Insyirah: Jurnal Ilmu Bahasa Arab dan Studi Islam
Vol. 8, No. 1, June 2025, p. 44-64
DOI: 10.26555/insyirah.v8i1.13163

Classpoint AI: Kualitas Empiris Pembuat Soal Otomatis dalam Assessmen Bahasa Arab

Classpoint AI: Empirical Quality of Automatic Item Test Generator in Arabic Foreign Language Assessment

[1]Risma Siti Istikomah*, [2]Mohamad Zaka Al Farisi, [3]Hikmah Maulani

[1]rismasiti@upi.edu, [2]zaka@upi.edu, [3]hikmahmaulani@upi.edu
[1,2,3]Universitas Pendidikan Indonesia, Indonesia

ARTICLE INFO

Article history
Received: 2 April 2025
Revised: 31 May 2025
Accepted: 4 June 2025

Keywords
Quality questions item,
Artificial Intelligence,
Classpoint,
Reading Comprehension,
Arabic Language Assessment.

*Corresponding Author

ABSTRACT

Artificial Intelligence (AI) plays a significant role in education in the modern era. A platform called Classpoint AI (CAI) has been widely utilized across various educational levels and disciplines worldwide. Numerous studies have reported positive responses toward using CAI as an interactive learning medium and an automatic question generator based on Bloom's Taxonomy. This study aims to analyze the quality of test items generated by CAI quantitatively. The accuracy of a test item must be empirically validated to ensure high-quality learning. A non-probability sampling technique, specifically total sampling, was employed to select 39 female students in Phase F as research participants. Data were collected through a test consisting of 14 multiple-choice questions created by CAI to assess reading comprehension proficiency in Arabic. An Ex-post facto method was used in this study, as the researcher did not modify the pre-programmed CAI system. Three software tools (SPSS, Anbuso, and Anates) were used to analyze the item quality, and teacher validation was also included. The findings revealed that most of the test items generated by CAI were invalid and exhibited low reliability. The questions were predominantly classified as easy. CAI-generated items demonstrated poor discrimination power but contained effective distractors. For foreign language learning, CAI-generated questions need to reconsider the complexity of the vocabulary used. Considering the widespread use of this platform among educators on an international scale, this study contributes a global perspective by highlighting the need for both developer evaluation and educator discernment in utilizing CAI.

1. Pendahuluan

Dewasa ini, pemanfaatan teknologi di dunia pendidikan masif digunakan. Sejak 2018, kemunculan Artificial Intelligence (AI) telah banyak diperbincangkan (Serdianus & Saputra, 2023). Beragam AI membantu guru dalam pembelajaran di kelas, salah satunya yaitu Classpoint AI (CAI). Platform CAI merupakan kecerdasan buatan berbasis aplikasi yang terintegrasi dengan Microsoft PowerPoint, dikembangkan sejak 2015 oleh Inknoe di Singapura. Pada awal 2023, versi kedua dirilis dengan fitur-fitur yang lebih canggih (Chau & Pham, 2023). CAI merupakan media pembelajaran interaktif dan dilengkapi dengan fitur pembuat soal otomatis berbasis gamifikasi. CAI masif dimanfaatkan guru di berbagai negara, terutama dalam pembelajaran bahasa asing. Di Saudi Arabia, penggunaan CAI memicu peningkatan prestasi akademik para pemelajar bahasa Inggris (Akram & Abdelrady, 2023). CAI mendorong minat belajar mahasiswa EFL di universitas independen Vietnam (Chau & Pham, 2023). Siswa EFL di Filipina mengalami peningkatan keterlibatan dan jiwa kompetitif setelah menggunakan CAI (Lugatoc, 2022). Di Indonesia, penerapan CAI terbukti meningkatkan empat keterampilan berbahasa Inggris terutama bagi siswa dengan tingkat kemahiran rendah (Mazlan dkk., 2023; Oktadela dkk., 2024). Kemudian, mahasiswa *non-native Arabic speakers* di Pusat Bahasa Universitas Yarmouk menunjukkan peningkatan signifikan di seluruh bidang pemahaman bacaan: pemahaman literal, inferensial, dan kritis-evaluatif dalam bahasa Arab (Hawamdeh dkk., 2025).

Sejauh ini, studi mengenai CAI dapat dipetakan menjadi tiga. Pertama, pemanfaatan CAI di berbagai negara yang meningkatkan minat dan prestasi pemelajar bahasa asing. Kedua, penggunaan CAI di kelas memberikan dampak positif bagi guru dan siswa. CAI membantu proses pembelajaran serta meningkatkan kemampuan pedagogik guru (Hadi dkk., 2024). Bahkan, guru berteknologi rendah sekalipun dapat dengan mudah mengakses CAI (Mazlan dkk., 2023). Media pembelajaran berbasis CAI terbukti dapat mengintegrasikan dengan sistem pendidikan yang ada (Nugraheni, 2024). CAI meningkatkan keterlibatan langsung antara siswa dan guru, sehingga mendorong pencapaian tujuan pendidikan (Abumosa, 2024). CAI meningkatkan kemampuan berpikir kreatif siswa (Muhiddin dkk., 2023). CAI memengaruhi antusiasme, semangat belajar, dan keaktifan siswa (Istiqomah, 2024; Muliani dkk., 2024). CAI meningkatkan rasa ingin tahu, kerjasama, berdiskusi, berpendapat, responsif, dan bertanggungjawab terhadap tugasnya (Fitriana, 2023). Penerapan CAI menunjukkan respon positif dalam aspek kegunaan, kognitif, dan psikomotorik siswa (Setiyanto, 2023; Wao dkk., 2022). CAI terbukti tidak bias gender dalam penerapannya (Aregbesola, 2025; Ramadhani & Unsiah, 2024). Ketiga, CAI sebagai media evaluasi

yang teruji dan layak digunakan menurut validasi ahli evaluasi (Sari & Pratiwi, 2023). Platform CAI memenuhi aspek kelayakan dan kepraktisan yang meliputi isi slide, variasi soal, dan media (Azmi dkk., 2024).

Berdasarkan studi terdahulu, belum ditemukan riset terhadap analisis kualitas instrumen evaluasi pada CAI sebagai pembuat soal otomatis, berupa uji validitas butir soal, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas distraktor. Maka, penelitian ini bertujuan untuk menganalisis kualitas butir soal yang dihasilkan oleh CAI secara kuantitatif. Dalam menganalisis mutu soal tes, suatu instrumen evaluasi dikatakan bermutu jika memenuhi syarat-syarat berikut: validitas, reliabilitas, objektivitas, praktisitas dan ekonomis (Arikunto, 2021). Selain itu, suatu tes berkualitas baik harus memiliki tingkat kesukaran soal, daya pembeda, dan analisis pengecoh soal yang baik (Amelia, 2016). Riset ini menutupi kekurangan studi terdahulu, karena mengkaji secara empris soal-soal buatan CAI berbasis Taksonomi Bloom dengan pendekatan statistika.

Riset sebelumnya terhadap kualitas butir soal yang dihasilkan oleh AI pembuat soal otomatis telah dilakukan pada platform OpExams (Zuhri dkk., 2024a) dan juga platform lainnya bernama QuestionWell AI (Zuhri dkk., 2024b). Mengingat masifnya pemanfaatan CAI, analisis kualitas butir soal buatan platform ini penting untuk dilakukan sebelum soal diujikan kepada siswa. Data hasil belajar yang akurat hanya akan diperoleh melalui butir soal tes yang bermutu sebagai penunjang kualitas pembelajaran itu sendiri (Susetyo, 2015). Meski di zaman modern ini, bermunculan AI yang dapat meringankan beban kerja manusia, penggunaannya tidak boleh dibiarkan begitu saja tanpa melalui pengujian yang memadai. Oleh karena itu, perlu dilakukan uji kelayakan baik oleh para ahli maupun melalui penelitian empiris berbasis pendekatan kuantitatif. Dengan penggunaan CAI di berbagai negara, riset ini diharapkan dapat memberi perspektif baru atas kontribusi teknologi terhadap dunia pendidikan secara global.

2. Metode

Dalam riset ini, peneliti menggunakan metode ex post facto. Metode ini meneliti hubungan antara sebab-akibat tanpa adanya intervensi atau manipulasi dari peneliti. Penelitian semacam ini dapat diterapkan ketika meneliti suatu program yang telah terjadi. Selain itu, penelitian dengan metode ex post facto sering dijumpai dalam riset bidang pendidikan dan berperan dalam pengambilan keputusan besar di ranah pendidikan (Sappaile, 2010).

Penelitian dilakukan di SMA Al-Ittihad Cianjur, Jawa Barat. Partisipan dalam penelitian ini adalah kelas yang mempelajari bahasa Arab yaitu para siswi fase F kelas XII IPS SMA Al-Ittihad Cianjur sebagai populasi. Dalam penelitian ini populasi tergolong relatif kecil, maka peneliti menggunakan teknik non-probability sampling berupa sampling jenuh. Menurut Sugiyono,

sampling jenuh merupakan keadaan di mana seluruh populasi berperan menjadi sampel penelitian. Hal ini dapat terjadi ketika jumlah populasi kurang dari 100 partisipan (Arikunto, 2013). Maka dari itu, dalam penelitian ini populasi yang berjumlah 39 orang, seluruhnya akan berperan menjadi sampel penelitian. Teknik ini dapat membantu peneliti melakukan generalisasi dengan menimalisir tingkat kesalahan yang ada (Kurniawan, 2015).

Teknik pengumpulan data dalam riset ini berupa tes objektif untuk menguji keterampilan membaca pemahaman bahasa Arab siswa. Tes membaca pemahaman merupakan teknik pengukuran untuk mengumpulkan informasi terhadap kemampuan seseorang dalam memahami suatu wacana (Syihabuddin, 2018). Maka dari itu, instrumen tes bersumber dari wacana. Wacana yang digunakan untuk membuat instrumen berasal dari buku Bahasa Arab SMA Kelas XII Kurikulum 2013 terbitan Gramasurya yang dijadikan sebagai buku sumber pembelajaran di kelas.

Instrumen tes yang digunakan adalah 14 butir soal berbahasa Arab berbentuk pilihan ganda yang dibuat secara otomatis oleh platform CAI. Adapun prosedur pembuatan soal otomatis menggunakan CAI dalam penelitian ini yaitu, 1) Peneliti mengunduh dan menginstal aplikasi CAI di website Classpoint, 2) *Sign in* dan pastikan CAI telah terintegrasi dengan Microsoft Powerpoint, 3) Peneliti memasukkan dua wacana yang berasal dari buku sumber, 4) CAI akan mengubah materi yang telah diinput ke dalam *slide* menjadi butir soal dengan menekan fitur *slideshow* lalu pada bagian *toolbar* gunakan fitur *Quiz AI*, 5) CAI memroses pembuatan soal berdasarkan bentuk soal dan tingkat Taksonomi Bloom yang dapat dipilih pada fitur *Options*, 6) Setelah itu, tekan tombol *Generate Question* untuk membuat butir soal yang diinginkan.

Instrumen penelitian yang digunakan peneliti untuk memeroleh data yaitu, 1) Classpoint AI (CAI), sebagai platform pembuat soal otomatis, 2) Lembar jawaban dan hasil penskoran tes, 3) Microsoft Excel untuk menyusun, mengorganisir, dan merekap hasil tes serta penghitung data analisis validitas dan reliabilitas 4) Software untuk membantu menganalisis kualitas butir soal buatan CAI, yaitu SPSS, Anbuso versi 8, Anates versi 4.

Dalam membuat instrumen evaluasi yang baik butir soal harus memiliki kesesuaian dengan indikator yang telah ditetapkan (Hazraini, 2017). Maka dari itu, sebelum paket soal diujikan kepada partisipan penelitian, soal-soal yang dihasilkan oleh CAI akan ditinjau terlebih dahulu kesesuaiannya dengan indikator yang ada. Sehingga soal yang akan diujikan selaras dengan tujuan pembelajaran.

Berdasarkan teori Anderson & Krathwohl klasifikasi capaian kognitif menurut dimensi Taksonomi Bloom Revisi terdiri dari LOTS yang memuat kemampuan mengingat (C1) dan

memahami (C2), MOTS terdiri dari kemampuan dan menerapkan (C3) dan menganalisis (C4), serta HOTS mencakup kemampuan menilai (C5) dan mencipta (C6) (Helmawati, 2019). Kemudian selaras dengan ini, Maulana dalam penelitiannya mengemukakan bahwa capaian pembelajaran bahasa Arab fase F dalam keterampilan membaca berada pada level MOTS yaitu minimal pada tingkatan memahami untuk aspek membaca-memirsa (Maulana, 2022). Dengan indikator sebagai berikut:

1. membaca huruf, kata dan kalimat serta teks bahasa Arab dengan lancar, cermat, dan tepat;
2. menentukan arti kosa kata dalam konteks kalimat tertentu;
3. menemukan fakta tersurat dalam teks;
4. menemukan makna tersirat dalam teks;
5. menemukan ide pokok dalam paragraf;
6. menghubungkan ide-ide yang terdapat dalam bacaan;
7. menyimpulkan ide pokok bacaan; dan
8. menjelaskan budaya dalam teks bacaan (Kemdikbudristek, 2022).

Setelah melalui tahap tes, peneliti akan melakukan penskoran pada lembar jawaban siswa dan merekapnya. Kemudian, analisis secara kuantitatif dilakukan untuk menganalisis data empirik pada penelitian ini. Data empirik diperoleh dari butir soal buatan CAI yang diujikan. Dalam riset ini, peneliti akan menganalisis 1) uji validitas butir soal, 2) uji reliabilitas, 3) tingkat kesukaran soal, 4) daya pembeda, dan 5) efektivitas distraktor.

3. Hasil dan Pembahasan

3.1. Validitas Butir Soal

Telaah validitas merupakan aspek penting yang harus dilakukan agar hasil tes dapat menggambarkan informasi siswa yang ril dan akurat (Laili, 2020). Validitas tes sangat dipengaruhi oleh validitas yang dimiliki oleh masing-masing butir soal yang membangunnya sebagai satu kesatuan yang tak terpisahkan. Dengan demikian, validitas butir soal ialah ketepatan yang dimiliki oleh sebutir soal untuk mengukur sesuatu yang harus diukur lewat soal tes tersebut (Hendriawan & Nurman, 2021). Jika soal dirancang untuk menguji keterampilan membaca pemahaman, maka butir-butir soal yang ditulis adalah untuk menanyakan pemahaman *testee* terhadap suatu wacana (Sumaningsih, 2015).

Setelah dilakukan penskoran pada lembar jawaban siswa, validitas dihitung dengan menginput data dikotomi. Jawaban benar direpresentasikan dengan nilai 1 dan jawaban salah direpresantasikan dengan nilai 0. Kemudian, penghitungan validitas butir soal dilakukan dengan

mengukur hubungan antara skor *testee* yang menjawab benar atau salah dengan skor total untuk setiap butir soalnya. Maka dari itu, teknik analisis uji validitas butir soal yang sebaiknya digunakan adalah teknik korelasi. Adapun dalam menganalisis data dikotomi yang memiliki dua kemungkinan jawaban, seperti benar-salah, ya-tidak, dan yang sejenis dapat menggunakan Korelasi Poin Biserial untuk setiap butir soalnya (Munip, 2017).

Dengan menggunakan rumus korelasi, maka akan diperoleh nilai koefisien korelasi yang memiliki dua kemungkinan. Jika nilainya lebih besar dari r tabel yaitu sebesar 0,316 (dengan signifikansi sebesar 5%) maka butir soal dinyatakan valid, jika sebaliknya maka soal dinyatakan tidak valid (invalid). Dalam menghitung validitas butir soal buatan CAI, peneliti dibantu dengan menggunakan software Anates dan SPSS dalam menganalisisnya.

Tabel 1. Uji Validitas Soal Buatan CAI

Butir Soal	Anates *(Poin Biserial)*	*SPSS (Pearson Product Moment)*	Keterangan
1	0,483	0,483	Valid
2	0,251	0,251	Invalid
3	0,229	0,229	Invalid
4	0,483	0,483	Valid
5	0,348	0,348	Valid
6	NAN	-	Invalid
7	0,241	0,241	Invalid
8	0,174	0,174	Invalid
9	-0,014	-0,014	Invalid
10	0,491	0,491	Valid
11	0,508	0,508	Valid
12	0,521	0,521	Valid
13	0,301	0,301	Invalid
14	0,153	0,153	Invalid

Uji validitas terhadap 14 soal pilihan ganda buatan CAI dianalisis dengan bantuan dua software berbeda. Analisis uji validitas berbasis software Anates menggunakan pendekatan Korelasi Point Biserial, sedangkan SPSS menggunkan pendekartan Korelasi Pearson Product Moment. Kedua hasil uji validitas tersebut menunjukan nilai yang identik terhadap koefisien korelasi yang dihasilkan.

Hasil uji validitas 14 soal pilihan ganda buatan CAI, menunjukan bahwa enam soal berkategori valid (43%). Soal-soal valid mengindikasikan keakuratan untuk mengukur kompetensi siswa terhadap tujuan pembelajaran yang hendak dicapai (Anshari dkk., 2024). Namun, soal dengan kategori valid, hanya mampu menyentuh nilai koefisien tertingginya yaitu sebersar 0,521 yang diinterpretasikan memiliki derajat validitas sedang karena nilai koefisiennya

berkisar antara 0,400 - 0,599 (Sukiman, 2012). Demikian pula, soal-soal valid lainnya yang rata-rata memiliki derajat validitas sedang. Soal-soal valid CAI belum dapat menyentuh derajat validitas berkategori tinggi seperti layaknya soal buatan manusia. Maka dari itu, untuk soal-soal tersebut, guru perlu mempertimbangkan adanya revisi atau pengembangan kembali.

Namun, perlu diakui bahwa delapan soal buatan CAI berkategori tidak valid (57%). Hal ini mengindikasikan soal belum dapat dijadikan tolak ukur keselarasan antara kompetensi siswa dengan tujuan pembelajaran. Perlu diperhatikan kembali mengenai hal-hal yang dapat memengaruhi validitas butir soal itu sendiri, terutama pada konteks pembelajaran bahasa asing. Kosa kata yang terlalu sulit dalam tes dapat menjadi faktor tidak validnya suatu soal (Syafiudin, 2020). Hal ini berpotensi membuat tujuan soal untuk mengukur tingkat pemahaman siswa terhadap suatu wacana menjadi tidak tepat sasaran.

Pada penelitian uji coba pembuat soal otomatis berbasis komputer dalam pembelajaran bahasa asing, siswa cenderung memilih jawaban benar jika sistem menyediakan soal dengan level kosakata setara dengan tingkat kemahiran siswa. Namun, khusus untuk soal yang menguji pemahaman wacana, siswa cenderung menjawab benar jika level kosakata pada soal lebih rendah daripada tingkat kemahiran siswa (Huang dkk., 2018). Hal ini membuktikan bahwa pembuat soal otomatis berbasis teknologi atau mesin masih perlu mempertimbangkan faktor kebahasaan agar siswa memahami konteks soal, terutama untuk pembelajaran bahasa asing.

3.2. Reliabilitas

Reliabilitas adalah metode untuk mempelajari, mengidentifikasi, dan mengestimasi keajegan skor tes. Tujuan uji reliabilitas adalah memastikan alat ukur yang digunakan dapat memberikan hasil konsisten pada pengukuran berulang dan dapat diandalkan (Anshari dkk., 2024). Dalam tes hasil belajar, kemampuan siswa yang sesungguhnya harus dapat diukur secara akurat melalui instrumen tes yang diujikan. Dalam penyusunan tes hasil belajar, sangat penting memerhatikan besaran nilai koefisien reliabilitas dari tes yang akan diujikan. Jika hasil dari skor tes pertama sama dengan hasil skor tes kedua, maka tes memiliki reliabilitas tinggi atau keduanya memiliki nilai korelasi yang tinggi (Setiyawan, 2014).

Untuk menganalisis reliabilitas soal berbasis tes objektif, formula yang dapat digunakan adalah teknik Kuder Richardson (KR). Pada teknik ini, prosedur yang digunakan cukup sederhana dan tidak memakan banyak waktu dari teknik analisis reliabilitas yang lain. Di mana untuk teknik yang lain, perlu adanya dua kali tes dalam jangka waktu tertentu, penambahan bentuk interval tes, proses penskoran tes yang dibagi dua, dan sejenisnya. Dalam prosedurnya, teknik Kuder Richardson hanya memerlukan satu kali tes dalam pengerjaan soal, lalu reliabilitas soal dapat

langsung dihitung menggunakan rumus KR (Widodo dkk., 2023). Maka dari itu, teknik KR menjadi yang paling diminati dan sering digunakan untuk tes objektif karena cenderung lebih sederhana dalam prosedurnya (Djiwandono, 2011).

$$r_{ii} = (\frac{k}{k-1})(\frac{S_t^2 - \Sigma_{piqi}}{S_t^2})$$

$k\ (jumlah\ soal) = 14$

$k - 1 = 13$

$S_t^2 (varians\ total) =$ 2,2656147272

$\Sigma_{piqi} =$ 1,747534517

$r_{ii} =$ 0,2462609937 dibulatkan menjadi 0,25

Setelah dilakukan uji reliabilitas menggunakan rumus KR20, akan dihasilkan nilai akhir yang disebut dengan koefisien korelasi reliabilitas tes (r_{ii}). Lalu, koefisien tersebut dapat diinterpretasikan dengan mengacu pada pembagian derajat reliabilitas menurut Guilford & Fruchter pada Tabel 2 berikut,

Tabel 2. Derajat Reliabilitas Tes

Koefisien Reliabilitas	Derajat Reliabilitas
0,81 – 1,00	Sangat Tinggi
0,61 – 0,80	Tinggi
0,41 – 0,60	Sedang
0,21 – 0,40	Rendah
0,00 – 0,20	Sangat Rendah

Berdasarkan hasil uji reliabilitas dengan rumus KR20, diperoleh koefisien reliabilitas sebesar 0,25 yang diinterpretasikan memiliki derajat reliabilitas rendah. Hal ini mengindikasikan soal tes buatan CAI tidak konsisten. Sama halnya dengan OpExams AI pembuat soal otomatis juga menghasilkan soal yang tidak reliabel (Zuhri dkk., 2024a). Tentu ini juga berkaitan dengan soal-soal buatan CAI yang didominasi oleh soal berkategori tidak valid menyebabkan rendahnya tingkat reliabilitas yang diperoleh soal buatan CAI. Reliabilitas sebuah soal menjadi pendukung terbentuknya validitas butir soal, sehingga sebuah soal yang valid biasanya reliabel (Sanusi & Aziez, 2021).

Selain faktor validitas, hal-hal lain yang dapat memengaruhi reliabilitas tes adalah (1) banyaknya butir soal (2) kelompok yang heterogen (3) objektivitas penskoran (4) metode reliabilitas yang digunakan (5) level kelompok dan tingkat kesukaran soal (6) homogenitas tes (Setiyawan, 2014). Nilai koefisien reliabilitas cenderung meningkat seiring banyaknya butir soal

tes yang diujikan, tetapi bukan berarti soal yang diujikan harus banyak tanpa memerhatikan batas ukuran dan indikator yang relevan dengan variabel (Anshari dkk., 2024). Dalam penelitian ini, jumlah soal yang dianalisis terbatas hanya 14 butir soal yang bersumber dari dua wacana. Hal ini mungkin menjadi salah satu faktor soal buatan CAI memiliki derajat reliabilitas rendah. Namun, jumlah soal yang terbatas ini telah dipilih secara representatif dan sesuai dengan indikator variabel yang diteliti, sehingga tetap memberikan gambaran yang relevan terhadap kemampuan CAI sebagai platform pembuat soal otomatis.

3.3. Tingkat Kesukaran Soal

Tingkat kesukaran butir soal merupakan perbandingan antara jumlah partisipan tes yang menjawab benar dengan jumlah keseluruhan pertisipan tes (Susetyo, 2015). Tingkat kesukaran soal direpresentasikan oleh Indeks Kesukaran Soal (IKS) yang dihitung untuk setiap butir soal dengan kisaran skor 0,00 – 1,00. Semakin kecil IKS yang diperoleh dari setiap butir soal, maka semakin sulit soal tersebut (Fatimah & Alfath, 2019). Berikut ini adalah Tabel 3 terkait IKS yang dapat dijadikan sebagai acuan.

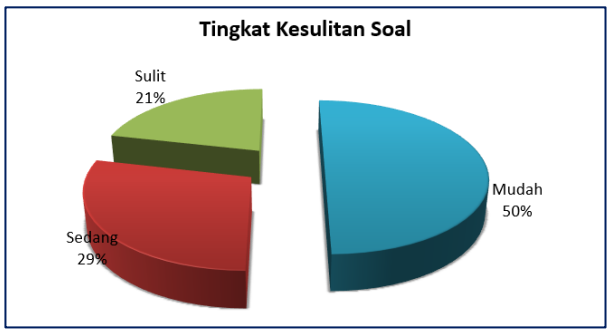
Tabel 3. Indeks Tingkat Kesukaran Soal

Indeks Tingkat Kesukaran Soal	Kategori
0,71 - 1,00	Mudah
0,31 - 0,70	Sedang
0,00 - 0,30	Sulit

Dalam menganalisis tingkat kesukaran soal buatan CAI, peneliti menggunakan software Anbuso yang disajikan pada Tabel 4,

Tabel 4. Analisis Tingkat Kesukaran Soal CAI

Butir Soal	Indeks Kesukaran Soal	Interpretasi Anbuso
1	0,949	Mudah
2	0,949	Mudah
3	0,974	Mudah
4	0,949	Mudah
5	0,641	Sedang
6	1,000	Mudah
7	0,692	Sedang
8	0,282	Sulit
9	0,026	Sulit
10	0,769	Mudah
11	0,846	Mudah
12	0,564	Sedang
13	0,333	Sedang
14	0,154	Sulit



Gambar 1. Diagram Persentase Tingkat Kesukaran Soal CAI (Anbuso)

Berdasarkan data tersebut, dari 14 soal buatan CAI sebanyak tujuh soal termasuk dalam kategori soal mudah, empat soal sedang, dan tiga soal sulit. Hasil analisis tingkat kesukaran soal menunjukan bahwa, meskipun CAI mengasilkan soal yang didominasi dengan kategori mudah (50%), CAI terbukti mampu menghasilkan soal dengan kategori sedang (29%) dan sulit (21%). Maka dari hasil analisis tersebut, didapatkan temuan bahwa CAI mampu membuat soal dengan beragam tingkat kesukaran. Selaras dengan hal ini, telaah ahli terhadap soal interaktif buatan CAI terbukti memiliki tingkat kesukaran yang beragam dan berdasarkan hierarki Taksonomi Bloom (Azmi dkk., 2024). Berbeda dengan QuestionWell AI yang seluruh soalnya memiliki tingkat kesukaran yang homogen, yaitu seluruhnya berkategori mudah (Zuhri dkk., 2024b).

Namun, keberagaman tingkat kesukaran soal perlu diiringi dengan keseimbangan sebaran soal untuk setiap kategorinya. Soal yang baik perlu memerhatikan proporsi tingkat kesukaran yang ideal demi terwujudnya pembelajaran yang optimal. Soal yang baik memiliki tingkat kesulitan yang seimbang, yaitu soal tidak terlalu mudah yang dapat dijawab benar seluruh *testee*, dan tidak terlalu sulit sehingga membuat hanya sedikit dari para *testee* yang mampu menjawab benar (Bano dkk., 2022). Menurut Arifin (2015) bahwa persentase tingkat kesukaran soal yang diujikan dapat mengacu pada rasio proporsi berikut:

1. Kategori soal sulit 25%, sedang 50%, dan mudah 25%
2. Kategori soal sulit 20%, sedang 60%, dan mudah 20%
3. Kategori soal sulit 15%, sedang 70%, dan mudah 15% (Arifin, 2015).

Dari hasil analisis tingkat kesukaran soal, didapatkan temuan bahwa proporsi soal sulit buatan CAI telah sesuai dengan proporsi soal sulit ideal. CAI menghasilkan persentase soal sulit sebesar 21% yang pada umumnya soal sulit memiliki rentang proporsi ideal sebesar 15% - 25%. Namun, terdapat indikasi ketidakseimbangan proporsi untuk soal berkategori mudah dan sedang. Pada proporsi soal yang ideal, seharusnya soal berkategori sedang lebih mendominasi daripada soal

mudah. Tapi pada soal-soal buatan CAI justru sebaliknya, proporsi soal mudah (50%) lebih mendominasi dibandingkan dengan soal sedang (29%).

Untuk memperkaya pemahaman analisis tingkat kesukaran soal buatan CAI ini, peneliti mencoba menggali perspektif guru. Berdasarkan telaah guru, soal buatan CAI memang memiliki tingkat kesukaran yang beragam mulai dari mudah sampai dengan sulit. Perbedaan tingkat kesukaran soal dapat terlihat jelas dari butir soal nomor 1-7 (wacana 1) yang berkategori mudah. Sedangkan pada soal nomor 8-14 (wacana 2) soal berkategori sulit yang cukup membingungkan siswa dalam menjawabnya.

Butir soal nomor 8-14 dianggap sulit oleh guru, CAI merancang soal tersebut untuk tingkat kognitif C4 (menganalisis), C5 (evaluasi), dan C6 (sintesis). Tentu ini menjadi penyebab soal terasa lebih sulit, jika dibandingkan dengan tujuh soal sebelumnya. Selaras dengan ini, ahli evaluasi menyoroti penggunaan kata kerja operasional (KKO) pada soal-soal buatan CAI untuk level C4, C5, dan C6 perlu ditinjau kembali (Sari & Pratiwi, 2023).

Dalam menyusun soal, penting untuk memerhatikan aspek bahasa dan kosa kata yang dipilih. Setiap butir soal yang diujikan harus menggunakan bahasa yang komunikatif, yaitu bahasa yang familiar dan sudah diajarkan kepada siswa (Erlina, 2022). Hal ini bertujuan agar siswa dapat memahami konteks soal dan menjawabnya. Kemudian, soal yang diberikan pada siswa juga perlu memerhatikan proporsi tingkat kesukaran soal yang seimbang (Nurjanah & Marlianingsih, 2015).

Sehingga, menurut telaah guru, soal-soal buatan CAI terutama dengan tingkat kognitif tinggi belum disertai pertimbangan terhadap tingkat kesulitan kosa kata yang digunakan, khususnya untuk pembelajaran bahasa Arab sebagai bahasa kedua (*non-native speaker*). Sebab, kosa kata yang digunakan terasa asing dan tidak familiar, seperti seolah-olah dibuat untuk penutur asli. Berbeda dengan soal buatan CAI berbahasa Indonesia yang dibuat untuk penutur jati, ahli menilai bahwa bahasa Indonesia yang digunakan sudah baik dan benar serta sesuai dengan tingkat perkembangan siswa (Azmi dkk., 2024). Selain itu, CAI juga masih perlu menyesuaikan sebaran tingkat kesukaran soal yang dihasilkan dengan proporsi ideal antara soal mudah, sedang, dan sulit.

Hal ini dapat terjadi karena pada CAI tidak terdapat fitur untuk memberikan *prompt* atau perintah untuk membuat soal sesuai dengan kebutuhan pembuat soal. Fitur yang disediakan terbatas pada tipe soal, level Taksonomi Bloom, dan bahasa yang digunakan untuk membuat pertanyaan. Padahal, *Prompt* dapat memaksimalkan kebutuhan pengguna dengan sesuai dan

spesifik. Selain itu, *prompt* juga berpotensi meningkatkan akurasi respon AI, mendorong kreativitas AI, dan memberikan data secara maksimal (Pujiati, 2024).

3.4. Daya Pembeda

Daya pembeda bertujuan untuk mengukur sejauh mana suatu tes dapat membedakan antara siswa yang mampu menguasai kompetensi dengan yang belum mampu menguasainya (Muhson dkk., 2015). Semakin tinggi koefisien daya pembeda butir soal, maka semakin akurat pula butir soal tersebut membedakan tingkat kompetensi siswanya. Dalam menghitung daya beda soal buatan CAI, peneliti menggunakan software Anbuso untuk menganalisisnya.

Tabel 5. Klasifikasi Daya Pembeda (Anbuso)

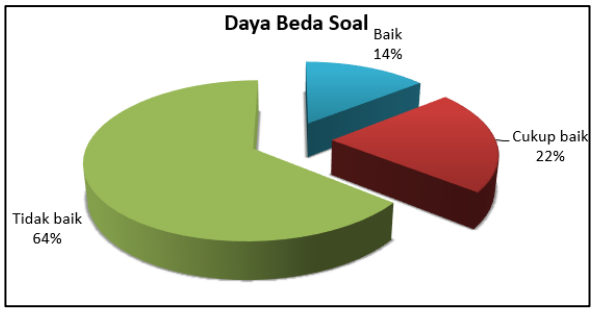
Indeks Daya Pembeda	Kategori
Daya Pembeda > 0.3	Baik
0.2 – 0.3	Cukup Baik
0.2 < Daya Pembeda	Tidak Baik

Sumber**:** (Muhson, 2017)

Hasil analisis daya pembeda soal buatan CAI dapat dilihat dalam Tabel 6, analisis dibantu dengan menggunakan software Anbuso.

Tabel 6. Hasil Analisis Daya Pembeda Soal CAI

Butir Soal	*Koefisien Daya Pembeda*	*Keterangan*
1	0.359	Baik
2	0.108	Tidak Baik
3	0.127	Tidak Baik
4	0.359	Baik
5	0.031	Tidak Baik
6	0.000	Tidak Baik
7	-0.067	Tidak Baik
8	-0.126	Tidak Baik
9	-0.118	Tidak Baik
10	0.236	Cukup Baik
11	0.298	Cukup Baik
12	0.220	Cukup Baik
13	-0.013	Tidak Baik
14	-0.088	Tidak Baik



Gambar 2. Diagram Daya Pembeda Soal CAI (Anbuso)

Hasil analisis daya beda menggunakan software Anbuso menunjukan bahwa soal-soal yang dibuat oleh CAI didominasi dengan soal yang memiliki daya beda tidak baik. Dari 14 soal yang dibuat CAI, sembilan soal memiliki daya beda tidak baik, tiga soal memiliki daya beda cukup baik, dan hanya dua soal yang memiliki daya beda baik. Daya beda dan tingkat kesukaran soal dapat dipengaruhi oleh efektivitas distraktor. Jika efektivitas distraktor buruk, maka daya beda dan tingkat kesukaran soal menjadi rendah, berlaku pula sebaliknya (Akhmadi, 2021).

Sebesar 14% soal memiliki kategori daya pembeda yang baik, yaitu pada soal nomor 1 dan 4. Kedua soal ini tergolong ke dalam soal mudah, dan *mayoritas* testee yang menjawab benar didominasi oleh kelompok atas. Namun, pada kedua soal ini, masih ditemukan adanya distraktor yang tidak berfungsi secara efektif. Hal ini menunjukkan bahwa kedua soal tersebut dapat membedakan siswa dalam memahami wacana.

Kemudian, untuk soal dengan daya beda berkategori cukup baik dengan persentase 22% terdapat pada tiga soal, yaitu soal nomor 10,11, dan 12. Soal nomor 10 dan 11 berkategori mudah dan soal nomor 12 berkategori sedang. Ketiga soal ini didominasi *testee* yang memilih jawaban benar berasal dari kelompok atas. Perbandingan antara kelompok atas dan bawah yang menjawab benar pada ketiga soal ini terlihat cukup signifikan. Selain itu, dari segi efektivitas distraktor yang terdapat pada ketiga soal ini mayoritas distraktornya terbukti efektif dalam membingungkan peserta uji.

Lalu untuk sisanya, sebanyak sembilan butir soal dinyatakan memiliki daya pembeda yang tidak baik dengan persentase 64%. Sembilan soal tersebut memiliki tingkat kesukaran yang bervariasi, mulai dari mudah sampai dengan sulit. Proporsi *testee* yang menjawab benar ataupun salah dari dua kelompok tersebut tidak terlihat secara signifikan dan hampir seimbang. Bahkan dalam beberapa soal, *testee* yang menjawab salah mayoritas berasal dari kelompok atas sehingga menyebabkan soal tersebut kehilangan fungsinya untuk menguji *testee*. Meskipun, pada efektivitas distraktornya didominasi oleh distraktor yang efektif. Namun, esensi soal yang diujikan menjadi hilang karena soal tidak mampu membedakan kemampuan siswa.

Butir soal dengan kategori daya pembeda baik (14%) dan cukup baik (22%) masih dapat dipertahankan. Kemudian untuk butir soal dengan kategori daya pembeda yang tidak baik (64%) perlu ditinjau agar dapat direvisi kembali. Namun, karena mayoritas butir soal yang dihasilkan memiliki kategori daya pembeda yang tidak baik, perlu adanya pertimbangan lebih lanjut untuk mengujikan soal buatan CAI ini.

Hasil penelitian menunjukkan bahwa soal buatan CAI khususnya pada soal pilihan ganda belum dapat dijadikan acuan sepenuhnya untuk mengukur kemampuan *testee* terhadap penguasaan materi yang telah diberikan. Dari 14 soal yang dibuat CAI, hanya lima soal yang layak diujikan dengan pertimbangan revisi, sedangkan sembilan soal sisanya tidak bisa digunakan sebagai instrumen evaluasi yang berkualitas. Sebab, guru akan kesulitan untuk mendapatkan data kemampuan siswa secara ril, karena soal buatan CAI terbukti belum memiliki kapabilitas untuk mengukurnya.

3.5. Efektivitas Distraktor

Di dalam soal objektif berbentuk pilihan ganda, terdapat lebih dari satu alternatif jawaban yang kemudian dapat disebut dengan opsi. Dari beberapa opsi yang disediakan, hanya ada satu jawaban dinyatakan benar yang disebut sebagai kunci jawaban. Sedangkan opsi lainnya disebut dengan distraktor (pengecoh). Efektivitas distraktor dapat dikatakan berfungsi jika paling tidak 5% dari partisipan tes memilih opsi tersebut dalam setiap butir soal yang diujikan (Uno & Koni, 2018). Hasil analisis efektivitas pengecoh soal buatan CAI dapat dilihat pada Tabel 7, analisis dibantu dengan software Anbuso.

Tabel 7. Efektivitas Pengecoh Soal Buatan CAI (Anbuso)

Butir Soal	*A*	*B*	*C*	*D*	*Kunci Jawaban*	*Jawaban Tidak Efektif*
1	0	0	2	37	D	AB
2	0	0	37	2	C	AB
3	38	0	1	0	A	BD
4	1	0	37	1	C	B
5	1	3	25	10	C	-
6	0	39	0	0	B	ACD
7	3	2	27	7	C	-
8	21	3	11	4	C	-
9	22	1	8	8	B	-
10	3	30	4	2	B	-
11	5	33	0	1	B	C
12	4	22	2	11	B	-
13	21	13	5	0	B	D
14	10	6	21	2	B	-

Hasil analisis efektivitas distraktor dengan menggunakan software Anbuso pada soal pilihan ganda buatan CAI menunjukkan bahwa, sebagian besar opsi yang menjadi distraktor pada soal-soal tersebut efektif dalam membingungkan *testee*. Hanya ada 12 distraktor yang tidak efektif (29%) dari 42 opsi distraktor selain kunci jawaban. Sisanya, sebanyak 30 distraktor efektif (71%) dalam membingungkan peserta uji. Hal ini berbeda dengan telaah ahli evaluasi dalam studi terdahulu, bahwa opsi-opsi yang dihasilkan oleh CAI perlu diperhatikan kembali (Sari & Pratiwi, 2023).

Sejalan dengan hal ini, berdasarkan telaah guru, distraktor yang dibuat oleh CAI membuat siswa bingung dalam menerjemahkan makna kosa kata kompleks, bukan pemahaman kontekstual yang mendalam terkait wacana yang diujikan. Akibatnya, siswa yang tidak memahami pertanyaan atau merasa tidak mampu menjawab dengan benar cenderung menebak secara cepat tanpa mempertimbangkan opsi jawaban secara mendalam (Wise & Kuhfeld, 2019). Hal ini mengindikasikan bahwa sebagian besar fungsi distraktor pada tes pilihan ganda buatan CAI belum tepat sasaran.

Berdasarkan hasil analisis secara kuantitatif, mayoritas distraktor buatan CAI dinilai efektif karena banyak *testee* terkecoh ketika memilih opsi tersebut. Namun, efektivitas ini tidak sepenuhnya mencerminkan kualitas distraktor yang baik. Sebab, opsi tersebut dipilih karena mayoritas distraktor membingungkan siswa dari segi kompleksitas bahasa. Dengan demikian, meskipun distraktor dinyatakan efektif secara statistik, substansi bahasa yang terlalu rumit justru menjadi faktor utama siswa terkecoh. Maka dari itu, untuk distraktor yang tidak efektif perlu diganti oleh guru dengan distraktor lain (Akhmadi, 2021).

Namun, penilaian kualitas soal tidak hanya berdasarkan pada seberapa banyak distraktor yang efektif. Melainkan, bergantung juga pada kualitas soal dari segi tingkat kesukaran, dan daya pembedanya. Maka, interpretasi terhadap hasil analisis kualitas soal menggunakan software Anbuso didasarkan oleh hasil kriteria pada Tabel 8 berikut,

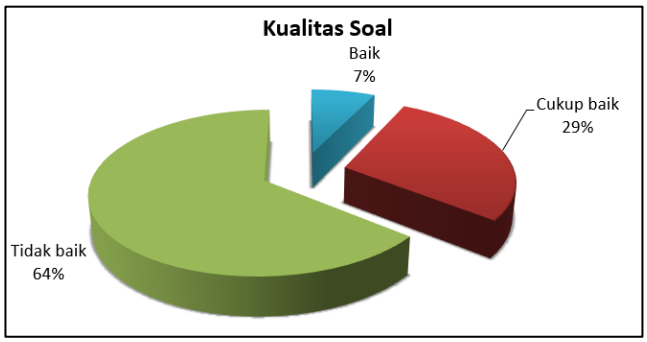
Tabel 8. Kriteria Kualitas Soal Anbuso

Kualitas Soal	Kriteria
Baik	Kualitas daya pembeda baik/cukup baik, Soal berkategori sedang, Seluruh distraktor efektif
Revisi Pengecoh	Kualitas daya pembeda baik/cukup baik, Soal berkategori sedang, Terdapat distraktor tidak efektif
Cukup Baik	Kualitas daya pembeda baik/cukup baik, Soal berkategori mudah/sulit
Tidak Baik	Kualitas daya pembeda tidak baik

Sumber**:** (Muhson, 2017)

Tabel 9. Kualitas Soal Buatan CAI (Anbuso)

Butir Soal	Kualitas Soal
1	Revisi Pengecoh
2	Tidak Baik
3	Tidak Baik
4	Revisi Pengecoh
5	Tidak Baik
6	Tidak Baik
7	Tidak Baik
8	Tidak Baik
9	Tidak Baik
10	Cukup Baik
11	Revisi Pengecoh
12	Baik
13	Tidak Baik
14	Tidak Baik



Gambar 3. Diagram Persentase Kualitas Soal Buatan CAI (Anbuso)

Dengan demikian, dari 14 soal yang dibuat oleh CAI, hanya ada satu butir soal yang memiliki kualitas soal yang baik (7%) maka soal ini boleh digunakan karena memiliki kualitas yang baik dari segi daya beda, tingkat kesukaran, efektivitas distraktor. Lalu, satu butir soal berkategori cukup baik dan tiga butir soal memerlukan revisi pengecoh (29%) maka soal ini boleh dipakai dengan melakukan revisi pada opsi yang tidak efektif. Terakhir, sembilan butir soal sisanya memiliki kualitas soal yang tidak baik (64%) yang berarti soal-soal ini memiliki daya beda yang tidak baik untuk dapat membedakan kemampuan *testee* dalam konteks membaca pemahaman bahasa Arab.

4. Simpulan

Meski studi sebelumnya telah membuktikan bahwa CAI layak sebagai media evaluasi berdasarkan telaah ahli. Namun, secara empiris berdasarkan pendekatan kuantitatif, kualitas soal-soal yang dihasilkan oleh CAI pembuat soal otomatis belum dapat memenuhi standar kualitas

butir soal yang baik. Soal buatan CAI didominasi oleh soal tidak valid dan memiliki derajat reliabilitas yang rendah. CAI mampu menghasilkan soal-soal dengan tingkat kesukaran yang beragam dan berbasis Taksonomi Bloom, tetapi mayoritas butir soal berkategori mudah. Sehingga proporsi tingkat kesukaran soal yang dihasilkan CAI tidak seimbang. Kemudian, soal-soal buatan CAI memiliki kualitas daya pembeda yang tidak baik untuk membedakan kemampuan membaca pemahaman siswa. Menurut telaah ahli pada studi sebelumnya, distraktor yang dihasilkan CAI untuk pembelajaran non-bahasa asing perlu ditinjau ulang efektivitasnya. Sedangkan dalam riset ini, secara kuantitatif mayoritas distraktor yang dihasilkan CAI dinilai efektif mengecoh *testee*.

Namun dari sudut pandang guru, CAI memang memerlukan peninjauan ulang khususnya terhadap kompleksitas bahasa yang digunakan di dalam soal dan opsi jawaban, karena terbukti membingungkan *testee* dari kompleksitas kosa kata, bukan dari segi analitis konteks soal dan jawaban. Berbeda dengan studi sebelumnya, ahli menelaah bahwa, bahasa yang digunakan CAI dalam tes interaktif non-bahasa asing sudah sesuai kaidah dan tingkat kemampuan siswa. Hal ini menjadi temuan baru, bahwa soal-soal buatan CAI terutama dengan tingkat kognitif C4, C5, C6 perlu mempertimbangkan tingkat kesulitan kosa kata yang digunakan, khususnya untuk pembelajaran bahasa Arab sebagai bahasa kedua (*non-native speaker*). CAI juga perlu mempertimbangkan tingkatan kognitif Taksonomi Bloom yang ditujukan untuk setiap jenjang atau fase yang sesuai dengan tingkat kemampuannya. Hal ini menyebabkan soal tidak merepresentasikan kemampuan kognitif peserta didik yang menjadi target evaluasi.

Penelitian ini memiliki beberapa keterbatasan. Pertama, ukuran sampel yang relatif kecil mungkin tidak dapat sepenuhnya merepresentasikan populasi yang lebih luas. Kedua, jumlah dan bentuk soal buatan CAI yang dianalisis dalam penelitian ini masih terbatas, sehingga mungkin belum dapat mencerminkan keseluruhan soal dan masih terdapat bentuk soal lain yang belum dieksplorasi. Selain itu, fokus penelitian ini terbatas pada keterampilan membaca pemahaman bahasa Arab, sehingga keterampilan berbahasa yang lain, seperti mendengarkan, berbicara, dan menulis, belum terakomodasi dalam penelitian ini.

Berdasarkan keterbatasan yang telah diuraikan, penelitian selanjutnya disarankan untuk menggunakan ukuran sampel yang lebih besar dan beragam, agar hasilnya lebih representatif. Penelitian berikutnya juga dapat memperluas jumlah dan jenis soal yang dianalisis, termasuk bentuk soal selain pilihan ganda yang tersedia pada CAI. Selain itu, disarankan untuk memperluas fokus penelitian, tidak hanya pada keterampilan membaca pemahaman bahasa Arab, tetapi juga mencakup keterampilan bahasa lainnya, sehingga dapat memberikan gambaran yang lebih komprehensif tentang pemanfaatan CAI sebagai platform pembuat soal otomatis. Kemudian,

penelitian selanjutnya juga dapat mengkaji lebih dalam terkait aspek kebahasaan soal-soal buatan CAI, khususnya untuk bahasa Arab bagi penutur asing.

Dengan demikian, soal-soal buatan CAI dalam menguji kemampuan membaca pemahaman bahasa Arab belum dapat dijadikan sebagai alternatif instrumen evaluasi final oleh guru. Perlu dilakukan pengembangan lebih lanjut terhadap model CAI pembuat soal otomatis. Penambahan fitur *prompt* untuk menginput tingkatan kognitif, jenjang pendidikan, indikator, dan tujuan pembelajaran yang disesuaikan dengan kebutuhan, menjadi saran yang dapat dipertimbangkan oleh pihak pengembang Classpoint AI (CAI) agar dapat menghasilkan soal yang lebih berkualitas dan tepat sasaran.

Sehingga CAI dapat menghasilkan soal berkualitas yang memperhitungkan aspek kebahasaan, tingkat kognitif sesuai jenjang pendidikan, serta prinsip-prinsip penyusunan soal yang baik dan benar. Diharapkan penelitian ini turut menyumbang sudut pandang baru terhadap penggunaan teknologi yang masif digunakan di dalam dunia pendidikan secara global. Di era modern ini, ketika penggunaan teknologi diiringi dengan kebijaksanaan guru dalam memanfaatkannya, tentu pendidikan berkualitas yang diharapkan akan tercapai.

Referensi

Abumosa, M. (2024). University Students' Perspectives on the Use of Interactive Presentation Technologies. *International Journal of Technology in Education and Science (IJTES)*, *8*(4), 645–667. https://doi.org/10.46328/ijtes.579

Akhmadi, M. (2021). Analisis Butir Soal Evaluasi Tema 1 Kelas 4 SDN Plumbungan Menggunakan Program Anates. *Ed-Humanistics : Jurnal Ilmu Pendidikan*, *6*(1), 799–806. https://doi.org/10.33752/ed-humanistics.v6i1.1464

Akram, H., & Abdelrady, A. H. (2023). Application of ClassPoint Tool in Reducing EFL Learners Test Anxiety: An Empirical Evidence from Saudi Arabia. *Journal of Computers in Education*, *10*(3), 529–547. https://doi.org/10.1007/s40692-023-00265-z

Amelia, M. A. (2016). Analisis Soal Tes Hasil Belajar High Order Thinking Skills (HOTS) Matematika Materi Pecahan untuk Kelas 5 Sekolah Dasar. *Jurnal Penelitian*, *20*, 123–131.

Anshari, M. I., Nasution, R., Irsyad, M., Alifa, A. Z., & Zuhriyah, I. A. (2024). Analisis Validitas dan Reliabilitas Butir Soal Sumatif Akhir Semester Ganjil Mata Pelajaran PAI. *Edukatif : Jurnal Ilmu Pendidikan*, *6*(1), 964–975. https://doi.org/10.31004/edukatif.v6i1.5931

Aregbesola, B. G. (2025). *Impact of Artificial Intelligent Class-Point on the Academic Achievement of Secondary School Students in Chemistry*. *2*(1), 1–11. https://researchvision.us/index.php/cognify/article/view/164

Arifin, Z. (2015). *Evaluasi Pembelajaran: Standar Penilaian menurut BNSP, Model Evaluasi, Instrumen Evaluasi, Penilaian Berbasis Kelas, Penilaian Portofolio, Analisis Kualitas Tes Refleksi Pelaksana Evaluasi*. Remaja Rosdakarya. https://books.google.co.id/books?id=V15NMwEACAAJ

Arikunto, S. (2013). *Metode Penelitian Kuantitatif Kualitatif dan R&D*. Alfabeta.

Arikunto, S. (2021). *Dasar-Dasar Evaluasi Pendidikan Edisi 3*. Bumi aksara.

Azmi, S., Sripatmi, Junaidi, & Wahidaturrahmi. (2024). Pengembangan Media Pembelajaran Interaktif Powerpoint Berbasis Classpoint pada Materi Matematika SMP. *Mandalika Mathematics and Education Journal*, *6*, 384. https://doi.org/10.29303/jm.v6i1.7267

Bano, V. O., Marambaawang, D. N., & Njoeroemana, Y. (2022). Analisis Kriteria Butir Soal Ujian Sekolah Mata Pelajaran IPA di SMP Negeri 1 Waingapu. *Ideas: Jurnal Pendidikan, Sosial, Dan Budaya*, *8*(1), 145. https://doi.org/10.32884/ideas.v8i1.660

Chau, T. H. T., & Pham, Q. V. B. (2023). EFL Learners' Perception of Class Point Tool Application in Enhancing their Satisfaction and Active Learning in Classroom. *Vietnam Journal of Education*, *7*(3), 302–312. https://doi.org/10.52296/vje.2023.309

Djiwandono, S. (2011). *Tes Bahasa Pegangan bagi Pengajar Bahasa*. PT. Indeks.

Erlina. (2022). Kaidah Penyusunan Tes Bahasa Arab (Pilihan Ganda). *El Jaudah: Jurnal Pendidikan Bahasa Dan Sastra Arab*, *3*(2), 82–98. https://doi.org/10.56874/ej.v3i2.1075

Fatimah, L., & Alfath, K. (2019). Analisis Kesukaran Soal, Daya Pembeda, dan Fungsi Distraktor. *Jurnal Komunikasi Dan Pendidikan Islam*, *8*(2), 1–14. https://doi.org/10.36668/jal.v8i2.115

Fitriana, N. (2023). Peningkatan Keaktifan Peserta Didik Melalui Media Persentasi Classpoint Dan Game Edukasi (Quizizz & Kahoot) Pada Pembelajaran Kimia. *ACTION : Jurnal Inovasi Penelitian Tindakan Kelas Dan Sekolah*, *3*(1), 35–41. https://doi.org/10.51878/action.v3i1.1982

Hadi, A., Zakaria, E., & Zulkarnain, D. (2024). Interactive Learning Media Training Using the Classpoint Application to Improve the Pedagogical Competence of Madrasah Ibtidaiyah Muslimat Nahdlatul Ulama Teachers in Palangka Raya. *Transformasi: Jurnal Pengabdian Masyarakat*, *20*(1), 28–38. https://doi.org/10.20414/transformasi.v20i1.8728

Hawamdeh, M. F., Khaled, M. M. B., Al-Barakat, A. A., & Alali, R. M. (2025). The Effectiveness of Classpoint Technology in Developing Reading Comprehension Skills among Non-Native Arabic Speakers. *International Journal of Information and Education Technology*, *15*(1), 39–48. https://doi.org/10.18178/ijiet.2025.15.1.2216

Hazraini, Hazraini. "Upaya Meningkatkan Kompetensi Guru Kelas Dalam Penyusunan Soal Pilihan Ganda Yang Baik dan Benar Melalui Pendampingan Berbasis KKG Semester Satu Tahun Pelajaran 2017/2018 di SD Negeri 40 Cakranegara." JUPE 2.2 (2017): 111-121.

Helmawati. (2019). *Pembelajaran dan Penilaian Berbasis HOTS*. Remaja Rosdakarya.

Hendriawan, D., & Nurman, M. (2021). *Evaluasi Pembelajaran Bahasa Arab*. Sanabil.

Huang, Y.-T., Chen, M. C., & Sun, Y. S. (2018). *Development and Evaluation of a Personalized Computer-aided Question Generation for English Learners to Improve Proficiency and Correct Mistakes*. https://doi.org/10.48550

Istiqomah, S. (2024). Penerapan Media Presentasi ClassPoint untuk Meningkatkan Hasil Belajar Mata Pelajaran Bahasa Inggris Peserta Didik di MTsN 9 Jombang. *Edu Aksara*, *3*(1), 1–16. https://doi.org/10.5281/zenodo.10842750

Kemdikbudristek. (2022). *Capaian Pembelajaran Bahasa Arab Fase F*. https://kurikulum.kemdikbud.go.id/file/cp/dasmen/34. CP Bahasa Arab.pdf

Kurniawan, T. (2015). Analisis Butir Soal Ulangan Akhir Semester Gasal Mata Pelajaran IPS Sekolah Dasar (Analysis of Odd Semester Final Test Items in Elementary School of Social Studies Subjects). *Journal of Elementary Education*, *4*(1), 2. https://journal.unnes.ac.id/sju/index.php/jee/article/view/7488

Laili, M. (2020). Ketepatan Kontruksi Butir Pilihan Ganda Bahasa Arab. *ALSUNIYAT: Jurnal Penelitian Bahasa, Sastra, Dan Budaya Arab*, *3*(2), 111–124. https://doi.org/10.17509/alsuniyat.v3i2.25272

Lugatoc, L. V. (2022). How ClassPoint Affects Learners in an English Class. *International Journal of Innovative Science and Research Technology*, *7*(11), 1569–1572. https://doi.org/10.5281/zenodo.7470871

Maulana, R. (2022). Analisis Capaian Pembelajaran Bahasa Arab dengan Taksonomi Bloom Revisi. *Jurnal PTK Dan Pendidikan*, *8*(2), 85–96. https://doi.org/10.18592/ptk.v8i2.7621

Mazlan, N. A., Tan, K. H., Othman, Z., & Wahi, W. (2023). ClassPoint Application for Enhancing Motivation in Communication among ESL Young Learners. *World Journal of English Language*, *13*(5), 520–526. https://doi.org/10.5430/wjel.v13n5p520

Muhiddin, N., Saenab, S., Muhiddin, M., & Ilham, I. (2023). Peningkatkan Kemampuan Berpikir Kreatif Peserta Didik Kelas VIII SMPN 18 Makassar melalui Penerapan Media Interaktif ClassPoint dengan Model Discovery Learning (Studi pada Materi Sistem Ekskresi). *Jurnal IPA Terpadu*, *1*, 634–641. https://doi.org/10.59562/semnasdies.v1i1.1139

Muhson, A. (2017). *Penggunaan Anbuso (Analisis Butir Soal) versi 8.0*. Universitas Negeri Yogyakarta.

Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2015). Kelayakan Anbuso sebagai Software Analisis Butir Soal bagi Guru. *Kependidikian*, *45*(2), 198–210. https://doi.org/10.21831/jk.v45i2.7499

Muliani, D. E., Azmi, K., Alius, M., Sulvayenti, A., & Amelia, L. (2024). The Influence of Classpoint Media on the Learning Motivation of Physics Education Study Program Students. *Kasuari: Physics Education Journal (KPEJ)*, *7*(1), 13–22. https://doi.org/10.37891/kpej.v7i1.484

Munip, A. (2017). *Penilaian Pembelajaran Bahasa Arab*. Fakultas Ilmu Tarbiyah dan Keguruan UIN Sunan Kalijaga Yogyakarta.

Nugraheni, A. (2024). *Trasformasi Media Artificial Intelegensi (AI) dengan Classpoint sebagai Aplikasi Pembelajaran Futuristik di Era Digital*. *8*(1), 1–10. http://dx.doi.org/10.53746/perspektif.v17i1.171

Nurjanah, & Marlianingsih, N. (2015). Analisis Butir Soal Pilihan Ganda Dari Aspek Kebahasaan. *Faktor Jurnal Ilmu Kependidikan*, *II*(1), 69–78. https://doi.org/10.30998/fjik.v2i1.377

Oktadela, R., Hadiyanti, P., & Elida, Y. (2024). The Implementation of Classpoint in Learning English: A Case Study. *Journal of English Language and Education*, *9*(3), 2024. https://doi.org/https://doi.org/10.31004/jele.v9i3.516

Pujiati. (2024). *Cara Membuat Prompt AI agar Hasil Optimal*. Dunia Dosen. https://duniadosen.com/cara-membuat-prompt-ai/

Ramadhani, F. S., & Unsiah, F. (2024). How Do EFL Students Across Gender Perceive Classpoint toward Their Motivation in Learning English? *ETERNAL (English Teaching Journal)*, *15*(1), 69–82. https://doi.org/10.26877/eternal.v15i1.338

Sanusi, R. N. A., & Aziez, F. (2021). Analisis Butir Soal Tes Objektif dan Subjektif untuk Keterampilan Membaca Pemahaman pada Kelas VII SMP N 3 Kalibagor. *Metafora: Jurnal Pembelajaran Bahasa Dan Sastra*, *8*(1), 99. https://doi.org/10.30595/mtf.v8i1.8501

Sappaile, B. I. (2010). Konsep Penelitian Ex-Post Facto. *Jurnal Pendidikan Matematika*, *1*(2), 105–113.

Sari, D., & Pratiwi, V. (2023). Pengembangan Instrumen Asesmen Berbasis HOTS (Higher Order Thinking Skill) Berbantuan Aplikasi Classpoint Pada Mata Pelajaran Layanan Lembaga Keuangan Syariah. *Jurnal Pendidikan Dan Kebudayaan (JURDIKBUD)*, *3*(2), 285–304. https://doi.org/10.55606/jurdikbud.v3i2.1915

Serdianus, S., & Saputra, T. (2023). Peran Artificial Intelligence Chatgpt dalam Perencanaan Pembelajaran di Era Revolusi Industri 4.0. *Masokan: Ilmu Sosial Dan Pendidikan*, *3*(1), 1–18. https://doi.org/10.34307/misp.v3i1.100

Setiyanto, S. (2023). Pandangan Mahasiswa Dalam Penggunaan Media Pembelajaran Interaktif Pada Mata Kuliah Dokumentasi Kebidananmenggunakan Classpoint. *Journal of Innovation And Future Technology (IFTECH)*, *5*(1), 69–78. https://doi.org/10.47080/iftech.v5i1.2463

Setiyawan, A. (2014). Faktor- faktor yang Mempengaruhi Reliabilitas Tes. *Jurnal An Nûr*, *VI*(2),

341–354.
Sukiman. (2012). *Pengembangan Sistem Evaluasi*. Insan Madani.
Sumaningsih, S. (2015). Kualitas Butir Soal UAS Bahasa Inggris Untuk Siswa Mts di Samarinda. *LINGUA: Jurnal Bahasa, Sastra, Dan Pengajarannya*, *12*(2), 223–232.
Susetyo, B. (2015). *Prosedur Penyusunan dan Analisis Tes untuk Penilaian Hasil Belajar Bidang Kognitif*. Refika Aditama.
Syafiudin. (2020). Validitas dan Reliabilitas Instrumen Penilaian pada Mata Pelajaran Bahasa Arab. *Jurnal Kajian Perbatasan Antarnegara, Diplomasi Dan Hubungan Internasional*, *3*(2), 106–118.
Syihabuddin, S. (2018). *Tes dan Evaluasi Pengajaran Bahasa*. UPI Press.
Uno, H., & Koni, S. (2018). *Assessment Pembelajaran*. Bumi Aksara.
Wao, Y. P., Priska, M., & Peni, N. (2022). Persepsi Mahasiswa Terhadap Penggunaan Media Pembelajaran Interaktif Classpoint Pada Mata Kuliah Zoologi Invertebrata. *Jurnal Inovasi Pembelajaran Biologi*, *3*(2), 76–87. https://doi.org/10.26740/jipb.v3n2.p76-87
Widodo, S., Ladyani, F., Asrianto, L. O., Rusdi, Khairunnisa, Lestari, S. M. P., Devrianya, A., Wijayanti, D. R., Hidayat, A., Dalfian, Nurcahyati, S., Sjahriani, T., Armi, Widya, N., & Rogayah. (2023). *Metodologi Penelitian*. CV Science Techno Direct.
Wise, S., & Kuhfeld, M. (2019). *What Happens When Test Takers Disengage? Understanding and Addressing Rapid Guessing (The Collaborative for Student Growth at NWEA Research Brief)*. https://www.nwea.org/uploads/2020/03/researchbrief-what-happens-when-test-takers-disengage-understanding-and-addressing-rapid-guessing-2019.pdf
Zuhri, N., Sopian, A., Sauri, S., & Nurbayan, Y. (2024). Analisis Validitas dan Reliabilitas Soal Bahasa Arab melalui Website OpExams Pembuat Soal Berbasis AI. *Jurnal Pendidikan Modern*, *9*(2), 87–91. https://doi.org/10.37471/jpm.v9i2.863
Zuhri, N. Z., Syihabuddin, S., & Tatang, T. (2024). Analisis Validitas, Reliabilitas, dan Tingkat Kesukaran Soal Bahasa Arab Tingkat SMP Berbasis Artificial Intelligence (AI) melalui Platform QuestionWell. *Jurnal Pendidikan Dan Pembelajaran Indonesia (JPPI)*, *4*(2), 693–704. https://doi.org/10.53299/jppi.v4i2.576