

Tren dan Karakteristik Psikometri Model Testlet pada Pengukuran Literasi Sains Pendidikan Menengah: Tinjauan Sistematis dan Proyeksi Masa Depan

Maulady An Nurul Islami^{1)*}, Djuniadi¹⁾, Wiwi Isnaeni¹⁾, Langlang Handayani¹⁾

¹⁾Penelitian dan Evaluasi Pendidikan, Sekolah Pascasarjana, Universitas Negeri Semarang

Corresponding Author: mauladyan_21@students.unnes.ac.id

ABSTRAK

Literasi sains merupakan salah satu kompetensi fundamental yang harus dikuasai pada keterampilan abad 21. Pengukuran literasi sains menuntut instrumen yang mampu memotret profil kemampuan kognitif tingkat tinggi. Penggunaan tes pilihan ganda konvensional dinilai kurang komprehensif, sementara asesmen berbasis kinerja seringkali kurang efisien. Model testlet yang terdiri dari sekumpulan butir soal dengan satu stimulus hadir sebagai solusi atas keterbatasan butir independen. Penelitian ini merupakan *Systematic Literature Review* (SLR) yang bertujuan untuk menganalisis tren penggunaan, karakteristik psikometri, serta memproyeksikan masa depan instrumen model testlet dalam pengukuran literasi sains. Penelusuran literatur melalui Google Scholar dan Crossref pada periode 2020-2026 menghasilkan 25 studi yang memenuhi kriteria inklusi setelah melalui proses identifikasi, penyaringan, dan penilaian kelayakan berdasarkan PRISMA 2020. Hasil analisis dan sintesis menunjukkan bahwa mayoritas instrumen testlet literasi sains menggunakan konteks kontekstual, Artificial Intelligence, dan deep learning. Berdasarkan analisis psikometri, Temuan studi menunjukkan bahwa tren pengukuran literasi sains beralih dari tes unidimensi diskrit menuju testlet yang dianalisis menggunakan *Rasch Testlet Model* (RTM) dan *Multidimensional Item Response Theory* (MIRT) untuk mengakomodasi *Local Item Dependence* (LID). Secara karakteristik, instrumen testlet terbukti menurunkan estimasi reliabilitas semu dan memberikan error standar yang lebih presisi dibandingkan model TRB tradisional. Proyeksi ke depan menunjukkan integrasi *Computerized Adaptive Testing* (CAT) berbasis testlet dan STEM berbasis kecerdasan buatan sebagai inovasi utama dalam lanskap asesmen sains.

Kata Kunci: Literasi Sains; Testlet; Psikometri; Systematic Literature Review

This is an open access article under the CC - BY license.



PENDAHULUAN

Pendidikan di Indonesia memiliki tujuan utama untuk mengembangkan karakter peserta didik secara menyeluruh. Sebagaimana yang telah diatur dalam UU Nomor. 20 Tahun 2003 Pasal 3 bahwa pendidikan nasional berfungsi untuk mengembangkan, membentuk karakter, dan membangun peradaban bangsa yang bermartabat. Menurut OECD (2023) yang sejalan dengan kajian Winarsi dan Ningsih (2025), literasi sains dalam kerangka PISA secara spesifik didefinisikan sebagai kemampuan individu dalam memanfaatkan pengetahuan sains, merumuskan pertanyaan ilmiah, serta menarik kesimpulan berbasis bukti. Penguasaan literasi sains menjadi tolak ukur kualitas penguasaan IPTEK berdampak terhadap peningkatan kualitas pembangunan sumber daya manusia, ekonomi, infrastruktur dan budaya, yang merupakan faktor pendukung capaian target SDGs sehingga dirancang untuk menjadi pondasi perkembangan ekonomi (Axmedovna & Tojimurodovna, 2021).

literasi sains menitikberatkan pada kemampuan peserta didik dalam memanfaatkan konsep keilmuan, keterampilan proses, dan integrasi nilai-nilai sains untuk mengambil keputusan dalam kehidupan sehari-hari (Roymana, 2022). Pembelajaran dan evaluasi literasi sains tidak sekadar bertumpu pada penguasaan konsep sains secara umum atau hafalan teoretis, melainkan menuntut integrasi keterampilan kognitif tingkat tinggi. Sebagaimana dijelaskan oleh Rahmasytoh dkk. (2026), pengembangan literasi sains memiliki keterikatan langsung dengan kemampuan berpikir kritis dan pemecahan masalah; di mana peserta didik diwajibkan untuk mampu menganalisis fenomena, mengevaluasi bukti empiris, serta menginterpretasi data secara saintifik sebelum mengambil keputusan. Rendahnya hasil tes literasi sains di Indonesia disebabkan oleh dua hal, yaitu peserta didik mempunyai dasar sains yang belum baik dan instrumen tes literasi sains yang telah dikembangkan belum cocok dengan kondisi yang ada di Indonesia (Rohmaya, 2016).

Tes uraian meskipun memiliki kelebihan dalam mengukur pemahaman mendalam dan kemampuan berpikir kritis, berbagai kelemahan yang terkait dengan subjektivitas penilaian, waktu koreksi yang lama, dan potensi ketidakadilan harus diperhatikan. Dalam penilaian instrumen bentuk esai atau uraian, unsur bluffing merupakan faktor penting yang harus diperhatikan (Hamdi, 2017). Penggunaan instrumen dalam bentuk-bentuk pilihan majemuk biasa (*multiple choice*) juga bukan solusi terbaik untuk menutupi kekurangan dari instrumen yang berbentuk uraian atau esai. Instrumen testlet dapat mengukur pengetahuan, proses berpikir peserta didik, serta memberikan stimulus memecahkan masalah (Gou & Li, 2024). Testlet merupakan sekumpulan butir soal pilihan majemuk (*multiple choice*) yang memiliki sumber informasi yang sama (Wainer et al., 2007). Menurut Ayan dan Pekmezci (2021) adalah salah satu jenis penilaian yang dapat digunakan untuk pemecahan masalah, pengukuran keterampilan, yang implementasinya lebih efisien dari segi waktu dan situasi dalam memberikan latihan kemampuan individual kepada peserta didik yang memiliki berbagai stimulus.

Secara pedagogis, format testlet sangat menguntungkan karena menyerupai proses kognitif penyelesaian masalah di dunia nyata, memungkinkan efisiensi waktu tes, dan dapat memotret multidimensi literasi sains meliputi kompetensi, pengetahuan (epistemik, konten, prosedural), serta sikap (Ningsih et al., 2024). Namun, tinjauan psikometri klasik maupun Teori Respons Butir (TRB) konvensional mensyaratkan asumsi independensi lokal (*local independence*). Independensi lokal menunjukkan sejauh mana keterkaitan jawaban terhadap suatu butir soal. Asumsi independensi lokal terpenuhi dalam penelitian ini apabila respon peserta tes terhadap satu butir testlet tidak mempengaruhi respon peserta tes terhadap butir soal yang lain. Menurut Gyamfi dan Acquaye (2023), independensi lokal dapat terdeteksi dengan membuktikan asumsi unidimensional. Independensi lokal menurut Hambleton & Swaminathan (1991) akan bermakna jika semua faktor yang mempengaruhi prestasi seperti kemampuan individu, lingkungan belajar, kesehatan, dan lain-lain cenderung konstan atau tetap sama, maka respon individu terhadap pasangan butir soal dalam instrumen akan independen secara statistik satu sama lain.

Penelitian ini bertujuan untuk mengidentifikasi dan menganalisis tren pengembangan instrumen penilaian literasi sains pemodelan testlet pada jenjang pendidikan menengah melalui *kajian systematic literature review*. Meskipun penggunaan model testlet semakin berkembang untuk mengukur kompetensi literasi sains berorientasi PISA, tinjauan literatur yang ada saat ini masih didominasi oleh evaluasi konseptual. Terdapat kesenjangan penelitian yang mendasar, dimana belum banyak studi sistematis yang membongkar bagaimana peneliti mengatasi anomali Local Item Dependence (LID) melalui pemodelan psikometri. Selain itu, eksplorasi terhadap inovasi penyusunan stimulus testlet seperti pemanfaatan narasi media deep learning maupun konteks kultural dalam membedah fenomena sains masih sangat minim dipetakan. Oleh karena itu, penelitian ini hadir untuk menjembatani kesenjangan tersebut dengan dalam memperkaya kerangka konseptual evaluasi literasi sains, serta manfaat praktis dalam menyusun alat ukur yang valid dan reliabel.

Keunikan studi ini terletak pada fokus spesifiknya terhadap karakteristik psikometri instrumen penilaian literasi sains pada jenjang pendidikan menengah. Pendekatan ini berbeda pada umumnya yang hanya meninjau aspek konseptual atau pengembangan indikator tanpa melakukan analisis mendalam terhadap validitas dan reliabilitas. Oleh karena itu, penelitian ini diharapkan dapat memberikan kontribusi baru dalam bidang evaluasi pendidikan pada konteks pengukuran literasi sains berbasis testlet pada peserta didik sekolah menengah. Di jenjang pendidikan menengah, penerapan instrumen berbasis testlet cenderung masih rumpang literatur tersebut, studi ini melakukan tinjauan sistematis terhadap rekam jejak riset dalam rentang waktu 2020 hingga 2026. Tinjauan ini bertujuan untuk membongkar secara komprehensif bagaimana karakteristik psikometri instrumen testlet dikalibrasi dalam pengukuran literasi sains dan bagaimana perkembangan serta proyeksi masa depan asesmen pendidikan ini.

METODE

Penelitian ini menggunakan desain *Systematic Literature Review* (SLR) dengan berpedoman *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA). Pencarian literatur dilakukan pada bulan 25 Mei- 20 Juni 2026 melalui dua pangkalan data akademik utama yaitu Crossref dan Google Scholar. Dalam strategi pencarian ini, Google Scholar difungsikan untuk memaksimalkan tingkat recall (keluasan jangkauan) guna menjangkau literatur *open access* dan publikasi nasional bereputasi secara komprehensif. Untuk mengantisipasi kelemahan kurasi otomatis pada Google Scholar, pangkalan data Crossref digunakan secara komplementer sebagai instrumen presisi. Penarikan metadata melalui Crossref memastikan bahwa artikel yang

diseleksi memiliki *Digital Object Identifier (DOI)* yang valid. Hal ini bertindak sebagai filter verifikasi yang ketat untuk menyingkirkan literatur abu-abu (*grey literature*) yang tidak melalui proses penelaahan sejawat (*peer review*) formal.

Prosedur SLR pada penelitian mengikuti pedoman PRISMA yang menekankan keterlacakan dari tahap identifikasi, penyaringan, kelayakan, hingga inklusi artikel (Johnson et al., 2022). Artikel yang bersumber dari kedua data base tersebut ditemukan 1057 publikasi dan ditemukan 104 duplikasi. Setelah melakukan penyaringan diperoleh 311 publikasi. Langkah selanjutnya, screening full text menyisihkan 39 publikasi, setelah di tinjau kembali diperoleh 25 publikasi. Instrumen penelitian berupa lembar ekstraksi data yang digunakan untuk mencatat informasi utama setiap artikel yang lolos seleksi. Data ekstraksi mencakup penulis & tahun, judul penelitian, metodologi, sampel/populasi, variabel utama, dan hasil penelitian, serta keterbatasan penelitian. Berikut tahapan SLR yang dilakukan oleh peneliti, yang secara umum dibagi menjadi tiga tahapan utama yaitu:

Perencanaan (*Planning*)

Rumusan Masalah

Langkah awal pada penelitian ini adalah merumuskan pertanyaan penelitian (Research Questions). Mengingat desain kajian ini lebih berfokus pada pemetaan sistematis (*systematic mapping*) terhadap kelayakan instrumen, maka perumusan pertanyaan penelitian disusun menggunakan kerangka PCC (Population, Concept, Context), yang secara metodologis lebih tepat untuk tinjauan tanpa kelompok intervensi. Batasan kerangka tersebut meliputi: (1) Population, difokuskan pada instrumen bagi peserta didik jenjang pendidikan menengah; (2) Concept, menelaah tren penggunaan serta karakteristik psikometri dari model testlet; dan (3) Context, dibatasi pada ruang lingkup asesmen pengukuran literasi sains. Proses formulasi pertanyaan dan ekstraksi kata kunci ini turut dioptimalkan menggunakan bantuan Artificial Intelligence

Protokol Review

Menyusun rencana kerja yang akan dilakukan, peneliti menentukan kriteria eksklusi dan inklusi. Kriteria inklusi mencakup: (a) artikel penelitian empiris atau kajian literatur komprehensif pada jurnal (*peer-reviewed*), (b) spesifik mengkaji psikometri dari model testlet pada ranah sains terpadu, biologi, fisika, atau kimia; (c) subjek populasi mencakup jenjang menengah (SMP/MTs, SMA/MA sederajat). Data diekstraksi berdasarkan entitas desain riset, model analisis psikometri yang digunakan, parameter kalibrasi butir, dan implikasi temuan.

Pelaksanaan (*Conducting*)

Pencarian Data

Tahap berikutnya, peneliti membuat *Search Query* yang akan digunakan untuk pencarian artikel berbantuan aplikasi Publish or Perish (PoP). Data yang digunakan berbasis data pada Google Scholar dan Crossref. Kata kunci pencarian yang digunakan merupakan kombinasi dari sintaks Boolean: ("*testlet model*" OR "*testlet effect*" OR "*testlet response theory*") AND ("*scientific literacy*" OR "*literasi sains*") AND ("*secondary education*" OR "*pendidikan menengah*" OR "*high school*") AND ("*psychometrics*" OR "*psikometri*" OR "*Rasch*").

Penyaringan (*Screening*)

Tahap penyaringan, dilakukan screening judul, screening abstrak, serta screening full teks menggunakan kriteria inklusi dan eksklusi. Penulis menggunakan aplikasi Zotero untuk membantu dalam proses penyaringan.

Ekstraksi dan Sintesis Data

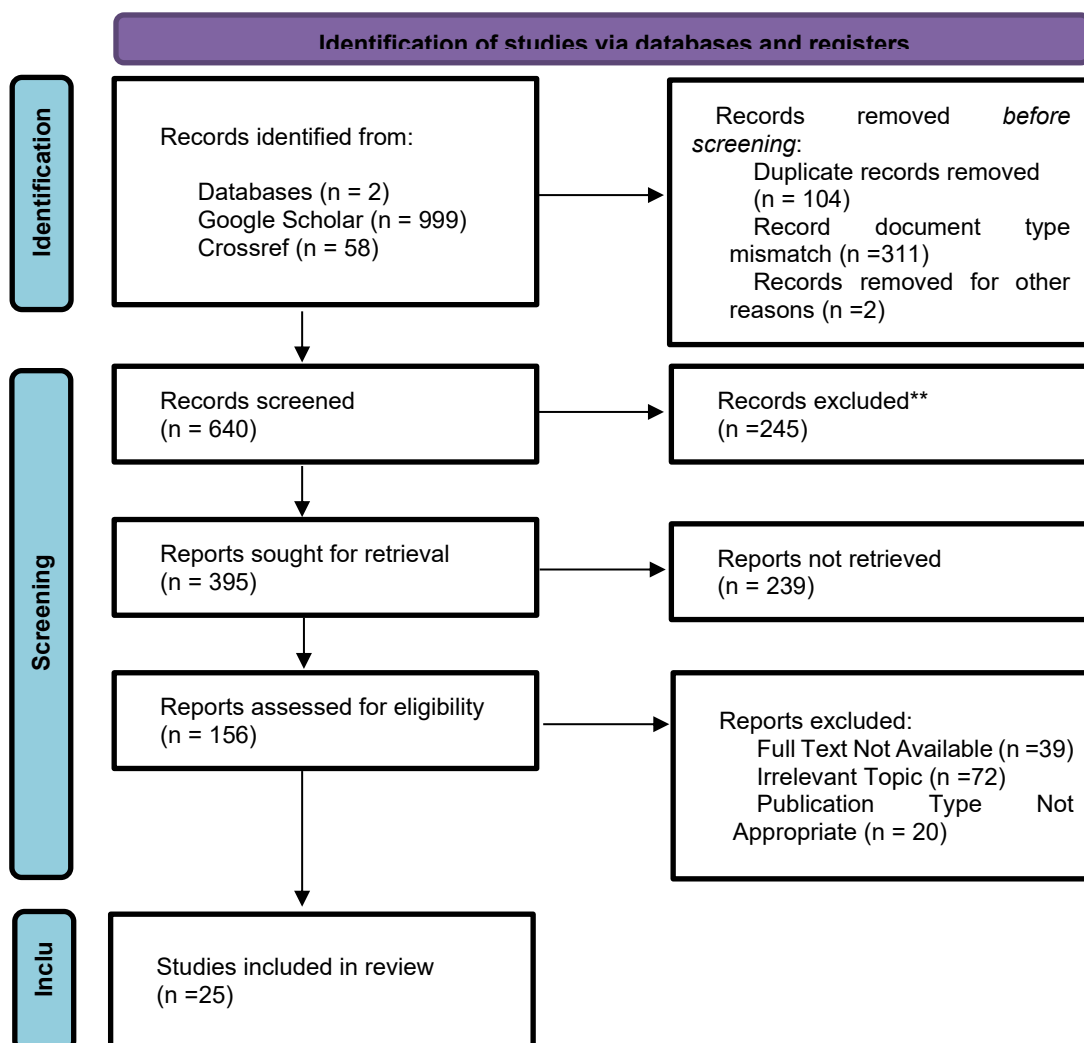
Peneliti mengambil data penting dari artikel yang terpilih, lalu dikelompokkan ke dalam tabel yang lebih ringkas dan dilanjutkan dengan menganalisis temuan.

Pelaporan (*Reporting*)

Pada tahap akhir, peneliti menyusun laporan akhir atau artikel ilmiah dengan mengikuti pedoman pelaporan yaitu *Preferred Reporting Items for Systematic Review and Meta Analyses* (PRISMA).

HASIL DAN PEMBAHASAN

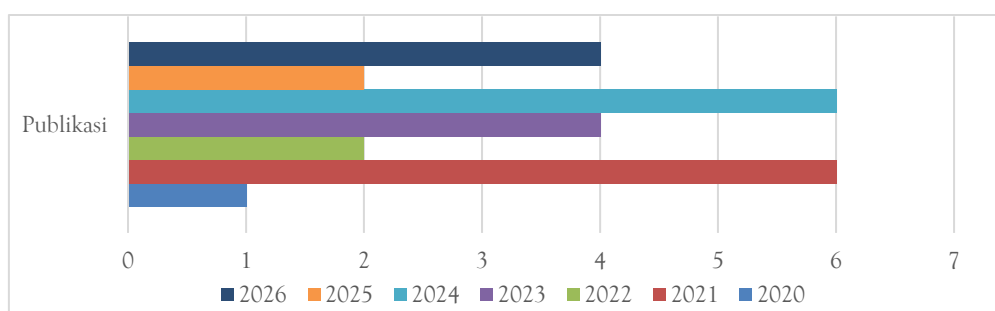
Penemuan artikel yang akan dianalisis dari proses identifikasi terhadap 999 database dari google scholar dan 58 dari database crossref. Selanjutnya dilakukan screening artikel menggunakan Zotero dan ditemukan 104 file duplikat. Gambaran bagan alir PRISMA sebagai berikut:



Gambar 1. Diagram Alir Metode PRISMA 2020

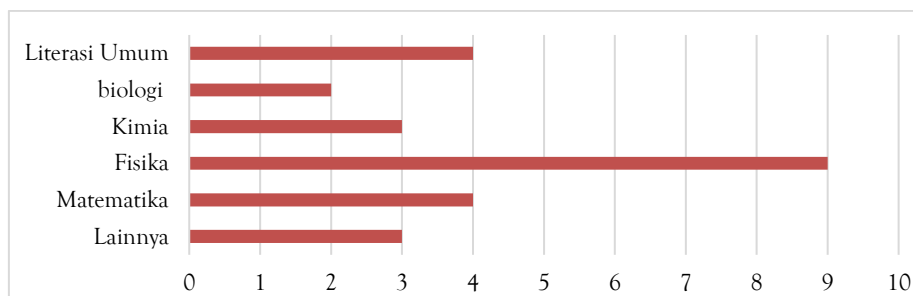
Tren penelitian penggunaan instrumen model testlet

Kebutuhan untuk mengukur literasi sains yang komprehensif mendorong peneliti untuk mulai mengadopsi testlet. Secara demografis, sasaran utama dari instrumen berbasis testlet ini adalah siswa Sekolah Menengah. Alasan utamanya adalah kompleksitas stimulus yang diberikan. Dalam konteks pendidikan di Indonesia, penelitian pengembangan instrumen testlet juga mulai marak seiring implementasi asesmen kompetensi minimum (AKM) yang konsepnya serupa dengan PISA. Analisis profil literasi sains siswa dan faktor yang memengaruhinya (Ni'mah, 2019).



Gambar 2. Distribusi Tahun Publikasi Penggunaan Model Testlet

Tren Penggunaan Testlet dalam mengukur literasi sains berdasarkan ekstraksi data, penggunaan format testlet untuk mengukur literasi sains mengalami lonjakan signifikan pasca tahun 2021 dan 2024. Hal ini sejalan dengan kemunculan kesadaran global akan perlunya instrumen psikometri yang kuat. banyak instrumen literasi sains jenjang SMA di Indonesia yang dikembangkan masih mengandalkan butir-butir independen berskala dikotomi, yang sering kali gagal merefleksikan keterampilan penalaran kritis (*higher order thinking skills/HOTS*). Namun, seiring dengan penetrasi kebijakan AKM dan standardisasi soal ala PISA, tren riset periode 2021-2026 memperlihatkan lonjakan penggunaan testlet.

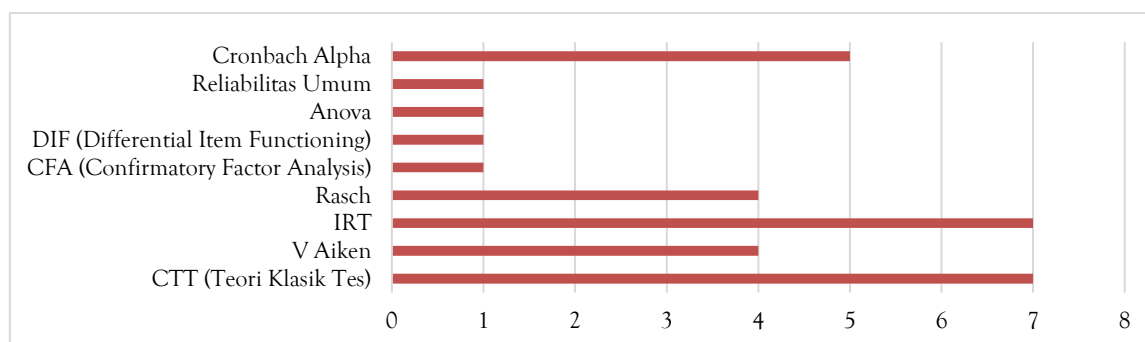


Gambar 3. Distribusi Bidang yang Diteliti

Pendidikan sains menduduki posisi yang esensial di mana stimulus berbasis teks sains, grafik pemanasan global, atau hasil eksperimen titrasi kimia menjadi instrumen utama. Habib (2023) menemukan bahwa pengintegrasian testlet di SMA mampu mengurangi *cognitive overload* peserta didik karena mereka tidak perlu membaca ulang konteks baru untuk setiap soal. Hal ini sejalan dengan premis bahwa instrumen yang valid harus mampu mencerminkan cara kerja sains secara autentik, bukan sekadar memori parsial. Jika ditinjau dari grafik gambar 2, terlihat bahwa bidang yang banyak mengintegrasikan testlet pada instrumen asesmen adalah fisika. Lingkungan pembelajaran yang menekankan pada HOTS lebih berpotensi mempersiapkan siswa menghadapi tantangan akademis dan sosial yang kompleks.

Di Indonesia, penerapan Kurikulum Merdeka juga memberikan penekanan kuat pada pengalaman belajar yang bermakna serta pengembangan kompetensi yang berorientasi pada HOTS (Faradella et al., 2024). Praktik pembelajaran masih lebih menekankan pada penghafalan prosedural daripada penalaran analitis. Akibatnya, siswa sering kali mengalami kesulitan saat dihadapkan pada persoalan fisika yang bersifat kontekstual dan melibatkan banyak langkah penyelesaian (Eichenlaub & Redish, 2019). Model penilaian testlet secara garis besar membantu pendidik untuk mengetahui kesulitan belajar fisika, karena materi fisika yang butuh penalaran tinggi, prosedur yang cukup panjang dan bersifat kontekstual.

Karakteristik psikometri penggunaan instrumen model testlet



Pada tinjauan psikometri ditemukan adanya pergeseran paradigma dalam memvalidasi instrumen literasi sains model testlet. Pada literatur periode 2020, sebagian besar pengembang instrumen masih menganalisis soal testlet dengan pendekatan Teori Tes Klasik (TTK). Selain itu, gambar 3 menunjukkan peningkatan penggunaan Teori Respons Butir (TRB). Beberapa studi mulai mengatasi LID dengan cara menghitung skor total dari setiap testlet, memperlakukan satu testlet sebagai satu polytomous item (butir politomus), lalu menganalisisnya menggunakan *Graded Response Model* (GRM) atau *Generalized Partial Credit Model* (GPCM). Pendekatan ini efektif mengeliminasi masalah LID karena dependensi lokal hanya terjadi antar butir, bukan antar testlet. Persoalan sentral yang banyak dielaborasi oleh para ahli psikometri pada kurun waktu ini adalah penanganan Testlet Effect

(Widyaningsih, 2020). Ketika butir-butir berbagi stem yang sama, varians ekstra muncul. berkurangnya jumlah parameter yang diestimasi ini terkadang dapat menurunkan informasi psikometris pengukuran (Zenisky et al., 2000).

Temuan dari berbagai literatur mengindikasikan adanya pergeseran paradigma dalam memvalidasi instrumen literasi sains model testlet. Beberapa studi terkini mulai mengatasi anomali Local Item Dependence (LID) dengan memperlakukan satu testlet sebagai satu butir politomus melalui pendekatan Graded Response Model (GRM) atau Generalized Partial Credit Model (GPCM) (Baghaei, 2023). Beberapa studi terkini mulai mengatasi anomali Local Item Dependence (LID) dengan memperlakukan satu testlet sebagai satu butir politomus melalui pendekatan Graded Response Model (GRM) atau Generalized Partial Credit Model (GPCM). Jika dibandingkan dengan instrumen yang dianalisis menggunakan Teori Tes Klasik (TTK) atau pendekatan diskrit konvensional yang mengabaikan independensi lokal, instrumen berpendekatan politomus ini dinilai jauh lebih efektif dalam mengeliminasi varians ekstra akibat berbagi stimulus yang sama (Testlet Effect) dan menghindari bias pada estimasi tingkat kesukaran butir. Namun demikian, masih terdapat kesenjangan dalam literatur; mayoritas kajian evaluasi psikometri asesmen literasi sains saat ini masih terbatas pada kalibrasi data simulasi atau skala lokal dengan sampel yang spesifik. Belum banyak riset empiris yang menguji dampak penerapan model testlet secara masif dalam membandingkan perbedaan fungsional pada beragam demografi peserta didik (Ha, 2022).

Melalui kajiannya terhadap tes pengetahuan hakikat sains, menegaskan bahwa jika efek LID tidak diakomodasi (misalnya tetap menggunakan Winsteps standar tanpa penskoran polikotomi (*partial credit model*), akan terjadi bias *item difficulty* (tingkat kesukaran butir), di mana soal seolah-olah terlihat lebih reliabel daripada kondisi aslinya. Independensi lokal mengacu pada kondisi di mana keberhasilan atau kegagalan dalam menyelesaikan satu soal hanya bergantung pada kemampuan individu, bukan pada faktor eksternal atau hubungan antar soal. Meskipun butir dalam satu kelompok testlet tidak independen, penting untuk memastikan bahwa skor antar kelompok testlet tetap independen (Hamdi, 2017). Independensi antar kelompok testlet adalah prasyarat yang penting dalam analisis IRT, hal ini dapat dibuktikan melalui uji fit model atau uji local independence. Jika independensi antar testlet terbukti, maka model ini dapat memberikan estimasi yang akurat tentang kemampuan peserta tes. Lestari (2024), menyatakan bahwa Tes pengetahuan hakikat sains yang terintegrasi dengan testlet dapat dikatakan hanya memuat satu dimensi. Dengan hasil tersebut dapat tarik kesimpulan bahwa tes pengetahuan hakikat sains dapat dinyatakan bersifat unidimensi.

Secara keseluruhan, hasil analisis terhadap kualitas psikometrik menunjukkan adanya kecenderungan penggunaan pendekatan psikometrik yang lebih komprehensif dan berbasis teori modern. Penerapan model Rasch mendominasi karena instrumen dianggap mampu memberikan pengukuran objektif dan bebas bias terhadap kemampuan peserta didik. Sementara EFA dan Cfa sebagai pelengkap untuk memperkuat struktur faktor dan memastikan bahwa konstruk yang diukur benar-benar mempresentasikan aspek literasi sains sesuai teori. Selain itu, teknik validasi isi seperti Aiken's V dan CVI tetap berada pada posisi penting pada tahap awal pengembangan, karena mampu menjamin kesesuaian indikator dengan kerangka konseptual yang digunakan. Kombinasi teknik klasik dan modern mencerminkan bahwa penelitian literasi sains tidak hanya berfokus pada hasil pengukuran, namun juga proses ilmiah membuat alat ukur yang akurat dan representatif. Pengutan validitas psikometrik menjadi aspek kunci dalam memperoleh kualitas hasil penelitian literasi sains dan memperkuat kontribusi terhadap pengembangan pendidikan berbasis teknologi di era transformasi digital.

Proyeksi Masa Depan Pengukuran Literasi Sains

Tabel 1. Tabel Kesenjangan dan Arah Penelitian

Kesenjangan	Arah Penelitian
Mayoritas instrumen masi berbasis Teori Tes Klasik (CTT) yang rentan bias akibat dependensi lokal antar-butir, serta kurangnya dukungan perangkat lunak untuk kalibrasi data berskala masif	Menerapkan <i>Graded Response Model</i> (GRM) dan Rasch Testlet Model terintegrasi cloud-computing guna mencapai stabilitas parameter dan presisi tingkat tinggi pada asesmen.
<i>Computerized testlet</i> saat ini belum memfasilitasi Higher Order Thinking Skill (HOTS) secara utuh, serta gagal mengukur korelasi antara dimensi kognitif dan afektif.	Merancang <i>computerize testlet</i> yang mengakomodasi evaluasi HOTS (penalaran) sekaligus mengukur hubungan dengan capaian pendidikan karakter secara holistik.

Kesenjangan	Arah Penelitian
Pengembangan testlet sering terisolasi pada topik spesifik dengan demografi purposive sampling yang sempit, sehingga membatasi kemampuan generalisasi instrumen.	Memperluas kerangka testlet ke berbagai sub disiplin sains dasar dengan melibatkan sampel yang lebih masif, inklusi, dan merepresentasikan demografi yang lebih luas.
Kajian literatur didominasi oleh evaluasi kelayakan psikometris instrumen, tanpa menguji dampak penggunaan testlet secara empiris terhadap intervensi pembelajaran di kelas.	Mengeksplorasi penggunaan testlet dalam lingkungan CAT untuk diagnosis <i>real time</i> , disertai studi longitudinal untuk memvalidasi sensitivitasnya terhadap efektivitas pembelajaran.

Berdasarkan perkembangan riset termutakhir menuju akhir tahun 2026, asesmen literasi sains terproyeksi mengalami otomasi tingkat lanjut. Aryadilah dkk. (2025) merumuskan tiga proyeksi utama. Pertama, pengintegrasian Computerized Adaptive Testing (CAT) dengan basis testlet. Testlet based CAT (t-CAT) diramalkan memiliki potensi yang signifikan untuk menjadi alternatif utama dari tes linier statis. Algoritma komputer tidak akan menyajikan soal per butir secara adaptif, melainkan memilih block testlet (bacaan beserta 3-5 pertanyaannya) yang paling selaras dengan tingkat kemampuan dari peserta didik. Sejalan dengan Nieminen et al., (2023), yang menekankan bahwa asesmen berbasis teknologi dapat meningkatkan efisiensi, transparansi, dan akurasi dalam pembelajaran sains berbasis inkuiri. Salah satu keuntungan utama penilaian berbasis web adalah kemampuannya untuk memberikan hasil dan umpan balik secara langsung (Sharmila et al., 2022).

Penggunaan *Artificial Intelligence* dalam konstruksi STEM literasi sains. Kemampuan AI untuk mengekstraksi data mentah dari fenomena alam (big data) menjadi paragraf naratif yang kompleks diyakini akan mempercepat produksi bank soal testlet secara masif, tentunya dengan tetap memerlukan validasi ahli. Novithania et al., (2024) menyatakan butir-butir soal pendukung dibuat memiliki tingkatan penyelesaian terhadap soal utama yang diberikan sehingga diharapkan dapat membantu pendidik untuk mendiagnosis kesulitan belajar siswa dengan efektif dan efisien.

Implikasi hasil penelitian dapat memperkuat landasan konseptual, metodologis, dan memberikan arah strategis bagi pengembangan kurikulum dan kebijakan pendidikan di era transformasi digital yang semakin kompleks. Mengukur literasi sains melalui pendekatan psikometrik dapat menjadi tolok ukur dalam merancang sistem pembelajaran yang lebih adaptif terhadap perubahan teknologi dan kebutuhan kompetensi abad ke-21. Dengan adanya instrumen yang terstandar dan teruji, lembaga pendidikan dapat secara sistematis menilai serta meningkatkan kemampuan literasi sains peserta didik dan pendidik secara berkelanjutan.

SIMPULAN

Tinjauan literatur sistematis ini menyimpulkan bahwa instrumen model testlet memiliki potensi yang sangat besar untuk mengukur kemampuan literasi sains secara efisien dan mendalam, menjembatani kekurangan tes pilihan ganda dan asesmen esai. Secara metodologis, ketergantungan butir lokal (LID) menjadi ancaman utama terhadap validitas skor jika hanya dianalisis menggunakan Teori Tes Klasik. Transisi menuju pemodelan politomus dalam kerangka Item Response Theory (IRT) dan adopsi Testlet Response Theory (TRT) ditunjukkan mampu mengatasi bias estimasi tersebut dan memberikan informasi kemampuan literasi sains peserta didik yang akurat. Rekomendasi bagi penelitian lanjutan adalah melakukan eksperimentasi empiris yang membandingkan parameter butir tes literasi sains berbasis testlet di Indonesia menggunakan model TRT dan model diagnostik kognitif. Karakteristik psikometri yang paling menonjol dari instrumen masa kini adalah keharusan mendeteksi dan memodelkan Local Item Dependence melalui varian Rasch Testlet Model atau MIRT agar koefisien reliabilitas tidak mengalami bias. Ke depannya, pertautan antara testlet based CAT dan otomasi kecerdasan buatan akan menjadi fondasi evaluasi pendidikan sains, para praktisi dan pengambil kebijakan untuk terus memperbarui kompetensi pengukuran psikometri mereka. Pengintegrasian deep learning, kontekstual atau isu-isu sosiosainifik, dan *Artifisial Intelligence* dapat diintegrasikan ke dalam instrumen penilaian model testlet agar menghasilkan instrumen yang valid, reliabel, adaptif, dan praktis.

Daftar Pustaka

Axmedovna, S. S., & Tojimurodovna, A. Y. (2021). PISA a Criteria for improving the quality of education. *Galaxy International Interdisciplinary Research Journal*, 9(05), 306–308.

- Ayan, C., & Pekmezci, F. B. (2021). The unit testlet dilemma: PISA sample. *International Journal of Assessment Tools in Education*, 8(3), 613–632. <https://doi.org/10.21449/ijate.948734>.
- Baghaei, P., & Christensen, K. B. (2023). Modelling local item dependence in C-Tests with the Loglinear Rasch Model. *Language Testing Journal*, 40(3), 820-827. doi: 10.1177/02655322231155109
- Depdiknas .2003. *Undang-undang RI No.20 tahun 2003 tentang sistem pendidikan nasional*.
- Eichenlaub, M., & Redish, E. F. (2019). Blending physical knowledge with mathematical form in physics problem solving. In G. Pospiech, M. Michelini, & B. S. Eylon (Eds.), *Mathematics in physics education* (pp. 101–120). Springer. https://doi.org/10.1007/978-3-030-04627-9_6
- Faradella, N. E., Wiyaka, W., & Egar, N. (2024). Teachers' strategies and challenges in conducting HOTS-based activities in Merdeka curriculum era. *Indonesian Journal of Education and Pedagogy*, 1(2), 75–90. <https://doi.org/10.61251/ijoe.v1i2.64>
- Gou, L., Zhou, W., & Li, X. (2024). Cognitive diagnosis testlet model for multiplechoice items. *Journal of Education and Behavioral Statistics*, 49(1). <https://doi.org/10.3102/10769986231165622>
- Gyamfi, A., & Acquaye, R. (2023). Parameters and models of item response theory (IRT): a review of literature. *Acta Educationis Generalis*, 13(3), 68–78. <https://doi.org/10.2478/atd-2023-0022>
- Ha, H. T. (2022). Test Format and Local Dependence of Items Revisited: A Case of Two Vocabulary Levels Test. *Brief Research Report*. doi: <https://doi.org.10.3389/fpsyg.2021.805450>
- Habib, Muhammad, dkk (2023). Pengembangan Instrumen Penilaian Model Testlet untuk Mengukur Kemampuan Berpikir Tingkat Tinggi pada Materi Ikatan Kimia. *Jurnal Ilmiah Ilmu Pendidikan*. 6(10). <https://doi.org/10.54371/jiip.v6i10.3015>
- Hambleton, R. K., & Swaminathan, H. (1991). *Item response theory principles and applications*. Kluwer Nijhoff Publishing.
- Hamdi, S. (2017). *Pengembangan instrumen tes matematika model testlet dengan permasalahan realistik untuk penilaian kelas di SD*. Universitas Negeri Yogyakarta.
- Johnson, E. E., O'Keefe, H., Sutton, A., & Marshall, C. (2022). The Systematic Review Toolbox: keeping up to date with tools to support evidence synthesis. *Systematic Reviews*, 11(1), 1–8. <https://10.1186/s13643-022-02122>
- Lestari, N. A., Susongko, P., & Hayati, M. N. (2024). Pengembangan Tes Pengetahuan Hakikat Sains pada Pelajaran IPA di SMP dengan Tema Cahaya dan Alat Optik. *PSEJ (Jurnal Pendidikan Sains Pancasakti)*, 9(2), 124-132. <https://doi.org/10.24905/psej.v9i2.224>
- Novithania, C. A., Maria, H.T., & Firdaus. (2024). Pengembangan Tes Diagnostik Testlet untuk Mendeteksi Kesulitan Belajar pada Materi Listrik Dinamis. *Jurnal Pendidikan dan Pengembangan*, 13(1), 121-126.
- Ni'mah, F. (2019). Research trends of scientific literacy in Indonesia: Where are we? *Jurnal Inovasi Pendidikan IPA*, 5(1), 23–30. <https://doi.org/10.21831/jipi.v5i1.20862>
- Ningsih, S.N., Istiyono, E., Widiastuti. An application of classical test theory for item characteristic analysis of chemical literacy instruments. *Jurnal Pendidikan Kimia Indonesia*. 7(2). 58-68. <https://doi.org/10.23887/jpki>.
- Ningsih, Sri, Nurfadillah. *Pengembangan Instrumen Pengukur Kompetensi Literasi Kimia Model Testlet pada Materi Ikatan Kimia Tingkat SMA*. Yogyakarta: UNY.
- Nieminen, J. H., Bearman, Margaret, & and Ajjawi, R. (2023). Designing the digital in authentic assessment: Is it fit for purpose? *Assessment & Evaluation in Higher Education*, 48(4), 529–543. <https://doi.org/10.1080/02602938.2022.2069674>
- OECD. (2023). *PISA 2022 results the state of learning and equity in education: Vol.I.* OECD. <https://doi.org/10.1787/53f23881-en>.

- Rahmasytoh, R., Yogica, R., & Fitri, R. (2026). Studi Literatur: Analisis Kualitatif Implementasi Problem-Based Learning (PBL) terhadap Kemampuan Berpikir Kritis dan Pemecahan Masalah Siswa pada Pembelajaran Biologi. *Jurnal Pendidikan MIPA*, 16(4), 1246-1254. <https://doi.org/10.37630/jpm.v16i4.4532>
- Rohmaya, N. (2022). Peningkatan Literasi Sains Siswa Melalui Pembelajaran IPA Berbasis Socioscientific Issues (SSI). *Jurnal Pendidikan MIPA*, 12(2), 107-117. <https://doi.org/10.37630/jpm.v12i2.553>
- Sharmila, S. K., Nikhila, M., Lakshmi Prasanna, J., Lavanya, M., & Bhavana, S. (2022). Unique Web-Based Assessment with a Secure Environment. *Proceedings of International Conference on Expert Clouds and Applications*, 725-736. DOI: 10.1109/ACCESS.2023.3339250
- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press
- Widyaningsih, S. W., Yusuf, I., Prasetyo, Z. K., & Istiyono, E. (2020). Online interactive multimedia oriented to HOTS through e-learning on physics material about electrical circuit. *Jurnal Pendidikan Indonesia*, 9(1), 1-14. <https://doi.org/10.23887/jpi-undiksha.v9i1.17667>
- Winarsi, D., & Ningsih, W. (2025). Pengaruh Penerapan Model Pembelajaran Project-Based Learning (PjBL) Terhadap Kemampuan Literasi Sains Peserta Didik. *Jurnal Pendidikan MIPA*, 15(3), 1161-1166. <https://doi.org/10.37630/jpm.v15i3.3220>
- Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2000). Effects of Local Item Dependence on the Validity of IRT Item, Test, and Ability Statistics. Paper presented at the Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.