Jurnal Pendidikan dan Pembelajaran Indonesia (JPPI)
p-ISSN: 2797-2879, e-ISSN: 2797-2860
Volume 4, nomor 2, 2024, hal. 693-704
Doi: https://doi.org/10.53299/jppi.v4i2.576

Analisis Validitas, Reliabilitas, dan Tingkat Kesukaran Soal Bahasa Arab Tingkat SMP Berbasis *Artificial Intelligence* (AI) melalui Platform QuestionWell

Najwa Zalfa Zuhri*, Syihabuddin, Tatang
Program Studi Pendidikan Bahasa Arab, Universitas Pendidikan Indonesia, Kota Bandung, Indonesia

*Coresponding Author: najwazalfa@upi.edu
Dikirim: 19-07-2024; Direvisi: 23-07-2024; Diterima: 25-07-2024

Abstrak: Penelitian ini bertujuan untuk menguji tingkat validitas, reliabilitas, dan tingkat kesukaran soal yang dihasilkan oleh kecerdasan buatan (AI) dengan menggunakan platform QuestionWell. Metode penelitian menggunakan pendekatan deskriptif kuantitatif dengan analisis validitas, reliabilitas, dan tingkat kesukaran soal secara empirik. Sampel penelitian terdiri dari 32 siswa kelas 8 SMP Bina Insan Mandiri. Soal yang diujikan sebanyak 20 butir soal berbentuk pilihan ganda sesuai dengan materi yang dipelajari siswa. Analisis data dilakukan menggunakan korelasi *Point Biserial* untuk validitas butir soal, KR-20 untuk reliabilitas, dan uji proporsi untuk tingkat kesukaran soal. Hasil penelitian menunjukkan bahwa sebanyak 60% soal yang dihasilkan oleh AI terbukti valid. Tingkat reliabilitas soal mencapai angka 0,743, menunjukkan konsistensi yang dapat diandalkan dalam pengukuran. Seluruh soal dikategorikan sebagai soal dengan tingkat kesukaran yang mudah. Berdasarkan pemeringkatan kompetensi membaca, seluruh butir soal termasuk pada tingkat dasar. Kelebihan pada pembuat soal ini adalah penggunaannya berpotensi dalam meningkatkan efisiensi dan membantu guru dalam membuat soal secara otomatis. Namun kekurangannya yaitu, guru perlu memastikan kembali soal yang akan diujikan sesuai dengan tujuan pembelajaran dan soal yang dihasilkan belum variatif dalam mengukur kompetensi membaca siswa. Dengan demikian, peneliti merekomendasikan untuk memandang AI sebagai sarana pendukung dalam penyusunan instrumen tes. Soal yang dihasilkan AI hanya sebagai bahan masukan bagi guru, bukan sebagai instrumen final. Peran guru dalam hal ini adalah berkolaborasi dengan AI untuk mengoptimalisasi peran AI dalam ranah pendidikan.

Kata Kunci: *Artificial Intelligence*; Bahasa Arab; QuestionWell; Uji Empiris

Abstract: This research aims to test the validity, reliability, and difficulty of questions produced by Artificial Intelligence (AI) using the QuestionWell platform. The research method uses a quantitative descriptive approach with empirical analysis of questions' validity, reliability, and difficulty level. The research sample consisted of 32 8th-grade students of Bina Insan Mandiri Junior High School. The questions tested were 20 items in the form of multiple-choice questions following the material studied by students. Data analysis used Point Biserial correlation for item validity, KR-20 for reliability, and proportion test for question difficulty. The results showed that 60% of the questions AI produced proved valid. The reliability level of the questions reached 0.743, indicating reliable consistency in measurement. All questions were categorized as questions with a manageable level of difficulty. Based on the reading competency ranking, all items were at the primary level. The advantage of this question generator is that its use has the potential to increase efficiency and assist teachers in creating questions automatically. However, the disadvantages are that teachers need to ensure the tested questions follow the learning objectives, and the questions generated still need to be varied in measuring students' reading competence. Therefore, researchers recommend viewing AI as a supporting tool in preparing test instruments. AI's questions are only input for teachers, not as the final instrument. In this case, the teacher's role is to collaborate with AI to optimize the role of AI in education.

Ciptaan disebarluaskan di bawah Lisensi Creative Commons Atribusi 4.0 Internasional

Keywords: Artificial Intelligence; Arabic Language; Empirical Testing; QuestionWell

PENDAHULUAN

Kesulitan guru dalam membuat soal merupakan tantangan yang nyata dalam dunia pendidikan (Winarti dkk., 2021). Kesulitan tersebut dapat disebabkan oleh berbagai faktor, salah satunya adalah kurangnya waktu. Sebagaimana yang dikemukakan oleh Ilmi (2021) bahwa guru memiliki kesibukan dalam mengerjakan administrasi guru, mengajar, mengoreksi tugas, membuat soal, dan berbagai kegiatan lainnya. Karena itu, beberapa pekerjaan seringkali tidak dapat terselesaikan saat jam kerja di sekolah, seperti yang dialami oleh salah satu guru pondok tahfiz di Kudus (Ilmi, 2021). Problem kesulitan dalam membuat soal juga dialami oleh guru di SDN 32 Cakranegara (Kurnia dkk., 2021). Kemudian masalah lain juga ditemukan pada guru sekolah dasar di Kota Magelang yang kerap kesusahan dan menghabiskan banyak waktu dalam membuat perangkat penilaian siswa, karena guru masih menggunakan alat evaluasi konvensional (Purnanto & Mahardika, 2017). Astuti (2020) juga menegaskan bahwa proses persiapan pembelajaran memakan waktu yang cukup banyak, termasuk pada persiapan evaluasi pembelajaran siswa (Astuti dkk., 2020).

Sejak tahun 2018, *Artificial Intelligence* (AI) atau kecerdasan buatan menjadi subjek hangat yang dibincangkan oleh publik (Serdianus, 2023). AI telah berkecimpung dalam sejumlah sektor, salah satunya pada sektor pendidikan (Zahara dkk., 2023). Guru sebagai pelaku pendidikan juga mendapatkan manfaat dari peran AI dalam memberikan dukungan dan meningkatkan kualitas pembelajaran (Serdianus, 2023). Kehadiran AI memberikan kontribusi dalam membantu guru untuk membuat soal-soal, seperti pada platform QuestionWell (Gapp, 2023).

QuestionWell merupakan platform berbasis web yang menggunakan kecerdasan buatan untuk menghasilkan pertanyaan-pertanyaan pilihan ganda berdasarkan teks, video, dan dokumen yang pengguna sisipkan pada laman yang tersedia. Gapp (2023) juga menegaskan bahwa QuestionWell dapat membantu pendidik menghasilkan konten yang bermutu untuk pengajaran. Sebagai pendidik, kegiatan evaluasi merupakan bagian integral dari proses pembelajaran yang perlu dilakukan guna mengukur pemahaman dan kemampuan siswa terhadap materi yang diajarkan (Ramadhani, 2019). Dalam konteks ini, pembuatan soal atau tes merupakan salah satu instrumen yang dapat digunakan oleh guru untuk menguji kemampuan siswa (Rahmi dkk., 2020).

Dengan adanya AI, tentu ada harapan dapat membantu mengefisienkan pekerjaan guru dalam membuat soal. Namun, seiring dengan segala kecanggihan dan keuntungan yang ditawarkan oleh AI, terdapat pula tantangan yang perlu disikapi dengan bijak (Serdianus, 2023). Dalam setiap tesnya guru perlu mempertimbangkan kualitas butir soal yang akan diujikan kepada siswa (Laili, 2020). Pendidik perlu memastikan keakuratan selama penggunaan kecerdasan buatan ini. Mengingat validitas soal menjadi aspek penting yang harus diperhatikan (Hayati & Lailatussaadah, 2016), terutama ketika kita berbicara tentang penggunaan AI dalam menghasilkan pertanyaan untuk menilai hasil pembelajaran siswa. Begitu juga, reliabilitas soal menjadi parameter krusial dalam menilai sejauh mana instrumen pengukuran dapat dipercaya (Sumintono, 2016). Selain itu, tingkat kesukaran juga menentukan kualitas suatu soal yang baik. Sebagaimana yang dikatakan oleh

Ciptaan disebarluaskan di bawah Lisensi Creative Commons Atribusi 4.0 Internasional

Nurhalimah dkk. (2022), bahwa tes yang baik tidak hanya memenuhi validitas dan reliabilitas, tetapi juga memiliki taraf kesukaran soal yang seimbang.

Penelitian terdahulu telah menyoroti terkait evaluasi pembelajaran, termasuk perihal validitas dan reliabilitas soal. Seperti pada penelitian yang dilakukan oleh Imaduddin dkk. (2022) tentang uji validitas dan reliabilitas soal pembelajaran bahasa Arab di sekolah dasar non-pesantren di Bekasi, menyatakan bahwa sistem penilaian harus dikembangkan dan diterapkan dengan baik dan keberhasilan penilaian ini tergantung pada keakuratan alat yang digunakan. Kemudian pada penelitian tentang keakuratan dan keterandalan sebuah alat tes pada pembelajaran bahasa Arab oleh Syaifudin (2020) menyatakan bahwa diperlukannya penilaian yang valid dan objektif untuk menentukan hasil belajar siswa yang baik. Selain itu, penelitian tentang validitas isi dan reliabilitas pada soal penilaian harian bahasa Arab di MI Manarul Islam Malang oleh Maghfirotunnisa dkk. (2023), menjelaskan bahwa analisis butir soal perlu dilakukan guru untuk meningkatkan mutu soal yang telah dibuat oleh guru. Penelitian-penelitian tersebut menunjukan bahwa kualitas soal menjadi perhatian pada proses evaluasi pembelajaran. Kemudian, soal yang digunakan pada penelitian terdahulu bukan soal yang dibuat oleh AI. Hal ini menjadi sebuah kesenjangan penelitian sebelumnya dengan penelitian kali ini.

Adapun penelitian terdahulu yang berhubungan dengan penerapan AI dalam dunia pendidikan di antaranya yaitu, telah dilakukan oleh Kawijaya (2023) pada penelitian tentang penggunaan *Artificial Intelligence Markup Language* (AIML) dalam menganalisis kekeliruan menulis terbimbing yang dinyatakan valid dan efektif membantu siswa dan guru mencapai tujuan pembelajaran bahasa Arab. Kemudian pada implementasi Quizziz berbasis AI pada Mata Kuliah Model Pembelajaran SD di Universitas Setia Budhi Rangkasbitung telah memberikan dampak positif bagi mahasiswa maupun dosen (Durrotunnisa & Nur, 2020). Afrita (2023) juga menyatakan bahwa penggunaan AI memiliki peluang besar untuk meningkatkan efisiensi dan efektivitas sistem pendidikan. Namun perlu adanya pengawasan yang bijak dalam penggunaannya agar kebermanfaatan AI dapat optimal bagi siswa maupun lembaga pendidikan (Afrita, 2023).

Terdapat penelitian serupa dengan penelitian kali ini yang telah dilakukan oleh Zuhri dkk. (2024) tentang analisis validitas dan reliabilitas butir soal pada soal yang dihasilkan oleh AI melalui platform OpExams. Kebaruan penelitian kali ini dengan penelitian sebelumnya terdapat pada analisis soal. Pada penelitian kali ini tidak hanya validitas dan reliabilitas butir soal, tetapi juga pada uji taraf kesukaran soal dan kesesuaian soal dengan tujuan membaca pada platform QuestionWell sebagai objek penelitian.

Berdasarkan penelitian terdahulu, penelitian ini masih sangat relevan. Maka dari itu, penelitian ini bertujuan untuk melanjutkan eksplorasi terhadap penggunaan AI dalam pembuatan soal pilihan ganda dengan fokus menguji tingkat validitas, reliabilitas, dan tingkat kesukaran butir soal yang dihasilkan AI melalui platform QuestionWell. Guna memberi wawasan tentang sejauh mana kecerdasan buatan dapat diandalkan dalam mendukung tugas guru, khususnya pada pembuatan soal. Dengan menelaah penggunaan AI dalam ranah pendidikan, riset ini diharapkan dapat berkontribusi terhadap perkembangan teknologi pendidikan dan sikap pendidik dalam menjalankan kegiatan pembelajaran dengan bantuan teknologi.

Ciptaan disebarluaskan di bawah Lisensi Creative Commons Atribusi 4.0 Internasional

METODE PENELITIAN

Penelitian ini menggunakan metode deskriptif kuantitatif dengan analisis validitas, reliabilitas, dan tingkat kesukaran terhadap butir soal yang dihasilkan oleh AI melalui platform QuestionWell. Hal ini sesuai dengan tujuan penelitian yang ingin menguji tingkat validitas, reliabilitas, dan tingkat kesukaran soal yang dihasilkan oleh AI. Bentuk soal yang diujikan merupakan tipe soal pilihan ganda. Subjek dalam penelitian ini adalah siswa kelas 8 SMP Bina Insan Mandiri sebanyak 32 orang. Pembuatan soal berdasarkan materi ajar pada buku *Durūsul Lugah Al-'Arabiyyah* jilid 1. Instrumen penelitian yang digunakan berupa soal pilihan ganda bahasa Arab yang dihasilkan oleh platform QuestionWell dan lembar jawaban siswa sebanyak 20 butir soal.

Adapun prosedur dalam penelitian, peneliti membuat soal bahasa Arab melalui platform QuestionWell dengan memasukkan materi pelajaran berbentuk teks cerita bahasa Arab. Kemudian, AI akan memroses pembuatan soal. Selanjutnya, soal yang dihasilkan oleh AI diujikan kepada siswa pada mata pelajaran bahasa Arab. Setelah itu, peneliti melakukan penskoran soal dan menganalisis kualitas butir soal terhadap 4 aspek: 1) Validitas 2) Reliabilitas 3) Tingkat kesukaran soal 4) Kesesuain dengan tujuan kompetensi membaca.

Analisis validitas butir soal menggunakan korelasi *Point Biserial* (Munip, 2017). Hal itu karena, data yang digunakan merupakan data dikotomi (Munip, 2017). Sebagaimana pada jawaban hanya memungkinkan jawaban benar dan salah. Kemudian (Setyo, 2015) mengemukakan bahwa dalam menghitung reliabilitas soal dapat menggunakan formula Kuder Richardson dan untuk menguji tingkat kesukaran soal dengan menghitung nilai P atau proporsi dengan membandingkan jumlah siswa yang menjawab dengan betul terhadap jumlah seluruh peserta yang mengikuti tes.

Penelitian ini menggunakan Microsoft Office Excel 2019 untuk menghitung validitas, reliabilitas, dan indeks kesukaran soal. Lalu, hasil hitung akan diinterpretasikan. Kemudian, hasil penelitian akan diperoleh sejauh mana AI dapat menghasilkan soal yang valid, reliabel, dan apakah tes tersebut memiliki tingkat kesukaran yang ideal.

HASIL DAN PEMBAHASAN

Uji coba soal bahasa Arab berbasis AI melalui platform QuestionWell dilakukan kepada siswa kelas 8 sebanyak 32 orang. Proses pembuatan soal dengan platform QuestionWell dimulai dengan menentukan teks ajar. Teks ajar yang digunakan berasal dari buku *Durūsul Lugah Al-'Arabiyyah* jilid 1 bab 21 dan bab 22. Selanjutnya, teks tersebut dimasukkan ke dalam kolom yang tersedia di laman QuestionWell. Setelah itu, pilih jumlah soal yang diinginkan dan klik tombol *"generate"*. AI akan melakukan proses pengolahan teks tersebut menjadi soal-soal berbentuk pilihan ganda yang terdiri dari 4 opsi jawaban. Tes ini berjumlah 20 butir soal yang terdiri dari 2 wacana, di mana masing-masing wacana terdiri dari 10 butir soal. Jumlah 10 butir soal merupakan jumlah yang ideal untuk menguji sebuah wacana (Syihabuddin, 2018). Setelah soal-soal selesai diproses, soal dapat diekspor ke dalam format Microsoft Office Word. Soal-soal yang telah dibuat akan diberikan kepada siswa untuk diujikan.

Setelah soal dikerjakan oleh siswa, langkah selanjutnya adalah melakukan penskoran terhadap jawaban yang diberikan oleh siswa. Setelah proses penskoran

Ciptaan disebarluaskan di bawah Lisensi Creative Commons Atribusi 4.0 Internasional

selesai, langkah berikutnya adalah menghitung validitas, reliabilitas, dan indeks kesukaran soal yang telah diujikan kepada siswa. Dengan melakukan perhitungan ini, peneliti dapat mengevaluasi seberapa baik soal-soal tersebut dalam mengukur kemampuan dan pengetahuan siswa secara objektif. Hasil uji validitas, reliabilitas, dan tingkat kesukaran soal dapat diuraikan sebagai berikut.

Hasil Analisis Validitas Tes

Validitas soal merupakan ukuran sejauh mana suatu soal dapat mengukur apa yang seharusnya diukur (Syaifuddin, 2020). Penilaian validitas butir soal cenderung mengarah kepada kesahihan penafsiran yang ditarik berdasarkan bukti-bukti pendukung, seperti tes (Muzaffar, 2016). Hal tersebut dapat dimaknai dengan validitas penafsiran hasil tes (Muzaffar, 2016). Validitas ini diukur setelah melakukan penilaian lembar jawaban siswa. Data nilai siswa disajikan dalam bentuk data dikotomi. Nilai 1 menunjukan siswa menjawab benar, sedangkan nilai 0 menunjukan siswa menjawab salah.

Rumus korelasi *Point Biserial* digunakan dalam menghitung validitas butir soal untuk mengukur hubungan antara jawaban benar atau salah pada suatu soal dengan skor total ujian. Interpretasi hasil validitas butir soal didasarkan pada nilai koefisien *Point Biserial* yang diperoleh. Jika nilai koefisien lebih besar dari nilai kritis (tabel), maka soal tersebut dinyatakan valid atau dapat juga diartikan apabila nilai r_{pbi} lebih besar dari r tabel dengan taraf signifikansi sebesar 5%, maka soal dapat dinyatakan valid, dan sebaliknya (Munip, 2017). Adapun hasil hitung memperoleh nilai koefisien yang disajikan pada tabel berikut.

Tabel 1. Analisis Validitas Butir Soal

Nomor Soal	Koefisien Korelasi *Poin Biserial* (r_{pbi})	R tabel (Taraf signifikansi 5%)	Kategori
1	0,566	0,349	Valid
2	0,195	0,349	Tidak Valid
3	0,790	0,349	Valid
4	0,138	0,349	Tidak Valid
5	0,790	0,349	Valid
6	0,096	0,349	Tidak Valid
7	0,067	0,349	Tidak Valid
8	0,790	0,349	Valid
9	0,000	0,349	Tidak Valid
10	0,637	0,349	Valid
11	-0,003	0,349	Tidak Valid
12	0,790	0,349	Valid
13	0,281	0,349	Tidak Valid
14	0,409	0,349	Valid
15	0,352	0,349	Valid
16	0,394	0,349	Valid
17	0,851	0,349	Valid
18	0,067	0,349	Tidak Valid
19	0,586	0,349	Valid
20	0,468	0,349	Valid

Ciptaan disebarluaskan di bawah Lisensi Creative Commons Atribusi 4.0 Internasional

Tabel 2. Hasil Uji Validitas Butir Soal

Nomor Soal	Jumlah Soal	Keterangan
1, 3, 5, 8, 10, 12, 14, 15, 16, 17, 19, dan 20	12	Valid
2, 4, 6, 7, 9, 11, 13, dan 18	8	Tidak Valid

Berdasarkan hasil tersebut, ditemukan sebanyak 12 soal dianggap valid, sedangkan 8 soal lainnya dinyatakan tidak valid. Hal ini menunjukkan bahwa 60% soal memiliki validitas yang baik, namun terdapat beberapa soal yang tidak perlu digunakan atau perlu diperbaiki untuk meningkatkan kualitas tes. Hasil validitas butir soal ini memberikan gambaran tentang seberapa baik soal-soal tersebut dalam mengukur kemampuan dan pengetahuan yang diuji. Dengan demikian, sebagian soal yang dihasilkan oleh QuestionWell sudah valid secara butir soal. Guru dapat menggunakan soal dengan tetap memperhatikan keakuratan setiap butirnya.

Adapun soal yang tidak valid sebanyak 40% memiliki kemungkinan faktor yang menyebabkan tidak validnya butir soal. Suseno (2014) menyebutkan terdapat 10 faktor yang mempengaruhi tidak validnya butir soal, di antaranya: 1) Instruksi pengerjaan tidak jelas. 2) Penggunaan bahasa yang sulit. 3) Terdapat bahasa yang multi tafsir. 4) Alokasi waktu pengerjaan tidak ideal. 5) Terlalu mudah menjawab soal. 6) Kualitas tes tidak memadai untuk mengukur hasil belajar. 7) Susunan tes tidak bagus. 8) Tes yang diujikan terlalu pendek. 9) Butir tes tersusun tidak runtut. 10) Pola jawaban mudah untuk ditebak. Berdasarkan faktor-faktor tersebut, prediksi yang menyebabkan soal tidak valid pada tes kali ini yaitu, terletak pada soal yang terlalu mudah dan tes yang terlalu pendek. Hal ini berkaitan dengan hasil analisis taraf kesukaran yang disajikan dalam pembahasan hasil tingkat kesukaran di bawah ini.

Hasil Analisis Reliabilitas Tes

Sebuah tes dapat disebut reliabel apabila tes digunakan dalam beberapa kali pelaksanaan tes pada sekumpulan subjek yang sama, kemudian menghasilkan hasil pengukuran yang cenderung sama, maka tes dapat dipercaya atau reliabel (Ida & Musyarofah, 2021). Dalam penelitian ini, reliabilitas soal diuji menggunakan rumus Kuder Richardson 20 (KR-20). Tes yang dinyatakan reliabel apabila memiliki nilai koefisien reliabilitas di atas 0,70 (Munip, 2017). Berikut hasil hitung uji reliabilitas dengan jumlah butir soal sebanyak 20 soal.

Tabel 3. Hasil Uji Reliabilitas Soal

Keterangan	Nilai
K (Jumlah Butir Soal)	20
$\frac{k}{k-1}$	1,053
$\sum pq$	0,966
Varian Skor Test	3,28
$1-\frac{\sum pq}{\sigma_A^2}$	0,706
Koefisien Reliabilitas	0,743

Hasil hitung dengan rumus KR-20 menunjukkan nilai koefisien reliabilitas sebesar 0,743. Berdasarkan kriteria interpretasi yang ditetapkan, dimana nilai koefisien lebih dari 0,70 dinyatakan soal reliabel. Maka dapat diartikan bahwa

Ciptaan disebarluaskan di bawah Lisensi Creative Commons Atribusi 4.0 Internasional

instrumen soal yang dihasilkan oleh AI pada penelitian ini memiliki reliabilitas yang tinggi. Hal ini mengindikasikan bahwa soal-soal pada tes tersebut memberikan hasil yang konsisten dan dapat diulang dalam situasi pengujian yang serupa. Sebagaimana yang diungkapkan oleh Saputra dkk. (2022), soal yang terbukti reliabel dapat digunakan berulang pada waktu yang berbeda kepada peserta tes yang berbeda.

Hasil reliabilitas yang tinggi, seperti yang ditemukan dalam penelitian ini, memberikan keyakinan bahwa AI mampu menghasilkan soal-soal dengan tingkat konsistensi yang menyerupai kualitas soal yang dibuat oleh manusia. Dengan demikian, keandalan AI dalam menghasilkan butir soal menjadi relevan dalam konteks pendidikan modern, di mana teknologi semakin memainkan peran dalam mendukung proses pembelajaran dan evaluasi secara efisien. Hal ini didukung oleh penelitian terdahulu yang mengkaji tentang pendidikan di era kecerdasan buatan yang mengemukakan bahwa AI memiliki potensi untuk berintegrasi dalam ranah pendidikan (Serdianus, 2023). Dengan tetap mengulas kualitas yang dihasilkan oleh teknologi tersebut (Serdianus, 2023).

Hasil Tingkat Kesukaran Tes

Uji tingkat kesukaran pada soal bertujuan untuk mengukur sejauh mana tingkat kesulitan pada setiap butir soal yang telah diberikan kepada siswa (Nurhalimah dkk., 2022). Setyo (2015) menjelaskan, menurut Witherington klasifikasi tingkat kesukaran pada butir soal dapat dibagi menjadi 3 tingkatan seperti pada tabel berikut.

Tabel 4. Tingkat Kesukaran Soal

Skala Nilai Proporsi	Kategori Tingkat Kesukaran
0,00-0,24	Sulit
0,25-0,74	Sedang
0,75-1,00	Mudah

Kategori tingkat kesukaran dapat diperoleh melalui hasil hitung perbandingan banyaknya siswa yang menjawab betul pada butir soal dengan jumlah seluruh siswa yang mengikuti tes. Kemudian nilai indeks kesukaran diinterpretasikan sesuai dengan Tabel 5. Analisis tingkat kesukaran memperoleh hasil sebagai berikut.

Tabel 5. Hasil Analisis Tingkat Kesukaran Soal

Butir Soal	Jumlah Benar	Indeks Kesukaran	Kategori
1	30	0,938	Mudah
2	31	0,969	Mudah
3	31	0,969	Mudah
4	30	0,938	Mudah
5	31	0,969	Mudah
6	31	0,969	Mudah
7	30	0,938	Mudah
8	31	0,969	Mudah
9	32	1,00	Mudah
10	30	0,938	Mudah
11	31	0,969	Mudah
12	31	0,969	Mudah
13	30	0,938	Mudah
14	29	0,906	Mudah

Ciptaan disebarluaskan di bawah Lisensi Creative Commons Atribusi 4.0 Internasional

15	30	0,938	Mudah
16	31	0,969	Mudah
17	30	0,938	Mudah
18	30	0,938	Mudah
19	29	0,906	Mudah
20	29	0,906	Mudah

Berdasarkan hasil uji tingkat kesukaran pada Tabel 5, menunjukan bahwa seluruh soal termasuk pada kategori soal yang mudah. Dimana nilai indeks kesukaran yang diperoleh berkisar dari 0,906 hingga 1,00. Sementara, sebuah tes yang ideal adalah tes yang memiliki keseimbangan antara tingkat sukar, sedang, dan mudah (Prastika, 2021). Dengan proporsi soal sulit 20%, sedang 50%, dan mudah 30% (Prastika, 2021). Dengan demikian, naskah tes ini memiliki proporsi tingkat kesulitan yang tidak seimbang. Sehingga, soal tidak dapat digunakan secara keseluruhan dalam satu naskah tes. Soal yang dapat digunakan sebanyak 30% atau 6 butir soal dari total keseluruhan tes.

Prediksi faktor yang mempengaruhi soal tersebut memiliki kategori mudah 100% yaitu, kemungkinan siswa sudah memahami dengan baik materi yang telah dipelajari atau soal tidak menuntut siswa untuk berpikir tingkat tinggi. Soal-soal yang dihasilkan menguji seputar tingkat pengetahuan informasi berkaitan dengan wacana yang tertera saja. Tidak pada ranah menganalisis, mengevaluasi, dan mengkreasikan suatu konsep.

Menurut teori yang dikemukakan oleh Arikunto (2015), soal dikategorikan ideal apabila soal tidak terlalu susah dan juga tidak terlalu mudah. Ia juga menjelaskan bahwa soal yang mudah tidak membuat siswa untuk berusaha lebih giat dalam memecahkan persoalan pada butir soal (Arikunto, 2015). Dengan demikian, tes pada penelitian ini belum termasuk pada kategori soal yang baik. Sehingga guru perlu memodifikasi, menghilangkan, atau menambahkan soal lain dengan tingkat kesukaran kategori sedang dan sulit untuk mencapai tes dengan taraf kesukaran yang seimbang.

Analisis Kesesuaian Tes dengan Tujuan Membaca

Tes yang diujikan kepada siswa adalah tes yang menekankan pada kemampuan membaca. Dalam tes terdapat 2 wacana yang diikuti dengan pertanyaan yang menguji pemahaman siswa terhadap wacana yang tertera. Kemampuan atau keterampilan membaca terbagi menjadi 2 yaitu, kemampuan membaca lantang dan membaca pemahaman (Syihabuddin, 2018). Pada penelitian ini, terbatas pada tes kemampuan membaca pemahaman. Aspek yang berkenaan dengan pemahaman bacaan yakni, pemahaman literal, pemahaman inferensial, dan pemahaman *elaborative* (Syihabuddin, 2018). Terdapat 4 tujuan tingkatan kemampuan membaca menurut Syihabuddin (2018) di antaranya: (1) membaca untuk memperoleh informasi khusus, (2) membaca untuk memperoleh kunci-kunci informasi pragmatik makna, (3) membaca dan menentukan parafrase yang baik, dan (4) membaca untuk menyatakan ide pokok paragraf.

Ciptaan disebarluaskan di bawah Lisensi Creative Commons Atribusi 4.0 Internasional

Tabel 6. Kompetensi Membaca pada Butir Soal

Nomor Soal	Butir Soal	Kompetensi Membaca *)			
		1	2	3	4
1	ما هي مهنة حمزة وأحمد ؟	✓			
2	ما هي مهنة حامد ؟	✓			
3	ما هي مهنة زوجة حامد؟	✓			
4	ما هو اسم أحد أبناء حامد؟	✓			
5	كم عدد الأبناء لدى حامد؟	✓			
6	ما هي الألوان التي تمتلكها زينب من المناديل؟	✓			
7	ما هي الألوان التي تمتلكها يوسف من الأقلام؟	✓			
8	كم عدد الأقلام التي يمتلكها يوسف؟	✓			
9	ماذا يمتلك طلحة بكثرة؟	✓			
10	كم عدد المفاتيح التي يمتلكها طلحة؟	✓			
11	كم عدد الأبواب في المدرسة؟	✓			
12	كم عدد النوافذ في الفصل؟	✓			
13	ما هي الأثاث الموجود في الفصل؟	✓			
14	من أين يأتي بيرم؟	✓			
15	من أين يأتي يوسف؟	✓			
16	من أين يأتي عمار؟	✓			
17	من أين يأتي محمد؟	✓			
18	كم عدد الطلاب في الفصل؟	✓			
19	ما هي البلاد التي يأتي منها الطلاب؟	✓			
20	ما اسم المعلم؟	✓			
	Jumlah	20	0	0	0

*)
1. Membaca untuk memperoleh informasi khusus.
2. Membaca untuk memperoleh kunci-kunci informasi pragmatik makna.
3. Membaca dan menentukan parafrase yang baik.
4. Membaca untuk menyatakan ide pokok paragraf.

Berdasarkan soal yang dihasilkan oleh AI pada Tabel 6, hasil analisis menunjukkan bahwa seluruh butir tes mengukur kompetensi pada tujuan membaca untuk mendapatkan informasi khusus. Butir-butir tersebut berhasil mengukur aspek pemahaman bacaan pada pemahaman literal. Pemahaman tersebut tidak menekankan kepada pembaca untuk mengaitkan apa yang ada dalam bacaan dengan hal-hal di luar bacaan (Mulyana dkk., 2023). Sehingga siswa hanya fokus pada rangkaian kalimat pada teks yang dibacanya. Sebagaimana soal yang tercantum pada Tabel 6, butir soal menanyakan hal-hal informatif yang ada di dalam bacaan, seperti bertanya tentang profesi, jumlah anak, warna suatu benda, jumlah suatu benda, dan lain-lain. Tingkatan membaca yang bertujuan untuk memperoleh informasi pada bacaan seperti ini termasuk pada kategori tingkat dasar (Syihabuddin, 2018). Hal ini mengindikasikan soal-soal tersebut belum mencakup keseluruhan kompetensi yang ditutur dalam kemampuan membaca. Butir soal yang berada pada kategori tingkat mendasar ini memungkinkan siswa mengerjakan soal dengan mudah. Sehingga hasil analisis taraf kesukaran soal pada Tabel 5 juga menunjukan seluruh soal merupakan kategori soal yang mudah.

Soal yang dihasilkan AI pada penelitian ini tergolong soal yang mudah dan kompetensi soal yang tidak bervariasi, tidaklah semata-mata karena kemampuan AI yang tidak mampu membuat soal yang bervariasi. Melainkan terdapat kemungkinan faktor di luar kendali AI yaitu, kode yang diperintahkan pada platform yang kurang

Ciptaan disebarluaskan di bawah Lisensi Creative Commons Atribusi 4.0 Internasional

kompleks. Seperti yang telah dibahas oleh Lutfiyatun dkk. (2023) dalam penelitiannya tentang pemanfaatan AI pada pembelajaran bahasa Arab, bahwa kelemahan AI yang perlu diperhatikan ialah AI cenderung hanya dapat menuntaskan tugas-tugas berdasarkan perintah atau kode yang diinput. Dengan demikian, proses ini dapat menjadi perhatian bagi guru untuk membuat kode yang lebih kompleks sebelum memproses soal pada platform berbasis AI yang akan digunakan.

KESIMPULAN

Platform QuestionWell pembuat soal berbasis AI pada penelitian kali ini menunjukkan potensi AI sebagai alat bantu yang bermanfaat bagi guru dalam meringankan beban kerja mereka, khususnya dalam pembuatan soal. Platform ini dalam menghasilkan soal pilihan ganda menunjukkan hasil yang positif dalam hal validitas dan reliabilitas butir soal. Sebanyak 60% soal yang dihasilkan terbukti valid, menunjukkan kemampuan AI dalam merumuskan pertanyaan yang relevan dengan materi yang disajikan. Selain itu, reliabilitas instrumen soal yang dihasilkan oleh AI tergolong cukup baik dengan nilai koefisien reliabilitas 0,743, mengindikasikan konsistensi soal dalam mengukur aspek yang dituju.

Namun, perlu diperhatikan dalam hal tingkat kesukaran soal. Seluruh soal yang dihasilkan memiliki kategori sebagai soal mudah. Hal ini menunjukkan ketidakseimbangan proporsi tingkat kesulitan yang ideal. Disarankan perlunya penyesuaian dalam menggunakan soal yang dihasilkan oleh AI, baik dengan memilih soal yang sesuai dengan tingkat kesulitan yang dibutuhkan ataupun dengan memodifikasi soal yang sudah ada untuk mencapai keseimbangan yang lebih baik. Kemudian, memberikan kode atau perintah yang lebih detail agar tes yang dihasilkan oleh AI cenderung sesuai dengan kebutuhan.

Dengan demikian, hasil dari penelitian ini dapat memberikan rekomendasi pada pengembangan teknologi AI dalam ranah pendidikan. Khususnya dalam pembuatan soal otomatis dengan tetap berada dalam pengawasan guru/ahli.

DAFTAR PUSTAKA

Afrita, J. (2023). Peran Artificial Intelligence dalam Meningkatkan Efisiensi dan Efektifitas Sistem Pendidikan. *COMSERVA: Jurnal Penelitian dan Pengabdian Masyarakat*, *2*(12), 3181–3187. https://doi.org/10.59141/comserva.v2i12.731

Arikunto, S. (2015). *Dasar-Dasar Evaluasi Pendidikan* (2 ed.). Jakarta: Bumi Aksara.

Astuti, D. P., Muslim, A., & Bramasta, D. (2020). Analisis Persiapan Guru Dalam Pelaksanaan Pembelajaran Matematika Di Kelas Iv Sd Negeri Jambu 01. *Jurnal Wahana Pendidikan*, *7*(2), 185. https://doi.org/10.25157/wa.v7i2.3676

Durrotunnisa, & Nur, H. R. (2020). Persepsi Mahasiswa PGSD terhadap Implementasi Quizizz sebagai Media Kuis Interaktif Berbasis Artificial Intelligence. *Jurnal Basicedu*, *5*(5), 3(2), 524–532. https://journal.uii.ac.id/ajie/article/view/971

Gapp, D. (2023). Open Artificial Intelligence Assistants for Educators. *Nurse Educator*, *48*(6), 303. https://doi.org/10.1097/NNE.0000000000001489

Ciptaan disebarluaskan di bawah Lisensi Creative Commons Atribusi 4.0 Internasional

Hayati, S., & Lailatussaadah, L. (2016). Validitas Dan Reliabilitas Instrumen Pengetahuan Pembelajaran Aktif, Kreatif Dan Menyenangkan (Pakem) Menggunakan Model Rasch. *Jurnal Ilmiah Didaktika*, *16*(2), 169. https://doi.org/10.22373/jid.v16i2.593

Ida, F. F., & Musyarofah, A. (2021). Validitas dan Reliabilitas dalam Analisis Butir Soal. *Al-Mu'Arrib: Journal of Arabic Education*, *1*(1), 34–44. https://doi.org/10.32923/al-muarrib.v1i1.2100

Ilmi, D. (2021). Problematika Guru dalam Pembelajaran Tahfiz dan Alternatif Solusinya (Studi Kasus di Pondok Tahfiz Yanbu'ul Qur'an Remaja Bejen Kudus). *Repository.iainkudus.ac.id*, 34.

Imaduddin, M. F., Maulani, H., & Taufik, I. H. (2022). Test the Validity and Reliability of Arabic Learning Questions. *Arabi : Journal of Arabic Studies*, *7*(2), 198–207. https://doi.org/10.24865/ajas.v7i2.523

Kawijaya, J. (2023). Penggunaan Artificial Intelligence Markup Language (AIML) Untuk Menganalisa Kesalahan Menulis Bahasa Arab di MTs Bilingual Batu. *Jurnal Jendela Pendidikan*, *3*(03), 352–362. https://doi.org/10.57008/jjp.v3i03.569

Kurnia, M., Zain, M. I., & Jaelani, A. K. (2021). Analisis Kesulitan Guru Dalam Menyusun Perangkat Pembelajaran Di Sdn 32 Cakranegara Tahun Ajaran 2020/2021. *Jurnal Ilmiah Mandala Education*, *7*(4), 315–320. https://doi.org/10.58258/jime.v7i4.2462

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

Lutfiyatun, E., Kurniati, D., & Fajriah, N. (2023). Pemanfaatan Artificial Intelligence (AI) Dalam Tarjamah dan Muhadatsah Di Perguruan Tinggi. *Balai Diklat Keagamaan Aceh*, *2*(2), 93–105. https://seulanga.kemenag.go.id/index.php/journal/article/view/136

Maghfirotunnisa, A., Nur Qomari, & Asbarin. (2023). Validitas Isi dan Reliabilitas Instrumen Penilaian Harian Mata Pelajaran Bahasa Arab MI Manarul Islam Malang. *Shaut al Arabiyyah*, *11*(1), 19–28. https://doi.org/10.24252/saa.v11i1.34368

Mulyana, K., Afryaningsih, Y., & Nurhaidah. (2023). Analisis Keterampilan Membaca Pemahaman Literal Peserta Didik Kelas V SD Negeri 1 Rasau Jaya. *Jurnal Ilmiah Multidisiplin*, *3*(1), 51–59.

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

Muzaffar, A. (2016). Validitas Tes dan Kualitas Butir Soal. *لسانـنـا (Lisanuna): Jurnal Ilmu Bahasa Arab dan Pembelajarannya*, *5*(1), 128–143. https://jurnal.ar-raniry.ac.id/index.php/lisanuna/article/view/859/674

Nurhalimah, S., Hidayati, Y., Rosidi, I., & Hadi, W. P. (2022). Hubungan Antara Validitas Item Dengan Daya Pembeda Dan Tingkat Kesukaran Soal Pilihan

Ciptaan disebarluaskan di bawah Lisensi Creative Commons Atribusi 4.0 Internasional

Ganda Pas. *Natural Science Education Research*, *4*(3), 249–257. https://doi.org/10.21107/nser.v4i3.8682

Prastika, Y. D. (2021). Pengaruh Validitas, Reliabilitas dan Tingkat Kesukaran Terhadap Kualitas Butir Soal Ekonomi Menggunakan Software Anates di SMKN 3 Bangkalan. *STKIP PGRI Bangkalan*, 1–11.

Purnanto, A. W., & Mahardika, A. (2017). Pelatihan Pembuatan Soal Interaktif Dengan Program Wondershare Quiz Creator Bagi Guru Sekolah Dasar Di Kota Magelang. *Warta LPM*, *19*(2), 141–148. https://doi.org/10.23917/warta.v19i2.2748

Rahmi, I., Syihabudin, & Sopian, A. (2020). Analisis Kualitas Tes Bahasa Arab Berbasis Higher Order Thinking Skill (HOTS). *Lisanuna*, *10*. https://doi.org/http://dx.doi.org/10.22373/ls.v10i1.7805

Ramadhani, D. A. (2019). Evaluasi Pengajaran Bahasa Arab Dengan Media Online Di Perguruan Tinggi. *Al Mi'yar: Jurnal Ilmiah Pembelajaran Bahasa Arab dan Kebahasaaraban*, *2*(1), 85. https://doi.org/10.35931/am.v2i1.105

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

Serdianus, S. (2023). Quo Vadis Pendidikan di Era Artificial Intelligence? *OSF Preprints*, 1–20. https://doi.org/10.31219/osf.io/sf7hc%0Ahttps://osf.io/sf7hc/

Setyo, B. (2015). *Prosedur Penyusunan & Analisis Tes untuk Penelitian Hasil Belajar Bidang Kognitif*. Bandung: Refika Aditama.

Sumintono, B. (2016). *Aplikasi Pemodelan Rasch pada asesmen pendidikan: Implementasi penilaian formatif (assessment for learning)*. *March*, 1–19. http://eprints.um.edu.my/15876/1/ITS_rasch_model_asesment_for_learning.pdf

Suseno, M. N. (2014). Pengembangan Pengujian Validitas Isi dan Validitas Konstrak: Interpretasi Hasil Pengujian Validitas. *Seminar Nasional Psikometri*, *01*, 1–7.

Syaifuddin. (2020). Validitas dan Reliabilitas Instrumen Penilaian Pada Mata Pelajaran Bahasa Arab. *Jurnal Kajian Perbatasan Antarnegara, Diplomasi dan Hubungan Internasional*, *3*(2), 106–118.

Syihabuddin. (2018). *Tes dan Evaluasi Pengajaran Bahasa*. Bandung: UPI Press.

Winarti, Harida, & Lestari, I. (2021). Deskripsi Kemampuan Guru Membuat Soal Berdasarkan Pada Kurikulum 2013 di Sekolah Menengah Atas Kabupaten Landak. *Jurnal Ilmiah Wahana Pendidikan*, *7*(2), 108–115. https://doi.org/10.5281/zenodo.4659018

Zahara, S. L., Azkia, Z. U., & Chusni, M. M. (2023). Implementasi Teknologi Artificial Intelligence (AI) dalam Bidang Pendidikan. *Jurnal Penelitian Sains dan Pendidikan (JPSP)*, *3*(1), 15–20. https://doi.org/10.23971/jpsp.v3i1.4022

Zuhri, N. Z., 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.

Ciptaan disebarluaskan di bawah Lisensi Creative Commons Atribusi 4.0 Internasional