KESESUAIAN *BLUEPRINT* DAN *ITEM ANALYSIS* SEBAGAI INDIKATOR VALIDITAS INSTRUMEN UJIAN MCQ

Nia Ayu Saraswati[1*], Miranti Dwi Hartanti[2], Ni Made Elva Mayasari[3], Rury Tiara Oktariza[4], Ratih Pratiwi[5]

[1]Departemen Pendidikan Kedokteran Fakultas Kedokteran Universitas Muhammadiyah Palembang
[2]Unit Pendidikan Kedokteran, Departemen Farmakologi Fakultas Kedokteran Universitas Muhammadiyah Palembang
[3]Unit Pendidikan Kedokteran, Departemen Kardiologi Fakultas Kedokteran Universitas Muhammadiyah Palembang
[4]Unit Pendidikan Kedokteran, Departemen Histologi Fakultas Kedokteran Universitas Muhammadiyah Palembang
[5]Unit Pendidikan Kedokteran, Departemen Obstetri dan Ginekologi Fakultas Kedokteran Universitas Muhammadiyah Palembang
*Email : nia_ayusaraswati@um-palembang.ac.id

ABSTRAK

Instrumen ujian pilihan ganda (*Multiple Choice Questions*/MCQ) merupakan metode penilaian ranah kognitif yang paling banyak digunakan dalam pendidikan kedokteran. Kualitas MCQ sangat dipengaruhi oleh kesesuaian antara perencanaan ujian melalui *blueprint* dan performa empiris butir soal yang dievaluasi melalui *item analysis*. Penelitian ini bertujuan untuk menganalisis hubungan antara *blueprint* ujian dan hasil *item analysis* sebagai salah satu indikator validitas instrumen MCQ. Penelitian ini menggunakan desain kuantitatif observasional deskriptif dengan pendekatan retrospektif. Data diperoleh dari dokumen *blueprint* ujian dan hasil ujian MCQ blok terakhir pada Program Studi Kedokteran Fakultas Kedokteran Universitas Muhammadiyah Palembang. Sebanyak 150 butir soal dianalisis menggunakan perangkat lunak statistik untuk memperoleh indeks kesukaran, daya beda, dan reliabilitas internal instrumen yang diukur menggunakan koefisien *Cronbach's alpha*. Perencanaan dalam *blueprint* soal dibandingkan dengan hasil *item analysis* setelah ujian. Hasil penelitian pada analisis setelah ujian menunjukkan bahwa sebagian besar butir soal berada pada kategori indeks kesukaran sedang (47,33%), dengan proporsi soal sukar dan mudah yang relatif seimbang. Analisis daya beda menunjukkan bahwa 70,67% butir soal memiliki daya beda baik (≥0,30), meskipun masih terdapat 29,33% butir soal dengan daya beda cukup hingga rendah. Reliabilitas internal instrumen ujian menunjukkan nilai *Cronbach's alpha* sebesar 0,91 yang tergolong sangat baik. Uji *Chi-Square* menunjukkan tidak terdapat perbedaan distribusi tingkat kesukaran soal antara *blueprint* dengan hasil *item analysis* setelah ujian (p=0.962). Namun, sebagian besar soal dalam *blueprint* mengalami pergeseran kategori tingkat kesukaran menjadi sedang dalam indeks kesukaran soal. Temuan ini menunjukkan bahwa meskipun validitas instrumen MCQ ini sudah baik dan distribusi tingkat kesukaran secara umum sesuai dengan *blueprint*, penentuan tingkat kesukaran oleh pembuat soal belum sepenuhnya mencerminkan indeks kesukaran yang aktual. Evaluasi dan perbaikan berkelanjutan terhadap butir soal tetap diperlukan untuk meningkatkan kualitas instrumen penilaian.

Kata kunci: blueprint, item analysis, MCQ, validitas instrumen, reliabilitas internal

ABSTRACT

Multiple Choice Questions (MCQs) are widely used as cognitive assessment tools in medical education due to their efficiency and objectivity. The quality of MCQs is strongly influenced by the alignment between examination planning through blueprinting and the empirical performance of test items evaluated using item analysis. This study aimed to examine the alignment between the examination blueprint and item analysis results as one of indicator of MCQ instrument validity. This study employed a descriptive observational quantitative design with a retrospective approach. Data were obtained from examination blueprint documents and MCQ test results from the most recent block examination conducted at the Medical Study Program, Faculty of Medicine, Universitas Muhammadiyah Palembang. A total of 150 MCQ items were analyzed using statistical software to calculate difficulty index, discrimination index, and internal reliability measured by Cronbach's alpha coefficient. The planned difficulty categories in the blueprint were compared with the item analysis results after the examination. The results demonstrated that most items were classified within the moderate difficulty category (47.33%), with relatively balanced proportions of difficult and easy items. Item discrimination analysis revealed that 70.67% of items showed good discrimination (≥0.30), although 29.33% of items demonstrated fair to poor discrimination. The internal reliability of the examination instrument was very high, with a Cronbach's alpha value of 0.91. Chi-square analysis showed no significant difference in the distribution of item analysis categories between

the blueprint and post-examination item analysis results. However, most items in the blueprint experienced a shift toward to the moderate category in the actual item difficulty index. These findings indicate that although the MCQ instrument demonstrated good psychometric quality and the overall distribution of difficulty levels was generally consistent with the blueprint, the determination of item difficulty by item writers did not fully reflect the actual difficulty index. Continuous evaluation and revision of items are recommended to enhance the overall quality of assessment instruments.

Keywords: blueprint, item analysis, MCQ, instrument validity, internal reliability

Submited : 16-04-2026 Revision : 25-05-2026 Accepted: 25-05-2026

Pendahuluan

Instrumen ujian pilihan ganda (*Multiple Choice Questions*/MCQ) merupakan metode penilaian ranah kognitif yang paling luas diterapkan dalam pendidikan kedokteran karena keunggulannya dalam hal efisiensi pelaksanaan dan objektivitas penilaian. MCQ merupakan instrumen yang efektif dalam mengukur berbagai level kognitif jika disusun berdasarkan prinsip konstruksi soal yang baik dan blueprint yang sistematis. Penggunaan blueprint dalam penilaian telah direkomendasikan secara luas untuk menjamin validitas isi (*content validity*) suatu instrumen evaluasi pendidikan.[1] MCQ mampu mengukur berbagai tingkat kemampuan kognitif berdasarkan Taksonomi Bloom, mulai dari penguasaan pengetahuan dasar hingga kemampuan aplikasi dan analisis. Meskipun demikian, mutu MCQ sangat ditentukan oleh kualitas konstruksi soal serta kesesuaiannya dengan tujuan pembelajaran dan *blueprint* ujian.[2,3]

Blueprint ujian berfungsi sebagai kerangka perencanaan yang memetakan capaian pembelajaran terhadap jumlah soal, tingkat kognitif, dan distribusi konten yang akan diujikan. *Blueprint* yang disusun secara sistematis memastikan bahwa kompetensi yang dinilai tercakup secara proporsional dalam instrumen ujian. Sementara itu, *item analysis* merupakan metode evaluasi pasca-ujian yang digunakan untuk menilai kualitas masing-masing butir soal, meliputi indeks kesukaran, daya pembeda, dan reliabilitas internal. Teknik ini telah banyak dimanfaatkan dalam penelitian terkait MCQ untuk menilai kualitas soal serta memberikan dasar rekomendasi perbaikan berkelanjutan terhadap bank soal.[4,5]

Dalam beberapa tahun terakhir, pendekatan berbasis bukti (*evidence-based assessment*) semakin menyoroti pentingnya penggabungan antara perencanaan instrumen dan evaluasi berbasis data dalam pendidikan kedokteran. *Blueprint* tidak hanya berfungsi sebagai alat untuk merancang distribusi materi, tetapi juga sebagai mekanisme untuk memastikan hubungan antara capaian pembelajaran, strategi pengajaran, dan metode penilaian. Penelitian terbaru menunjukkan bahwa penggunaan *blueprint* yang terstruktur secara signifikan memperkuat validitas isi dan mengurangi potensi bias dalam penyusunan soal MCQ, terutama dalam kurikulum berbasis kompetensi.[6,7]

Selain itu, metode psikometri modern menegaskan bahwa kualitas instrumen penilaian tidak hanya bergantung pada validitas isi, tetapi juga harus didukung oleh bukti validitas internal melalui analisis statistik. *Item analysis* memainkan peran penting dalam proses evaluasi karena mampu memberikan data kuantitatif tentang performa setiap item, termasuk efektivitas distraktor dan konsistensi respons peserta. Penelitian terbaru menunjukkan bahwa mengintegrasikan *blueprint* dan *item analysis* dapat secara berkelanjutan meningkatkan kualitas bank soal serta mendukung pengambilan keputusan akademik yang lebih tepat.[8,9]

Lebih jauh lagi, dalam pendidikan kedokteran modern yang membutuhkan akuntabilitas tinggi, kualitas alat penilaian sangat berpengaruh terhadap mutu lulusan. Instrumen yang tidak valid dapat mengakibatkan pengambilan keputusan yang keliru terkait kompetensi mahasiswa.[10] Berbagai penelitian sebelumnya menegaskan pentingnya evaluasi kualitas MCQ, mengingat instrumen yang tidak dianalisis secara empiris dapat tampak berfungsi secara administratif tetapi tidak mampu membedakan tingkat kemampuan peserta didik secara akurat. Studi lain juga menekankan perlunya evaluasi butir soal secara berkesinambungan guna meningkatkan efektivitas dan ketepatan sistem penilaian.[11]

Hingga saat ini FK Universitas Muhammadiyah Palembang telah menyelenggarakan ujian MCQ secara sumatif dalam setiap ujian akhir blok. Proses penilaian ini dimulai dengan perencanaan pembuatan *blueprint* oleh penulis soal pada

setiap instrumen MCQ mencakup pemetaan kompetensi yang dinilai, level kognitif taksonomi Bloom, dan proporsi tingkat kesukaran soal. Soal MCQ yang telah tersusun juga ditentukan nilai batas lulusnya dengan melakukan *standard setting*. Kemudian instrumen MCQ diujikan kepada mahasiswa, dan dilakukan *item analysis post-examination*. Namun hingga kini, belum pernah dilakukan analisis kesesuaian *blueprint* dengan hasil *item analysis*.

Tujuan dari penelitian ini adalah mengkaji kesesuaian antara perencanaan *blueprint* ujian dan temuan *item analysis* sebagai salah satu dasar penentuan validitas instrumen ujian MCQ.

Metode Penelitian

Penelitian ini menerapkan desain kuantitatif observasional deskriptif dengan pendekatan retrospektif. Penelitian dilaksanakan pada Program Studi Kedokteran di Fakultas Kedokteran Universitas Muhammadiyah Palembang. Data penelitian bersumber dari ujian pilihan ganda (*Multiple Choice Questions*/MCQ) blok terakhir yang telah dilaksanakan dan terdokumentasi secara lengkap dalam arsip institusi. Pendekatan retrospektif digunakan dengan memanfaatkan data autentik hasil ujian yang sebelumnya telah digunakan dalam proses evaluasi akademik.

Subjek penelitian berupa dokumen *blueprint* ujian serta data hasil ujian MCQ dari blok yang dianalisis. *Blueprint* ujian memuat informasi mengenai distribusi jumlah soal, domain kompetensi, tingkat kognitif berdasarkan Taksonomi Bloom, dan proporsi tingkat kesukaran soal. Data hasil ujian selanjutnya dianalisis menggunakan perangkat lunak statistik untuk memperoleh parameter *item analysis*, yang meliputi indeks kesukaran (*difficulty index*), daya pembeda (*discrimination index*), serta reliabilitas internal instrumen ujian yang diukur menggunakan koefisien *Cronbach's alpha*.

Instrumen analisis dalam penelitian ini mengacu pada kriteria standar evaluasi butir soal. Indeks kesukaran dikategorikan optimal apabila nilai p berada pada rentang 0,30–0,70. Daya pembeda dinilai baik apabila memiliki nilai ≥0,30, yang menunjukkan kemampuan butir soal dalam membedakan mahasiswa dengan tingkat kemampuan tinggi dan rendah. Reliabilitas internal instrumen ujian dinyatakan memadai apabila nilai *Cronbach's alpha* mencapai ≥0,70.

Tahapan analisis diawali dengan pemetaan setiap butir soal MCQ terhadap *blueprint* ujian berdasarkan tingkat kesukaran yang telah direncanakan. Selanjutnya dilakukan perhitungan analisis menggunakan perangkat lunak statistik. Kriteria indeks kesukaran dan daya beda yang digunakan dalam penelitian ini mengacu pada standar evaluasi pendidikan yang telah banyak digunakan dalam penelitian psikometri dan pendidikan kedokteran.[12] Hasil *item analysis* kemudian dibandingkan dengan distribusi soal dalam *blueprint* untuk menilai tingkat kesesuaian kesukaran soal yang dibuat penulis soal dengan tingkat kesulitan soal aktual setelah ujian.

Hasil Penelitian

Analisis kualitas butir soal dilakukan dengan membandingkan kesesuaian *blueprint* ujian dan hasil *item analysis*. Setiap butir soal dipetakan ke dalam *blueprint* berdasarkan domain kompetensi dan level kognitif yang telah ditetapkan. Selanjutnya dilakukan perhitungan indeks kesukaran, daya beda, dan reliabilitas internal menggunakan perangkat lunak statistik. Indeks kesukaran diklasifikasikan menjadi sukar (<0,30), sedang (0,30–0,70), dan mudah (>0,70). Distribusi indeks kesukaran menunjukkan bahwa sebagian besar butir soal berada pada kategori sedang, yang mengindikasikan tingkat kesukaran yang proporsional sesuai dengan tujuan

pembelajaran. Hasil ini mendukung kesesuaian antara perencanaan ujian dalam blueprint dan performa empiris butir soal berdasarkan item analysis.

Tabel 1. Distribusi Indeks Kesukaran Butir Soal MCQ

Kategori Indeks Kesukaran	Jumlah Soal	Persentase (%)
Sukar (<0,30)	36	24,00
Sedang (0,30-0,70)	71	47,33
Mudah (>0,70)	43	28,67
Total	**150**	**100,00**

Tabel 1 menunjukkan distribusi kesukaran butir soal MCQ. Sebagian besar butir soal menunjukkan kategori sedang, sedangkan proporsi soal sukar dan mudah relatif seimbang .

Tabel 2. Distribusi Daya Beda Butir Soal MCQ

Kategori Daya Beda	Nilai Daya Beda	Jumlah Soal	Persentase (%)
Baik	≥0,30	106	70,67
Cukup	0,20-0,29	33	22,00
Rendah	<0,20	11	7,33
Total		**150**	**100,00**

Berdasarkan kriteria *corrected item-total correlation*, butir soal dengan nilai daya beda ≥0,30 dikategorikan baik. Hasil analisis menunjukkan 70,67% butir soal memiliki daya beda baik, yang menunjukkan kemampuan *item* dalam membedakan mahasiswa berkemampuan tinggi dan rendah secara memadai. Namun, masih terdapat 29,33% butir soal dengan daya beda cukup hingga rendah. Butir soal dengan daya beda rendah berpotensi tidak memberikan kontribusi optimal terhadap validitas internal instrumen dan perlu direvisi atau dieliminasi pada penggunaan selanjutnya. Temuan ini mengindikasikan bahwa meskipun *blueprint* telah dirancang dengan baik, kualitas diskriminatif setiap butir soal masih bervariasi dan memerlukan evaluasi lanjutan.

Meskipun sebagian besar butir soal telah sesuai dengan *blueprint* dari segi domain kompetensi dan level kognitif, tidak seluruh butir menunjukkan daya beda yang optimal, sehingga menegaskan pentingnya evaluasi empiris melalui *item analysis* sebagai pelengkap validitas isi yang dijamin oleh *blueprint*.

Tabel 3. Reliabilitas Internal Instrumen Ujian MCQ

Parameter	Nilai
Jumlah Butir Soal	150
Jumlah Peserta Ujian	151
Cronbach's alpha	0,91
Kategori Reliabilitas	**Sangat Baik**

Reliabilitas internal instrumen ujian MCQ dianalisis menggunakan koefisien *Cronbach's alpha*. Hasil analisis menunjukkan bahwa nilai reliabilitas internal ujian berada pada kategori sangat baik. Nilai *Cronbach's alpha* yang diperoleh memenuhi kriteria minimal reliabilitas yang direkomendasikan untuk instrumen penilaian pendidikan, sehingga menunjukkan konsistensi internal antarbutir soal yang memadai.

Nilai *Cronbach's alpha* sebesar 0,91 menunjukkan bahwa instrumen ujian memiliki konsistensi internal yang sangat baik. Hal ini mengindikasikan bahwa butir soal dalam ujian tersebut secara keseluruhan mengukur konstruk yang relatif homogen dan dapat digunakan secara andal untuk menilai capaian pembelajaran mahasiswa.

Reliabilitas internal yang baik menunjukkan bahwa secara keseluruhan instrumen ujian memiliki konsistensi yang memadai, meskipun hasil *item analysis* menunjukkan adanya beberapa butir soal dengan daya beda rendah yang memerlukan perbaikan lebih lanjut.

Penulis soal telah menentukan proporsi tingkat kesulitan soal instrumen MCQ dalam *blueprint*. Distribusi tingkat kesukaran soal hasil *item analysis* pasca-ujian dengan *blueprint* terlihat dalam Tabel 4.

Tabel 4. Distribusi Kesesuaian Tingkat Kesulitan Soal Blueprint dan Hasil *Item analysis*

		Hasil Item Analisis			
		Mudah n(%)	**Sedang n(%)**	**Sulit n(%)**	**Total n(%)**
Blueprint	**Mudah n(%)**	25 (16,7)	38 (25,3)	18 (12,0)	81 (54,0)
	Sedang n(%)	12 (8,0)	23 (15,3)	12 (8,0)	47 (31,3)
	Sulit n(%)	6 (4,0)	10 (6,7)	6 (4,0)	22 (14,7)
Total		**43 (28,7)**	**71 (47,3)**	**36 (24,0)**	**150 (100)**

Pearson Chi-square: $X^2=0.608$, $df=4$, $p=0.962$

Dalam pada Tabel 4 terlihat banyak soal dalam kategori mudah dalam *blueprint* yang dibuat penulis soal ternyata menjadi kategori sedang dalam hasil *item analysis.* Bahkan soal yang diprediksi sulit oleh pembuat soal, sebagian besar masuk dalam kategori sedang. Tingkat kesukaran soal kategori sedang (47,3%) mendominasi hampir semua kelompok kategori kesukaran soal pada *blueprint*. Namun, hasil uji *Chi-square* menunjukkan $p>0.05$, maka tidak terdapat perbedaan signifikan antara kategori kesukaran soal yang direncanakan pembuat soal pada *blueprint* dengan hasil *item analysis* setelah ujian.

Pembahasan

Hasil analisis menunjukkan bahwa sebagian besar butir soal MCQ berada pada kategori tingkat kesukaran sedang (0,30–0,70), sementara proporsi soal yang termasuk kategori sukar dan mudah relatif lebih kecil. Distribusi ini mengindikasikan bahwa instrumen ujian memiliki tingkat kesulitan yang proporsional dengan tujuan pembelajaran dan tidak terlalu berat atau terlalu ringan bagi peserta didik. Klasifikasi tingkat kesukaran soal seperti yang digunakan dalam penelitian ini adalah metodologi yang umum diterapkan dalam *item analysis*. Indeks kesukaran (*difficulty index*) mengukur persentase responden yang menjawab item dengan benar, dan nilai dalam kisaran menengah secara empiris dianggap menunjukkan keseimbangan yang baik antara soal yang terlalu mudah dan yang terlalu sulit. Studi sebelumnya pada kajian *item analysis* juga menemukan pola yang serupa: mayoritas *item* dalam ujian kesehatan memiliki tingkat kesukaran dalam rentang yang dapat diterima, dan hanya sebagian kecil item yang tergolong terlalu mudah atau terlalu sukar. Hal ini menunjukkan bahwa instrumen secara umum mampu mengukur kemampuan secara optimal tanpa bias distribusi ekstrem dalam kesukaran soal.[13]

Selain itu, kondisi di mana sebagian besar soal berada pada kategori sedang sering dikaitkan dengan kemampuan soal dalam membedakan level kemampuan mahasiswa

secara efektif. Literatur *item analysis* menyatakan bahwa *item* dengan tingkat kesukaran menengah cenderung memiliki daya beda yang lebih kuat, sehingga lebih efektif dalam memisahkan peserta dengan kemampuan tinggi dan rendah, dibandingkan dengan item yang terlalu mudah atau terlalu sukar. Dengan demikian, distribusi seperti yang ditunjukkan pada Tabel 1. tidak hanya mencerminkan kesesuaian dengan *blueprint*, namun juga mendukung validitas internal instrumen di mana kualitas psikometrik soal tetap terjaga. Temuan ini konsisten dengan studi yang menekankan pentingnya kesesuaian antara perencanaan instrumen penilaian dan performa empiris item dalam ujian MCQ. Penelitian sebelumnya menunjukkan bahwa penggunaan *item analysis* sebagai evaluasi kuantitatif memberikan bukti penting tentang apakah butir soal telah memenuhi kriteria kualitas, termasuk distribusi tingkat kesulitan yang seimbang, yang pada gilirannya memperkuat validitas instrumen secara keseluruhan.[14]

Meskipun distribusi indeks kesukaran ini mendukung kesesuaian antara *blueprint* dengan performa soal, perlu diingat bahwa tingkat kesukaran hanyalah satu dari beberapa indikator kualitas *item*. Evaluasi lanjutan dengan melihat daya beda dan fungsi pengecoh (*distractor*) juga penting untuk memastikan bahwa soal tidak hanya memiliki tingkat kesukaran yang tepat tetapi juga mampu membedakan peserta secara optimal dan meminimalkan jawaban tebak. Integrasi ketiga aspek ini (kesukaran, daya beda, dan kualitas pengecoh) merupakan praktik terbaik dalam item analysis untuk menjamin bahwa instrumen ujian benar-benar valid dan reliabel secara psikometrik.[11]

Temuan bahwa sebagian besar soal berada dalam kategori kesukaran sedang sesuai dengan prinsip konstruksi tes klasik. Prinsip ini menyatakan bahwa distribusi tingkat kesukaran yang optimal akan meningkatkan sensitivitas instrumen dalam mendeteksi variasi kemampuan peserta. Studi terbaru menunjukkan bahwa soal dengan tingkat kesukaran moderat memberikan kontribusi terbesar terhadap peningkatan reliabilitas tes dibandingkan dengan soal yang ekstrem. Hal ini disebabkan oleh soal moderat yang memiliki varians skor yang lebih tinggi, sehingga meningkatkan daya ukur instrumen.[15,16]

Selain itu, distribusi tingkat kesukaran yang seimbang menunjukkan bahwa proses perencanaan telah memperhitungkan keseimbangan tingkat kognitif. Dalam kurikulum berbasis kompetensi, penting untuk menjaga keseimbangan antara soal level rendah dan tinggi agar evaluasi tidak hanya mengukur kemampuan mengingat, tetapi juga kemampuan analisis dan penerapan klinis.[17]

Analisis daya beda butir soal merupakan indikator penting dalam *item analysis* yang mencerminkan kemampuan setiap soal untuk membedakan antara peserta ujian dengan kemampuan tinggi dan rendah. Indeks diskriminasi (*discrimination index*) umumnya dihitung berdasarkan korelasi *item*-total atau perbandingan performa kelompok atas dan bawah, dan nilai ≥0,30 sering dianggap menunjukkan kemampuan diskriminatif yang baik. Hal ini sejalan dengan praktik dalam literatur pendidikan dan psikometri di mana soal dengan nilai daya beda ≥0,30 dianggap efektif dalam membedakan peserta ujian berdasarkan tingkat pemahaman mereka terhadap isi materi yang diuji.[18]

Hasil yang ditunjukkan pada Tabel 2. mengindikasikan bahwa 70,67 % butir soal memiliki daya beda baik (≥0,30), yang memperkuat temuan bahwa mayoritas soal mampu secara memadai membedakan performa mahasiswa berkemampuan tinggi dan rendah. Hal ini mencerminkan kualitas instrumen yang relatif kuat dari aspek diskriminatif, yang merupakan salah satu komponen penting dalam menilai validitas internal ujian. Temuan serupa dilaporkan dalam studi *item analysis* di pendidikan kesehatan, di mana proporsi soal dengan indeks diskriminasi baik dominan (sekitar 60–

80 %) menunjukkan bahwa mayoritas butir soal tersebut dapat digunakan secara efektif untuk mengevaluasi kemampuan peserta ujian.[13]

Namun demikian, terdapat 29,33 % butir soal dengan daya beda cukup hingga rendah (nilai < 0,30). Item-item ini berpotensi kurang efektif dalam membedakan peserta ujian yang berkemampuan tinggi dan rendah, sehingga memerlukan evaluasi dan perbaikan lebih lanjut. Literatur metodologi pengukuran pendidikan menyatakan bahwa soal dengan daya beda rendah sering disebabkan oleh konstruk item yang kurang tajam, distraktor yang tidak berfungsi optimal, atau tingkat kesukaran soal yang terlalu ekstrem (terlalu mudah atau terlalu sukar), yang mengurangi kemampuan item tersebut untuk mendeteksi perbedaan kemampuan secara akurat.[19]

Fenomena adanya variasi dalam nilai daya beda meskipun *blueprint* telah dirancang dengan baik menggambarkan bahwa kesesuaian *blueprint* terhadap perencanaan konten tidak otomatis menjamin kualitas diskriminatif setiap *item* soal. *Blueprint* diperlukan untuk memastikan cakupan konten dan keseimbangan level kognitif, tetapi evaluasi empiris melalui *item analysis* tetap diperlukan untuk menilai performa aktual setiap *item*. Studi sebelumnya juga menekankan kombinasi antara perencanaan blueprint yang matang dan evaluasi item analysis sebagai pendekatan komprehensif dalam validasi instrumen ujian.[20]

Variasi daya beda yang ditemukan dalam penelitian ini juga berkaitan dengan kualitas distraktor pada setiap butir soal. Distraktor yang tidak efektif dapat menurunkan kemampuan diskriminatif item karena tidak mampu menarik respons dari peserta dengan kemampuan rendah. Penelitian terbaru menunjukkan bahwa efektivitas distraktor merupakan faktor utama dalam meningkatkan daya beda soal pilihan ganda. Item dengan distraktor yang seragam dan masuk akal cenderung memiliki nilai diskriminasi yang lebih tinggi.[21,22]

Selain itu, tingginya tingkat ketidaksesuaian pada beberapa item juga bisa disebabkan oleh ketidaksesuaian antara tingkat kognitif yang direncanakan dalam blueprint dan pelaksanaannya dalam soal. Ini menunjukkan bahwa meskipun *blueprint* telah disusun secara baik, proses pembuatan soal tetap memerlukan pelatihan khusus dan review dari rekan untuk memastikan kualitas soal tetap terjaga.[23]

Oleh karena itu, meskipun mayoritas butir soal menunjukkan daya beda yang baik, *item* dengan daya beda rendah perlu direvisi atau dihapus dari bank soal untuk meningkatkan kualitas evaluasi secara keseluruhan. Revisi dapat mencakup peninjauan ulang stem soal, kualitas distraktor, dan relevansi konten terhadap tujuan pembelajaran. Dengan demikian, integrasi antara blueprint dan item analysis akan memperkuat validitas internal dan konstruksi instrumen ujian MCQ.

Analisis reliabilitas internal menggunakan koefisien *Cronbach's alpha* merupakan salah satu pendekatan psikometrik yang paling umum digunakan untuk menilai konsistensi internal suatu instrumen penilaian, termasuk ujian pilihan ganda. *Cronbach's alpha* menghitung sejauh mana butir-butir soal dalam sebuah tes saling berkorelasi dan mengukur konstruk yang sama secara konsisten dalam satu kesatuan instrumen. Nilai koefisien ini berkisar antara 0 sampai 1, di mana nilai yang lebih tinggi menunjukkan tingkat konsistensi internal yang lebih baik antar butir dalam instrumen. Secara umum, nilai *Cronbach's alpha* ≥0,70 dianggap reliabel, nilai ≥0,80 dikategorikan baik, dan nilai ≥0,90 sering dipandang sangat baik hingga *excellent* dalam konteks pengukuran pendidikan dan psikometri. Hasil penelitian menunjukkan bahwa instrumen ujian MCQ memiliki nilai *Cronbach's alpha* sebesar 0,91, yang berada pada kategori sangat baik menurut pedoman umum penilaian reliabilitas. Nilai ini menunjukkan bahwa butir-butir soal dalam instrumen memiliki konsistensi internal yang tinggi, artinya

jawaban peserta terhadap berbagai *item* soal saling berkaitan secara konsisten dan diperkirakan mengukur konstruk kemampuan akademik yang sama secara stabil. Reliabilitas internal yang tinggi seperti ini umumnya dianggap menunjang keandalan skor tes dalam menilai kompetensi peserta ujian, sehingga keputusan akademik yang diambil berdasarkan skor tersebut dapat dipertanggungjawabkan secara ilmiah. Temuan serupa juga dilaporkan dalam studi lain yang mengevaluasi reliabilitas internal instrumen pendidikan, di mana nilai Cronbach's alpha di atas ambang minimal menunjukkan bahwa instrumen penilaian dapat menghasilkan hasil yang konsisten dan kurang dipengaruhi oleh error pengukuran. Studi ini menegaskan bahwa penerapan Cronbach's alpha sebagai ukuran internal consistency reliability merupakan praktik yang luas dipakai dalam penelitian pendidikan untuk memastikan bahwa skor-skor yang dihasilkan dari item-item tes dapat dipercaya secara statistik.[24] Meskipun koefisien reliabilitas internal yang tinggi mencerminkan stabilitas dan keterkaitan antaritem, penting untuk diingat bahwa reliabilitas internal tidak serta-merta menjamin bahwa semua butir soal memiliki kualitas psikometrik yang optimal secara individual. Reliabilitas yang tinggi tidak selalu menjamin validitas instrumen, namun merupakan prasyarat penting dalam pengukuran yang baik.[25] Dalam penelitian ini masih terdapat sejumlah item dengan daya beda rendah, yang berarti beberapa butir soal kurang efektif membedakan performa peserta ujian. Oleh karena itu, reliabilitas internal yang tinggi sebaiknya diinterpretasikan bersama dengan hasil *item analysis* lainnya seperti indeks kesukaran dan indeks diskriminasi (daya beda) untuk memberikan gambaran yang lebih komprehensif mengenai kualitas instrumen secara keseluruhan. Dengan demikian, reliabilitas internal yang sangat baik pada instrumen ujian MCQ ini mendukung klaim bahwa instrumen tersebut cukup konsisten untuk digunakan dalam menilai capaian pembelajaran mahasiswa.

Kesesuaian Tingkat kesulitan soal yang dirancang dalam *blueprint* oleh penulis soal cukup realistis karena tidak ada perbedaan bermakna dengan hasil item analysis setelah ujian, seperti yang tergambar dalam Tabel 4. Hal ini mencerminkan penilaian penulis soal relatif presisi dengan hasil performa mahasiswa dan kualitas soal. Namun, sebagian soal mengalami pergeseran kategori tingkat kesukaran terutama menuju kategori sedang saat dilakukan *item analysis* setelah ujian. Hal ini menunjukkan tidak selalu tingkat kesulitan yang dibuat penulis soal dapat menjadi prediktor kualitas psikometrik soal. Hasil tingkat kesulitan soal dalam *item analysis* sangat dipengaruhi performa mahasiswa dan kualitas soal yang mencakup kualitas stem, distraktor, dan ada atau tidaknya *clue* dalam soal. Soal-soal yang mengalami pergeseran kategori dapat menjadi bahan evaluasi berkelanjutan terhadap kualitas butir soal, yang akan memperkaya kualitas bank soal. Evaluasi kualitas butir soal ini diperlukan agar alat ukur tetap valid dan reflektif terhadap kemampuan peserta, khususnya ketika ada butir dengan karakteristik psikometrik yang kurang ideal.

Simpulan dan Saran

Penelitian ini menunjukkan bahwa hasil *item analysis* dan kesesuaian antara *blueprint* ujian dan hasil *item analysis* dapat digunakan sebagai indikator dalam menilai validitas instrumen ujian MCQ. Sebagian besar butir soal telah sesuai dengan *blueprint* dari segi domain kompetensi dan level kognitif, serta didukung oleh distribusi indeks kesukaran yang proporsional dan nilai reliabilitas internal yang sangat baik (*Cronbach's alpha* = 0,91). Temuan ini mengindikasikan bahwa instrumen ujian memiliki konsistensi internal yang tinggi dan secara umum mampu mengukur capaian pembelajaran mahasiswa secara andal, meskipun masih terdapat variasi kualitas pada tingkat butir soal individual, khususnya terkait daya beda.

Berdasarkan hasil tersebut, disarankan agar penggunaan *blueprint tetap* dipertahankan sebagai dasar perencanaan ujian, serta dilengkapi dengan pelaksanaan *item analysis* secara rutin setelah setiap ujian. Butir soal dengan daya beda rendah perlu dievaluasi dan direvisi untuk meningkatkan kualitas diskriminatif instrumen. Perubahan tingkat kategori kesukaran soal pada hasil item analysis dapat ditindaklanjuti sebagai dasar evaluasi terhadap penulis soal untuk dapat melakukan penilaian ulang terhadap soal tersebut. Penelitian selanjutnya diharapkan dapat mengembangkan evaluasi validitas instrumen secara lebih komprehensif, termasuk validitas konstruk atau validitas kriteria, serta melibatkan lebih banyak blok atau institusi pendidikan kedokteran guna memperkuat generalisasi temuan dan mendukung pengembangan sistem penilaian yang bermutu dan berkelanjutan.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Fakultas Kedokteran Universitas Muhammadiyah Palembang atas dukungan dan fasilitasi yang diberikan dalam pelaksanaan penelitian ini, khususnya dalam penyediaan data dan izin penggunaan dokumen akademik yang diperlukan. Apresiasi juga disampaikan kepada seluruh pihak yang telah berkontribusi secara langsung maupun tidak langsung sehingga penelitian ini dapat terlaksana dengan baik.

Daftar Pustaka

1. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003 Sep 27;37(9):830–7. doi:10.1046/j.1365-2923.2003.01594.x
2. Rodriguez MC. Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. Educational Measurement: Issues and Practice. 2005 Jun 9;24(2):3–13. doi:10.1111/j.1745-3992.2005.00006.x
3. Norcini J, Anderson MB, Bollela V, Burch V, Costa MJ, Duvivier R, et al. 2018 Consensus framework for good assessment. Med Teach. 2018 Nov 2;40(11):1102–9. doi:10.1080/0142159X.2018.1500016
4. Permasutha MB, Rahayu GR, Giri MKW, Pradiptha DAGF. Multiple-Choice Questions in Basic Biomedical Science Module. Jurnal Pendidikan dan Pengajaran. 2024 May 17;57(1):47–56. doi:10.23887/jpp.v57i1.63314
5. Eleragi AMS, Miskeen E, Hussein K, Rezigalla AA, Adam MIE, Al-Faifi JA, et al. Evaluating the multiple-choice questions quality at the College of Medicine, University of Bisha, Saudi Arabia: a three-year experience. BMC Med Educ. 2025;25(1). doi:10.1186/s12909-025-06700-2
6. Tsegaye BS, Asemu MM, Hailu HB. Construct validity and reliability of Amharic version of DASS-21 scale among Ethiopian Defense University College of Health Science students. BMC Health Serv Res. 2024 Aug 9;24(1):914. doi:10.1186/s12913-024-11267-7
7. Almarzooq ZI, Lopes M, Kochar A. Virtual Learning During the COVID-19 Pandemic. J Am Coll Cardiol. 2020 May;75(20):2635–8. doi:10.1016/j.jacc.2020.04.015
8. Mila Nu Nu Htay, M Ganesh Kamath, Anand K M, Sandheep Sugathan, Eunice Ong Luyee, Soumendra Sahoo. Overview of Reliability and Validity of Assessments in Medical Education. International Journal of Transformative Health Professions Education. 2025 Jan 26;1(1):14–8. doi:10.71354/tr76gs42
9. Eleragi AMS, Miskeen E, Hussein K, Rezigalla AA, Adam MIE, Al-Faifi JA, et al. Evaluating the multiple-choice questions quality at the College of Medicine,

University of Bisha, Saudi Arabia: a three-year experience. BMC Med Educ. 2025 Feb 13;25(1):233. doi:10.1186/s12909-025-06700-2
10. Harden RM. Outcome-Based Education: the future is today. Med Teach. 2007 Jan 3;29(7):625–9. doi:10.1080/01421590701729930
11. Sadeghi P, Pourabbas A, Dehghani G, Katebi K. Quantitative and qualitative item analysis of exams of basic medical sciences departments of Tabriz University of Medical Sciences in 2023. BMC Med Educ. 2025;25(1). doi:10.1186/s12909-025-07539-3
12. Haladyna TM, Downing SM, Rodriguez MC. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. Applied Measurement in Education. 2002 Jul;15(3):309–33. doi:10.1207/S15324818AME1503_5
13. Gebremichael MW, Baraki B, Mehari MA, Assalfew B. Item analysis of multiple choice questions from assessment of health sciences students, Tigray, Ethiopia. BMC Med Educ. 2025;25(1). doi:10.1186/s12909-025-06904-6
14. Kumar D, Jaipurkar R, Shekhar A, Sikri G, Srinivas V. Item analysis of multiple choice questions: A quality assurance test for an assessment tool. Med J Armed Forces India. 2021;77. doi:10.1016/j.mjafi.2020.11.007
15. Cappelleri JC, Jason Lundy J, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. Clin Ther. 2014 May;36(5):648–62. doi:10.1016/j.clinthera.2014.04.006 PubMed PMID: 24811753.
16. Kim YH, Kim BH, Kim J, Jung B, Bae S. Item difficulty index, discrimination index, and reliability of the 26 health professions licensing examinations in 2022, Korea: a psychometric study. J Educ Eval Health Prof. 2023 Nov 22;20:31. doi:10.3352/jeehp.2023.20.31
17. Husnain A, Khan A, Khan MU, Hussain FN. Subjective quality of multiple choice questions used in undergraduate courses in orthopedics and other specialties. Pak J Med Sci. 2020 Oct 19;36(7). doi:10.12669/pjms.36.7.2864
18. Abdel-Hameed AA, Al-Faris EA, Alorainy IA, Al-Rukban MO. The criteria and analysis of good multiple choice questions in a health professional setting. Saudi Med J. 2005;26(10).
19. Rezigalla AA, Eleragi AMESA, Elhussein AB, Alfaifi J, ALGhamdi MA, Al Ameer AY, et al. Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. BMC Med Educ. 2024;24(1). doi:10.1186/s12909-024-05433-y
20. Maulina N, Novirianthy R. Item Analysis and Peer-Review Evaluation of Specific Health Problems and Applied Research Block Examination. Jurnal Pendidikan Kedokteran Indonesia: The Indonesian Journal of Medical Education. 2020 Jul 28;9(2):131. doi:10.22146/jpki.49006
21. Rezigalla AA, Eleragi AMESA, Elhussein AB, Alfaifi J, ALGhamdi MA, Al Ameer AY, et al. Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. BMC Med Educ. 2024 Apr 24;24(1):445. doi:10.1186/s12909-024-05433-y
22. Chauhan GR, Chauhan BR, Vaza J V, Chauhan PR. Relations of the Number of Functioning Distractors With the Item Difficulty Index and the Item Discrimination Power in the Multiple Choice Questions. Cureus. 2023 Jul;15(7):e42492. doi:10.7759/cureus.42492 PubMed PMID: 37644928.
23. Touissi Y, Hjiej G, Hajjioui A, Ibrahimi A, Fourtassi M. Does developing multiple-choice Questions Improve Medical Students' Learning? A Systematic

Review. Med Educ Online. 2022 Dec;27(1):2005505. doi:10.1080/10872981.2021.2005505 PubMed PMID: 34969352.

24. Arbeni W, Windiani A, Sihotang DSB, Anggraini N, Wulandari S, Nugroho A. Test Reliability Analysis in Educational Evaluation: A Quantitative Approach to Consistency and Validity. Holistic Science. 2025 Feb 15;5(1):59–64. doi:10.56495/hs.v5i1.838
25. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. 2011 Jun 27;2:53–5. doi:10.5116/ijme.4dfb.8dfd