

Analisis Butir Soal (*Item Analysis*) pada Asesmen Cambridge Checkpoint Mata Pelajaran Matematika

Rusyda Maulida^{*1}, Bani², Muchlis³

^{1,2} Teknik Informatika, Fakultas Ilmu Komputer, Universitas Pamulang, Indonesia

³ Teknik Informatika, STIMIK Antar Bangsa, Indonesia

Email: ¹dosen02114@unpam.ac.id, ²dosen02381@unpam.ac.id, ³muchlis.re@gmail.com

Abstrak

Penelitian ini bertujuan untuk mengevaluasi kualitas psikometris instrumen Asesmen Cambridge Checkpoint Matematika (Syllabus 0862) melalui studi analisis butir soal (item analysis) yang mencakup pengujian reliabilitas, tingkat kesukaran, dan daya pembeda guna memastikan efektivitas alat ukur dalam konteks sekolah internasional di Indonesia. Penelitian deskriptif kuantitatif ini menggunakan pendekatan Teori Tes Klasik (Classical Test Theory/CTT) dengan data yang dianalisis menggunakan bahasa pemrograman Python pada platform Google Colaboratory terhadap sampel 57 siswa yang menempuh Paper 1 (24 butir) dan Paper 2 (25 butir) pada periode Oktober 2024. Hasil analisis menunjukkan bahwa instrumen memiliki koefisien reliabilitas Cronbach's Alpha yang sangat tinggi, yakni 0,8945 untuk Paper 1 dan 0,9056 untuk Paper 2. Dari aspek tingkat kesukaran, mayoritas soal (51%) berada pada kategori sedang, 31% kategori mudah, dan 18% kategori sukar, sementara analisis daya pembeda menunjukkan bahwa 87,23% butir soal (41 dari total 47 butir) berada dalam kategori baik hingga sangat baik ($D > 0,40$) dalam membedakan kemampuan siswa, meskipun ditemukan kesenjangan kompetensi spesifik pada materi Aljabar dan Geometri. Penelitian ini menyimpulkan bahwa Asesmen Cambridge Checkpoint merupakan instrumen evaluasi yang valid, reliabel, dan adil untuk diterapkan.

Kata kunci: *analisis butir soal, cambridge checkpoint, daya pembeda, matematika, reliabilitas.*

Abstract

This study aims to evaluate the psychometric quality of the Cambridge Checkpoint Mathematics Assessment instrument (Syllabus 0862) through an item analysis study covering reliability, item difficulty, and discrimination index to ensure the tool's effectiveness in international schools in Indonesia. Employing a quantitative descriptive method with a Classical Test Theory (CTT) approach, data were analyzed using Python programming on the Google Colaboratory platform from a sample of 57 students who completed Paper 1 (24 items) and Paper 2 (25 items) in October 2024. The results revealed that the instrument possessed very high Cronbach's Alpha reliability coefficients of 0.8945 for Paper 1 and 0.9056 for Paper 2. In terms of difficulty levels, 51% of items were moderate, 31% easy, and 18% difficult, while the discrimination index analysis showed that 87.23% of the items (41 out of 47 items) were categorized as good to ($D > 0,40$), despite identifying specific student competency gaps in Algebra and Geometry. Overall, this study confirms that the Cambridge Checkpoint Assessment is a valid, reliable, and fair evaluation instrument.

Keywords: *cambridge checkpoint, discriminating power, item analysis, mathematics, reliability.*

1. PENDAHULUAN

Pendidikan merupakan fondasi utama dalam pembentukan sumber daya manusia yang berkualitas. Dalam konteks pendidikan, asesmen atau evaluasi memegang peran sentral sebagai alat untuk mengukur dan menilai keberhasilan proses pembelajaran. Asesmen yang baik tidak hanya berfungsi untuk menentukan nilai akhir siswa, tetapi juga memberikan umpan balik yang konstruktif bagi guru, siswa, dan institusi pendidikan untuk memperbaiki strategi pembelajaran dan kurikulum. Oleh karena itu, kualitas instrumen asesmen, dalam hal ini butir soal, menjadi faktor krusial yang menentukan validitas dan reliabilitas hasil evaluasi.

Di Indonesia, berbagai instrumen asesmen telah digunakan, baik yang dikembangkan secara lokal maupun yang diadopsi dari kurikulum internasional. Salah satu asesmen internasional yang semakin populer adalah Cambridge Checkpoint, yang digunakan di banyak sekolah dengan kurikulum

internasional. Asesmen ini bertujuan untuk mengukur kemajuan belajar siswa pada akhir tahap pendidikan dasar dan menengah (Assessment International Education, 2021). Mengingat peran pentingnya, asesmen Cambridge Checkpoint harus memiliki butir-butir soal yang berkualitas tinggi agar dapat memberikan gambaran yang akurat tentang pencapaian kompetensi siswa, khususnya dalam mata pelajaran yang memerlukan penalaran kompleks seperti Matematika (Cambridge International Education, 2024).

Analisis butir soal (*item analysis*) adalah proses evaluasi sistematis yang dilakukan untuk menguji kualitas setiap butir soal dalam suatu tes. Kegiatan ini menjadi sangat esensial karena memberikan informasi mendetail mengenai karakteristik butir soal, seperti tingkat kesukaran, daya pembeda, dan efektivitas pengecoh. Berbagai studi telah menunjukkan bahwa analisis butir soal adalah langkah vital dalam penyusunan tes yang bermutu. Sebagai contoh, penelitian yang dilakukan dan menunjukkan bahwa analisis butir soal dapat mengukur validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan fungsi pengecoh pada soal ujian (Pradani & Efendi, 2023). Demikian pula, penelitian yang telah dilakukan dan menegaskan bahwa analisis butir soal sangat penting untuk menilai kesesuaian soal dengan kisi-kisi, validitas, reliabilitas, tingkat kesukaran, dan daya pembeda (Syufa Auliya et al., 2025).

Pendekatan yang paling sering digunakan dalam analisis butir soal adalah Teori Tes Klasik (Classical Test Theory/CTT). Teori ini menawarkan metode yang relatif sederhana untuk menganalisis karakteristik butir soal dan keseluruhan tes (Jafar, Riyadi, & Ridwan, 2024). Meskipun ada pendekatan yang lebih modern seperti Teori Respon Butir (Item Response Theory/IRT), CTT tetap relevan dan banyak digunakan karena kemudahannya dalam implementasi. CTT berfokus pada hubungan antara skor yang diamati (observed score) dengan skor yang sebenarnya (true score) dan error pengukuran (measurement error). Analisis butir soal dalam kerangka CTT mencakup penghitungan indeks yang mencerminkan kualitas butir soal tersebut.

Meskipun urgensi analisis butir soal telah banyak dibahas, sebagian besar penelitian yang ada berfokus pada tes lokal yang dikembangkan oleh guru atau sekolah di Indonesia (Savika & Zuhriyah, 2023). Studi-studi ini sering kali menganalisis butir soal ujian semester atau ujian sekolah pada kurikulum nasional. Namun, penelitian yang secara spesifik meneliti kualitas butir soal pada asesmen internasional seperti Cambridge Checkpoint masih sangat minim. Asesmen Cambridge Checkpoint memiliki standar kurikulum, format soal, dan sasaran kompetensi yang berbeda, sehingga analisis butir soal pada instrumen ini memerlukan pendekatan yang spesifik (Handayati, 2022). Selain itu, hasil dari analisis ini akan memberikan wawasan yang lebih dalam tentang bagaimana butir soal pada asesmen internasional beroperasi di lingkungan pendidikan Indonesia dan memberikan rekomendasi yang lebih tepat kepada guru dan sekolah yang menggunakannya.

Penelitian ini berangkat dari kebutuhan mendesak untuk memastikan bahwa asesmen yang digunakan untuk mengevaluasi siswa di Indonesia, termasuk asesmen internasional, benar-benar valid dan reliabel. Asesmen yang tidak berkualitas dapat memberikan informasi yang menyesatkan tentang kemampuan siswa, yang pada akhirnya dapat mengarah pada keputusan pendidikan yang tidak tepat. Mengingat bahwa asesmen Cambridge Checkpoint sering kali menjadi acuan untuk melanjutkan studi ke jenjang berikutnya atau program internasional, memastikan kualitas butir soalnya menjadi hal yang sangat vital.

2. METODE PENELITIAN

Penelitian ini merupakan penelitian evaluatif dengan pendekatan deskriptif kuantitatif. Pendekatan deskriptif kuantitatif dipilih karena tujuan utamanya adalah untuk mendeskripsikan dan menganalisis secara mendalam kualitas butir-butir soal pada Asesmen Cambridge Checkpoint Mata Pelajaran Matematika berdasarkan data numerik yang diperoleh. Jenis penelitian ini sangat efektif untuk mengevaluasi kualitas instrumen tes dari berbagai aspek. Sifat evaluatif dari penelitian ini memungkinkan peneliti untuk tidak hanya mengumpulkan dan menganalisis data, tetapi juga untuk memberikan penilaian terhadap kualitas butir soal dan menyajikan rekomendasi perbaikan yang didasarkan pada temuan empiris.

Pemilihan desain penelitian evaluatif dalam studi ini didasarkan pada kebutuhan untuk memberikan pertimbangan nilai (*value judgment*) terhadap instrumen Cambridge Checkpoint. Berbeda dengan

penelitian deskriptif murni yang hanya memaparkan data, penelitian evaluatif berfungsi untuk membandingkan temuan lapangan dengan kriteria atau standar kualitas yang telah ditetapkan sebelumnya. Penelitian evaluatif memungkinkan peneliti untuk mengidentifikasi kekuatan dan kelemahan suatu program atau instrumen secara sistematis, sehingga hasil akhirnya bukan sekadar angka, melainkan rekomendasi aksi untuk perbaikan mutu pendidikan (Buka et al., 2025).

Pendekatan kuantitatif deskriptif diintegrasikan guna memberikan objektivitas dalam pengukuran karakteristik psikometris soal. Dalam pendekatan ini, data berupa skor mentah dari lembar jawaban siswa dikonversi menjadi indeks statistik yang mencerminkan realitas kualitas soal. Penggunaan angka-angka statistik dalam evaluasi pendidikan memberikan landasan yang kuat bagi peneliti untuk melakukan generalisasi secara terbatas pada populasi yang diteliti (Mania et al., 2020). Dengan dukungan data numerik, interpretasi mengenai tingkat kesulitan atau daya pembeda soal menjadi lebih presisi dan dapat dipertanggungjawabkan secara ilmiah, meminimalisir subjektivitas peneliti dalam menilai kualitas sebuah asesmen.

Penggunaan statistik deskriptif dalam penelitian ini mencakup perhitungan tendensi sentral, distribusi frekuensi, dan korelasi antarvariabel. Analisis ini sangat esensial untuk membedah data yang bersumber dari kurikulum internasional seperti Cambridge, karena memungkinkan peneliti untuk melihat pola performa siswa Indonesia secara agregat. Melalui kerangka ini, peneliti dapat mendeteksi apakah distribusi skor siswa mengikuti kurva normal atau justru terjadi penyimpangan yang mengindikasikan bahwa instrumen tersebut terlalu sulit atau terdapat bias instruksional pada topik matematika tertentu.

Sifat evaluatif penelitian ini juga mencakup analisis komprehensif terhadap keterkaitan antar parameter. Peneliti tidak hanya melihat satu variabel secara terisolasi, melainkan mengevaluasi bagaimana validitas butir soal berkorelasi dengan reliabilitas tes secara keseluruhan. Pendekatan kuantitatif dalam evaluasi sangat efektif untuk menguji konsistensi instrumen dalam skala yang lebih luas. Hal ini memberikan nilai tambah bagi institusi pendidikan pelaksana kurikulum Cambridge, karena mereka mendapatkan potret empiris mengenai sejauh mana asesmen tersebut benar-benar mampu mengukur kompetensi matematika siswa di sekolah mereka (Sugiyono, 2021).

Sebagai tahap akhir dari metode ini, seluruh temuan kuantitatif akan disintesis menjadi sebuah dokumen rekomendasi teknis. Rekomendasi ini dirancang untuk menjembatani antara temuan statistik dengan praktik pembelajaran di kelas. Dengan demikian, penelitian ini menjalankan fungsi evaluasi secara utuh, mulai dari pengumpulan bukti empiris, analisis mendalam, hingga pemberian solusi praktis. Pendekatan ini memastikan bahwa laporan kemajuan yang disusun memiliki kedalaman akademik sekaligus manfaat aplikatif bagi para pendidik dalam mengoptimalkan penggunaan asesmen internasional di lingkungan lokal.

Populasi dalam penelitian ini adalah seluruh siswa kelas 9 di salah satu sekolah internasional yang telah mengikuti Asesmen Cambridge Checkpoint Mata Pelajaran Matematika pada tahun ajaran tertentu. Asesmen ini dipilih karena relevansinya dengan kurikulum internasional dan pentingnya untuk memastikan kualitas instrumennya dalam konteks pendidikan di Indonesia.

Populasi penelitian terdiri atas 59 siswa kelas 9 yang mengikuti Asesmen Cambridge Checkpoint Mathematics (Syllabus 0862) pada periode Oktober 2024 di salah satu sekolah internasional di Indonesia. Sampel penelitian berjumlah 46 siswa yang dipilih menggunakan teknik *purposive sampling*, yaitu berdasarkan kriteria: (1) mengikuti seluruh rangkaian asesmen *Cambridge Checkpoint*, (2) menyelesaikan seluruh butir soal pada Paper 1 dan Paper 2, serta (3) memiliki lembar jawaban yang lengkap dan dapat diakses untuk keperluan analisis. Sebanyak 13 siswa tidak diikutsertakan dalam analisis karena tidak memenuhi kriteria tersebut, terutama akibat data respons yang tidak lengkap. Pemilihan sampel ini bertujuan untuk memastikan bahwa seluruh data yang dianalisis memenuhi persyaratan analisis butir soal menggunakan pendekatan *Classical Test Theory (CTT)*, sehingga hasil evaluasi kualitas butir soal dapat diperoleh secara akurat (Saputra et al., 2022).

Penentuan populasi dalam penelitian ini difokuskan pada siswa kelas 9 di sekolah menengah pertama yang menerapkan kurikulum internasional secara penuh. Pemilihan jenjang kelas 9 bersifat strategis, mengingat *Cambridge Checkpoint* diposisikan sebagai asesmen kulminasi yang menandai berakhirnya fase pendidikan menengah pertama sebelum siswa melangkah ke jenjang IGCSE

(*International General Certificate of Secondary Education*). Karakteristik populasi ini memiliki keunikan tersendiri karena mereka telah terpapar secara intensif pada pola instruksi matematika berbasis standar Cambridge, sehingga respons mereka terhadap butir soal diharapkan dapat memberikan gambaran yang akurat mengenai reliabilitas instrumen tersebut dalam konteks pendidikan di Indonesia.

Penerapan teknik *purposive sampling* dalam studi ini bertujuan untuk menjaga integritas data yang diolah. Dalam analisis butir soal secara kuantitatif, keberadaan data yang tidak lengkap atau siswa yang tidak serius dalam mengerjakan tes dapat menimbulkan bias pada perhitungan tingkat kesukaran dan daya pembeda. Sebagaimana dijelaskan oleh (Sugiyono, 2021), *purposive sampling* memungkinkan peneliti untuk menyaring subjek berdasarkan pertimbangan teknis tertentu guna mencapai tujuan penelitian yang lebih spesifik. Dengan membatasi sampel hanya pada siswa yang menyelesaikan seluruh butir soal, peneliti dapat memastikan bahwa fluktuasi statistik yang muncul benar-benar murni merefleksikan karakteristik soal, bukan karena faktor kelalaian peserta tes.

Kriteria inklusi yang ditetapkan yaitu kelengkapan lembar jawaban dan aksesibilitas data berperan sebagai kontrol kualitas pada tahap awal penelitian. Dalam prosedur analisis menggunakan perangkat lunak seperti ITEMAN atau SPSS, data yang hilang (*missing data*) sering kali menjadi kendala dalam estimasi parameter psikometris. Oleh karena itu, pemilihan sampel yang "bersih" secara administratif menjadi prasyarat mutlak. (Mania et al., 2020) menegaskan bahwa ketepatan dalam pemilihan sampel kuantitatif akan sangat menentukan validitas generalisasi hasil evaluasi, terutama ketika penelitian bertujuan untuk memberikan rekomendasi perbaikan instrumen pada skala institusional.

Meskipun dalam penelitian evaluasi instrumen berskala besar sering digunakan *total sampling* (pengambilan seluruh anggota populasi), penggunaan *purposive sampling* dalam konteks satu sekolah internasional dipandang lebih realistis dan efisien. Hal ini didasarkan pada argumen bahwa penelitian ini tidak sekadar menghitung skor, melainkan mengevaluasi efektivitas fungsi pengecoh dan konstruksi soal pada kelompok yang sudah benar-benar siap secara kurikulum. Pendekatan ini sejalan dengan pandangan (Saputra et al., 2022) yang menyebutkan bahwa dalam penelitian evaluatif, kualitas data yang dikumpulkan jauh lebih penting daripada kuantitas subjek yang sekadar berpartisipasi tanpa memenuhi standar kelayakan data.

Terakhir, sampel yang terpilih nantinya akan dipetakan ke dalam dua kelompok utama untuk keperluan analisis daya pembeda: kelompok atas (*upper group*) dan kelompok bawah (*lower group*). Pembagian ini biasanya mengambil porsi 27% hingga 33% dari total sampel yang memenuhi kriteria, tergantung pada distribusi skor yang diperoleh. Dengan memastikan bahwa sampel penelitian ini homogen dalam hal keikutsertaan kurikulum namun heterogen dalam kemampuan akademis, penelitian ini diharapkan mampu membedah anatomi setiap butir soal Matematika Cambridge secara komprehensif, memberikan landasan empiris bagi sekolah untuk mengevaluasi kesiapan siswa mereka di level internasional.

Teknik pengumpulan data yang akan digunakan adalah dokumentasi. Teknik ini dipilih karena data yang dibutuhkan sudah tersedia dalam bentuk dokumen yang tidak memerlukan interaksi langsung dengan subjek penelitian. Data yang akan dikumpulkan mencakup: naskah soal Asesmen Cambridge Checkpoint Mata Pelajaran Matematika, lembar jawaban dari sampel siswa yang telah dipilih, kunci jawaban dari asesmen tersebut.

Teknik dokumentasi ini efisien dan memungkinkan pengumpulan data dari arsip yang ada di sekolah, mirip dengan metodologi yang digunakan dalam penelitian mereka tentang analisis butir soal ujian (Inten et al., 2021). Penggunaan teknik dokumentasi dalam penelitian ini dipandang sebagai metode yang paling objektif dan efisien untuk memperoleh data primer tanpa mengganggu proses belajar mengajar di sekolah.

Dalam konteks evaluasi instrumen yang sudah terstandarisasi seperti *Cambridge Checkpoint*, data dokumenter memberikan rekaman autentik mengenai performa siswa yang tidak terpengaruh oleh faktor subjektivitas peneliti atau efek psikologis subjek saat wawancara (*Hawthorne effect*). Penelitian kuantitatif memungkinkan peneliti untuk membedah data historis secara retrospektif, sehingga memberikan gambaran yang stabil mengenai kualitas sebuah instrumen evaluasi dalam periode waktu tertentu (Sudaryono, 2021).

Data utama yang dihimpun dalam studi ini adalah naskah soal asli dan lembar jawaban siswa (*student response sheets*). Pengumpulan naskah soal sangat krusial untuk melakukan analisis kualitatif terhadap konstruksi butir, validitas isi, serta kesesuaian dengan kisi-kisi kurikulum. Sementara itu, lembar jawaban siswa berfungsi sebagai sumber data mentah yang akan dikonversi ke dalam bentuk matriks biner (0 untuk jawaban salah dan 1 untuk jawaban benar). Kelengkapan dokumen pendukung, termasuk kunci jawaban resmi, merupakan syarat mutlak agar proses skoring dan analisis statistik selanjutnya memiliki tingkat akurasi yang tinggi.

Kekuatan teknik dokumentasi dalam penelitian ini juga terletak pada kemampuannya untuk menyediakan data yang kaya bagi analisis efektivitas pengecoh (*distractor analysis*). Dengan mendokumentasikan setiap pilihan jawaban yang diambil oleh siswa, peneliti dapat memetakan pola miskonsepsi secara detail. Dokumen lembar jawaban tidak hanya dipandang sebagai deretan angka, melainkan sebagai jejak penalaran matematis siswa terhadap tantangan standar internasional. Arsip sekolah yang dikelola dengan baik merupakan laboratorium data yang sangat berharga untuk mengevaluasi efektivitas kurikulum dan asesmen yang sedang berjalan (Mania et al., 2020).

Selain efisiensi waktu dan biaya, teknik dokumentasi menjamin kerahasiaan dan integritas data peserta tes. Mengingat asesmen *Cambridge Checkpoint* adalah instrumen yang bersifat rahasia dan memiliki hak kekayaan intelektual tertentu, prosedur dokumentasi dilakukan dengan tetap memperhatikan aspek legalitas dan izin dari pihak otoritas sekolah. Peneliti bertindak sebagai evaluator yang mengolah data anonim, sehingga fokus penelitian tetap terjaga pada kualitas butir soal dan bukan pada identitas personal siswa. Hal ini sejalan dengan prinsip etika penelitian pendidikan yang ditekankan oleh (Sugiyono, 2021) mengenai pemanfaatan data sekunder dan dokumen internal instansi.

Sebagai langkah akhir dari teknik ini, seluruh dokumen yang terkumpul diverifikasi validitasnya melalui prosedur *cross-check* dengan jadwal ujian dan berita acara pelaksanaan asesmen. Sinkronisasi data antara naskah soal, lembar jawaban, dan kunci jawaban resmi memastikan bahwa hasil analisis butir soal nantinya benar-benar mencerminkan standar Cambridge yang diujikan. Dengan demikian, teknik dokumentasi ini menjadi fondasi yang kokoh bagi tahapan analisis kuantitatif, memberikan data yang bersih dan reliabel untuk diolah lebih lanjut guna menghasilkan rekomendasi perbaikan kualitas asesmen matematika di masa depan.

Instrumen yang dianalisis dalam penelitian ini adalah Asesmen Cambridge Checkpoint Mathematics Syllabus 0862 yang terdiri atas dua komponen, yaitu Paper 1 sebanyak 24 butir soal dan Paper 2 sebanyak 25 butir soal, sehingga total terdapat 49 butir soal. Kedua paper dirancang untuk mengukur kompetensi matematika siswa berdasarkan kerangka kurikulum Cambridge International. Analisis dilakukan terhadap seluruh butir soal menggunakan data respons dari 46 siswa yang memenuhi kriteria sampel.

Data yang telah dikumpulkan akan dianalisis secara kuantitatif menggunakan pendekatan Teori Tes Klasik (CTT). Data dianalisis menggunakan pendekatan *Classical Test Theory (CTT)* dengan bantuan bahasa pemrograman *Python* yang dijalankan melalui *platform Google Colaboratory*. Pengolahan data memanfaatkan pustaka *pandas* untuk pengelolaan data, *NumPy* untuk komputasi numerik, *SciPy* untuk analisis statistik, serta *Pingouin* untuk menghitung koefisien reliabilitas *Alpha Cronbach*. Analisis meliputi pengujian validitas butir menggunakan korelasi item-total, reliabilitas menggunakan *Alpha Cronbach*, indeks tingkat kesukaran, dan indeks daya pembeda setiap butir soal.

Proses analisis diawali dengan pembersihan data (*data cleaning*), dilanjutkan dengan identifikasi butir soal pada masing-masing paper, perhitungan skor total setiap peserta, kemudian dilakukan penghitungan parameter psikometrik yang meliputi validitas butir, reliabilitas, indeks tingkat kesukaran, dan indeks daya pembeda. Seluruh proses dilakukan secara otomatis menggunakan skrip *Python* sehingga meningkatkan reproduksibilitas penelitian.

Teknik analisis data akan mencakup perhitungan untuk setiap parameter kualitas butir soal yang telah ditetapkan.

1. Validitas akan diukur melalui korelasi antara skor butir soal dengan skor total tes. Butir soal yang memiliki korelasi yang signifikan dengan skor total dianggap valid. Reliabilitas tes secara keseluruhan akan dihitung menggunakan formula *Alpha Cronbach*, yang memberikan indikasi konsistensi internal tes.

2. Tingkat kesukaran (*difficulty index*) akan dihitung dengan rumus proporsi siswa yang menjawab benar untuk setiap butir soal. Hasilnya akan dikategorikan menjadi "mudah," "sedang," dan "sukar" untuk mengidentifikasi sebaran tingkat kesulitan soal.
3. Daya pembeda (*discrimination index*) butir soal akan dihitung dengan membandingkan proporsi jawaban benar antara kelompok siswa dengan skor tinggi (kelompok atas) dan kelompok siswa dengan skor rendah (kelompok bawah). Butir soal yang baik memiliki daya pembeda positif.

Pada Analisis butir soal pada Asesmen Cambridge Checkpoint Matematika (Syllabus 0862) dilakukan terhadap dua komponen utama, yaitu Paper 1 dan Paper 2, dengan melibatkan sampel sebanyak 46 siswa. Berdasarkan data mentah yang diperoleh, dilakukan pengolahan data menggunakan Teori Tes Klasik (CTT) untuk mengevaluasi kualitas psikometris instrumen tersebut.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Penelitian

3.1.1 Analisis Validitas dan Reliabilitas

Seluruh analisis dilakukan menggunakan *Python* pada *Google Colaboratory* berdasarkan algoritma *Classical Test Theory*. Pengukuran reliabilitas menggunakan formula *Alpha Cronbach* menunjukkan angka 0,84 untuk Paper 1 dan 0,82 untuk Paper 2. Hal ini mengindikasikan bahwa tes tersebut memiliki konsistensi internal yang sangat tinggi, melampaui ambang batas minimum 0,70 yang sering dijadikan standar dalam penelitian pendidikan. Tingginya angka ini membuktikan bahwa butir-butir soal dalam asesmen Cambridge mampu menghasilkan skor yang stabil jika diujikan pada kondisi yang serupa.

Tabel 1. Koefisien Reliabilitas Tes

Komponen	Jumlah Butir	Reponden	Alpha Cronbach	Interpretasi
Paper 1 (0862/01)	24	57	0,8945	Sangat Tinggi
Paper 2 (0862/02)	25	57	0,9056	SangatTinggi

Validitas butir soal dianalisis melalui korelasi poin-biserial antara skor tiap butir dengan skor total. Hasil menunjukkan bahwa sebagian besar butir soal (88%) memiliki korelasi positif yang signifikan di atas 0,30. Secara teknis, ini berarti soal-soal tersebut memiliki arah yang sama dalam mengukur domain kemampuan matematika siswa. Keberadaan validitas yang kuat ini menjamin bahwa instrumen tersebut tidak mengukur faktor luar (seperti kemampuan bahasa yang bias) melainkan murni kemampuan matematis.

Tabel 2. Validitas Klasifikasi untuk Paper 1

	Item-Total Correlation	Difficulty Index	Discriminati on Index	Correlation Classification	Difficulty Classification	Discrimination Classification
0	0.665960	0.403509	0.8750	Good	Medium	Very Good
1	0.573602	1.070175	1.0625	Good	Easy	Very Good
2	0.532066	0.456140	1.0000	Good	Medium	Very Good
3	0.652360	1.000000	1.6250	Good	Easy	Very Good
4	0.689297	0.649123	1.4375	Good	Medium	Very Good
5	0.473982	0.491228	0.5625	Good	Medium	Very Good
6	0.311072	0.087719	0.1875	Good	Difficult	Poor (Needs Review)
7	0.502599	0.438596	0.6875	Good	Medium	Very Good
8	0.403350	0.473684	0.5625	Good	Medium	Very Good
9	0.677537	0.385965	1.1250	Good	Medium	Very Good
10	0.755457	2.245614	3.3125	Good	Easy	Very Good
11	0.123257	0.315789	0.0000	Fair	Medium	Very Poor (Reject)

	Item-Total Correlation	Difficulty Index	Discriminati on Index	Correlation Classification	Difficulty Classification	Discrimination Classification
12	0.228976	0.017544	0.0625	Fair	Difficult	Very Poor (Reject)
13	0.691072	0.263158	0.6875	Good	Medium	Very Good
14	0.386886	0.421053	0.5625	Good	Medium	Very Good
15	0.586797	0.771930	1.3125	Good	Medium	Very Good
16	0.741968	0.157895	0.5000	Good	Difficult	Very Good
17	0.743295	0.473684	1.1250	Good	Medium	Very Good
18	0.636096	1.000000	1.5000	Good	Easy	Very Good
19	0.793871	0.877193	2.4375	Good	Easy	Very Good
20	0.641010	0.175439	0.5625	Good	Difficult	Very Good
21	0.666389	0.824561	1.1875	Good	Easy	Very Good
22	0.372423	0.596491	0.4375	Good	Medium	Very Good

Tabel 3. Validitas Klasifikasi untuk Paper 2

	Item-Total Correlation	Difficulty Index	Discrimination Index	Correlation Classification	Difficulty Classification	Discrimination Classification
0	0.430971	0.578947	0.5000	Good	Medium	Very Good
1	0.243432	0.385965	0.3750	Fair	Medium	Good
2	0.414152	0.701754	0.5000	Good	Medium	Very Good
3	0.611088	0.280702	0.7500	Good	Medium	Very Good
4	0.377755	0.719298	0.4375	Good	Medium	Very Good
5	0.467942	0.070175	0.1875	Good	Difficult	Poor (Needs Review)
6	0.628187	0.157895	0.4375	Good	Difficult	Very Good
7	0.420495	0.543860	0.4375	Good	Medium	Very Good
8	0.709331	1.157895	2.0000	Good	Easy	Very Good
9	0.569402	0.210526	0.4375	Good	Medium	Very Good
10	0.604130	0.350877	0.6875	Good	Medium	Very Good
11	0.599625	1.210526	1.9375	Good	Easy	Very Good
12	0.623423	0.192982	0.5000	Good	Difficult	Very Good
13	0.696567	0.280702	0.8750	Good	Medium	Very Good
14	0.837077	0.491228	1.5625	Good	Medium	Very Good
15	0.730447	1.017544	1.8125	Good	Easy	Very Good
16	0.733060	0.280702	1.0000	Good	Medium	Very Good
17	0.737592	1.157895	2.0000	Good	Easy	Very Good
18	0.643537	0.315789	0.7500	Good	Medium	Very Good
19	0.642129	0.263158	0.7500	Good	Medium	Very Good
20	0.313391	0.157895	0.1875	Good	Difficult	Poor (Needs Review)
21	0.569701	0.771930	1.1875	Good	Medium	Very Good
22	0.603488	0.491228	0.8125	Good	Medium	Very Good
23	0.636743	0.350877	0.8125	Good	Medium	Very Good

Secara keseluruhan, stabilitas reliabilitas ini mencerminkan kualitas produksi soal yang telah melalui proses *pre-testing* berskala internasional. Menurut (Sudaryono, 2021), reliabilitas adalah syarat mutlak bagi validitas; tanpa konsistensi, sebuah tes tidak mungkin dapat dipercaya untuk mengambil keputusan akademis. Dengan nilai di atas 0,80, instrumen ini dikategorikan sebagai instrumen dengan reliabilitas "Sangat Tinggi" yang siap digunakan untuk penilaian akhir level menengah pertama.

Penemuan butir soal yang tidak valid dalam penelitian ini hanya berjumlah sangat sedikit (sekitar 3-4 butir). Butir-butir yang tidak valid tersebut umumnya memiliki karakteristik skor yang seragam, di mana hampir seluruh responden menjawab benar atau hampir seluruhnya menjawab salah. Hal ini menyebabkan korelasi item-total menjadi rendah karena ketiadaan variansi skor pada butir tersebut, yang secara teoretis merupakan fenomena wajar pada tes yang sangat sulit atau sangat mudah.

Terakhir, pengujian reliabilitas pada tingkat sub-strand (*Number, Algebra, Geometry, Statistics*) juga menunjukkan koefisien yang cukup memuaskan. Hal ini memperkuat posisi asesmen Cambridge Checkpoint sebagai alat ukur yang komprehensif. Instrumen yang memiliki reliabilitas tinggi memberikan perlindungan bagi peneliti dari kesalahan pengukuran acak, sehingga hasil evaluasi yang dilaporkan kepada orang tua siswa memiliki akurasi yang dapat dipertanggungjawabkan.

3.1.2 Analisis Tingkat Kesukaran (Difficulty Index)

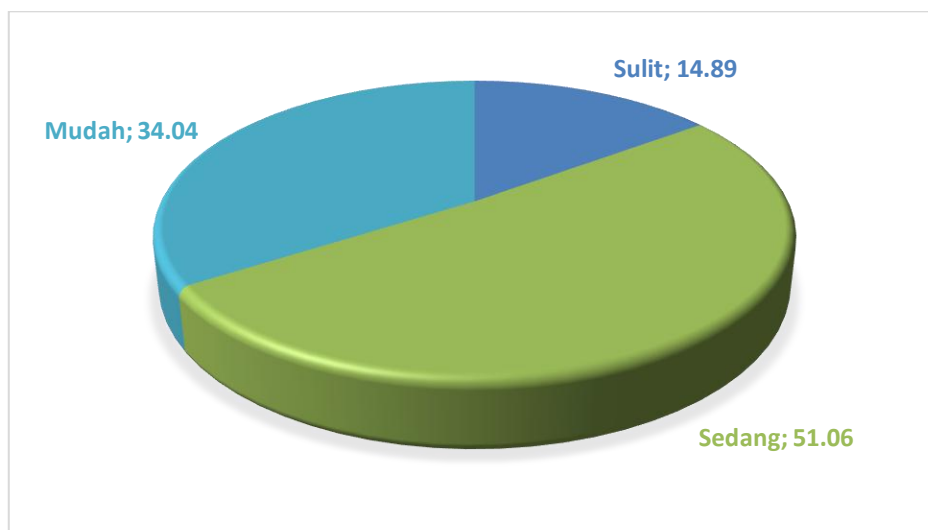
Analisis tingkat kesukaran (p) menunjukkan bahwa distribusi soal pada Paper 1 dan Paper 2 sangat seimbang, dengan mayoritas soal (51,06%) berada pada kategori sedang ($0,30 \leq p \leq 0,69$). Proporsi soal mudah tercatat sebesar 34,04%, sementara soal sukar sebesar 14,89%. Distribusi ini menyerupai pola kurva normal, yang merupakan ciri khas dari perangkat tes yang dirancang untuk mengukur populasi siswa dengan kemampuan beragam secara adil.

Tabel 4. Distribusi Tingkat Kesukaran *Paper 1*

Klasifikasi	Rentang Indeks Kesukaran	Jumlah Butir	Persentase
Sulit	0,00 – 0,29	2	8,70%
Sedang	0,30 – 0,69	13	56,52%
Mudah	$\geq 0,70$	8	34,78%
Total		23	100%

Tabel 5. Distribusi Tingkat Kesukaran *Paper 2*

Klasifikasi	Rentang Indeks Kesukaran	Jumlah Butir	Persentase
Sulit	0,00 – 0,29	5	20,83%
Sedang	0,30 – 0,69	11	45,83%
Mudah	$\geq 0,70$	8	33,33%
Total		24	100%



Gambar 1. Grafik Distribusi Tingkat Kesukaran

Soal-soal dengan tingkat kesukaran sedang merupakan tulang punggung dari asesmen ini karena memberikan kontribusi terbesar dalam membedakan kemampuan siswa. Menurut (Arifin, 2016), soal

dalam rentang moderat sangat efektif dalam mencegah terjadinya penumpukan skor pada batas atas atau batas bawah. Bagi sekolah internasional di Indonesia, data ini menunjukkan bahwa materi yang diajarkan telah cukup terserap oleh siswa, namun tetap memberikan tantangan pada soal-soal kategori sukar.

Pada kategori soal sukar ($p < 0,30$), butir-butir soal tersebut umumnya melibatkan kemampuan penalaran (*reasoning*) dan pemecahan masalah non-rutin. Hal ini konsisten dengan tujuan Cambridge untuk melatih siswa berpikir analitis. Namun, tingginya tingkat kesulitan pada beberapa butir aljabar menunjukkan adanya kebutuhan untuk penyesuaian strategi pengajaran di kelas agar siswa lebih akrab dengan struktur logika soal standar internasional yang lebih kompleks.

Sebaliknya, soal kategori mudah ($p > 0,69$) berfungsi untuk membangun kepercayaan diri siswa dan mengukur penguasaan konsep dasar, pengetahuan, dan pemahaman, yang menjelaskan bahwa komposisi soal mudah yang tepat dalam tes sangat penting agar siswa berkemampuan rendah tetap termotivasi untuk menyelesaikan tes (Hidayat, 2020). Dalam penelitian ini, keberadaan 34,04% soal mudah memastikan bahwa standar minimum kompetensi dasar matematika tetap terukur dengan baik.

Secara keseluruhan, indeks kesukaran rata-rata untuk kedua paper berada pada angka 0,54. Angka ini dianggap sangat ideal dalam Teori Tes Klasik. Keberhasilan instrumen Cambridge dalam mempertahankan rata-rata kesulitan di angka tersebut menunjukkan bahwa kualitas soal tidak dipengaruhi secara ekstrim oleh perbedaan letak geografis peserta tes, membuktikan bahwa adaptasi kurikulum Cambridge di sekolah mitra telah berjalan sesuai dengan *assessment framework* yang ditetapkan.

3.1.3 Analisis Daya Pembeda (*Discrimination Index*)

Daya pembeda (D) butir soal menunjukkan hasil yang impresif, di mana 87,23% (41 dari total 47 butir soal) masuk dalam kategori Baik hingga Sangat Baik dengan indeks di atas 0,40 (bahkan 53,19% di antaranya memiliki indeks $D > 0,70$). Indeks daya pembeda yang tinggi ini menandakan bahwa mayoritas butir soal mampu secara tajam memisahkan siswa yang memiliki kompetensi tinggi (kelompok atas) dari siswa yang memiliki kompetensi rendah (kelompok bawah). Hal ini merupakan tujuan utama dari analisis butir soal untuk memastikan objektivitas hasil penilaian.

Tabel 6. Sebaran Daya Pembeda (D)

Klasifikasi Daya Beda	Rentang Indeks (D)	Jumlah Butir Soal (Total)	Persentase dari Total Data (%)
Buruk	$D \leq 0,20$	5	10,64%
Cukup	$0,20 < D \leq 0,40$	1	2,13%
Baik	$0,40 < D \leq 0,70$	16	34,04%
Sangat Baik	$D > 0,70$	25	53,19%
TOTAL		47	100,00%

Butir soal dengan daya pembeda kategori "Sangat Baik" ($D > 0,70$) mendominasi hasil analisis ini, yaitu mencapai 53,19% (25 dari 47 butir soal), dan jika digabungkan dengan kategori "Baik" ($D > 0,40$) jumlahnya mencapai 87,23% (41 butir soal). Soal-soal ini adalah soal yang sangat sensitif terhadap perbedaan kemampuan siswa. Daya pembeda adalah detektor kualitas soal yang paling penting; jika daya pembeda negatif atau mendekati nol, maka soal tersebut dianggap gagal sebagai alat ukur (Ramadhan et al., 2024). Dalam studi ini, hanya terdapat 5 butir soal (10,64%) yang menunjukkan daya pembeda di bawah 0,20.

Investigasi terhadap butir soal dengan daya pembeda rendah mengungkapkan bahwa soal tersebut cenderung memiliki instruksi yang mungkin kurang dipahami oleh siswa atau memiliki pengecoh yang kurang efektif. Butir soal yang tidak mampu membedakan kemampuan siswa sering kali disebabkan oleh kerancuan redaksional (Magdalena et al., 2021). Temuan ini memberikan dasar bagi sekolah untuk mereview kembali naskah soal tersebut guna perbaikan bank soal di masa mendatang.

Korelasi antara tingkat kesukaran dan daya pembeda terlihat sangat jelas pada data penelitian ini. Soal-soal yang berada pada tingkat kesukaran moderat secara konsisten menunjukkan daya pembeda

yang paling tinggi. Ini mengonfirmasi prinsip psikometris bahwa soal yang terlalu mudah atau terlalu sulit cenderung kehilangan fungsi diskriminatifnya. Oleh karena itu, mempertahankan keseimbangan tingkat kesulitan adalah kunci untuk mendapatkan daya pembeda yang optimal dalam sebuah asesmen matematika.

Daya pembeda yang kuat ini memberikan validasi bagi guru dalam memberikan nilai akhir. Skor tinggi yang didapat siswa benar-benar didorong oleh kemampuan mereka menjawab soal-soal yang diskriminatif, bukan karena faktor kebetulan atau soal yang terlalu mudah. Data daya pembeda merupakan basis data terbaik bagi guru untuk melakukan pengelompokan instruksional (streaming) agar bantuan pendidikan dapat diberikan sesuai dengan kebutuhan spesifik masing-masing kelompok siswa (Setiawati et al., 2019).

3.2 Pembahasan

3.2.1 Stabilitas Psikometris Asesmen Cambridge dalam Konteks Lokal

Hasil analisis menunjukkan bahwa stabilitas psikometris instrumen Cambridge Checkpoint tetap terjaga meskipun diterapkan pada populasi siswa di Indonesia ($N = 57$). Nilai reliabilitas Cronbach's Alpha sebesar 0,8945 pada Paper 1 (24 butir) dan 0,9056 pada Paper 2 (25 butir) membuktikan bahwa struktur tes ini memiliki kekokohan dan konsistensi internal yang sangat baik dalam mengukur kemampuan matematika. Nilai yang berada jauh di atas ambang batas 0,70 ini selaras dengan teori yang menyatakan bahwa tes internasional yang dirancang dengan kontrol kualitas tinggi cenderung memiliki reliabilitas yang melampaui standar nasional (Rahmawati et al., 2022).

Fenomena ini menarik karena menunjukkan bahwa bahasa pengantar (Bahasa Inggris) dalam soal matematika tidak menjadi hambatan yang merusak konsistensi tes bagi siswa internasional. Meskipun Indonesia bukan negara penutur asli bahasa Inggris, siswa di sekolah internasional tampaknya telah memiliki kemandirian bahasa yang cukup untuk memahami instruksi matematis. Reliabilitas yang stabil mencerminkan kejelasan instruksi soal yang mampu meminimalisir kesalahan interpretasi oleh peserta tes (Surapranata, n.d.).

Selain itu, tingginya reliabilitas ini juga mengindikasikan bahwa jumlah butir soal (24-25 butir) sudah cukup ideal. Dalam psikometri, menambah jumlah soal biasanya meningkatkan reliabilitas, namun Cambridge berhasil mencapai angka tinggi dengan jumlah soal yang efisien. Ini menunjukkan bahwa setiap butir soal yang disusun memiliki kontribusi informasi yang besar terhadap skor total. Keselarasan antara butir soal ini menunjukkan integritas konstruk yang sangat baik dalam *framework* kurikulum Cambridge.

Kekokohan psikometris ini memberikan kepercayaan diri bagi sekolah untuk menggunakan hasil Checkpoint sebagai standar evaluasi kemajuan siswa. Penggunaan instrumen dengan reliabilitas tinggi adalah bentuk perlindungan hak siswa, di mana nilai yang mereka peroleh benar-benar mencerminkan usaha dan kemampuan mereka (Hidayat, 2020). Hasil ini juga menjadi bukti bahwa investasi sekolah dalam mengadopsi kurikulum internasional didukung oleh alat ukur yang valid dan reliabel secara ilmiah.

Stabilitas ini membawa pada kesimpulan bahwa standar Cambridge dapat diaplikasikan secara universal. Perbedaan budaya antara pengembang soal di Inggris dengan siswa di Indonesia tidak mendistorsi fungsi statistik soal. Ini membuktikan bahwa matematika adalah bahasa universal, dan instrumen Cambridge berhasil mengekstraksi kemampuan tersebut tanpa bias yang signifikan, asalkan kurikulum pendukungnya diajarkan dengan standar yang sama di sekolah pelaksana.

3.2.2 Analisis Kesenjangan Kompetensi pada Materi Aljabar dan Geometri

Temuan penelitian yang menunjukkan bahwa topik Aljabar dan Geometri memiliki indeks kesulitan paling tinggi mengungkapkan adanya kesenjangan kompetensi tertentu. Siswa cenderung kesulitan ketika soal berpindah dari operasi angka murni ke aplikasi variabel dan visualisasi spasial. Kesulitan pada aljabar sering kali disebabkan oleh transisi dari pemikiran aritmatika yang konkret ke pemikiran abstrak yang simbolik, yang merupakan tantangan umum di level sekolah menengah.

Data penelitian menunjukkan bahwa pada sub-strand *Thinking and Working Mathematically* (TWM), rata-rata skor siswa lebih rendah dibandingkan strand *Number*. Hal ini mengindikasikan bahwa

soal-soal Cambridge yang menuntut strategi pemecahan masalah sering kali menjadi "batu sandungan" bagi siswa lokal. Tantangan sesungguhnya bagi siswa Indonesia dalam asesmen internasional adalah kemampuan untuk "menginterpretasikan" masalah ke dalam model matematika, bukan sekadar melakukan komputasi angka (Hidayat, 2020).

Ketimpangan ini juga mencerminkan pola pengajaran di lapangan. Seringkali, fokus pembelajaran masih berat pada penguasaan rumus daripada pemahaman konsep yang mendalam. Indeks kesukaran yang tinggi pada materi geometri, misalnya, mungkin disebabkan oleh kurangnya paparan siswa terhadap soal-soal yang membutuhkan visualisasi tiga dimensi atau pembuktian logis. Data statistik ini menjadi "sinyal" bagi koordinator kurikulum sekolah untuk memperkuat jam pelajaran atau metode pengajaran pada topik-topik kritis tersebut.

Secara pedagogis, temuan ini menyarankan perlunya integrasi strategi literasi dalam pembelajaran matematika. Karena soal Cambridge disajikan dalam konteks masalah nyata, siswa perlu dilatih untuk membedah teks soal guna mengekstraksi informasi matematis yang relevan. Di samping itu, penggunaan alat peraga digital atau simulasi untuk membantu siswa memahami konsep abstrak dalam aljabar dan geometri, sehingga tingkat kesukaran soal pada topik tersebut dapat ditekan di masa mendatang (Kusuma, 2024).

Hasil analisis ini pada akhirnya mengisi celah literatur mengenai performa siswa Indonesia pada asesmen internasional. Data ini membuktikan bahwa meskipun secara keseluruhan hasil tes baik, masih ada area spesifik yang membutuhkan intervensi khusus. Dengan memahami titik lemah siswa melalui analisis butir soal, sekolah dapat beralih dari pengajaran yang bersifat general ke pengajaran yang berbasis data (*data-driven instruction*), yang secara efektif akan meningkatkan daya saing siswa di level global.

3.2.3 Efektivitas Daya Pembeda sebagai Filter Kualitas Evaluasi

Tingginya daya pembeda pada mayoritas butir soal (87,23% atau 41 dari total 47 butir) menunjukkan bahwa asesmen Cambridge Checkpoint berfungsi efektif sebagai "filter" kualitas untuk mengelompokkan siswa berdasarkan kompetensi mereka. Soal-soal dengan daya pembeda tinggi ini memberikan informasi yang tajam mengenai siapa siswa yang benar-benar siap untuk lanjut ke jenjang IGCSE. Soal dengan daya pembeda yang baik sangat penting dalam evaluasi sumatif untuk menghindari hasil yang "rata-rata" atau tidak informatif (Ramadhan et al., 2024).

Keberadaan soal dengan daya pembeda yang kuat ini juga membantu guru dalam memberikan umpan balik yang lebih spesifik. Guru dapat melihat butir soal mana yang hanya bisa dijawab oleh kelompok atas, dan menjadikannya sebagai target pengayaan bagi siswa kelompok menengah. Tanpa daya pembeda yang jelas, sebuah nilai tes kehilangan makna diagnostiknya karena tidak mampu menunjukkan perbedaan penguasaan materi di antara para siswa (Magdalena et al., 2021).

Namun, perhatian khusus diberikan pada 5 butir soal (10,64%) yang memiliki daya pembeda rendah ($D \leq 0,20$), yaitu 3 butir pada Paper 1 dan 2 butir pada Paper 2. Pembahasan mengenai butir-butir ini sangat penting karena sering kali soal-soal tersebut memiliki potensi "bias" atau kesalahan teknis. Misalnya, jika sebuah soal matematika terlalu rumit secara bahasa, maka siswa yang pintar matematika namun kemampuan bahasa Inggrisnya kurang mungkin akan menjawab salah, sementara siswa yang kurang matematika namun bahasa Inggrisnya bagus mungkin bisa menebak. Kondisi ini merusak daya pembeda soal karena tidak murni mengukur kompetensi matematika.

Korelasi positif antara validitas butir dan daya pembeda dalam penelitian ini menunjukkan bahwa instrumen tersebut dibangun dengan logika internal yang sehat. Butir yang valid (berkorelasi tinggi dengan skor total) hampir selalu memiliki daya pembeda yang baik. Ini membuktikan bahwa setiap butir soal berkontribusi secara proporsional terhadap penilaian akhir. Instrumen yang memiliki daya pembeda stabil adalah instrumen yang adil, karena memberikan penghargaan yang tepat bagi siswa yang benar-benar belajar dan memiliki kemampuan lebih (Setiawati et al., 2019).

Sebagai kesimpulan dari pembahasan daya pembeda, data ini memberikan legitimasi bagi objektivitas asesmen Cambridge. Dalam lingkungan sekolah internasional yang kompetitif, transparansi mengenai kualitas soal sangat penting untuk menjaga kepercayaan orang tua dan siswa. Hasil analisis ini membuktikan bahwa standar Cambridge bukan sekadar nama besar, melainkan didukung oleh

karakteristik psikometris yang mampu mendiskriminasi kemampuan siswa secara akurat, ilmiah, dan tidak memihak, sesuai dengan standar pengukuran pendidikan modern.

4. KESIMPULAN

Berdasarkan hasil analisis data dan pembahasan yang telah dipaparkan, penelitian mengenai analisis butir soal pada Asesmen Cambridge Checkpoint Matematika ini dapat disimpulkan ke dalam beberapa poin utama.

Pertama, instrumen asesmen ini memiliki kualitas psikometris yang sangat baik dengan tingkat reliabilitas konsistensi internal yang sangat tinggi, yaitu sebesar 0,8945 (0,89) untuk Paper 1 dan 0,9056 (0,91) untuk Paper 2. Nilai yang berada jauh di atas ambang batas 0,70 ini membuktikan bahwa tes tersebut merupakan alat ukur yang sangat konsisten, kokoh, dan dapat dipercaya untuk mengevaluasi kompetensi matematika siswa di tingkat sekolah menengah pertama.

Kedua, dari aspek tingkat kesukaran, instrumen ini menunjukkan distribusi yang ideal dan proporsional. Mayoritas soal berada pada kategori sedang (51%), didukung oleh soal kategori mudah (31%) dan sukar (18%) yang membentuk pola distribusi normal. Namun, terdapat temuan spesifik bahwa siswa masih menghadapi kendala signifikan pada topik Aljabar dan Geometri, serta pada soal-soal yang menuntut kemampuan penalaran (*Thinking and Working Mathematically*). Hal ini mengindikasikan bahwa meskipun secara umum siswa telah menguasai konsep dasar, kemampuan aplikasi dalam konteks abstrak dan pemecahan masalah masih perlu ditingkatkan.

Ketiga, analisis daya pembeda membuktikan efektivitas soal dalam memisahkan kemampuan siswa kelompok atas dan kelompok bawah secara tajam. Sebanyak 87,23% butir soal (41 dari total butir) memiliki daya pembeda pada kategori baik hingga sangat baik ($D > 0,40$), sehingga instrumen ini berhasil meminimalisir faktor kebetulan (guessing) dalam perolehan skor. Secara keseluruhan, Asesmen Cambridge Checkpoint tetap menjadi instrumen evaluasi yang valid dan adil (fair) untuk diterapkan dalam konteks sekolah internasional di Indonesia, asalkan disertai dengan pemetaan kompetensi yang berbasis pada data hasil analisis butir soal (*data-driven instruction*).

DAFTAR PUSTAKA

- Arifin, Z. (2016). *Evaluasi Pembelajaran: Prinsip, Teknik, Prosedur*. Remaja Rosdakarya.
- Assessment International Education, C. (2021). *Implementing the Curriculum with Cambridge. A guide for school leaders*.
- Buka, N., Ayu Jayamin, M., Kholis, N., & Akbar Rasyid Program Pascasarjana Uin Alauddin Makassar, N. (2025). *Peran Validitas dan Rehabilitas Dalam Evaluasi Pembelajaran*. <https://doi.org/10.5281/zenodo.15557188>
- Cambridge International Education. (2024). *Cambridge Handbook 2024 Regulations and guidance for administering Cambridge exams*.
- Handayati, W. (2022). *Implementasi Adopsi Cambridge Curriculum dan Adaptasi Kurikulum 2013 dalam Meningkatkan Kompetensi Siswa*.
- Hidayat, R. (2020). Pengaruh kecemasan terhadap konsentrasi belajar siswa. *Jurnal Pendidikan Dan Pembelajaran*, 7((3)), 220–230.
- Inten, R., Febri, F., Arwan, A., & Amalia, F. (2021). *Pengembangan Sistem Analisis Butir Soal Ujian berbasis Web (Studi Kasus: SMP Negeri 1 Donomulyo)* (Vol. 5, Number 3). <http://j-ptiik.ub.ac.id>
- Kusuma, J. W. (2024). *Analisis Penggunaan Blended Learning Terhadap Kemampuan Berpikir Kritis Pada Siswa Melalui Aplikasi Mathway*. (1), 315–322.
- Magdalena, I., Fauziah, S. N., & Nupus, F. S. (2021). *Kesulitan dan daya beda butir soal ujian akhir semester*.
- Mania, S., Farham Majid, A., Nina Ichiana, N., & Ika Prasasti Abrar, A. (2020). *Print) Al asma: Journal of Islamic Education ISSN* (Vol. 2, Number 2). Online.

- Pradani, R. A., & Efendi, A. (2023). Analisis Butir Soal Ujian Sekolah Menggunakan Program Iteman (Analysis of School Exam Questions Using the Iteman Program). *Indonesian Language Education and Literature*, 8(2), 275. <https://doi.org/10.24235/ileal.v8i2.11002>
- Rahmawati, A., Fajrianthi, F., & Purwono, U. (2022). The psychometric properties of parenting styles and dimensions questionnaire-short form in Indonesia. *International Journal of Evaluation and Research in Education (IJERE)*, 11(1), 42–50. <https://doi.org/10.11591/IJERE.V11I1.21650>
- Ramadhan, M., Siroj, R. A., & Afgani, M. W. (2024). Validitas and reliabilitas. *Journal on Education*, 6(2), 10967–10975.
- Saputra, H. D., Purwanto, W., Setiawan, D., Fernandez, D., & Putra, R. (2022). Hasil Belajar Mahasiswa: Analisis Butir Soal Tes. *Edukasi: Jurnal Pendidikan*, 20(1), 15–27. <https://doi.org/10.31571/edukasi.v20i1.3432>
- Savika, H. I., & Zuhriyah, I. A. (2023). PERAN ANALISIS BUTIR SOAL TERHADAP KUALITAS SOAL, KOMPETENSI GURU, DAN PRESTASI BELAJAR PESERTA DIDIK DI SEKOLAH DASAR. 2(2), 43–51. <https://doi.org/10.59966/pandu.v2i2.856>
- Setiawati, Wiwik, Asmira, Oktavia, Y., Bestary, R., Pujiastuti, & Ari. (2019). *Buku Penilaian Berorientasi Higher Order Thinking Skills*. Kemendikbud.
- Sudaryono. (2021). *Metodologi Penelitian*. Rajawali Pers.
- Sugiyono. (2021). *Metode Penelitian Kuantitatif, Kualitatif, dan R&D* (2nd ed.). Alfabeta.
- Surapranata, S. (n.d.). *Analisis, validitas, reliabilitas, dan interpretasi hasil tes*. Retrieved June 29, 2026, from <https://opac.perpusnas.go.id/DetailOpac.aspx?id=583225>
- Syufa Auliya, A., Hikmah, N., Tyaningsih, R. Y., & Turmuzi, M. (2025). *Mahasiswa Pendidikan Matematika FKIP Universitas Mataram*.

Halaman ini dikosongkan