



Analisis psikometrik instrumen asesmen PAI: Optimalisasi kualitas butir soal melalui parameter daya pembeda

Anggi Nurmalita*, Muhammad Luthfi Dzulfikar, Ahmad Hikam Mutha'aliin, Tryana Hafilda Dewi, Salsabila Mustopa, Tarsono

Universitas Islam Negeri Sunan Gunung Djati Bandung, Indonesia

*angginurmalita47@gmail.com

Abstract

The quality of evaluation instruments in Islamic Religious Education (PAI) is often hindered by the dominance of Low Order Thinking Skills items, which fail to objectively differentiate the spectrum of students' abilities. This study aims to optimize the quality of the PAI Final Semester Assessment (AAS) instrument through an empirical analysis of the discrimination power parameter. Using a descriptive quantitative approach, this study involved all 7th-grade students at SMP Pasundan 1 Bandung (N = 30) through a total sampling technique. Response data from 30 multiple-choice items were analyzed using IBM SPSS Statistics 26 with the extreme group procedure (upper and lower 27%). The reliability test results showed very high internal consistency, with a Cronbach's Alpha coefficient of 0.978. Findings from the discrimination power analysis identified that 83.3% (25 items) were categorized as excellent, but 10% (3 items) were categorized as poor, featuring low to negative indices. This discrimination power malfunction was triggered by an excessively low difficulty level (too easy) and the failure of distractor functions. Based on these findings, it is recommended that PAI teachers completely reconstruct the ineffective items by revising the distractor wordings to make them more homogeneous, or by replacing them with items requiring higher cognitive levels to avoid assessment bias. Overall, this study concludes that digital psychometric testing is a crucial instrument in item bank curation to ensure discriminant validity and assessment fairness in the digital era.

Keywords: Item Analysis; End-of-Semester Assessment; Discrimination Power; Islamic Religious Education; Instrument Reliability.

Abstrak

Kualitas instrumen evaluasi dalam Pendidikan Agama Islam (PAI) sering kali terhambat oleh dominasi butir soal tingkat rendah (*Low Order Thinking Skills*) yang gagal mendiferensiasi spektrum kemampuan peserta didik secara objektif. Penelitian ini bertujuan untuk mengoptimalisasi kualitas instrumen Asesmen Akhir Semester (AAS) PAI melalui analisis parameter daya pembeda secara empiris. Menggunakan pendekatan kuantitatif deskriptif, penelitian ini melibatkan seluruh peserta didik kelas VII di SMP Pasundan 1 Bandung (N= 30) melalui teknik total sampling. Data respons dari 30 butir soal pilihan ganda dianalisis menggunakan IBM SPSS Statistics 26 dengan prosedur kelompok ekstrem (atas dan bawah) sebesar 27%. Hasil uji reliabilitas menunjukkan konsistensi internal yang sangat tinggi dengan koefisien *Cronbach's Alpha* sebesar 0,978. Temuan pada analisis daya pembeda mengidentifikasi bahwa 83,3% (25 butir) soal berkategori sangat baik, namun terdapat 10% (3 butir) soal berkategori jelek dengan indeks rendah hingga negatif. Malfungsi daya beda tersebut dipicu oleh tingkat kesukaran yang terlalu rendah (*too easy*) dan kegagalan fungsi pengecoh (*distractor*). Berdasarkan temuan tersebut, guru PAI direkomendasikan untuk melakukan rekonstruksi total terhadap butir

soal yang tidak efektif melalui revisi redaksi pengecoh agar lebih homogen atau menggantinya dengan soal yang memiliki level kognitif lebih tinggi guna menghindari bias penilaian. Secara keseluruhan, penelitian ini menyimpulkan bahwa pengujian psikometrik digital merupakan instrumen krusial dalam kurasi bank soal untuk menjamin validitas diskriminan dan keadilan asesmen di era digital.

Kata kunci: Analisis Butir Soal; Asesmen Akhir Semester; Daya Pembeda; Pendidikan Agama Islam; Reliabilitas Instrumen.

Pendahuluan

Evaluasi merupakan komponen integral dalam sistem pendidikan yang berfungsi untuk mengukur ketercapaian tujuan pembelajaran serta efektivitas proses instruksional. Dalam konteks Pendidikan Agama Islam (PAI), evaluasi tidak hanya bertujuan mengukur ranah kognitif, tetapi juga memetakan kedalaman pemahaman konsep keagamaan peserta didik. Pendidikan memegang peranan penting dalam membentuk generasi bangsa yang berkarakter dan kompetitif, di mana guru menjadi aktor utama yang harus memiliki kompetensi dalam merencanakan, melaksanakan, serta mengevaluasi pendidikan secara bermutu di sekolah (Tarsono, 2024). Namun, efektivitas evaluasi sangat bergantung pada kualitas psikometrik instrumen yang digunakan; instrumen yang lemah akan menghasilkan data bias yang gagal merepresentasikan kemampuan peserta didik yang sebenarnya.

Profesionalisme guru dalam melakukan evaluasi tidak hanya menuntut kemampuan administratif, tetapi juga kompetensi pedagogik dalam memastikan setiap butir soal memiliki kualitas psikometrik yang dapat dipertanggungjawabkan. Hal ini krusial karena pembuatan soal tidak boleh dilakukan secara instan tanpa melalui tahapan analisis validitas dan reliabilitas empiris guna mengasah kemampuan kognitif peserta didik secara akurat (Azzahra, Sumarni, & Putranta, 2024). Instrumen yang lemah akan menghasilkan data bias yang gagal merepresentasikan kompetensi religiusitas dan kognitif siswa secara autentik. Di era digital, tuntutan terhadap akurasi evaluasi semakin meningkat seiring berkembangnya pemanfaatan teknologi dalam penyelenggaraan asesmen pendidikan. Guru dituntut tidak hanya mampu menyusun instrumen evaluasi, tetapi juga memanfaatkan perangkat lunak analisis butir soal untuk menguji karakteristik psikometrik, seperti validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh. Melalui analisis tersebut, instrumen evaluasi dapat disempurnakan secara berkelanjutan sehingga mampu menghasilkan pengukuran yang lebih objektif, adil, dan akurat sebagai dasar pengambilan keputusan dalam proses pembelajaran.

Namun, dominasi aspek hafalan dalam evaluasi PAI saat ini menjadi tantangan nyata bagi guru di era *Society 5.0* yang menuntut kemampuan berpikir kritis dan kreatif. Untuk menjawab tantangan tersebut, implementasi pendekatan *Higher Order Thinking Skills* (HOTS) menjadi krusial dalam mendesain pembelajaran yang aktif, inovatif, dan mampu mengasah ketajaman analisis peserta didik terhadap materi keagamaan (Sukardi, 2008). Dalam konteks analisis instrumen, meskipun pendekatan *Item Response Theory* (IRT) menawarkan analisis karakteristik butir yang sangat mendalam melalui pemetaan kemampuan spesifik siswa yang kompleks, penggunaan *Classical Test Theory* (CTT) yang dioptimalkan melalui perangkat lunak statistik tetap menjadi pilihan

yang sangat relevan, praktis, dan akurat bagi pendidik dalam mengevaluasi efektivitas soal pada skala kelas atau satuan pendidikan (Fernanda & Hidayah, 2020). Integrasi bantuan komputasi digital dalam CTT memungkinkan identifikasi reliabilitas dan daya pembeda yang presisi tanpa mengabaikan aspek kemudahan interpretasi bagi praktisi pendidikan.

Meskipun pendekatan analisis psikometrik modern telah tersedia, kesenjangan (*research gap*) yang sering ditemukan dalam praktik evaluasi PAI adalah penyusunan butir soal yang cenderung bersifat rutin tanpa analisis empiris yang memadai. Banyak instrumen asesmen PAI di sekolah masih didominasi oleh aspek hafalan atau *Low Order Thinking Skills* (LOTS). Kondisi ini diperparah dengan karakteristik materi PAI yang sering kali memiliki pilihan jawaban (*distractor*) homogen, di mana seluruh opsi terlihat benar secara normatif-religius. Akibatnya, instrumen tersebut gagal membedakan antara peserta didik yang menguasai kompetensi secara mendalam dengan mereka yang hanya memahami materi secara superfisial atau sekadar menebak berdasarkan intuisi moral. Fenomena ini menuntut pemahaman terhadap hakikat daya pembeda soal sebagai parameter psikometrik yang mampu memisahkan (*discriminate*) peserta didik berkemampuan tinggi dengan peserta didik berkemampuan rendah secara objektif dan adil.

Analisis daya pembeda memiliki fungsi dan tujuan strategis dalam meningkatkan kualitas asesmen secara berkelanjutan. Secara fungsional, parameter ini menjamin bahwa setiap butir soal memiliki validitas diskriminan dalam mengukur dimensi kemampuan yang diinginkan, sekaligus memberikan dasar objektif bagi pendidik dalam melakukan kurasi bank soal. Dengan demikian, tujuan utama dari evaluasi berbasis daya beda ini adalah untuk mewujudkan keadilan asesmen (*assessment fairness*). Hal ini memastikan bahwa skor yang diperoleh siswa benar-benar merepresentasikan kompetensi nyata, bukan sekadar artefak dari faktor kebetulan (*guessing factor*) maupun ambiguitas redaksi soal yang sering kali merugikan peserta didik berkemampuan tinggi.

Optimalisasi kualitas instrumen evaluasi PAI melalui analisis parameter psikometrik kini menjadi sangat mendesak karena dua faktor fundamental. Pertama, kompleksitas materi keislaman yang mencakup aspek konseptual dan kontekstual menuntut redaksi soal yang presisi; tanpa daya beda yang kuat, instrumen cenderung terjebak pada polarisasi tingkat kesukaran yang ekstrem terlalu mudah atau terlalu sulit bagi seluruh spektrum kemampuan siswa. Kedua, ketiadaan analisis empiris menyebabkan pendidik kehilangan dasar objektif dalam melakukan kurasi bank soal, yang pada jangka panjang akan mendegradasi standar kualitas penilaian pendidikan agama. Oleh karena itu, akurasi instrumen evaluasi ini merupakan fenomena kritis yang menuntut pembuktian secara empiris guna menjamin keadilan dan objektivitas asesmen (Saputra & Mayurida, 2025). Integrasi teknologi melalui analisis psikometrik digital diharapkan mampu mentransformasi proses evaluasi PAI dari sekadar rutinitas administratif menjadi instrumen penjamin mutu pembelajaran yang lebih saintifik dan akuntabel.

Terdapat sejumlah studi yang telah membahas evaluasi kualitas instrumen asesmen PAI, namun masih memiliki keterbatasan fokus. Pertama, penelitian yang dilakukan oleh Azzahra dkk. (2024) dan Rahayu serta Putranta (2025) cenderung berfokus pada

analisis validitas dan kesesuaian butir soal secara makro menggunakan pendekatan *Item Response Theory* (IRT) Model Rasch (Azzahra dkk., 2024). Kedua, Penelitian lain umumnya masih berfokus pada diagnosis statistik dasar, seperti analisis tingkat kesukaran dan kualitas pengecoh, tanpa melakukan rekayasa yang lebih spesifik untuk meningkatkan keandalan instrumen secara menyeluruh. sebagaimana ditemukan dalam karya Rishan dan Sulaiman (Keislaman, Sma, & Padang, 2023). Perbedaannya, penelitian-penelitian tersebut belum mengeksplorasi bagaimana kelemahan statistik yang terdeteksi dapat dikonversi menjadi sebuah rancangan perbaikan (optimalisasi) secara empiris. Belum ada studi yang secara spesifik membedah daya beda soal menggunakan aplikasi SPSS 26 sebagai jembatan antara sekadar evaluasi diagnostik menuju peningkatan presisi pengukuran kompetensi siswa.

Berdasarkan kesenjangan tersebut, penelitian ini menghadirkan kebaruan (*novelty*) melalui pengembangan kerangka evaluasi “Optimalisasi Psikometrik Instrumen PAI”. Berbeda dengan model evaluasi deskriptif pasif yang hanya menampilkan persentase kelayakan atau penolakan butir soal, penelitian ini menawarkan pendekatan transformatif dengan memfungsikan kembali parameter daya pembeda sebagai instrumen rekayasa butir soal. Dalam pendekatan ini, daya pembeda tidak lagi dipahami sekadar sebagai angka korelasi, tetapi dimanfaatkan sebagai alat untuk mengidentifikasi anomali pemahaman, menyeleksi sensitivitas butir, serta memurnikan instrumen pengukuran. Penelitian ini bertujuan untuk menganalisis secara mendalam bagaimana strategi optimalisasi tersebut diterapkan secara tepat guna menghasilkan instrumen asesmen PAI yang kuat secara statistik dan mampu mengukur pemahaman keberagaman peserta didik secara akurat.

Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode deskriptif yang bertujuan untuk mengevaluasi kualitas instrumen evaluasi melalui analisis parameter psikometrik. Desain penelitian difokuskan pada analisis daya pembeda butir soal sebagai indikator kualitas asesmen (Kiki Fajariani & Erlia Hanum, 2025). Populasi penelitian ini adalah seluruh peserta didik kelas VII di SMP Pasundan 1 Bandung yang terdistribusi ke dalam tiga kelas paralel. Pengambilan sampel dilakukan dengan teknik total sampling, di mana seluruh anggota populasi yang mengikuti Asesmen Akhir Semester (AAS) dijadikan sebagai partisipan untuk mendapatkan gambaran kualitas instrumen yang komprehensif. Karakteristik partisipan mencakup keberagaman tingkat pemahaman kognitif yang merepresentasikan populasi siswa kelas VII di sekolah tersebut secara utuh.

Instrumen penelitian berupa data sekunder yang berasal dari respons jawaban peserta didik terhadap naskah AAS PAI tahun ajaran 2025/2026. Instrumen ini terdiri dari 30 butir soal pilihan ganda. Validitas isi (*content validity*) instrumen dilakukan melalui tahap *expert judgment* oleh guru senior dan sejawat di SMP Pasundan 1 Bandung untuk memastikan kesesuaian antara butir soal dengan tujuan pembelajaran (TP) dan domain kognitif yang diukur. Teknik pengumpulan data dilakukan melalui dokumentasi digital hasil jawaban siswa yang diekstraksi dari platform asesmen sekolah.

Analisis data dilakukan dengan membagi responden ke dalam kelompok atas (*upper group*) dan kelompok bawah (*lower group*) menggunakan proporsi 27% (Christy Matius & Gunawan, 2022). Validasi data dilakukan menggunakan *software* IBM SPSS *Statistics* 26 guna meminimalisir kesalahan manual dan memastikan akurasi indeks daya beda yang dihasilkan. Pertimbangan etis dijunjung tinggi dengan menerapkan prosedur anonimitas pada seluruh identitas partisipan.

Tabel 1. Analisis Data Instrumen PAI di SMP Pasundan 1 Bandung

Komponen Data	Deskripsi
Jumlah Butir Soal	30 Butir Soal Pilihan Ganda
Lokasi Penelitian	SMP Pasundan 1 Bandung
Partisipan	Seluruh Peserta Didik Kelas VII (3 Kelas Paralel)
Perangkat Lunak Statistik	IBM SPSS <i>Statistics</i> 26
Fokus Analisis	Indeks Daya Pembeda (D)

Peneliti mengumpulkan data analitik kuantitatif terhadap 30 butir soal menggunakan teknik total sampling untuk menghindari bias dalam pemilihan data. Hasil analisis awal menunjukkan bahwa instrumen ini memiliki sebaran data yang lengkap dari tiga kelas paralel, yang memungkinkan dilakukan pengelompokan peserta didik ke dalam kategori kelompok atas dan kelompok bawah untuk menguji keakuratan daya pembeda setiap butir soal secara empiris.

Prosedur analisis data dilakukan secara digital menggunakan perangkat lunak IBM SPSS *Statistics* 26. Tahapan analisis meliputi:

1. Uji Reliabilitas: Menggunakan koefisien *Cronbach's Alpha* untuk melihat konsistensi internal instrumen.
2. Pembentukan Kelompok Ekstrem: Responden diurutkan berdasarkan skor total dari tertinggi ke terendah, kemudian dibagi menjadi dua kelompok yaitu kelompok atas (*upper group*) dan kelompok bawah (*lower group*) dengan proporsi masing-masing sebesar 27% (n= 30) merujuk pada prosedur Kelley (Christy Matius & Gunawan, 2022).
3. Prosedur penghitungan daya pembeda dalam penelitian ini dilakukan dengan membagi responden ke dalam dua kelompok ekstrem, yaitu kelompok atas (*upper group*) dan kelompok bawah (*lower group*). Merujuk pada teori evaluasi standar, proporsi yang digunakan untuk masing-masing kelompok adalah sebesar 27% dari total peserta didik (Christy Matius & Gunawan, 2022).

Secara matematis, penentuan indeks daya pembeda (D) pada setiap butir soal mengikuti formula:

$$D = \frac{BA - BB}{N}$$

Keterangan:

D : Indeks Daya Pembeda

BA : Jumlah peserta didik kelompok atas yang menjawab benar

BB : Jumlah peserta didik kelompok bawah yang menjawab benar

N : Jumlah peserta pada salah satu kelompok (atas atau bawah)

Kriteria interpretasi indeks daya pembeda diklasifikasikan ke dalam empat kategori: Sangat Baik (0,71 - 1,00), Baik (0,41 - 0,70), Cukup (0,21 - 0,40), dan Jelek (< 0,20).

Pemanfaatan SPSS terbukti memberikan akurasi yang lebih tinggi dalam mengidentifikasi butir soal yang tidak efektif. Butir soal yang terlalu mudah atau terlalu sukar di luar rentang tersebut direkomendasikan untuk direvisi karena dapat menyebabkan instrumen kehilangan kemampuannya dalam membedakan spektrum kemampuan peserta didik secara tajam (Qadir, Huda, 2024).

Hasil dan Pembahasan

A. Analisis realibilitas instrumen

Sebelum dilakukan analisis terhadap indeks daya pembeda pada setiap butir soal, terlebih dahulu dilakukan uji reliabilitas untuk mengukur konsistensi internal instrumen. Pengujian ini menggunakan teknik *Cronbach's Alpha* dengan bantuan perangkat kerja IBM SPSS *Statistics* 26. Penggunaan perangkat lunak dalam analisis statistik, baik melalui IBM SPSS maupun program pendukung lainnya seperti QUEST, terbukti memberikan akurasi yang lebih tinggi dalam memvalidasi instrumen dibandingkan metode manual. Integrasi teknologi ini sangat membantu peneliti dalam meminimalisir kesalahan manusia (*human error*) serta mempercepat identifikasi terhadap butir soal yang tidak efektif atau tidak memenuhi kriteria model yang ditetapkan (Azzahra dkk., 2024). Hasil pengujian terhadap 30 butir soal Asesmen Akhir Semester (AAS) PAI menunjukkan nilai sebagai berikut:

Tabel 2. Hasil Uji Reabilitas Instrumen PAI

Reliability Statistics	
Cronbach's Alpha	N of Items
.978	30

Berdasarkan Tabel 2, nilai *Cronbach's Alpha* yang diperoleh sebesar 0,978. Angka ini menunjukkan bahwa instrumen memiliki tingkat reliabilitas yang sangat tinggi (di atas ambang batas 0,70), sehingga layak untuk dianalisis lebih lanjut pada level butir soal. Secara empiris, jumlah alternatif jawaban dalam tes objektif pilihan ganda memiliki pengaruh signifikan terhadap indeks reliabilitas dan daya pembeda instrumen. Semakin banyak jumlah alternatif jawaban yang disediakan dengan pengecoh yang homogen, maka semakin tinggi pula tingkat reliabilitas serta kemampuan soal dalam membedakan kelompok peserta didik yang berkemampuan tinggi dan rendah (Alwi, 2015).

B. Sebaran indeks daya pembeda instrumen PAI

Setelah instrumen dinyatakan reliabel secara keseluruhan, tahap berikutnya adalah mengidentifikasi indeks daya pembeda pada masing-masing butir soal. Indeks daya pembeda merupakan parameter psikometrik yang menentukan efektivitas butir soal dalam memisahkan peserta didik kelompok atas dengan peserta didik kelompok bawah. Berdasarkan hasil pengolahan data, klasifikasi indeks daya pembeda (D) dirangkum dalam tabel berikut:

Tabel 3. Analisis Data Instrumen PAI di SMP Pasundan 1 Bandung

Kategori	Indeks (D)	Frekuensi Soal	Nomor Soal	Persentase (%)
Sangat Baik	0,71 - 1,00	25	1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 13, 14, 15, 16, 17, 18, 20, 21, 22, 24, 25, 26, 27, 28, 29, 30	83,3%
Baik	0,41 - 0,70	2	10, 17	6,7%
Cukup	0,21 - 0,40	0	-	
Jelek	< 0,20	3	12, 19, 23	10%
TOTAL		30		100%

Berdasarkan Tabel 3, terlihat bahwa mayoritas butir soal memiliki kualitas yang sangat tinggi dalam membedakan kemampuan peserta didik. Sebanyak 25 butir soal (83,3%) menempati kategori Sangat Baik dengan indeks daya pembeda di atas 0,71. Hal ini mengindikasikan bahwa sebagian besar soal yang disusun telah mampu mengukur kompetensi esensial materi PAI secara tajam dan akurat. Pada kategori Baik, terdapat 2 butir soal (6,7%) yaitu nomor 10 dan 17. Meskipun indeksnya tidak setinggi kelompok pertama, kedua soal ini tetap dinyatakan valid dan efektif untuk digunakan dalam asesmen karena memiliki nilai korelasi item-total di atas ambang batas minimal 0,30. Menariknya, dalam instrumen ini tidak ditemukan soal dengan kategori Cukup, yang menunjukkan bahwa instrumen ini memiliki kecenderungan polarisasi yang jelas antara soal yang berkualitas tinggi dan soal yang tidak efektif.

Namun demikian, terdapat 3 butir soal (10%) yang masuk dalam kategori Jelek, yakni soal nomor 12, 19, dan 23. Ketiga soal tersebut memiliki indeks daya pembeda yang sangat rendah, bahkan menyentuh angka negatif (misalnya butir 23 dengan -0,025). Rendahnya indeks pada ketiga butir soal ini disebabkan oleh tingkat kesukaran yang terlalu rendah, di mana hampir seluruh peserta didik baik dari kelompok atas maupun kelompok bawah mampu menjawab soal dengan benar. Dalam konteks evaluasi pendidikan, butir soal yang masuk kategori jelek ini direkomendasikan untuk dihapus atau direvisi total guna menjaga reliabilitas keseluruhan instrumen. Secara keseluruhan, dengan nilai reliabilitas *Cronbach's Alpha* sebesar 0,978, instrumen AAS PAI ini dinyatakan sangat layak untuk digunakan, setelah melakukan eliminasi pada butir-butir soal yang tidak valid.

C. Faktor-faktor yang memengaruhi daya pembeda soal

Berdasarkan temuan data pada Tabel 2, terdapat beberapa faktor fundamental yang memengaruhi variasi indeks daya pembeda pada instrumen PAI di SMP Pasundan 1 Bandung. Pertama, heterogenitas kemampuan peserta didik menjadi faktor kunci dalam keberhasilan pengujian ini. Pengambilan sampel dari tiga kelas paralel (VII A, VII B, dan VII C) memberikan sebaran kemampuan yang cukup luas, sehingga memfasilitasi butir soal untuk memisahkan kelompok atas dan kelompok bawah secara kontras. Sejalan dengan teori evaluasi, semakin beragam tingkat penguasaan kompetensi dalam suatu sampel, semakin akurat instrumen dalam menunjukkan daya bedanya.

Kedua, kualitas pengecoh (*distractor*) pada setiap butir soal memegang peranan krusial. Pada 25 butir soal yang berkategori sangat baik, pengecoh berfungsi secara efektif karena dipilih oleh peserta didik dari kelompok bawah, sementara kelompok atas

cenderung memilih kunci jawaban yang tepat. Sebaliknya, pada butir soal nomor 12, 19, dan 23 yang berkategori jelek, fungsi pengecoh tidak berjalan secara optimal. Hal ini disebabkan oleh tingkat kesukaran soal yang terlalu rendah (*too easy*), sehingga tidak terjadi proses eliminasi pilihan jawaban oleh peserta didik, yang pada akhirnya menghilangkan daya pembeda soal tersebut.

Ketiga, ambiguitas redaksi dan bahasa dalam butir soal. Adanya nilai negatif pada butir soal nomor 23 (-0,025) mengindikasikan adanya potensi kesalahan teknis, baik dalam perumusan kalimat soal maupun kunci jawaban yang membingungkan. Kondisi ini menyebabkan peserta didik yang memiliki kemampuan tinggi (kelompok atas) justru terkecoh oleh pilihan jawaban yang tersedia, sedangkan peserta didik kelompok bawah menjawab benar karena faktor kebetulan (*guessing*). Oleh karena itu, kejernihan bahasa dan kesesuaian materi esensial PAI menjadi prasyarat mutlak dalam meningkatkan kualitas psikometrik instrumen.

Terakhir, integrasi teknologi dalam analisis statistik melalui penggunaan IBM SPSS *Statistics* 26 sangat membantu peneliti dalam mengidentifikasi faktor-faktor di atas secara objektif. Penggunaan fitur *Scale if Item Deleted* memungkinkan peneliti untuk melihat implikasi setiap butir soal terhadap reliabilitas keseluruhan, sehingga faktor-faktor yang merusak kualitas soal dapat segera dieliminasi untuk menjaga mutu asesmen akhir semester.

Temuan penelitian ini menunjukkan bahwa instrumen Asesmen Akhir Semester (AAS) PAI kelas VII di SMP Pasundan 1 Bandung memiliki kualitas psikometrik yang sangat unggul. Nilai reliabilitas sebesar 0,978 mengindikasikan bahwa instrumen ini memiliki konsistensi internal yang sangat kuat. Penggunaan perangkat lunak statistik IBM SPSS 26 terbukti memberikan akurasi yang lebih tinggi dalam memvalidasi instrumen dibandingkan metode manual, yang sejalan dengan pandangan bahwa integrasi teknologi dalam evaluasi pendidikan mampu meminimalisir kesalahan manusia (*human error*) serta mempercepat identifikasi butir soal yang tidak efektif. Hasil daya pembeda yang dominan pada kategori "Sangat Baik" (83,3%) mengonfirmasi teori bahwa konstruksi soal yang selaras dengan tujuan pembelajaran esensial akan menghasilkan polarisasi skor yang tajam antara kelompok peserta didik atas dan bawah. Fenomena ini serupa dengan temuan literatur yang menyatakan bahwa instrumen terstruktur berfungsi sebagai alat ukur kognitif objektif. Namun, terdapat anomali pada butir soal nomor 12, 19, dan 23 yang memiliki indeks daya pembeda rendah hingga negatif; hal ini menunjukkan adanya *ceiling effect* di mana soal terlalu mudah sehingga kehilangan fungsi diferensiasinya. Berdasarkan analisis kualitatif terhadap butir soal yang masuk kategori "Jelek", ditemukan bahwa tingkat kesukaran yang terlalu rendah menjadi faktor utama hilangnya daya diskriminasi soal. Sebagai contoh, pada butir soal nomor 12 yang berbunyi:

"Pengertian zikir jika ditinjau dari segi bahasa adalah... (a) Berdoa kepada Allah Swt.; (b) Mengingat Allah Swt.; (c) Memuji Allah Swt.; (d) Mengucapkan nama Allah Swt."

Soal tersebut hanya mengukur ranah kognitif tingkat rendah (C1-Mengingat) yang bersifat hafalan definisi. Pilihan jawaban (*distractor*) yang tersedia cenderung memiliki kedekatan makna yang sangat tinggi namun sangat umum, sehingga hampir seluruh

peserta didik baik dari kelompok atas maupun kelompok bawah dapat dengan mudah menebak jawaban yang benar (opsi B). Kondisi ini menciptakan *ceiling effect*, di mana instrumen gagal memetakan perbedaan kemampuan antara peserta didik yang memiliki pemahaman mendalam dengan mereka yang hanya mengandalkan hafalan superfisial. Temuan mengenai adanya butir soal yang memiliki tingkat kesukaran terlalu rendah atau 'sangat mudah' mengindikasikan bahwa soal tersebut kurang layak dalam memberikan tantangan kognitif yang memadai bagi peserta didik. Hal ini sejalan dengan perlunya variasi tingkat kesulitan agar instrumen mampu memetakan kemampuan siswa secara proporsional antara kategori sedang, sulit, maupun mudah (Komala, Chandra, & Ubaidillah, 2021).

Guna menindaklanjuti temuan soal yang tidak efektif, diperlukan upaya rekonstruksi butir soal dengan meningkatkan level kognitif dari sekadar hafalan (*Low Order Thinking*) menuju aspek penalaran yang lebih kompleks. Sebagai contoh, perbaikan pada butir soal nomor 12 dapat diarahkan untuk menguji pemahaman konsep secara kontekstual sebagai berikut:

"Dalam kehidupan sehari-hari, seseorang sering kali mengucapkan kalimat '*Subhanallah*' saat melihat keindahan alam, namun hatinya tidak merasa tunduk kepada keagungan pencipta. Berdasarkan hakikat zikir secara bahasa, mengapa tindakan tersebut belum memenuhi pengertian zikir yang sempurna? (a) Karena zikir harus diucapkan dengan suara keras; (b) Karena zikir secara bahasa berarti 'mengingat' yang melibatkan kesadaran hati, bukan sekadar lisan; (c) Karena zikir hanya boleh dilakukan di dalam masjid; (d) Karena zikir secara bahasa berarti memuji dengan alasan tertentu."

Dengan bentuk soal rekonstruksi di atas, peserta didik dituntut untuk melakukan analisis hubungan antara teori (bahasa) dengan praktik (perilaku), sehingga daya pembedanya akan meningkat secara signifikan. Secara teoretis, variasi indeks daya pembeda ini dipengaruhi oleh beberapa faktor determinan yang saling berkaitan. Variasi indeks daya pembeda ini dipengaruhi oleh beberapa faktor signifikan. Pertama, heterogenitas kemampuan peserta didik dari tiga kelas paralel (VII A, VII B, dan VII C) memungkinkan pemisahan skor yang kontras. Dalam teori evaluasi, semakin beragam kemampuan sampel, semakin tinggi peluang butir soal menunjukkan daya diskriminasinya. Kedua, efektivitas pengecoh (*distractor*) pada 25 butir soal berkategori sangat baik berhasil memerangkap peserta didik kelompok bawah, sementara pada soal nomor 12, 19, dan 23, pengecoh gagal berfungsi karena tingkat kesukaran yang terlalu rendah.

Hal ini membuktikan bahwa kualitas teknis konstruksi pilihan jawaban sangat memengaruhi tajam atau tumpulnya daya beda sebuah soal (Marengke, 2019). Ketiga, nilai negatif pada butir 23 (-0,025) mengindikasikan potensi ambiguitas redaksi yang menyebabkan peserta didik kelompok atas terkecoh, sementara kelompok bawah menjawab benar karena faktor kebetulan. Faktor bahasa dan kejernihan kalimat merupakan variabel krusial; ambiguitas menyebabkan peserta didik kelompok atas (berkemampuan tinggi) cenderung berpikir terlalu kompleks atau terkecoh, sementara kelompok bawah menjawab benar hanya karena faktor kebetulan (*guessing*). Oleh karena itu, daya pembeda tidak hanya bergantung pada materi, tetapi juga pada teknis penulisan soal yang bebas dari bias bahasa.

Implikasi praktis dari temuan ini menekankan pentingnya bagi pendidik PAI di SMP Pasundan 1 Bandung untuk melakukan kurasi bank soal berbasis data secara berkala guna menjamin keadilan asesmen. Secara teoretis, hasil ini memperkuat argumen bahwa validitas isi harus didukung oleh validitas empiris melalui analisis statistik untuk menjaga mutu evaluasi pembelajaran. Keterbatasan penelitian ini terletak pada cakupan sampel yang terbatas pada satu institusi dan fokus pada data kuantitatif objektif tanpa melibatkan data kualitatif seperti wawancara untuk mendalami miskonsepsi siswa. Oleh karena itu, penelitian masa depan disarankan untuk mengeksplorasi penggunaan kecerdasan buatan (AI) dalam memprediksi tingkat kesukaran soal serta melakukan analisis kualitatif terhadap alasan jawaban peserta didik untuk mendapatkan pemahaman yang lebih komprehensif.

Kesimpulan

Kesimpulan dari penelitian ini menunjukkan bahwa penggunaan teknologi analisis data dalam evaluasi Pendidikan Agama Islam (PAI) menjadi unsur penting untuk menjamin objektivitas dan keadilan penilaian. Temuan utama dalam riset ini membuktikan bahwa sebagian besar butir soal yang disusun memiliki ketajaman yang sangat baik dalam membedakan kemampuan peserta didik. Hal ini memberikan bukti bahwa penyusunan soal yang sistematis dan selaras dengan tujuan pembelajaran dapat meminimalisir kesalahan pengukuran, sehingga hasil belajar siswa dapat terpotret secara lebih akurat. Secara teoretis, hasil penelitian ini memberikan pemahaman tentang pentingnya melakukan uji statistik sebagai pendamping pemeriksaan materi dalam pengembangan soal-soal PAI. Secara praktis, guru-guru di SMP Pasundan 1 Bandung disarankan untuk tidak hanya mengandalkan perkiraan saat menyusun soal, tetapi mulai menggunakan hasil analisis data untuk memperbaiki bank soal secara berkelanjutan. Munculnya beberapa butir soal yang tidak efektif dalam pengujian ini menjadi pengingat bahwa kejelasan bahasa dan ketepatan tingkat kesulitan soal adalah hal yang mutlak diperhatikan.

Daftar Pustaka

- Alwi, I. (2015). Pengaruh Jumlah Alternatif Jawaban Tes Obyektif Bentuk Pilihan Ganda Terhadap Reliabilitas, Tingkat Kesukaran Dan Daya Pembeda. *Jurnal Ilmiah Faktor Exacta*, 3(2), 184–193. Retrieved from https://journal.lppmunindra.ac.id/index.php/Faktor_Exacta/article/view/19/18
- Azzahra, N. S., Sumarni, S., & Putranta, H. (2024). Analisis Validitas dan Realibilitas Kualitas Soal Pilihan Ganda Asesmen Kompetensi Minimum (AKM) Mata Pelajaran Pendidikan Agama Islam Menggunakan *QuranicEdu: Journal of ...*, 4(1), 85–94. Retrieved from <https://jurnalannur.ac.id/index.php/QuranicEdu/article/view/681>
- Christy Matius, A., & Gunawan, W. (2022). Validitas dan Reliabilitas Perceived Future Employability Scale untuk Dewasa Muda. *Psikologika: Jurnal Pemikiran Dan Penelitian Psikologi*, 27(1), 23–46. <https://doi.org/10.20885/psikologika.vol27.iss1.art3>
- Fernanda, J. W., & Hidayah, N. (2020). *Classical Test Theory dan Rasch Model*. 2(1), 49–60.
- Keislaman, J., Sma, D. I., & Padang, N. (2023). *Analisis Butir Soal Tipe Multiple Choice Ujian Akhir Semester Ganjil Kelas X Tahun 2022/2023 Mata Pelajaran*

- Pendidikan Agama Islam Di Sma Negeri 1 Padang*. 5, 981–997.
- Kiki Fajariani, & Erlia Hanum. (2025). Evaluasi Kualitas Instrumen Tes Pilihan Ganda pada Mata Pelajaran IPA Kelas V Sekolah Dasar Berdasarkan Analisis Butir Soal. *Jurnal Pengabdian Masyarakat Dan Riset Pendidikan*, 4(2), 9224–9229. <https://doi.org/10.31004/jerkin.v4i2.2984>
- Komala, E., Chandra, E., & Ubaidillah, M. (2021). Analisis Validitas Butir Dan Ketepatan Target Ukur Instrumen. *Jurnal Pendidikan Biologi*, 12, 187–201.
- Marengke, M. (2019). Persepsi Konseptualitas Perakitan Butir Soal. *Foramadiahi: Jurnal Kajian Pendidikan Dan Keislaman*, 10(2), 190. <https://doi.org/10.46339/foramadiahi.v10i2.122>
- Qadir, Huda, H. (2024). Analisis Butir Tes: Tingkat Kesukaran, Daya Pembeda, Eektivitas Pengecoh. *Cv.Eureka Media Aksara*, 3(3), 105–134.
- Rahayu, D. S., & Putranta, H. (2025). Analisis Kualitas Butir Soal AKM Literasi dan Numerasi Fiqih dengan Model Rasch Berbasis Software Quest. *Jurnal Pendidikan Agama Islam*, 04, 121–130.
- Saputra, A., & Mayurida. (2025). Pengembangan Instrumen Evaluasi (Analisis Pengembangan Penilaian Satuan Pendidikan MTs/SMP). *Jurnal Pendidikan Dan Keagamaan*, 2(4), 313–323.
- Tarsono. (2024). Peran Guru PAI dalam Pengembangan Alat Evaluasi yang Mendukung Proses Pembelajaran yang Berkualitas. *Keislaman Dan Ilmu Pendidikan*, 6, Nomor 1, 318–325.