
Analisis Kualitas Butir Soal Pilihan Ganda PAI Berbasis Artificial Intelligence: Komparasi ChatGPT dan Gemini AI

Muhammad Isa Anshari¹⁾*, Fauziah Qhotrun Nada²⁾, Mohammad Irsyad³⁾

¹⁾Program Studi Pendidikan Agama Islam, Fakultas Tarbiyah, IAI Darussalam Martapura

^{2,3)}Pascasarjana Pendidikan Agama Islam, UIN Maulana Malik Ibrahim Malang

*Muhammad Isa Anshari

Email: muhammad.isa@iaidarussalam.ac.id

fauziahqhotrun99@gmail.com

irsyadmuhammm@gmail.com

Abstrak

Penelitian ini bertujuan menganalisis dan membandingkan kualitas psikometrik butir soal pilihan ganda Pendidikan Agama Islam (PAI) yang dihasilkan ChatGPT dan Gemini AI ditinjau dari validitas, reliabilitas, tingkat kesukaran, dan daya pembeda. Desain penelitian kuantitatif deskriptif-komparatif dengan teknik analisis butir soal diujikan pada 54 siswa kelas VIII (kelas 8C dan 8G) SMP Negeri 1 Gandusari. Kedua platform diberi prompt terstruktur identik yang mensyaratkan komposisi 30% LOTS, 40% MOTS, dan 30% HOTS. Hasil menunjukkan 11 dari 20 butir ChatGPT (55%) dan 8 dari 20 butir Gemini (40%) dinyatakan valid, tanpa perbedaan proporsi yang signifikan ($p = 0,527$). Reliabilitas ChatGPT tergolong rendah ($\alpha = 0,416$) sedangkan Gemini tidak reliabel ($\alpha = -0,217$). Seluruh butir dari kedua platform berkategori mudah, dengan perbedaan signifikan pada tingkat kesukaran ($p = 0,002$). Daya pembeda kedua platform tidak berbeda signifikan ($p = 0,174$). Disimpulkan bahwa AI generatif dapat mempercepat penyusunan naskah soal, tetapi belum dapat menjamin kualitas psikometrik yang memadai tanpa telaah ahli dan uji coba empiris.

Kata kunci: Kecerdasan Artifisial, ChatGPT, Gemini AI, Analisis Butir Soal, Soal Pilihan Ganda

Abstract

This study aims to analyze and compare the psychometric quality of Islamic Religious Education (PAI) multiple-choice items generated by ChatGPT and Gemini AI in terms of validity, reliability, difficulty level, and discrimination index. Using a quantitative descriptive-comparative design with item analysis techniques, the study was conducted with 54 eighth-grade students (classes 8C and 8G) at SMP Negeri 1 Gandusari. Both platforms received identical structured prompts requiring a composition of 30% LOTS, 40% MOTS, and 30% HOTS. Results show 11 of 20 ChatGPT items (55%) and 8 of 20 Gemini items (40%) were deemed valid, with no significant difference in the proportion ($p = 0.527$). ChatGPT reliability was low ($\alpha = 0.416$) while Gemini was unreliable ($\alpha = -0.217$). All items from both platforms fell into the easy category, with a significant difference in difficulty level ($p = 0.002$). The discrimination index of both platforms did not differ significantly ($p = 0.174$). The study concludes that generative AI can accelerate the drafting of test items but cannot yet guarantee adequate psychometric quality without expert review and empirical testing.

Keywords: Artificial Intelligence, ChatGPT, Gemini AI, Item Analysis, Multiple Choice Questions

PENDAHULUAN

Perkembangan kecerdasan artifisial (*Artificial Intelligence*) generatif dalam beberapa tahun terakhir membuka babak baru dalam praktik pendidikan, termasuk dalam penyusunan instrumen evaluasi pembelajaran. Ketersediaan model bahasa besar (*large language model*) seperti ChatGPT dan Gemini AI memungkinkan pendidik menyusun butir soal dalam hitungan detik, sesuatu yang secara konvensional membutuhkan waktu dan keahlian penyusunan soal yang tidak sedikit (Suharman, 2022). Kedua platform tersebut kini banyak dibandingkan penggunaannya dalam berbagai konteks pembelajaran, termasuk dalam menjawab pertanyaan konseptual maupun menyusun materi ajar (Azizah & Mansyur, 2024).

Namun demikian, kemudahan menghasilkan soal secara instan tidak serta menjamin kualitas soal tersebut secara psikometrik. Persoalan ini menjadi lebih krusial pada mata pelajaran

Pendidikan Agama Islam (PAI), karena kekeliruan konten sekecil apa pun berpotensi berdampak pada pemahaman keagamaan peserta didik. Di sisi lain, tuntutan kurikulum saat ini mengharuskan guru menyusun soal dengan sebaran level kognitif yang proporsional, mencakup *Lower Order Thinking Skills* (LOTS), *Middle Order Thinking Skills* (MOTS), dan *Higher Order Thinking Skills* (HOTS) berdasarkan Taksonomi Bloom Revisi (Zakiah & Lestari, 2019). Pada praktiknya, banyak guru mengalami kesulitan menyusun soal HOTS secara mandiri, mulai dari menganalisis Kompetensi Dasar, merumuskan stimulus kontekstual, hingga menyusun butir soal yang benar-benar mengukur penalaran tingkat tinggi (Rakyal Aini, Karma, dan Affandi 2023). Kondisi inilah yang mendorong sebagian pendidik beralih memanfaatkan AI generatif sebagai alat bantu penyusunan soal HOTS melalui teknik prompt terstruktur.

Beberapa penelitian terdahulu telah menguji kualitas soal yang dihasilkan AI generatif. Dalam penelitian (Zuhri, 2024) menganalisis validitas, reliabilitas, dan tingkat kesukaran soal Bahasa Arab yang dihasilkan platform AI QuestionWell pada 32 siswa SMP, dan menemukan bahwa 60% soal terbukti valid dengan reliabilitas sedang (0,743), tetapi seluruh butir soal berkategori mudah. Pada ranah perbandingan platform AI, (Azizah & Mansyur, 2024) membandingkan kapabilitas ChatGPT dan Gemini dalam menjawab pertanyaan konseptual fisika dan menemukan ChatGPT unggul dalam kejelasan, akurasi, dan konsistensi jawaban dibandingkan Gemini. Namun, penelitian pembandingan tersebut umumnya menilai kualitas jawaban naratif AI atas suatu pertanyaan, bukan menilai kualitas psikometrik butir soal yang dihasilkan AI ketika benar-benar diujikan kepada peserta didik. Penelitian tentang kualitas instrumen tes buatan guru maupun soal PAI konvensional telah banyak dilakukan menggunakan kerangka analisis butir (Anshari et al., 2024), tetapi penerapan kerangka yang sama untuk membandingkan dua platform AI generatif pada mata pelajaran PAI belum banyak ditemukan.

Berdasarkan uraian tersebut, teridentifikasi kesenjangan penelitian (research gap): pertama, sebagian besar studi komparasi ChatGPT dan Gemini menilai kualitas jawaban naratif AI, bukan kualitas psikometrik soal yang dihasilkannya; kedua, studi validasi soal berbasis AI yang telah ada umumnya menguji satu platform saja tanpa pembandingan; ketiga, belum ditemukan penelitian yang secara khusus menguji soal PAI berbasis AI generatif dengan kerangka analisis butir soal klasik yang lengkap sekaligus membandingkan dua platform AI mainstream. Penelitian ini berupaya mengisi kesenjangan tersebut dengan menguji secara empiris kualitas butir soal pilihan ganda PAI hasil ChatGPT dan Gemini AI menggunakan prompt terstruktur yang identik, dan mengujikannya pada kelompok responden yang sama sehingga variabel karakteristik peserta didik dapat dikendalikan. Kebaruan penelitian ini terletak pada desain komparatif antara dua platform AI generatif pada satu mata pelajaran yang sarat nilai yaitu Pendidikan Agama Islam, dengan analisis statistik uji beda yang eksplisit menjawab ada tidaknya perbedaan kualitas butir soal antarplatform.

Berdasarkan rumusan masalah ketiga, hipotesis penelitian dirumuskan sebagai berikut: H_0 : tidak terdapat perbedaan yang signifikan pada tingkat kesukaran, daya pembeda, dan proporsi butir valid antara soal pilihan ganda PAI yang dihasilkan ChatGPT dan Gemini AI. H_1 : terdapat perbedaan yang signifikan pada salah satu atau lebih dari ketiga aspek tersebut.

Berdasarkan uraian di atas, penelitian ini bertujuan untuk: (1) menganalisis kualitas soal pilihan ganda PAI yang dihasilkan ChatGPT ditinjau dari validitas, reliabilitas, tingkat kesukaran, dan daya pembeda; (2) menganalisis kualitas soal pilihan ganda PAI yang dihasilkan Gemini AI ditinjau dari keempat aspek yang sama; dan (3) menganalisis perbedaan kualitas soal pilihan ganda PAI yang dihasilkan ChatGPT dan Gemini AI.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif deskriptif dengan desain komparatif melalui teknik analisis butir soal (*item analysis*) berdasarkan teori tes klasik (*classical test theory*). Desain penelitian bersifat intra-subjek (*within-subject comparative design*), karena

kelompok responden yang sama mengerjakan dua set instrumen yang berbeda, yaitu soal hasil ChatGPT dan soal hasil Gemini AI, sehingga variabel karakteristik peserta didik relatif terkendali ketika membandingkan kualitas kedua set soal (Anwar, 2021).

Populasi penelitian ini adalah seluruh siswa kelas VIII SMP Negeri 1 Gandusari. Sampel penelitian berjumlah 54 siswa, terdiri atas 25 siswa kelas VIII C dan 29 siswa kelas VIII G, yang ditentukan menggunakan teknik sampel jenuh (total sampling), yaitu seluruh siswa pada kedua kelas tersebut dijadikan sampel penelitian.

Instrumen penelitian berupa dua set soal pilihan ganda mata pelajaran PAI, masing-masing berjumlah 20 butir, yang dihasilkan menggunakan ChatGPT dan Gemini AI. Kedua platform diberi prompt terstruktur yang identik pada materi PAI *Iman kepada Kitab-Kitab Allah*. Prompt tersebut mensyaratkan komposisi tiga level kognitif berdasarkan Taksonomi Bloom Revisi, yaitu 30% *Lower Order Thinking Skills* (LOTS: C1–C2), 40% *Middle Order Thinking Skills* (MOTS: C3), dan 30% *Higher Order Thinking Skills* (HOTS: C4–C6) yang wajib disertai stimulus kontekstual. Prompt yang digunakan dalam menghasilkan kedua set soal tersebut diadaptasi dari materi Webinar Edutech #26: “Pemanfaatan Kecerdasan Artifisial untuk Efisiensi Pembuatan Soal Pembelajaran”, yang diselenggarakan sebagai bagian dari pengembangan kompetensi dalam pemanfaatan kecerdasan artifisial untuk penyusunan soal pembelajaran.

Analisis data dilakukan melalui lima tahap. Pertama, uji validitas butir menggunakan korelasi product moment Pearson, dengan membandingkan nilai r -hitung terhadap r -tabel pada taraf signifikansi 5% ($df = n-2 = 52$; r -tabel = 0,2681); butir dinyatakan valid apabila r -hitung $\geq r$ -tabel. Kedua, uji reliabilitas menggunakan koefisien Cronbach’s Alpha. Ketiga, tingkat kesukaran dihitung dari proporsi jawaban benar (mean) tiap butir, dengan kategori 0,00–0,30 sukar, 0,31–0,70 sedang, dan 0,71–1,00 mudah. Keempat, daya pembeda dihitung menggunakan corrected item-total correlation, dengan kategori $\geq 0,40$ baik sekali, 0,30–0,39 baik, 0,20–0,29 cukup, dan $< 0,20$ buruk; nilai negatif menandakan butir bermasalah. Kelima, untuk menjawab rumusan masalah ketiga dilakukan uji beda antara set soal ChatGPT dan Gemini AI. Uji normalitas Shapiro-Wilk dan uji homogenitas Levene dilakukan terlebih dahulu untuk menentukan jenis uji beda yang sesuai: apabila data terdistribusi normal digunakan independent sample t-test, apabila tidak normal digunakan uji Mann-Whitney U. Perbedaan proporsi butir valid diuji menggunakan uji Fisher’s Exact karena ukuran sel yang kecil. Seluruh pengolahan data dilakukan dengan bantuan SPSS.

HASIL DAN PEMBAHASAN

Hasil Uji Validitas Butir Soal

Uji validitas merupakan uji yang dilakukan untuk mengetahui keabsahan/ketepatan/kecermatan suatu item pertanyaan dalam mengukur variabel yang diteliti. Suatu item pertanyaan disebut valid, apabila mampu melakukan pengukuran sesuai dengan apa yang seharusnya diukur (Kurniawan & Puspitaningtyas, 2016). Hasil uji validitas butir soal ChatGPT dan Gemini AI disajikan pada Tabel 1 dan Tabel 2.

Tabel 1. Hasil Uji Validitas Butir Soal ChatGPT

Butir Soal	Hasil Uji		Keputusan	Butir Soal	Hasil Uji		Keputusan
	r hitung	r tabel			r hitung	r tabel	
1	0,412	0,2681	Valid	11	0,56	0,2681	Valid
2	0,236	0,2681	Tidak Valid	12	-0,057	0,2681	Tidak Valid
3	0,566	0,2681	Valid	13	0,331	0,2681	Valid
4	0,34	0,2681	Valid	14	0,202	0,2681	Tidak Valid
5	-0,566	0,2681	Tidak Valid	15	0,202	0,2681	Tidak Valid
6	0,402	0,2681	Valid	16	0,566	0,2681	Valid
7	0,612	0,2681	Valid	17	0,421	0,2681	Valid
8	0,563	0,2681	Valid	18	0,167	0,2681	Tidak Valid
9	-0,119	0,2681	Tidak Valid	19	0,28	0,2681	Valid
10	0,170	0,2681	Tidak Valid	20	-0,132	0,2681	Tidak Valid

Tabel 2. Hasil Uji Validitas Butir Soal Gemini AI

Butir Soal	Hasil Uji		Keputusan	Butir Soal	Hasil Uji		Keputusan
	r_{hitung}	r_{tabel}			r_{hitung}	r_{tabel}	
1	0,094	0,2681	Tidak Valid	11	0,233	0,2681	Tidak Valid
2	-0,108	0,2681	Tidak Valid	12	0,278	0,2681	Valid
3	-0,026	0,2681	Tidak Valid	13	0,030	0,2681	Tidak Valid
4	0,324	0,2681	Valid	14	-0,217	0,2681	Tidak Valid
5	0,249	0,2681	Tidak Valid	15	0,417	0,2681	Valid
6	0,462	0,2681	Valid	16	-0,385	0,2681	Tidak Valid
7	0,221	0,2681	Tidak Valid	17	0,045	0,2681	Tidak Valid
8	0,179	0,2681	Tidak Valid	18	0,434	0,2681	Valid
9	0,436	0,2681	Valid	19	0,335	0,2681	Valid
10	0,537	0,2681	Valid	20	0,147	0,2681	Tidak Valid

Hasil uji validitas menunjukkan 11 dari 20 butir soal (55%) hasil ChatGPT dinyatakan valid, sedangkan pada Gemini AI hanya 8 dari 20 butir (40%) yang valid. Secara deskriptif ChatGPT menghasilkan proporsi butir valid yang lebih tinggi, namun uji Fisher's Exact menunjukkan perbedaan proporsi tersebut tidak signifikan secara statistik ($p = 0,527$). Artinya, dari sisi validitas empiris, kedua platform AI menghasilkan soal dengan kualitas yang secara statistik setara, meski keduanya sama-sama menyisakan sekitar setengah dari seluruh butir yang tidak valid. Temuan ini sejalan dengan hasil pengujian soal berbasis AI QuestionWell yang juga menunjukkan tidak seluruh butir soal AI lolos uji validitas empiris (Zuhri, 2024), memperkuat dugaan bahwa validitas isi yang tampak logis pada teks soal AI tidak otomatis berbanding lurus dengan validitas empiris ketika diujikan kepada peserta didik nyata. Hal ini menegaskan pentingnya telaah ahli (*expert judgment*) atas validitas isi sebelum soal AI digunakan secara operasional (Mulyatiningsih, 2011).

Hasil Uji Reliabilitas Butir Soal

Konsep reliabilitas adalah sejauh mana hasil suatu pengukuran dapat dipercaya. Sedangkan suatu instrumen dikatakan reliabel (andal) jika jawaban seseorang terhadap pertanyaan adalah konsisten atau stabil dari waktu ke waktu. Uji reliabilitas digunakan untuk mengetahui konsistensi alat ukur, apakah alat pengukur yang digunakan dapat diandalkan dan tetap konsisten jika pengukuran tersebut diulang (Widiyanto, 2018). Hasil uji Reliabilitas butir soal ChatGPT dan Gemini AI disajikan pada Tabel 3.

Tabel 3. Hasil Uji Reliabilitas Instrumen Soal

Platform	Cronbach's Alpha	Jumlah Butir	Kategori
ChatGPT	0,416	20	Rendah
Gemini AI	-0,217	20	Tidak Reliabel

Koefisien reliabilitas Cronbach's Alpha soal ChatGPT sebesar 0,416, tergolong kategori rendah, sedangkan soal Gemini AI menghasilkan nilai negatif (-0,217). Nilai alpha negatif secara statistik menandakan bahwa rata-rata kovarians antarbutir soal bernilai negatif, sebuah kondisi yang menurut Tavakol dan Dennick biasanya terjadi ketika terdapat butir-butir yang saling bertentangan secara psikometrik dalam mengukur konstruk yang sama, atau ketika instrumen tidak bersifat unidimensional (Tavakol & Dennick, 2011). Dengan kata lain, hasil ini mengindikasikan bahwa soal Gemini AI, sebagaimana disusun dan diujikan dalam penelitian ini, belum dapat diandalkan sebagai instrumen pengukuran yang konsisten. Dibandingkan dengan reliabilitas soal ulangan harian konvensional yang disusun guru, yang pada menunjukkan kategori reliabilitas yang jauh lebih baik, hasil ini menunjukkan bahwa soal buatan AI generatif dalam penelitian ini belum mampu menandingi konsistensi psikometrik soal buatan guru yang telah melalui proses penyusunan konvensional (Hilaliyah et al., n.d.). Nilai reliabilitas yang rendah pada kedua platform mengindikasikan perlunya proses revisi butir dan pengujian ulang sebelum soal-soal ini layak digunakan untuk pengambilan keputusan penilaian yang sesungguhnya.

Hasil Uji Tingkat Kesukaran Butir Soal

Tingkat kesukaran butir soal adalah proporsi peserta tes menjawab benar terhadap butir soal tersebut. Soal yang baik adalah soal yang tidak terlalu mudah atau tidak terlalu sukar. Soal yang terlalu mudah tidak merangsang siswa untuk mempertinggi usaha memecahkannya. Sebaliknya soal yang terlalu sukar akan menyebabkan mahasiswa menjadi putus asa dan tidak mempunyai semangat untuk mencoba lagi karena di luar jangkauannya (Amrina, 2022). Hasil uji tingkat kesukaran butir soal ChatGPT dan Gemini AI disajikan pada Tabel 4 dan Tabel 5.

Tabel 4. Hasil Uji Tingkat Kesukaran Butir Soal ChatGPT

Butir Soal	Mean (Output SPSS)	interpretasi	Butir Soal	Mean (Output SPSS)	interpretasi
1	0,93	Mudah	11	0,91	Mudah
2	0,93	Mudah	12	0,98	Mudah
3	0,96	Mudah	13	0,91	Mudah
4	0,93	Mudah	14	0,96	Mudah
5	0,98	Mudah	15	0,96	Mudah
6	0,96	Mudah	16	0,98	Mudah
7	0,96	Mudah	17	0,87	Mudah
8	0,78	Mudah	18	0,94	Mudah
9	0,96	Mudah	19	0,91	Mudah
10	0,98	Mudah	20	0,91	Mudah

Tabel 5. Hasil Uji Tingkat Kesukaran Butir Soal Gemini AI

Butir Soal	Mean (Output SPSS)	interpretasi	Butir Soal	Mean (Output SPSS)	interpretasi
1	0,91	Mudah	11	0,80	Mudah
2	0,91	Mudah	12	0,81	Mudah
3	0,76	Mudah	13	0,89	Mudah
4	0,87	Mudah	14	0,87	Mudah
5	0,85	Mudah	15	0,80	Mudah
6	0,78	Mudah	16	0,98	Mudah
7	0,78	Mudah	17	0,96	Mudah
8	0,96	Mudah	18	0,91	Mudah
9	0,80	Mudah	19	0,94	Mudah
10	0,93	Mudah	20	0,89	Mudah

Tabel 4 dan Tabel 5 menunjukkan seluruh 20 butir soal ChatGPT dan seluruh 20 butir soal Gemini AI berkategori mudah, meskipun kedua platform telah diinstruksikan secara eksplisit melalui prompt untuk menyusun 30% soal HOTS (C4–C6) yang semestinya menghasilkan tingkat kesukaran sedang hingga sukar. Uji Mann-Whitney U menunjukkan perbedaan tingkat kesukaran antarplatform signifikan secara statistik ($U = 315$; $p = 0,002$), dengan rerata proporsi jawaban benar pada soal ChatGPT (0,935) lebih tinggi yang berarti soal tersebut secara empiris lebih mudah dibandingkan soal Gemini AI (0,870). Temuan bahwa seluruh butir soal berkategori mudah, terlepas dari label kognitif yang disematkan dalam prompt, konsisten dengan temuan (Zuhri, 2024) yang juga mendapati seluruh butir soal QuestionWell berkategori mudah. Pola ini mengindikasikan adanya kesenjangan antara desain kognitif yang diminta melalui instruksi prompt dan tingkat kesukaran empiris yang sesungguhnya dialami peserta didik. Kondisi ini memperkuat argumen (Aini et al., 2023) bahwa penyusunan soal HOTS yang otentik memerlukan lebih dari sekadar label kognitif tertulis, melainkan juga kecermatan dalam merancang stimulus dan pengecoh sesuatu yang belum sepenuhnya dapat digantikan oleh instruksi prompt semata.

Hasil Uji Daya Pembeda Butir Soal

Daya pembeda adalah kemampuan sesuatu soal untuk membedakan antara siswa yang pandai dengan siswa yang berkemampuan rendah. Angka yang menunjukkan besarnya daya pembeda disebut indeks Diskriminasi (D). Indeks diskriminasi berkisar antara 0,00 samapai 1,00

(Zein, 2012). Hasil uji daya beda butir soal ChatGPT dan Gemini AI disajikan pada Tabel 6 dan Tabel 8.

Tabel 6. Hasil Uji Daya Pembeda Butir Soal ChatGPT

No.	Corrected Item-Total Correlation	Interpretasi	No.	Corrected Item-Total Correlation	Interpretasi
1	0,71	Baik Sekali	11	0,75	Baik
2	0,50	Baik Sekali	12	0,04	Baik Sekali
3	0,83	Baik	13	0,32	Cukup
4	0,50	Baik Sekali	14	0,39	Cukup
5	0,61	Baik	15	0,32	Cukup
6	0,83	Baik	16	0,61	Baik
7	0,83	Baik Sekali	17	0,32	Cukup
8	0,53	Baik Sekali	18	0,48	Baik
9	0,32	Cukup	19	0,64	Baik
10	-0,06	Buruk	20	0,77	Baik

Tabel 7. Hasil Uji Daya Pembeda Butir Soal Gemini AI

No.	Corrected Item-Total Correlation	Interpretasi	No.	Corrected Item-Total Correlation	Interpretasi
1	0,71	Baik Sekali	11	0,49	Baik
2	0,36	Cukup	12	0,36	Cukup
3	0,21	Cukup	13	0,07	Buruk
4	0,60	Baik	14	0,70	Baik Sekali
5	0,44	Baik	15	0,31	Cukup
6	0,55	Baik	16	0,55	Baik
7	0,55	Baik	17	0,03	Buruk
8	0,24	Cukup	18	0,36	Cukup
9	0,29	Cukup	19	0,21	Cukup
10	0,68	Baik	20	0,53	Baik

Soal ChatGPT menunjukkan rerata daya pembeda 0,512 dengan rentang -0,06 hingga 0,83, sedangkan soal Gemini AI menunjukkan rerata 0,412 dengan rentang 0,03 hingga 0,71. Uji independent sample t-test menunjukkan tidak terdapat perbedaan daya pembeda yang signifikan antara kedua platform ($t = 1,385$; $p = 0,174$). Meski demikian, perlu dicatat adanya satu butir soal ChatGPT (butir nomor 10) dengan daya pembeda negatif (-0,06), yang mengindikasikan peserta didik berkemampuan rendah justru cenderung lebih sering menjawab benar dibandingkan peserta didik berkemampuan tinggi pada butir tersebut sebuah tanda butir soal yang bermasalah dan perlu direvisi atau dibuang (Nurhalimah et al., 2022). Secara umum, kesetaraan daya pembeda antara kedua platform menunjukkan bahwa dari sisi kemampuan membedakan peserta didik berkemampuan tinggi dan rendah, ChatGPT dan Gemini AI menghasilkan kualitas yang relatif setara.

Hasil dan Pembahasan Uji Beda Kualitas Butir Soal ChatGPT dan Gemini AI

Tabel 8. Ringkasan Uji Beda Kualitas Butir Soal ChatGPT dan Gemini AI

Aspek yang Dibandingkan	Uji Statistik	Nilai Statistik	Sig. (p)	Kesimpulan
Tingkat kesukaran	Mann-Whitney U	$U = 315$	0,002	Berbeda signifikan
Daya pembeda	Independent t-test	$t = 1,385$	0,174	Tidak berbeda signifikan
Proporsi butir valid	Fisher's Exact	$OR = 2,04$	0,527	Tidak berbeda signifikan

Secara ringkas, hasil uji beda menunjukkan pola yang bervariasi antaraspek: (1) tingkat kesukaran soal ChatGPT dan Gemini AI berbeda secara signifikan, dengan Gemini menghasilkan soal yang relatif lebih sulit meski tetap berkategori mudah; (2) daya pembeda kedua platform tidak berbeda signifikan; dan (3) proporsi butir valid kedua platform juga tidak berbeda signifikan. Dengan demikian, hipotesis nol (H_0) ditolak sebagian: perbedaan signifikan hanya ditemukan pada aspek tingkat kesukaran, sedangkan pada aspek daya pembeda dan proporsi validitas H_0 diterima. Temuan ini menunjukkan bahwa keunggulan satu platform AI atas platform lainnya dalam menyusun soal PAI bersifat parsial dan bergantung pada aspek

psikometrik yang ditinjau, sehingga generalisasi bahwa satu platform lebih baik secara menyeluruh daripada platform lainnya tidak didukung oleh data penelitian ini.

Secara keseluruhan, temuan penelitian ini menunjukkan bahwa AI generatif seperti ChatGPT dan Gemini AI memang mampu mempercepat proses penyusunan naskah soal pilihan ganda PAI, termasuk mengikuti instruksi format dan proporsi level kognitif yang diminta. Namun, kecepatan tersebut tidak berbanding lurus dengan jaminan kualitas psikometrik. Reliabilitas yang rendah bahkan negatif pada Gemini AI serta ketidaksesuaian antara label kognitif HOTS dengan tingkat kesukaran menunjukkan bahwa soal hasil AI generatif belum layak digunakan secara langsung tanpa proses telaah dan uji coba lebih lanjut, sejalan dengan temuan pada instrumen HOTS bidang lain yang juga menekankan pentingnya pengujian empiris sebelum instrumen digunakan (Suprpto et al., 2020). Guru PAI yang memanfaatkan AI untuk menyusun soal disarankan tetap melakukan: (1) telaah ahli terhadap validitas isi dan ketepatan dalil keagamaan; (2) uji coba empiris skala kecil sebelum digunakan sebagai instrumen penilaian resmi; dan (3) revisi butir-butir yang terbukti tidak valid atau memiliki daya pembeda negatif. Dengan demikian, AI generatif lebih tepat diposisikan sebagai alat bantu awal (*starting point*) dalam penyusunan soal, bukan sebagai pengganti proses analisis butir soal yang mapan dalam evaluasi pendidikan (Sudijono 2011).

KESIMPULAN

Penelitian ini menunjukkan bahwa kualitas psikometrik soal PAI yang dihasilkan ChatGPT dan Gemini AI tergolong belum memadai untuk digunakan secara langsung. ChatGPT menghasilkan validitas sedang (55% valid) dengan reliabilitas rendah ($\alpha = 0,416$), sedangkan Gemini menghasilkan validitas rendah (40% valid) dengan reliabilitas tidak memadai ($\alpha = -0,217$). Seluruh butir berkategori mudah, berbeda signifikan pada aspek tingkat kesukaran ($p = 0,002$) tetapi tidak berbeda signifikan pada daya pembeda ($p = 0,174$) maupun proporsi validitas ($p = 0,527$). Perbedaan signifikan hanya pada tingkat kesukaran mengindikasikan bahwa keunggulan satu platform bersifat parsial. Disarankan pendidik PAI yang memanfaatkan AI melakukan (1) telaah ahli terhadap validitas isi, (2) uji coba empiris sebelum penggunaan resmi, dan (3) revisi butir-butir tidak valid atau bermasalah. AI generatif lebih tepat diposisikan sebagai alat bantu awal dalam penyusunan soal, bukan pengganti proses analisis butir soal yang mapan. Penelitian selanjutnya disarankan memperluas sampel, melibatkan validitas isi ahli, dan membandingkan platform AI lain.

REFERENSI

- Aini, R., Karma, I. N., & Affandi, L. H. (2023). Kesulitan Guru dalam Menyusun Soal Evaluasi Berbasis Higher Order Thinking Skills dalam Pembelajaran Kurikulum 2013. *Educatio*, 9(4), 2062–2069. <https://doi.org/10.31949/educatio.v9i4.6035>
- Amrina, Z. (2022). Evaluasi Pendidikan. In PT. Bumi Aksara. LPPM Universitas Bung Hatta.
- 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*, 06(01).
- Anwar, Z. (2021). *Belajar Psikologi Eksperimen*. Psychology Forum.
- Azizah, M. A., & Mansyur, J. (2024). Perbandingan Kapabilitas Respons Chatgpt Dan Gemini Terhadap Pertanyaan Konseptual Tentang Optik. *Jurnal Pendidikan Fisika Tadulako Onlin*, 12(April), 15–25.
- Hilaliyah, T., Sultan, U., Tirtayasa, A., Sultan, U., Tirtayasa, A., & Guru, S. B. (n.d.). *ASESMEN SUMATIF BUATAN GURU MELALUI SOAL UJI COBA PADA MATERI BAHASA INDONESIA KELAS V SEKOLAH*. 7(2), 659–668.
- Kurniawan, A. W., & Puspitaningtyas, Z. (2016). *Metode penelitian kuantitatif*. PANDIVA Buku.
- Mulyatiningsih, E. (2011). *Riset Terapan Bidang Pendidikan dan Teknik*. UNY Press.
- Nurhalimah, S., Hidayati, Y., Rosidi, I., & Hadi, W. P. (2022). Hubungan antara validitas item dengan

- daya pembeda dan tingkat kesukaran soal pilihan ganda pas. *Natural Science Education Research*, 4(3), 249–257.
- Suharman, W. (2022). Pemanfaatan ChatGPT dalam Dunia Pendidikan. *Education Journal : Journal Education Research and Development*, 7(2), 158–166.
- Suprpto, E., Sumiharsono, R., & Ramadhan, S. (2020). The Analysis of Instrument Quality to Measure the Students ' Higher Order Thinking Skill in Physics Learning. *Journal of TURKISH SCIENCE EDUCATION*, 17(4), 520–527. <https://doi.org/10.36681/tused.2020.42>
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2(1), 53–55. <https://doi.org/10.5116/ijme.4dfb.8dfd>
- Widiyanto, J. (2018). *Evaluasi Pembelajaran (Sesuai dengan Kurikulum 2013)*. UNIPMA PRESS.
- Zakiah, L., & Lestari, I. (2019). *Berpikir Kritis dalam Konteks Pembelajaran*. ERZATAMA KARYA ABADI.
- Zein, M. (2012). *Evaluasi Pembelajaran Matematika*. Daulat Riau.
- Zuhri, N. Z. (2024). *Analisis Validitas , Reliabilitas , dan Tingkat Kesukaran Soal Bahasa Arab Tingkat SMP Berbasis Artificial Intelligence (AI) melalui Platform QuestionWell*. 4, 693–704.