subjek, dan metode analisis beragam, diperlukan kajian literatur untuk memetakan kualitas instrumen serta cakupan bidang ilmu dan jenjang pendidikan. Kajian ini menganalisis 30 artikel periode 2021–2026 berdasarkan model pengembangan, bentuk instrumen, subjek uji coba, aspek kualitas, dan metode analisis, dan diharapkan menjadi acuan bagi peneliti dan guru dalam mengembangkan instrumen HOTS yang valid, reliabel, dan layak.

B. Metode Penelitian

Penelitian ini menggunakan metode kajian literatur dengan pendekatan deskriptif kualitatif untuk mengkaji prosedur pengembangan, aspek kualitas, dan metode pengujian instrumen tes Higher Order Thinking Skills (HOTS) dalam evaluasi

pembelajaran. Artikel ilmiah ditelusuri melalui SINTA dan Google Scholar dengan kata kunci “pengembangan instrumen tes HOTS” dan “instrumen penilaian HOTS”. Artikel dipilih jika membahas analisis kebutuhan atau pengembangan instrumen tes/penilaian berbasis HOTS, terbit pada 2021–2026, serta memuat fokus penelitian, metode, dan hasil penelitian, sehingga diperoleh 30 artikel. Data dianalisis berdasarkan tahun terbit, bidang ilmu, jenjang pendidikan, model pengembangan, dan aspek pengujian kualitas melalui perbandingan antarartikel.

C. Hasil Penelitian dan Pembahasan

Berdasarkan penelusuran literatur, diperoleh 30 artikel nasional periode 2021–2026 yang relevan dengan pengembangan instrumen tes HOTS. Artikel dianalisis berdasarkan model pengembangan, penyusunan soal, pengujian kualitas, metode, dan hasil penelitian untuk mengidentifikasi kecenderungan serta kekurangan penelitian terdahulu.

Tabel 1. Daftar Artikel

No	Nama Peneliti	Tahun	Desain dan Metode Pengambilan Data	Hasil Penelitian
1	Hamidah & Wulandari	2021	R&D Sugiyono (7 dari 10 langkah); 30 PG C4–C6 via Quizizz; 3 validator; 34 siswa SMK XII OTKP Ipiems Surabaya; Iteman	Validasi ahli 94%; 25 dari 30 butir valid (r tabel 0,3291); reliabilitas 0,77; kesukaran 4 mudah, 25 sedang, 1 sulit; 2 butir daya beda kurang; respon siswa 89,5%
2	Litna et al.	2021	R&D Mardapi sampai merakit tes; 25 PG; 7 pakar (CVR Lawshe); 125 siswa (kelompok kecil) dan 250 siswa (kelompok besar) kelas XI, empat SMAN Manggarai Timur; analisis klasik dan IRT 1 parameter	CVR 1; 20 dari 25 butir diterima; reliabilitas 0,803; daya beda 16 baik, 7 sangat baik; 3 butir pengecoh tidak efektif
3	Azmi et al.	2021	R&D 4D; 25 PG two-tier; 60 siswa kelas XI SMAN 8 Semarang; analisis klasik	Valid dan reliabel; KR-20 0,735; respon siswa 94,67%
4	Amrina et al.	2021	Deskriptif kuantitatif; dosen Pendidikan IPS dan 56 mahasiswa Pendidikan Ekonomi FKIP Universitas Sriwijaya angkatan 2019/2020; wawancara, dokumentasi, angket; analisis persentase	80% soal UAS masih berorientasi MOTS (40% C3, 20% C4); 92,2% mahasiswa menyadari pentingnya instrumen berbasis HOTS untuk melatih analisis masalah sosial
5	Merta et al.	2022	R&D model Mardapi; 30 butir HOTS dan instrumen motivasi belajar; 60 siswa kelas V SD; uji coba satu kali.	CVR 1; korelasi butir 0,32–0,85 (semua valid); reliabilitas 0,918; kesukaran 6 mudah, 21 sedang, 3 sukar; daya beda 13 baik, 13 sedang, 4 cukup
6	Hanafi et al.	2022	R&D ADDIE; 3 uraian; 4 validator; uji kecil 10 siswa, uji besar 20 siswa kelas IX MTs	Validasi ahli 80–90%; validitas butir 0,574–0,706; alpha 0,77; kesukaran sedang; daya beda baik (2), sedang (1)
7	Wahyuningsih et al.	2022	Design Based Research (4 tahap); 15 uraian SPLDV; validasi ahli materi, ahli bahasa, 7 guru dari 4 SMP Kab. Sorong; tanpa uji coba siswa; SMP kelas VIII	Ahli materi 4,4; ahli bahasa 5; guru 4,3 (sangat baik); dinyatakan valid dan praktis; reliabilitas, kesukaran, daya beda tidak dilaporkan
8	Arsiana et al.	2022	R&D Borg & Gall disederhanakan (5 tahap); 20 PG tema 8 (Bumiku) C4–C5–C6; 4 dosen ahli (CVR Lawshe); 82 siswa kelas VI Gugus III Gerokgak, Buleleng; Excel	CVR semua butir valid; rerata r 0,473 (r tabel 1% 0,286), semua valid; KR-20 0,82; daya beda rerata 0,43 (1 sangat baik, 15 baik, 4 cukup); kesukaran rerata 0,502 (1 mudah, 17 sedang, 2 sukar)
9	Suarti	2022	R&D Borg & Gall disederhanakan (7 langkah); 40 PG tematik kelas IV (tema 8 subtema 3); 3 dosen ahli (Aiken); uji awal 28 siswa, uji lapangan total 126 siswa, Gugus I Bebandem	Aiken's V 0,67; 0,83; 1,00 (35 soal diterima tanpa revisi, 5 dengan revisi); 40 butir valid; KR-20 0,914 (uji awal); uji lapangan: kesukaran 32,5% mudah, 57,5% sedang, 10% sukar; daya beda 7 sangat baik, 10 baik, 23 cukup; pengecoh berfungsi 97,5%
10	Elsa et al.	2022	R&D evaluasi formatif Tessmer; 15 PG; 2	Validitas rerata 3,375 (valid); kepraktisan siswa 94,17% dan guru 87,14%;

			validator; 3 siswa one-to-one dan 6 siswa small group	reliabilitas, kesukaran, daya beda tidak dilaporkan
11	Yuparing et al.	2023	R&D Sugiyono 10 langkah; 5 PG dan 5 uraian singkat; 2 validator; uji produk 6 siswa, uji pemakaian 26 siswa	Validasi 95,5%; praktikalitas siswa 83%, guru 87%; 8 soal sedang, 2 mudah; daya beda 0,41–0,72; reliabilitas tidak dilaporkan
12	Jumrah et al.	2023	R&D Tessmer; 10 uraian; 3 validator; Aiken dan Rasch (Winsteps); 28 siswa kelas VI	Aiken's V 0,775; praktikalitas 74,67% dan 80,00%; reliabilitas person 0,95, butir 0,17, alpha 0,96; tanpa DIF
13	Saragih et al.	2023	R&D; adaptasi 10 langkah Rasch dan 6 langkah soal HOTS; 10 PG; 4 ahli; 36 siswa kelas XI; Ministep	Aiken's V 0,92; 10 butir valid (MNSQ, ZSTD, Pt Mean Corr); reliabilitas item 0,86
14	Rahayu et al.	2023	Evaluasi formatif; 10 uraian; partial credit model; SD kelas IV; 30 siswa (small group) dan 60 siswa (field test)	CVR 1; kepraktisan 91%; reliabilitas 0,79; daya beda 7 baik, 3 cukup; kesukaran 3 mudah, 6 sedang, 1 sulit
15	Rizki et al.	2024	R&D 4D; 20 PG; 36 siswa kelas X SMAN 5 Magelang	10 dari 20 butir valid; alpha 0,674; kesukaran 15% mudah, 65% sedang, 20% sulit; validitas konstruk 3,61
16	Fransiska et al.	2024	4D tanpa disseminate; 3 uraian C4–C6; 2 validator; 6 siswa (keterbacaan), 23 siswa (uji lapangan) SMPN 1 Silat Hulu kelas VII	Keterbacaan 73,5%; r_xy 0,816–0,910 (valid); reliabilitas 0,92; kesukaran sedang (2), sukar (1); daya beda cukup (1), baik (2)
17	Fitriani et al.	2024	R&D 8 tahap; 15 soal (5 benar-salah, 5 PG, 5 uraian); mahasiswa Pendidikan Akuntansi; Anates	Daya beda baik; 8 dari 10 objektif dan 3 dari 5 uraian berkesukaran sedang; respon 65,4–84,6%
18	Miladanta et al.	2024	R&D tahapan Gall dimodifikasi (5 tahap); 24 esai (paket A dan B); 48 siswa kelas XII	Validitas rerata 0,61; reliabilitas 0,87 dan 0,80; kesukaran rerata 0,6; 12 dari 24 soal layak
19	Wandira et al.	2025	R&D 4D; 15 PG two-tier C4–C6; 2 dosen dan 1 guru; 200 siswa kelas XI (SMAN 10 dan SMAN 3 Pontianak); Rasch (Winsteps) dan EFA	Aiken's V 0,97–0,98; alpha 0,80, person reliability 0,74, item reliability 0,90; 13 butir fit; kesukaran 2 sangat mudah, 7 mudah, 3 sulit, 3 sangat sulit; daya beda 1 jelek, 3 cukup, 11 baik
20	Sembiring & Asmin	2025	R&D 4D dimodifikasi menjadi 3D; 5 uraian SPLTV; 3 validator; uji I 27 siswa, uji II 26 siswa kelas X	Validasi ahli 3,84 (sangat valid); uji I rerata 76,71, 4 dari 5 butir valid, ketuntasan klasikal 66,67%; uji II rerata 87,83, 5 butir valid, ketuntasan klasikal 92,30%; peningkatan 11,12; reliabilitas tidak dilaporkan
21	Amrina et al.	2025	R&D Borg & Gall; PG dan uraian ekonomi; That Quiz; 15 siswa kelas X SMAN 10 Palembang	Validasi ahli materi 98,61%, bahasa 87,5%, media 96,56%; 17 PG dan 4 uraian valid dan reliabel; praktikalitas 72%
22	Mariani et al.	2025	R&D Sugiyono 7 tahap; bahasa Indonesia, SMP kelas VII	25 dari 30 butir valid; reliabilitas 0,78; validasi ahli 94%
23	Dewy et al.	2025	Kuantitatif deskriptif; 80 PG HOTS (C4–C6); 3 ahli; 40 mahasiswa Pendidikan Teknik Elektro Unimed (purposive)	67 dari 80 butir valid (83,75%); KR-20 0,83; kesukaran sedang 95,24% (60 dari 63 butir); 56 butir layak
24	Nadia Safira et al.	2026	ADDIE; C4–C6; pecahan senilai, SD kelas III; WordWall	Validasi ahli 75% (layak); reliabilitas, kesukaran, daya beda tidak dilaporkan

25	Jannata et al.	2026	Borg & Gall disederhanakan (5 tahap); 30 butir; 1 dosen dan 6 guru PJOK; 87 siswa SD di DIY; EFA/PCA dengan R	Aiken's V 0,837; KMO 0,63; 3 faktor; 3 butir dibuang (27 butir final); reliabilitas, kesukaran, daya beda tidak dilaporkan
26	Triwiguna et al.	2026	R&D ADDIE; 40 PG berkonteks dunia kerja; 3 ahli; 24 siswa kelas XI SMKN 3 Singaraja	Aiken's V 0,92; KR-20 0,92; kesukaran 32,5% mudah, 52,5% sedang, 15% sukar; daya beda 72,5% sangat baik, 27,5% baik; pengecoh efektif
27	Ermansyah et al.	2026	R&D 4D; 5 esai (15 subsoal) C4–C6; 29 siswa kelas XI IPA SMAS Adhyaksa 1 Jambi; Aiken, Pearson, alpha	Aiken's V 0,88; alpha 0,735; 8 butir memenuhi semua kriteria
28	Swandewi et al.	2026	Oriondo-Wilson; 30 uraian; 2 validator (Gregory); 25 siswa kelas X SMAN 1 Sawan	Validitas isi 0,86; 30 butir valid (r hitung > 0,396); reliabilitas 0,92
29	Jumarin et al.	2026	R&D ADDIE; 50 butir; 2 ahli materi, 2 ahli evaluasi, 3 guru; 100 siswa kelas VIII SMPN 6 Muaro Jambi (total sampling)	Validasi ahli materi 93,33%, evaluasi 91,66%; 50 butir valid (r 0,186–0,651); alpha 0,833; kesukaran 15 mudah, 22 sedang, 13 sukar; respon guru dan siswa baik–sangat baik
30	Batubara et al.	2026	R&D Borg & Gall (5 dari 10 tahap); 25 PGSejarah Indonesia Masa Kolonial; validasi ahli instrumen, ahli evaluasi, dosen sejarah (Aiken, 0,58–0,91); 30 mahasiswa Pendidikan Sejarah IPTS	3 dari 25 butir tidak valid (r tabel 0,361); KR-20 0,5899 (cukup); kesukaran rerata 0,6168 (3 sukar, 17 sedang, 5 mudah); daya beda rerata 0,2728 (5 jelek, 7 cukup, 11 baik, 2 baik sekali)

Penelitian pengembangan instrumen HOTS berlangsung konsisten pada 2021–2026 dengan empat sampai tujuh artikel per tahun (empat pada 2021, enam pada 2022, empat pada 2023, empat pada 2024, lima pada 2025, dan tujuh pada 2026). Matematika paling banyak diteliti (10 artikel), disusul IPA (8 artikel), sedangkan ilmu sosial, kejuruan/teknik, tematik SD, PJOK, dan bahasa Indonesia masing-masing satu sampai empat. Berdasarkan jenjang, SD (8), SMP/MTs (6), dan SMA (10) hampir seimbang, sedangkan SMK dua dan perguruan tinggi empat artikel. Sebagian besar penelitian memadukan validasi ahli dengan uji coba empiris. Distribusi artikel berdasarkan model pengembangan disajikan pada Tabel 2.

Tabel 2. Distribusi Artikel Berdasarkan Model Pengembangan.

Model Pengembangan	Jumlah	Rentang Tahun	Fokus Yang Dominan
Borg & Gall / Sugiyono	9	2021-2026	IPS dan Humaniora (3)
4D	6	2021-2026	IPA (4)
ADDIE	4	2022-2026	Matematika (2)
Tessmer / evaluasi formatif	3	2022-2023	Matematika (3)
Mardapi	2	2021-2022	Beragam (matematika dan IPA)
Lainnya (Design Based Research, Oriondo-Wilson, 10 langkah Rasch, analisis uji coba)	4	2022-2026	IPA (2)
R&D tanpa model tercantum	1	2024	Akuntansi (1)
Analisis Kebutuhan (deskriptif kuantitatif)	1	2021	IPS (1)
Total	**30**		

Berdasarkan Tabel 2, model Borg & Gall atau Sugiyono paling dominan (9 artikel) dan umumnya disederhanakan menjadi lima sampai tujuh langkah (Hamidah & Wulandari, 2021; Suarti, 2022; Mariani et al., 2025; (Arsiana et al., 2022; Miladanta et al., 2024; Jannata et al., 2026; Batubara et al., 2026), sedangkan (Yuparing et al., 2023) mempertahankan sepuluh langkah. Model 4D dipakai pada enam artikel untuk IPA (Azmi et al., 2021; Rizki et al., 2024; Wandira et al., 2025; Ermansyah et al., 2026; serta matematika (Fransiska et al., 2024; Sembiring & Asmin, 2025) dengan tahap disseminate ditiadakan oleh (Fransiska et al., 2024; Sembiring & Asmin, 2025). ADDIE dipakai pada

empat artikel (Hanafi et al., 2022; Safira et al., 2026; Triwiguna et al., 2026; Jumarin et al., 2026). Tessmer atau evaluasi formatif pada tiga artikel Elsa et al. (2022), Jumrah et al. (2023), dan Rahayu et al. (2023). dan Mardapi pada dua artikel Litna et al. (2021) dan Merta et al. (2022). Model lain mencakup Design Based Research (Wahyuningsih et al., 2022), prosedur Oriondo-Wilson (Swandewi et al., 2026), gabungan langkah Rasch dan penyusunan soal HOTS Saragih et al. (2023), dan delapan tahap penyusunan tes (Fitriani et al., 2024) sedangkan Dewy et al. (2025) memakai pendekatan kuantitatif deskriptif tanpa model pengembangan. Rincian jumlah butir, subjek, dan hasil uji tiap artikel dirangkum pada Tabel 1.

Sebagian instrumen memanfaatkan aplikasi, yaitu Quizizz (Hamidah & Wulandari, 2021; Mariani et al., 2025) iSpring Suite (Fitriani et al., 2024), That Quiz (Amrina et al., 2025), dan WordWall (Safira et al., 2026). Pendidikan IPS (Amrina et al., 2021), pendidikan akuntansi (Fitriani et al., 2024), Elektronika Dasar (Dewy et al., 2025), dan Pendidikan Sejarah (Batubara et al., 2026).

Hampir semua artikel menyimpulkan instrumennya valid dan layak, tetapi kelengkapan bukti kualitasnya berbeda. Hanafi et al. (2022) dan Triwiguna et al. (2026) melaporkan validitas, reliabilitas, kesukaran, dan daya pembeda sekaligus, sedangkan Elsa et al. (2022) dan (Safira et al., 2026) hanya melaporkan validasi ahli dan kepraktisan. Analisis butir yang lebih mendalam dilakukan, misalnya, dengan Rasch pada Wandira et al. (2025) dan Jumrah et al. (2023) dan pada beberapa artikel sebagian butir tetap gugur meskipun instrumen dinyatakan layak.

Pengembangan instrumen HOTS pada 2021–2026 berlangsung pada banyak mata pelajaran dan jenjang dengan matematika dan IPA sebagai fokus utama, dan tidak lagi berhenti pada penyusunan soal C4–C6, tetapi disertai pengujian empiris butir, sejalan dengan Supriyati (2023) Dominasi Borg & Gall atau Sugiyono (9 artikel), disusul 4D dan ADDIE, sejalan dengan (Novyandini et al., 2021) dan (Wardani & Siregar, 2023). Namun, jumlah tahap Borg & Gall berkisar dari lima sampai sepuluh, dan sebagian penelitian tidak menempuh seluruh tahap:Fransiska et al. (2024) tanpa disseminate, Sembiring &

Asmin (2025) memangkas 4D menjadi tiga tahap, Litna et al. (2021) berhenti pada merakit tes, dan Elsa et al. (2022) berhenti pada tahap small group. Dengan demikian, label model yang sama belum tentu berarti prosedur yang sama, sehingga hasil antarartikel sulit dibandingkan dan direplikasi.

Validitas isi dikuantifikasi dengan empat cara: Aiken's V (Suarti, 2022; Jumrah et al., 2023; Saragih et al., 2023; Wandira et al., 2025; Jannata et al., 2026; Ermansyah et al., 2026; Triwiguna et al., 2026), rumus Gregory (Swandewi et al., 2026), CVR (Litna et al., 2021; Merta et al., 2022; Rahayu et al., 2023), serta rerata skor atau persentase kelayakan (Hamidah & Wulandari, 2021; Azmi et al., 2021; Elsa et al., 2022; Hanafi et al., 2022; Yuparing et al., 2023; Sembiring & Asmin, 2025; Amrina et al., 2025; Jumarin et al., 2026). Jumlah penilainya berkisar dari tiga orang (Jumrah et al., 2023; Wandira et al., 2025; Triwiguna et al., 2026) sampai tujuh orang (Jannata et al., 2026), dan rerata Aiken's V dari 0,775 (Jumrah et al., 2023) hingga 0,97–0,98 (Wandira et al., 2025). Perbedaan ini membuat nilai validitas antarartikel tidak dapat dibandingkan secara langsung.

Koefisien reliabilitas berkisar dari 0,59 (Batubara et al., 2026) hingga 0,96 (alpha Cronbach pada Jumrah et al., 2023). dengan nilai antara lain 0,674 (Rizki et al., 2024), 0,735 (Azmi et al., 2021), 0,77 (Hanafi et al., 2022), 0,80 (Wandira et al., 2025), 0,833 (Jumarin et al., 2026), dan 0,918 (Merta et al., 2022). sejalan dengan koefisien 0,91 pada Arriah & Romba (2023). Namun, label "reliabel" mencakup kualitas yang berbeda: Jumrah et al. (2023) melaporkan alpha 0,96 bersama reliabilitas butir Rasch hanya 0,17, sehingga satu angka reliabilitas yang tinggi belum tentu menggambarkan kestabilan butir. Persentase kepraktisan atau respon pengguna bervariasi dari 65,4% (Fitriani et al., 2024) hingga 94,67% (Azmi et al., 2021).

Label "layak" sering berdampingan dengan gugurnya sebagian butir: Litna et al. (2021) menerima 20 dari 25 butir, Rizki et al. (2024) 10 dari 20 valid, Miladanta et al. (2024) 12 dari 24 layak, Dewy et al. (2025) 56 dari 80 layak, dan Ermansyah et al. (2026) delapan dari 15 subsoal sedangkan Azmi et al. (2021) masih mendapati delapan dari 25 butir tidak valid pada uji skala besar. Pelaporan juga tidak selalu

konsisten. Jumarin et al. (2026) menyatakan seluruh 50 butir valid dengan korelasi terendah 0,186, padahal pada 100 responden nilai itu mendekati atau berada di bawah r tabel (sekitar 0,195), dan jumlah butir layak pada Wandira et al. (2025) dan Dewy et al. (2025) berbeda antarbagian artikel. Kelayakan menurut ahli karenanya belum menjamin kualitas butir secara empiris, sehingga analisis butir dan pelaporan yang transparan tetap diperlukan.

Pendekatan berbasis respons butir baru dijumpai pada sebagian kecil artikel, yaitu analisis item respon pada Litna et al. (2021), Rasch pada Jumrah et al. (2023), Saragih et al. (2023), dan Wandira et al. (2025), serta partial credit model pada Rahayu et al. (2023). Jannata et al. (2026) menyebut analisisnya IRT dengan R, tetapi prosedur yang dilaporkan berupa analisis faktor eksploratori dan PCA, sehingga istilah “komponen” dan “faktor” perlu dibedakan.

Ukuran sampel uji coba berkisar dari 3 siswa pada tahap one-to-one Elsa et al. (2022) hingga 250 siswa pada uji kelompok besar Litna et al. (2021). Fransiska et al. (2024) dan Hanafi et al. (2022) hanya menguji tiga butir pada 23 siswa dan 20 siswa, sehingga reliabilitas 0,92 dan 0,77 perlu ditafsirkan hati-hati. Sampel sekecil ini, termasuk 24-29 siswa pada Triwiguna et al. (2026), Swandewi et al. (2026), dan Ermansyah et al. (2026) membatasi kestabilan estimasi reliabilitas dan analisis butir.

Kajian ini juga menemukan beberapa celah. Matematika dan IPA masih mendominasi, sedangkan ilmu sosial dan humaniora hanya diwakili empat artikel Amrina et al. (2021), Fitriani et al. (2024), Amrina et al. (2025), dan Batubara et al. (2026). dan jenjang perguruan tinggi juga hanya empat artikel. Tidak semua artikel melaporkan seluruh aspek kualitas: Safira et al. (2026) dan Elsa et al. (2022) hanya melaporkan validasi ahli dan kepraktisan, Sembiring & Asmin (2025) serta Yuparing et al. (2023) tidak memuat reliabilitas, Swandewi et al. (2026) tidak memuat kesukaran dan daya pembeda, dan Jannata et al. (2026) tidak memuat reliabilitas, kesukaran, maupun daya pembeda. Di sisi lain, tampak kecenderungan mengaitkan soal dengan konteks lokal, seperti batik Pandeglang (Hanafi et al., 2022), lahan basah (Elsa et al., 2022), dan

tenun ikat Sintang (Fransiska et al., 2024), dan dunia kerja (Triwiguna et al., 2026), serta memanfaatkan aplikasi pembelajaran.

Secara umum, pengembangan instrumen tes HOTS di Indonesia sudah berkembang luas, tetapi kualitas bukti empirisnya belum merata. Sebagian besar instrumen dinyatakan valid dan layak, namun cara mengukur validitas, jumlah penilai, tahapan model, dan metode analisis berbeda-beda sehingga nilai antarartikel sulit dibandingkan. Kelayakan menurut ahli perlu dilengkapi analisis butir yang transparan (reliabilitas, tingkat kesukaran, dan daya pembeda), sampel uji coba yang memadai, serta penelitian pada ilmu sosial-humaniora dan perguruan tinggi.

D. Kesimpulan

Dari 30 artikel yang dikaji, pengembangan instrumen tes HOTS menggunakan berbagai model, dengan Borg & Gall atau Sugiyono paling dominan, diikuti 4D dan ADDIE. Instrumen umumnya memenuhi kriteria validitas dan kelayakan, tetapi pelaporan reliabilitas, tingkat kesukaran, dan daya pembeda belum merata, dan perbedaan metode analisis serta jumlah subjek uji coba menjadi keterbatasan. Kajian ini memberi gambaran kecenderungan pengembangan dan pengujian instrumen HOTS, meskipun terbatas pada 30 artikel. Penelitian selanjutnya perlu memperluas cakupan bidang ilmu dan jenjang serta melengkapi pengujian kualitas instrumen agar menghasilkan alat evaluasi yang lebih valid dan reliabel.

DAFTAR PUSTAKA

Amrina, D. E., Deskoni, D., & Mardetini, E. (2021). Analisis kebutuhan instrumen penilaian berbasis HOTS pada mata kuliah pendidikan IPS. *JURNAL PROFIT: Kajian Pendidikan Ekonomi Dan Ilmu Ekonomi*, *8*(1), 11–18. https://doi.org/10.36706/jp.v8i1.13119

Amrina, D. E., Hasmidyani, D., Mardetini, E., & Fitriyanti. (2025). Pengembangan instrumen penilaian berbasis HOTS berbantu aplikasi That Quiz. *Jurnal Edueco*, *8*(1), 263–273. https://doi.org/10.36989/didaktik.v9i2.1057

Arriah, F., & Romba, S. (2023). Development of a numeracy literacy ability test instrument based on high order thinking skills (hots) in junior high school mathematics learning. *Desimal: Jurnal Matematika*, *6*(1), 51–62. https://doi.org/10.24042/djm

Arsiana, K. E. A., Yudiana, K., & Antara, P. A. (2022). The Assessment Instrument for Higher Order Thinking Skills in Learning Theme 8 (Bumiku) for Sixth Grade

of Elementary School. *Journal for Lesson and Learning Studies*, *5*(2), 267–273.

Avina, Y., & Winarsih. (2020). Pengembangan instrumen penilaian sebagai contoh paket soal Higher Order Thinking Skills (HOTs) materi pencemaran lingkungan kelas X SMA. *BioEdu: Berkala Ilmiah Pendidikan Biologi*, *9*(1), 217–223.

Azmi, N. L., Nurhayati, S., Priatmoko, S., & Wardani, S. (2021). Pengembangan instrumen tes untuk mengukur hots peserta didik pada materi laju reaksi. *Journal of Chemistry In Education*, *10*(1), 45–52. http://journal.unnes.ac.id/sju/index.php/chemined

Batubara, U. N., Siregar, E., Ramadhani, A., & Nasution, P. (2026). Pengembangan instrumen penilaian E-HOTS dalam meningkatkan kritis analitis mahasiswa program studi pendidikan sejarah. *Jurnal Education and Development*, *14*(1), 534–540.

Dewy, M. S., Mulyana, D., Zulkarnain, S. A. B., Isnaini, M., Simamora, Y., & Imannuel, M. F. (2025). Analisis uji coba instrumen tes Berbasis Higher Order Thinking Skills (HOTS) pada Mata Kuliah Elektronika Dasar. *Jurnal Teknologi Pendidikan (JTP)*, *18*(2), 117–126. https://doi.org/10.24114/jtp.v18i2.70213

Elsa, L. Y., Ansori, H., & Budiarti, I. (2022). Pengembangan instrumen penilaian berbasis HOTS dengan konteks lahan basah pada persamaan garis lurus kelas VIII SMPN 2 Banjarmasin. *Jurmadikta (Jurnal Mahasiswa Pendidikan Matematika)*, *2*(3), 66–74. https://doi.org/10.20527/jurmadikta.v2i3.1454

Ermansyah, A. J. R., Afrida, Asmiyunda, Epinur, & Kriswantoro. (2026). Pengembangan instrumen penilaian Higher Order Thinking Skills untuk materi bentuk molekul berdasarkan Teori Vsepr. *Jurnal Riset Dan Inovasi Pendidikan Sains (JRIPS)*, *5*(2), 218–229. https://doi.org/10.36085/jrips.v5i2.10880

Fitriani, Pratikto, H., Rahayu, W. P., & Prabowo, A. E. (2024). Pengembangan instrumen assesment pembelajaran HOTS menggunakan iSpring Suite. *Research and Development Journal Of Education*, *10*(2), 695–707.

Fransiska, F., Hamdani, H., Siregar, N., Bistari, B., & Rustam, R. (2024). Pengembangan instrumen tes HOTS pada materi persegi dan persegi panjang berbasis motif tenun Ikat Sintang. *Variabel*, *7*(1), 14–22. https://doi.org/10.26737/var.v7i1.4838

Hamidah, M., & Wulandari, S. (2021). Pengembangan instrumen penilaian berbasis hots menggunakan aplikasi “quizizz. *Efisiensi : Kajian Ilmu Administrasi*, *18*(1), 105–124. https://jurnal.uny.ac.id/index.php/efisiensi/article/view/72745/pdf

Hanafi, M., Syamsuri, S., & Mutaqin, A. (2022). Pengembangan instrumen soal Higher Order Thinking Skills (Hots) Matematika berdasarkan Brookhart konteks motif Batik Pandegelang pada Siswa MTs. Media Pendidikan Matematika, 10(1), 43–59. https://doi.org/10.33394/mpm.v10i1.5207

Jannata, T., Kusuma, M., & Fajaruddin, S. (2026). Konstruksi instrumen tes HOTS permainan bola besar di sekolah dasar:

Analisis Item Response Theory dengan R. *Wiyata Dharma: Jurnal Penelitian Dan Evaluasi Pendidikan*, *14*(1), 1–17. https://doi.org/10.30738/wd.v14i1.22413

Jumarin, M. A. B., Munar, H., Hadinata, R., & Suardika, I. K. (2026). Pengembangan instrumen penilaian berbasis Higher Order Thinking Skill pembelajaran pendidikan jasmani olahraga dan kesehatan. *Jambura Health and Sport Journal*, *8*(1), 1–8. https://doi.org/10.37311/jhsj.v8i1.32461

Jumrah, Rukli, & Sulfasyah. (2023). Pengembangan Instrumen Tes Berbasis HOTS dengan pendekatan pengukuran rasch pada pelajaran matematika topik bangun ruang untuk siswa Sekolah Dasar. *Jurnal Basicedu*, *7*(1), 11–27.

Litna, K. ., Mertasari, N. M. ., & Sudhirta, G. (2021). Pengembangan instrumen tes Higher Order Thinking Skills (HOTS) Matematika SMA Kelas X. *Jurnal Penelitian Dan Evaluasi Pendidikan Indonesia*, *11*(1), 10–20. https://ejournal2.undiksha.ac.id/index.php/jurnal_ep/article/view/278

Mariani, T., Hutagalung, S. M., & Adisaputera, A. (2025). Pengembangan instrumen penilaian teks deskripsi berbasis hots berbantuan aplikasi Quizizz di Kelas VII SMP Amalia Medan. *Bahastra: Jurnal Pendidikan Bahasa Dan Sastra Indonesia*, *9*(2), 184–191. https://app.dimensions.ai/details/publication/pub.1193636102%0Ahttps://jurnal.uisu.ac.id/index.php/Bahastra/article/download/9543/pdf

Merta, I., Putrayasa, I., & Wibawa, I. (2022). Pengembangan instrumen penilaian Higher Order Thinking Skills (Hots) dan motivasi belajar pada muatan IPA Kelas V Sekolah Dasar. *Jurnal Penelitian Dan Evaluasi Pendidikan Indonesia*, *12*(1), 25–36. https://doi.org/10.23887/jpepi.v12i1.925

Miladanta, A., Nuryantini, A., Farida, I., & Cahyanto, T. (2024). Pengembangan instrumen Higher Order Thinking Skills (HOTS) materi alat optik melalui validitas, reliabilitas, daya pembeda, dan tingkat kesukaran. *Jurnal Penelitian Sains Dan Pendidikan*, *4*(2), 148–158.

Novyandini, H., Amilda, & Pratiwi, R. (2021). Development of assessment instrument based on Higher Order Thinking Skill (HOTS) in reaction rate material. *International Education Conference (IEC) FITK*, *1*(1), 264–270.

Rahayu, N. K. L. A., Widiartini, N. K., & Divayana, D. G. H. (2023). Pengembangan Instrumen Higher Order Thinking Skill (Hots) Matematika. *Jurnal Penelitian Dan Evaluasi Pendidikan Indonesia*, *13*(1), 17–26. https://doi.org/10.23887/jpepi.v13i1.1761

Rahman, F., Al-Musyarraf, M., Moulana, M., & Pradita, E. (2026). Teknik pengembangan instrumen penilaian berbasis HOTS mengacu pada Taksonomi Anderson & Krathwohl: Studi Kepustakaan 2015–2025. *Jurnal Jendela Pendidikan*, *6*(1). https://www.ejournal.jendelaedukasi.id/index.php/JJP/article/view/6

Rizki, D., Almay, F. A. P., & Ramadani, S. D. (2024). Pengembangan instrumen penilaian Higher Order Thinking Skills (HOTS) Pada Materi Ekologi di SMA. *Biodik: Jurnal Ilmiah Pendidikan Biologi*, *10*(4),

691–702. https://doi.org/10.22437/biodik.v10i4.36741

Safira, N., Sukmawarti, & Hidayat. (2026). Pengembangan instrumen penilaian berorientasi Hots Berbantuan Aplikasi Wordwall Pembelajaran Matematika SD. *Jurnal Penelitian Pendidikan Indonesia*, *3*(3), 105–110.

Saragih, I., Damanik, S. A., Darmana, A., & Suyanti, R. D. (2023). Pengembangan Instrumen Tes Berbasis Hots Pada Materi Asam-Basa Menggunakan Rasch Model. *UNESA Journal of Chemical Education*, *12*(3), 217–224.

Sembiring, S. N. B., & Asmin. (2025). Pengembangan instrumen tes HOTS untuk meningkatkan kemampuan berpikir kritis matematis siswa melalui model pembelajaran problem based learning. *Jurnal Pendidikan Matematika*, *9*(1), 200–210.

Suarti, N. W. (2022). Pengembangan instrumen penilaian tematik berbasis HOTS untuk peserta didik kelas IV sekolah dasar. *Indonesian Journal of Educational Development*, *2*(4), 534–548.

Supriyati, Y. (2023). the Development of Assessment Instrument of Reasoning Ability on Hots for Physics Learning. *JISAE: Journal of Indonesian Student Assessment and Evaluation*, *9*(2), 188–198. https://doi.org/10.21009/jisae.v9i2.39933

Swandewi, N. K. C., Arnyana, I. B. P., & Bestari, I. A. P. (2026). Pengembangan instrumen Soal HOTS pada Materi Virus untuk Murid Kelas X SMA. *Jurnal Pendidikan Biologi Undiksha*, *13*(2), 187–202. https://doi.org/10.23887/jjpb.v13i2.115843

Triwiguna, I. N. S., Ratnaya, I. G., & Adiarta, A. (2026). Pengembangan dan validasi instrumen penilaian berbasis hots untuk pengukuran kemampuan teoritis intalasi listrik pada pendidikan vokasi. *Jurnal Pendidikan Teknik Elektro Undiksha*, *15*(1), 1–13. https://doi.org/10.23887/jjpte.v15i1.115029

Wahyuningsih, R. T., Astutik, H. S., & Triono, M. (2022). Pengembangan instrumen Higher Order Thinking Skill (Hots) dalam pembelajaran Matematika Di SMP. *Theorema: The Journal Education of Mathematics*, *3*(1), 20–26. https://doi.org/10.36232/theorema.v3i1.2606

Wandira, D. A., Hamdani, & Firdaus. (2025). Pengembangan instrumen tes untuk Mengukur Higher Order Thinking Skills (HOTS) peserta Didik pada Materi Suhu dan Kalor SMA. *Jurnal Education and Development Institut Pendidikan Tapanuli SelatanTapanuli Selatan*, *13*(1), 147–153. https://journal.ipts.ac.id/index.php/ED/article/download/6527/3490

Wardani, N., & Siregar, T. J. (2023). The Development of HOTS (Higher-Order Thinking Skills) assessment instrument based on numeration literacy in mathematics for junior high school. *Indonesian Journal of Science and Mathematics Education*, *6*(2), 238–252. https://doi.org/10.24042/ijsme.v6i2.17433

Yuparing, F. R., Wiyono, B. B., & Sutadji, E. (2023). Instrumen asesmen Higher Order Thinking Skills (HOTS) pada mata pelajaran matematika kelas IV SD Negeri 4 Tanggung. *Jurnal Pendidikan Indonesia*, *4*(04), 311–323.