Jurnal Inovasi Pendidikan MH Thamrin p-ISSN: 2549-3310 e:ISSN: 2623-2901
Vol 7 (1), Maret 2023 Hal 46-60

Scoping Review: *Item Analysis* Pada *Multiple Choice Questions* (MCQs) dalam Pembelajaran

Sri Yunita[1], Yasinta Dewi Kristianti[1, 2], Novia Martin[1], Nuryudica Harefa[1], Elly Yana[1,3], Patricia Zenitha Aritonang[1], Dessy Nur Safitri[1], Ari Indra Susanti[4], Qorinah Estiningtyas Sakilah Adnani[4]

[1] Magister Kebidanan, Fakultas Kedokteran, Universitas Padjadjaran, Bandung, Indonesia
[2] Universitas Mohammad Husni Thamrin Jakarta
[3] Dinas Kesehatan Provinsi Kepulauan Bangka Belitung
[4] Departemen Ilmu Kesehatan Masyarakat, Fakultas Kedokteran, Universitas Padjadjaran, Bandung, Indonesia

***Correspondence author*:**Yasinta Dewi Kristianti**,** email: yasinta22001@mail.unpad.ac.id

DOI : https://doi.org/10.37012/jipmht.v7i1.1671

Abstrak

Salah satu komponen penting dari proses belajar mengajar adalah penilaian dan evaluasi. Ini adalah cara untuk mengetahui batas kemampuan, dan perkembangan hasil pembelajaran mahasiswa serta tingkat efektivitas pengajaran dosen. Pertanyaan multiple choice (MCQs) atau pertanyaan pilihan berganda adalah salah satu jenis penilaian dan evaluasi yang sangat populer. Agar ujian tetap sesuai dengan penilaian yang diinginkan, standar yang berlaku untuk pembuatan instrumen MCQ harus dipatuhi. Diharapkan review ini dapat menambah referensi untuk meningkatkan pengetahuan tentang pembuatan instrumen evaluasi MCQ dengan berfokus pada analisis hasil item analisis pada instrumen evaluasi MCQ. Struktur yang digunakan oleh Arksey dan O'Malley terdiri dari lima langkah. Menurut hasil, terdapat tujuh artikel yang diperoleh dari proses pencarian. Terdapat empat tema yang ditemukan dalam hasil analisis. Mereka adalah sebagai berikut: menentukan kualitas bagian soal; elemen yang mempengaruhi tingkat kesulitan soal; menjamin validitas dan kredibilitas bagian soal; dan menggunakan metode analisis untuk menghasilkan distraktor. Dalam pemilihan evaluasi pembelajaran dalam pilihan berganda (MCQ), ketelitian diperlukan dalam pembuatan soal dan opsi jawaban yang menggunakan item analysis. Keterbatasan metode *Scoping Review* dibandingkan dengan penelitian lainnya adalah terdapat 5 tahapan penyelesaian dan 1 tahapan *optional* serta pemanfaatan metode yang dilakukan untuk menentukan *theoretical framework* factor.

Kata Kunci: *item analysis, multiple choice question, reliabilitas, review, validitas*

Abstract

Assessment and evaluation are critical components of the teaching and learning process for determining students' abilities, progress, and development and assessing educators' teaching effectiveness. Multiple choice questions (MCQs) are one sort of assessment and evaluation approach that is often utilized. Standards in manufacturing MCQ instruments must be maintained so that the test remains consistent with the desired assessment. This scoping review focuses on the examination of findings relating item analysis on the MCQs assessment instrument in order to add references in developing knowledge in the development of MCQs instruments. The framework utilized is Arksey and O'Malley's, which has five stages. The search method yielded seven articles, according to the results. The analysis reveals four themes: evaluating item quality; factors affecting question difficulty level; assuring item validity and reliability; and applying an analysis system to generate distractors. As a result, accuracy is required when creating questions and answer possibilities utilizing item analysis, particularly in MCQs. The limitations of the Scoping Review method compared to other studies are that there are 5 completion stages and 1 optional stage as well as the use of the method used to determine theoretical framework factors.

Keywords: item analysis, multiple choice questions, reliability, validity

PENDAHULUAN

Penilaian dan evaluasi dalam proses belajar mengajar merupakan salah satu bagian penting dalam mengukur batas kemampuan dan perkembangan peserta didik serta menilai efektivitas pengajaran oleh pendidik. Proses evaluasi pembelajaran diaplikasikan untuk mengukur luaran sebuah program secara sistematik dan juga pencapaian hasil belajar mahasiswa secara individual di dalam kelas maupun saat praktik klinik (Billings & Halstead, 2020). Penilaian atau tes merupakan komponen kritis dalam pendidikan profesi kesehatan yang terus berkembang dengan pesat (Schuwirth & van der Vleuten, 2018). Begitu pula dengan pendidikan kebidanan yang menjadi dasar dalam mempersiapkan bidan kompeten agar mampu memberikan asuhan terstandar dan sesuai dengan kebaruan ilmu pengetahuan (Luyben et al., 2017).

Assessment atau penilaian didefinisikan oleh *American Research Association* sebagai proses sistematik untuk mengukur atau mengevaluasi karakteristik atau performa dari individu, program, atau wujud lainnya dengan tujuan untuk menarik sebuah kesimpulan (Yudkowsky et al., 2020). Penilaian secara waktu terbagi menjadi dua, yaitu penilaian formatif dan sumatif. Konsep penilaian formatif dan sumatif ini diperkenalkan pada tahun 1960an untuk mengidentifikasi peran evaluasi program dalam pengembangan bahan kurikulum baru (Dolin et al., 2018). Penilaian formatif dilakukan ketika proses belajar sedang berjalan, sedangkan penilaian sumatif dilaksanakan di akhir program pendidikan (Billings & Halstead, 2020).

Penilaian dan evaluasi memiliki banyak jenis metode, salah satu penilaian yang cukup lazim digunakan adalah *multiple choice questions* (MCQs) atau pertanyaan pilihan berganda. MCQs seringkali digunakan dalam penilaian formatif, dan menjadi salah satu komponen penilaian sumatif. Penilaian dan evaluasi yang menggunakan MCQs cukup tinggi pada bidang pendidikan medis dan keperawatan (D'Sa & Visbal-Dionaldo, 2017). Instrumen penilaian MCQs lebih mudah dilaksanakan dibandingkan dengan instrumen penilaian lainnya, akan tetapi kekuatan instrumen tergantung dengan pertanyaan yang ditanyakan (Coughlin & Featherstone, 2017). Instrumen tes harus dibangun dengan baik agar dapat merefleksikan fungsi kognitif yang diinginkan (D'Sa & Visbal-Dionaldo, 2017). Instrumen MCQs yang dibangun dengan baik mampu menilai level kognitif lebih tinggi dalam taksonomi Bloom seperti interpretasi data, sintesis data, dan aplikasi pengetahuan lebih dari hanya mengingat kembali materi ajar (Elgadal & Mariod, 2021).

Standar dalam pembuatan instrumen MCQs perlu dijaga agar tes tetap sesuai dengan penilaian yang diinginkan. *Item analysis* sangat penting untuk memperbaiki butir soal yang

akan digunakan kembali pada tes selanjutnya dan dapat digunakan untuk mengeliminasi item yang menyesatkan (Quaigrain & Arhin, 2017). *Scoping review* ini berfokus pada analisis temuan mengenai *item analysis* pada instrumen evaluasi MCQs dimana metode penilaian ini sangan sering dilakukan oleh bidang pendidikan sehingga diperlukan kajian agar dapat menambah referensi dalam meningkatkan pengetahuan dan keterampilan dalam membangun instrumen MCQs.

METODE

Studi ini menggunakan metode *scoping review*, yaitu pendekatan yang telah lazim dikenal dalam dunia kesehatan untuk meninjau atau memetakan beberapa bukti penelitian. Beberapa tujuan dari *scoping review* antara lain: memeriksa keluasan bidang penelitian; menentukan nilai melalui tinjauan sistematis secara lengkap; meringkas dan menyebarkan hasil penelitian; serta mengidentifikasi *gap* atau kesenjangan dari penelitian yang telah ada (Arksey & O'Malley, 2005; Levac et al., 2010). Pada penulisan artikel ini, penulis menggunakan *framework* yang dikembangkan oleh Arksey dan O'Malley pada tahun 2005, terdiri dari 5 (lima) tahap sebagai berikut: (1) identifikasi pertanyaan penelitian; (2) identifikasi studi yang relevan; (3) seleksi artikel penelitian; (4) *charting* data; (5) menyusun, meringkas, dan melaporkan hasil.

a. Identifikasi pertanyaan penelitian

Pada *scoping review* ini, penulis ingin mengetahui tentang *item analysis* pada pertanyaan pilihan berganda sebagai alat penilaian validitas dan reliabilitas soal. Pada penyusunan pertanyaan penelitian, penulis menggunakan framework PEO (Arksey & O'Malley, 2005). Pertanyaan penelitian *scoping review* adalah "Bagaimana *item analysis* pada *multiple choice question* untuk menilai validitas dan reliabilitas soal?"

Table 1. *Framework*

P (*Population and Problem*)	E (*Exposure*)	O (*Outcome*)
Multiple choice question	*Item analysis*	*Assessment* OR *Validity* OR *Reliability*

Selain menentukan pertanyaan review, penulis mempersempit ruang lingkup artikel yang dicari dengan menetapkan kriteria inklusi dan eksklusi agar dapat terarah dan sesuai dengan fokus penelitian (Levac et al., 2010). Adapun kata kunci yang digunakan dalam pencarian artikel berdasarkan penyusunan pencarian Boolean. Kata kunci pada *review* ini adalah item

AND analysis AND MCQ AND *assesment* OR *validity* OR *reliability*.

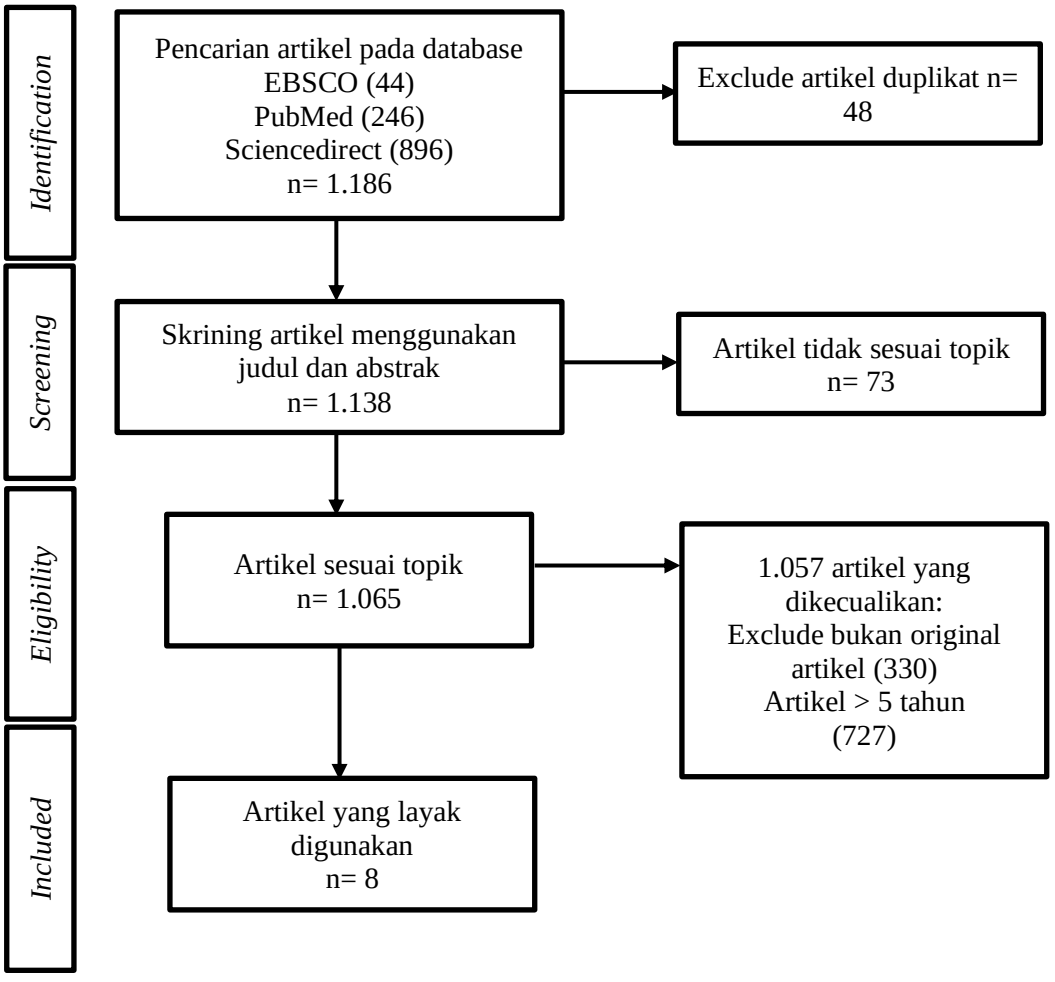
Table 2. Kriteria inklusi dan eksklusi

Kriteria Inklusi	Kriteria Eksklusi
a. Artikel lima tahun terakhir (2018-2023) b. *Original article* c. Penelitian kuantitatif dan kualitatif d. Artikel berbahasa Inggris	a. Artikel *review* b. Artikel *commentary* c. Buku

b. Identifikasi studi yang relevan

Pencarian artikel menggunakan tiga *database* yaitu: PubMed, EBSCO, dan Science Direct. Dari hasil pencarian menggunakan kata kunci ditemukan 246 artikel dari PubMed, 44 artikel dari EBSCO, dan 896 artikel dari Science Direct.

Gambar 1. PRISMA *Flowchart*



c. *Charting Data*

Tabel 3. *Charting Data*

No	Author(S)/ Year/Title	Country	Purpose	Research Design and Methode	Result
1	(Nojomi & Mahmoudi, 2022) *Assessment Of Multiple-Choice Questions By Item Analysis For Medical Students' Examinations*	Tehran, Iran	Untuk mengetahui analisis butir soal-soal MCQ pada tes kepaniteraan mahasiswa kedokteran umum	Studi cross-sectional dilakukan pada sampel 1202 soal MCQ yang mencakup teori di Fakultas Kedokteran Universitas Ilmu Kedokteran Iran selama tahun keempat kepaniteraan mahasiswa kedokteran pada semester kedua tahun 2019. Nilai ujian mereka diperoleh dari kantor ujian fakultas kedokteran, Deputi Pendidikan Kedokteran Umum.	Dari 1202 soal MCQs, menurut indeks kesulitan, sebagian besar pertanyaan (666) dianggap dapat diterima (55,39%). Dari segi indeks diskriminasi (DI), 530 (44,09%) soal memiliki koefisien diskriminasi rata-rata. Selain itu, 215 (17,88%) memiliki DI negatif atau buruk dan memerlukan revisi atau eliminasi dari bank tes. Dari 1202 soal pilihan ganda, 669 (50,7%) dirancang pada tingkat kognitif yang lebih rendah (taksonomi I), 174 (14,5%) termasuk dalam taksonomi II, dan 419 (34,8%) dari pertanyaan memiliki taksonomi III. Selain itu, menurut kelemahan strukturan dari Milman Cheklist bahwa kelemahan structural yang paling umum adalah kurangnya pilihan negatif untuk *stem* 1127 (93,8%), sedangkan pilihan vertical 376 (31,3%) adalah yang paling tidak umum.
2	(Bhattacherjee et al., 2022) *Evaluation Of Multiple-Choice Questions By Item Analysis, From An Online Internal*	Bengal, India	Untuk mengetahui analisis butir-butir soal MCQ dari penilaian internal mahasiswa kedokteran semester 6	Sembilan puluh delapan siswa MBBS pada semester 6 hadir untuk penilaian internal pada 14 Agustus 2020, melalui mode online menggunakan *Google Form*. Ada 60 soal pilihan ganda "jenis tanggapan tunggal" yang masing-masing	Enam puluh soal pilihan ganda dengan 240 pilihan (60 opsi benar dan 180 distraktor) dianalisis. Rata-rata nilai yang dicapai adalah 42,92± (standar deviasi [SD] 5,07). Dif I, DI, dan DE masing-masing adalah 47,95± (SD 16,39), 0,12± (SD 0,10), dan 18,42± (SD 15,35). Soal-soal yang dapat

	Assessment Of 6th Semester Medical Students In A Rural Medical College, West Bengal			terdiri dari 1 tanda tanpa tanda negatif untuk jawaban yang salah. Waktu yang diberikan adalah 80 menit. Soal pilihan ganda dibuat oleh semua guru di departemen. Semua soal pilihan ganda memiliki *stem* tunggal, satu jawaban yang benar (kunci), dan tiga alternatif yang salah (pengalih perhatian). Setiap item dianalisis indeks kesukaran (Dif I), indeks diskriminasi (DI), dan efisiensi pengecoh (DE).	dikategorikan sulit ditemukan sebanyak 15%, sedangkan soal-soal yang tergolong mudah sebanyak 46,67%. Item dengan DI buruk adalah 70% dan 21,66% adalah diskriminasi yang dapat diterima. Diskriminasi negatif ditunjukkan oleh 6,67% item. Korelasi negatif yang sangat lemah ditemukan antara Dif I dan DI. Dari total 180 distraktor, 51,66% distraktor tidak berfungsi. 1 NFD dan 2 NFD ditemukan di masing-masing 35% item. 16,67% item memiliki ketiga distraktor sebagai NFD, sedangkan hanya 13,33% item yang tidak memiliki NFD.
3	(D. Kumar et al., 2021) *Item Analysis Of Multiple Choice Questions: A Quality Assurance Test For An Assessment Tool*	Pune, India	Mengidentifikasi soal MCQ dengan menggunakan analisis item dari tiga tes yang diikuti mahasiswa sarjana kedokteran dan sarjana bedah (MBBS)	Tiga penilaian internal terhadap 150 siswa MBBS tahun pertama dilakukan yang mencakup 90 soal MCQ. Setiap MCQ terdiri dari *stem* dan 04 pilihan dengan hanya satu respon yang benar dan tiga distraktor. Jawaban yang benar diberi nilai satu dan jawaban yang salah diberi nilai nol. Tidak ada tanda negatif. Setiap item dianalisis untuk empat indeks yaitu DIF I, DI, DE dan reliabilitas konsistensi internal.	Dari total 90 soal MCQ, sebagian besar yaitu 74 (82%) soal memiliki tingkat kesulitan yang baik/dapat diterima dengan rata-rata DIF I sebesar 55,32±7.4 (rata-rata±SD), sedangkan tujuh (8%) terlalu sulit dan sembilan (10%) terlalu mudah. Sebanyak 72 (80%) item memiliki DI sangat baik hingga dapat diterima dan 18 (20%) memiliki DI buruk dengan rata-rata DI keseluruhan 0,31±0,12. Ada korelasi lemah yang signifikan antara DIF I dan DI (r⁄40,140, p <.0001). DE rata-rata adalah 32,35 ±31,3 dengan 73% distraktor fungsional secara keseluruhan. Ukuran reliabilitas item tes dengan Cronbach alpha adalah 0,85 dan Kuder-Richardson Formula 20 adalah 0,71 yang baik.

					Kesalahan standar pengukuran adalah 1,22.
4.	(Applegate et al., 2019) *The Effect of Option Homogeneity in Multiple-Choice Items*	Amerika Serikat	Penelitian ini melakukan uji empiris pengaruh homogenitas opsi pada item pilihan ganda pada pemeriksaan lisensi profesional untuk menentukan prediktabilitas dan besarnya perubahan	Eksperimen ini dilakukan dengan menggunakan item dari ujian lisensi profesional nasional yang besar. Statistik item ditetapkan dengan menyemai item eksperimen dengan item reguler yang diberikan selama tes adaptif terkomputerisasi. Item eksperimental tidak diberikan secara adaptif dan tidak dapat dibedakan oleh peserta ujian. Tanggapan item eksperimental didokumentasikan tetapi tidak dihitung sebagai bagian dari skor tes. Setidaknya 533 tanggapan dicatat untuk setiap item dalam penelitian ini.	Lima puluh item awalnya dimasukkan dalam setiap sampel; Namun, karena masalah administrasi, ada tanggapan yang tidak memadai untuk analisis untuk beberapa item. Hasil untuk 47 item dari sampel yang mirip dengan yang tidak mirip dan untuk 49 item dari sampel yang tidak mirip dengan yang serupa dilaporkan. Minimal 533 tanggapan dicatat untuk setiap kalibrasi item, dengan rata-rata 580 respons per item. Data analisis menunjukkan tidak ada pengaruh yang konsisten pada kesulitan soal, diskriminasi, kesesuaian dengan pengukuran model, atau waktu respons yang terkait dengan tidak adanya atau adanya homogenitas pilihan. Sementara hasilnya negatif, mereka mempertanyakan pedoman yang ditetapkan dalam pengembangan item. Sebuah hipotesis diajukan untuk menjelaskan mengapa efek ini ditemukan dalam beberapa penelitian tetapi tidak yang lain.
5.	(Pan & Jiang, 2022) *Examining Distractor Qualities Of Pediatrics Subject Tests From a National Assesment*	Brussel, Belgia	Untuk menyempurnakan jumlah distraktor dalam penilaian berdasarkan analisis distraktor	Dari tes mata pelajaran pediatri di tingkat nasional, tanggapan mentah dari 44.332 pemeriksaan terhadap sembilan belas pertanyaan pilihan ganda dianalisis, sehingga kualitas distraktor dievaluasi melalui	Analisis item tradisional menunjukkan bahwa sebagian besar item memiliki sifat psikometrik yang dapat diterima, dan dua item ditandai dengan tingkat kesulitan dan diskriminasi item yang rendah. Analisis distraktor menunjukkan bahwa sekitar sepertiga item

				metode tradisional dan lanjutan seperti indeks korelasi kanonik. Selain itu, studi simulasi dilakukan untuk menyelidiki dampak menghilangkan nomor distraktor pada reliabilitas.	memiliki distraktor yang berfungsi buruk berdasarkan frekuensi pilihan yang relatif rendah (<5%) dan ukuran efek kecil dari diskriminasi distraktor. studi simulasi juga menegaskan bahwa penyusutan angka distraktor menjadi 4 dapat dilakukan.
6.	(Harti et al., 2021) *All India AYUSH post graduate entrance exam 2019 e AYURVED A MCQ item analysis*	Bengaluru, India	Untuk menganalisis soal pilihan ganda AIAPGET 2019 aliran Ayurveda.	Ujian ini berbasis komputer yang dilakukan di 25 pusat di seluruh India. Kertas pertanyaan memiliki 100 soal pilihan ganda dengan 1 jawaban yang benar dan 3 pengecoh untuk setiap item (Pernyataan masalah).	Soal AIAPGET 2019 aliran Ayurveda memiliki Indeks Kesulitan 37,32 ± 16,11 Indeks Diskriminatif 0,46 ± 0,27 dan Indeks Distraktor 89 ± 17,8. Kesimpulan: Analisis kami menunjukkan bahwa meskipun ideal, soal-soal cenderung mengarah ke sisi kesulitan.
7.	(A. P. Kumar et al., 2023) *A Novel Approach To Generate Distractors For Multiple Choice Questions*	Amerika Serikat	Penelitian yang diusulkan menyajikan sebuah sistem yang disebut DIGEN (DIstractor GENeration) yang menghasilkan distraktor untuk yang diberikan MCQ, jadi ada kebutuhan untuk mengevaluasi setiap item secara terpisah.	Experiment	DIGEN mengambil teks tidak terstruktur serta pertanyaan pilihan ganda dengan kunci sebagai sumber wajib bersama dengan ontologi yang dapat menjadi sumber opsional untuk menghasilkan distraktor secara otomatis dalam domain teknis. Distractors yang dihasilkan telah dievaluasi berdasarkan Item Response Theory, yang menunjukkan hasil yang menjanjikan.

HASIL DAN PEMBAHASAN

a. Hasil

1. Karakteristik

Hasil *review* artikel yang diperoleh dari negara maju yaitu Amerika Serikat yang mendapatkan 2 artikel dan 1 artikel dari negara Belgia. Sementara artikel dari negara berkembang yaitu India yang mendapatkan 4 artikel dan 1 dari negara Iran. Semua artikel ini didominasi oleh negara India sebagai negara berkembang dengan menggunakan studi penelitian kuantitatif.

2. Tematik

Hasil *review* ditemuakan beberapa tema yang sesuai dengan fokus *review*, sebagai berikut.

Tabel 4. Hasil Analisis *Scoping Review*

No	Tema	Sub Tema
1	Menentukan Kualitas Butir Soal	a) *Difficulty index* [1, 2] b) *Discrimination index* [1, 2, 3, 4] c) *Item analysis* [1, 2, 3, 4, 5, 7] d) *Multiple-choice questions* (2, 8)
2	Faktor Yang Mempengaruhi Tingkat Kesukaran Soal.	a) *Item writing* [5] b) *Item difficulty* [1, 2, 3, 4, 7] c) *Discrimination index* [1, 2, 3, 4] d) Daya Pembeda [1, 3, 7, 8]
3	Memastikan Validitas & Reliabilitas Butir Soal	a) *Discrimination index* [1, 2, 3, 4] b) *Distractor effectiveness* [1, 3, 7, 8]
4	Sistem Analisis Untuk Menghasilkan Distraktor	a) *Ontology* [8] b) DIGEN(DIstractor GENeration) [8] c) *Distractor generation* [8] d) Pilihan ganda AIAPGET 2019 aliran Ayurveda. [7]

b. Pembahasan

Multiple choice question (MCQ)

MCQ memiliki kekuatan dan kelemahan. Salah satu kelemahan MCQ adalah tidak dapat menilai aspek afektif, tetapi dapat menilai aspek kognitif dan psikomotorik juga (Padamjeet et al., 2018). Selain itu, keuntungan pilihan ganda adalah ramah pengguna, yang berarti dapat digunakan dengan cepat di kelas yang ramai. Analisis soal dapat menunjukkan validitas dan

reliabilitas tes. Sangat sulit untuk mengetahui apakah mahasiswa telah mempelajari konsep yang diujikan. Setiap pertanyaan memiliki indeks kesukaran yang lebih rendah. Indeks kesulitan dan diskriminasi tetap negatif, meskipun ujian AIAPGET 2019 menunjukkan nilai negatif (Sandeep Prakash & Dattatraya Hanumantrao, 2020).

Pertanyaan pilihan ganda adalah metode penilaian formatif dan sumatif yang paling umum. Pemahaman, penerapan, analisis, sintesis, dan evaluasi adalah semua komponen evaluasi soal pilihan ganda. Membangun soal pilihan ganda yang berbeda adalah tantangan, serta menghasilkan kesimpulan statistik untuk mengevaluasi seberapa baik soal pilihan ganda dan faktor distraktornya. Ini adalah alat penilaian yang bermanfaat bagi penguji dan calon peserta ujian (Hingorjo & Jaleel, 2012).

Mengidentifikasi dan memilih distraktor yang masuk akal dan dapat diandalkan adalah bagian penting dari pengembangan soal pilihan ganda dalam pembelajaran karena membantu mengendalikan tingkat kesulitan pertanyaan (Zhang & VanLehn, 2021). Indeks Kesulitan (DI), yang digunakan untuk mengevaluasi tingkat kesulitan dalam pilihan ganda, Indeks Diskriminasi (DI), yang digunakan untuk membedakan mahasiswa dengan kemampuan yang lebih baik dan lebih rendah, dan Indeks Efek Distractor (DE), yang merupakan parameter yang paling umum digunakan. Parameter ini dievaluasi untuk memastikan bahwa soal pilihan ganda asli dan dapat diandalkan sehingga dapat disimpan, diubah, atau dibuang selama proses pengembangannya. Item analysis akan membantu menyusun soal pilihan ganda untuk ujian masuk tingkat nasional, meskipun pengulangan soal pilihan ganda tidak disukai (Harti et al., 2021).

Menurut penelitian yang dilakukan di India yang bertujuan untuk menganalisis soal pilihan ganda AIAPGET 2019, aliran Ayurveda menunjukkan Indeks Kesulitan 37,32 ± 16,11, Indeks Diskriminatif 0,46 ± 0,27, dan Indeks Distraktor 89 ± 17,8. Seperti yang ditunjukkan oleh hasil analisis, masalah cenderung mengarah ke sisi yang lebih sulit meskipun kondisinya ideal.

Dalam soal pilihan ganda, indeks kesulitan item dan diskriminasi item berada dalam kisaran yang dapat diterima. Namun, perlu dicatat bahwa perlu direkomendasikan untuk membuang atau merevisi soal-soal yang mudah dan indeks diskriminasi negatif atau buruk dengan memberikan pelatihan kepada dosen untuk memperbaiki bank soal, karena soal-soal yang mudah dan buruk dapat mempengaruhi dalam pengukuran penilaian pencapaian pembelajaran yang akan berdampak terhadap hasil penilaian pembelajaran (Nojomi & Mahmoudi, 2022).

Disarankan agar analisis item dilakukan tidak hanya untuk semua asesmen berbasis MCQ tetapi juga untuk jenis pertanyaan asesmen lainnya, seperti pertanyaan fitur kunci dan masalah manajemen pasien. Oleh karena itu, metode analitik komparatif dapat digunakan sebagai alat penilaian ujian klinis tambahan untuk menilai pengetahuan, keterampilan, dan kinerja mahasiswa kedokteran. Oleh karena itu, ujian pertanyaan esai modifikasi harus dievaluasi secara teratur untuk memastikan validitas dan kredibilitasnya (Nojomi & Mahmoudi, 2022).

Item Analysis

Menciptakan tes dengan tingkat kesulitan yang spesifik adalah salah satu tantangan terus-menerus dalam proses pengembangan penilaian pendidikan dan psikologi profesional . Setiap penilaian memiliki tiga ciri utama: konten dan konstruk (menilai apa yang ingin dinilai), validitas (menilai apa yang ingin dinilai), dan reliabilitas (menilai seberapa baik skor satu jawaban yan benar). Setiap penilaian didasarkan pada kemampuan pemecahan masalah yang luar biasa (Kiat et al., 2018).

Item Analysis secara umum dapat sebagai alat statistic yang digunakan untuk menilai kinerja siswa dalam pembelajaran menggunakan suatu tes, membantu mengidentifikasi butir-butir soal yang buruk dan menentukan akar penyebab dari kinerja yang kurang tersebut sehingga dapat memastikan penilaian kompetensi siswa yang efektif dan akurat. *Item Analysis* juga dipandang sebagai alat penilaian tingkat efisiensi pembelajaran bagi siswa dalam pembelajaran di bidang akademik (Christian et al., 2017).

Sebuah *Item Analysis* biasanya mencakup kesulitan item, perbedaan item, dan reliabilitas tes (misalnya, *alfa Cronbach*). Secara definisi kesulitan soal sebagai persentase benar (atau *p-value*), artinya persentase peserta ujian yang mendapatkan soal dengan benar dalam sampel. Indeks diskriminasi item meliputi korelasi antara item secara total) dan indeks diskriminasi item, artinya ada perbedaan rasio jawaban yang benar pada sepertiga atas dan bawah peserta ujian (Pan & Jiang, 2022).

Item analysis pada evaluasi MCQ yang dilakukan pada mahasiswa didapatkan hasil bahwa soal-soal yang dikategorikan sulit ditemukan sebanyak 15%, sedangkan soal-soal yang tergolong mudah sebanyak 46,67%. Item dengan indeks diskriminasi soal buruk adalah 70% dan 21,66% adalah diskriminasi yang dapat diterima. Diskriminasi negatif ditunjukkan oleh 6,67% item. Setelah di kaji ada beberapa faktor penyebab antara lain pemilihan soal oleh penguji tersebut tertulis didalam silabus dan soal yang diberikan terlalu mudah atau ambigu, juga memiliki kunci jawaban yang salah dan antara satu pilihan dengan pilihan lain hampir

memiliki kemiripan, sehingga mahasiswa yang menjawab mengetahui jawaban dari pilihan yang berbeda (Bhattacherjee et al., 2022).

Item Analysis menilai kualitas setiap item (pertanyaan) dalam hal tingkat kesulitannya (Indeks Kesulitan), kemampuannya untuk membedakan antara yang berkinerja tinggi dan berkinerja rendah (Indeks Diskriminasi) dan seberapa benar pengecoh digunakan dalam setiap item (Pengalih perhatian). Hasil analisis item membantu meningkatkan kekuatan soal pilihan ganda, karena item yang baik dapat dipertahankan sedangkan item yang buruk direvisi, diganti, atau dihilangkan (Applegate et al., 2019).

Pada dasarnya, dalam menyelidiki suatu properti item yaitu pada konsep validitasnya, sehingga penting sekali untuk mengevaluasi kualitas item dan *distractor* untuk menjaga kualitas penilaian standar. Beberapa cara mengevaluasi kualitas *distractor* dengan: 1. Menghitung frekuensi pilihan *distractor* dan rasio seleksi yang meningkat, 2. Menghitung ukuran efek untuk mendeteksi diskriminatif *distractor* menggunakan *canonical correlation* di setiap masing-masing item (RCC) (Pan & Jiang, 2022).

Canonical correlation pada masing-masing item (RCC) untuk mengevaluasi ukuran efek distraktor. Sebagai salah satu jenis metode analisis multivariat, korelasi kanonik dapat digunakan untuk menggambarkan hubungan antara satu set variabel multivariat (yaitu, matriks pilihan *distractor* dari setiap item) dan satu variabel *continue* (yaitu, skor total peserta ujian yang melakukan tidak memecahkan item yang ditargetkan), menunjukkan kekuatan diskriminatif dari pengecoh.

Melalui hasil penelitiannya menyimpulkan bahwa analisis butir soal merupakan alat yang berharga karena dapat membantu pendidikuntuk mempertahankan soal pilihan ganda yang valid dan membuang butir-butir soal yang tidak layak digunakan. Hal ini juga membantu dalam meningkatkan keterampilan guru dalam menyusun dan mengidentifikasi butir-butir soal yang penting untuk diujikan kepada peserta didik (Kaur et al., 2016).

Studi Experiment Kumar (2023), penelitian yang diusulkan menyajikan sebuah sistem yang disebut DIGEN (DIstractor GENeration) yang menghasilkan *distractor* dalam MCQs secara otomatis. DIGEN menggunakan teks tidak terstruktur dan pertanyaan pilihan ganda dengan kata kunci dan ontologi tertentu. *Distractor* yang dihasilkan telah dievaluasi berdasarkan *Item Response Theory*, yang menunjukkan hasil yang meyakinkan.

Faktor ini tidak hanya membantu mengurangi tebakan acak tetapi juga membedakan tingkat kognitif di antara peserta ujian yang berbeda. *Distractor* menjadi masuk akal jika untuk pertanyaan yang diberikan memenuhi persyaratan, seperti yang dinyatakan dalam Zhang dan

VanLehn (2021): (a) *distractor* salah; (b) *distractor* secara semantik terkait dengan kunci, dan (c) *distractor* memberikan distraksi sehingga kunci dapat diidentifikasi hanya dengan beberapa pemahaman tentang konsep yang ditanyakan di stem.

Untuk menentukan apakah tes pilihan ganda tersebut memenuhi kriteria tes yang baik, kualitas tes harus dievaluasi. Tes yang berkualitas adalah tes yang mampu menggambarkan kemampuan dan kondisi peserta didik. Selain menentukan kualitas, hasil evaluasi tes ini dapat memberikan informasi tentang kelebihan dan kekurangan soal dalam tes tersebut. Dengan demikian, dapat diketahui soal mana yang bisa disimpan untuk digunakan lagi, diperbaiki, atau bahkan dibuang. Dalam praktiknya, tes pilihan ganda ini dapat dievaluasi dengan melakukan analisis butir soal atau item *analysis* (Arikunto, 2018).

Keterbatasan *Review*

Keterbatasan *review* ini adalah memasukan artikel berbahasa inggris dan terbatasnya jurnal - jurnal yang berkaitan dengan topik bahasan, sehingga penulis cukup kesulitan dalam melakukan *review*.

SIMPULAN

Berdasarkan penjelasan di atas dapat di tarik kesimpulan bahwa *Item analysis:*

1. Merupakan prosedur yang relative sederhana yang menyediakan metode untuk menganalisis pengamatan, indateterprestasi dan pengetahuan yang dicapai oleh siswa dan informasi mengenai reliabilitas serta validitas suatu butir/tes (Date et al., 2019).
2. Merupakan pengukuran yang di lakukan terhadap *Item Difficulty*, *Distractor*, dan *Item Discrimination.*
3. Merupakan alat yang yang berharga karena dapat membantu mengevaluasi sebuah tes/ butir item yang digunakan

Bahwa diperlukannya ketelitian dalam membuat soal dan opsi jawaban dengan menggunakan *item analysis* khususnya dalam pemilihan evaluasi pembelajaran dalam pilihan berganda (MCQ), dengan memperhatikan kualitas setiap item (pertanyaan) dalam hal tingkat kesulitannya (Indeks Kesulitan), kemampuannya untuk membedakan antara yang berkinerja tinggi dan berkinerja rendah (Indeks diskriminasi) dan seberapa benar pengecoh digunakan dalam setiap item (pengalih perhatian). Dengan demikian, MCQ dapat digunakan sebagai alat ukur yang tepat, sesuai, dan efektif untuk menilai hasil belajar mahasiswa di masa depan. Diperlukannya penambahan *database* pencarian artikel yang lebih banyak dalam melakukan

telaah jurnal untuk dapat menambah referensi dalam penulisan sehingga dapat lebih mempertajam dalam pembahasan penelitian.

REFERENSI

Applegate, G. M., Sutherland, K. A., Becker, K. A., & Luo, X. (2019). The Effect of Option Homogeneity in Multiple-Choice Items. *Applied Psychological Measurement*, *43*(2), 113–124. https://doi.org/10.1177/0146621618770803

Arikunto, S. (2018). *Dasar-dasar Evaluasi Pendidikan* (Edisi 3). PT Bumi Aksara.

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology: Theory and Practice*, *8*(1), 19–32. https://doi.org/10.1080/1364557032000119616

Bhattacherjee, S., Mukherjee, A., Bhandari, K., & Rout, A. (2022). Evaluation of multiple-choice questions by item analysis, from an online internal assessment of 6 th semester medical students in a rural medical college, West Bengal. *Indian Journal of Community Medicine*, *47*(1), 92. https://doi.org/10.4103/ijcm.ijcm_1156_21

Billings, D. M., & Halstead, J. A. (2020). *Teaching in Nursing, 6th Edition*. Elsevier.

Christian, D. S., Prajapati, A. C., Rana, B. M., & Dave, V. R. (2017). Evaluation of multiple choice questions using item analysis tool: a study from a medical institute of Ahmedabad, Gujarat. *International Journal Of Community Medicine And Public Health*, *4*(6), 1876. https://doi.org/10.18203/2394-6040.ijcmph20172004

Coughlin, P. A., & Featherstone, C. R. (2017). How to Write a High Quality Multiple Choice Question (MCQ): A Guide for Clinicians. *European Journal of Vascular and Endovascular Surgery*, *54*(5), 654–658. https://doi.org/10.1016/j.ejvs.2017.07.012

D'Sa, J. L., & Visbal-Dionaldo, M. L. (2017). Analysis of Multiple Choice Questions: Item Difficulty, Discrimination Index and Distractor Efficiency. *International Journal of Nursing Education*, *9*(3), 109. https://doi.org/10.5958/0974-9357.2017.00079.4

Date, A. P., Borkar, A. S., Badwaik, R. T., Siddiqui, R. A., Shende, T. R., & Dashputra, A. V. (2019). Item analysis as tool to validate multiple choice question bank in pharmacology. *International Journal of Basic & Clinical Pharmacology*, *8*(9), 1999. https://doi.org/10.18203/2319-2003.ijbcp20194106

Dolin, J., Black, P., Harlen, W., & Tiberghien, A. (2018). *Exploring Relations Between Formative and Summative Assessment* (pp. 53–80). https://doi.org/10.1007/978-3-319-63248-3_3

Elgadal, A. H., & Mariod, A. A. (2021). Item Analysis of Multiple-choice Questions (MCQs): Assessment Tool For Quality Assurance Measures. *Sudan Journal of Medical Sciences*. https://doi.org/10.18502/sjms.v16i3.9695

Harti, S., Mahapatra, A. K., Gupta, S. K., & Nesari, T. (2021). All India AYUSH post graduate entrance exam 2019 – AYURVEDA MCQ item analysis. *Journal of Ayurveda and Integrative Medicine*, *12*(2), 356–358. https://doi.org/10.1016/j.jaim.2021.01.013

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *JPMA. The Journal of the Pakistan Medical Association*, *62*(2), 142–147. http://www.ncbi.nlm.nih.gov/pubmed/22755376

Kaur, M., Singla, S., & Mahajan, R. (2016). Item analysis of in use multiple choice questions in pharmacology. *International Journal of Applied and Basic Medical Research*, *6*(3), 170. https://doi.org/10.4103/2229-516X.186965

Kiat, J. E., Ong, A. R., & Ganesan, A. (2018). The influence of distractor strength and response order on MCQ responding. *Educational Psychology*, *38*(3), 368–380. https://doi.org/10.1080/01443410.2017.1349877

Kumar, A. P., Nayak, A., Manjula Shenoy, K., Goyal, S., & Chaitanya. (2023). A novel approach to generate distractors for Multiple Choice Questions. *Expert Systems with Applications*, *225*(April). https://doi.org/10.1016/j.eswa.2023.120022

Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G., & Srinivas, V. (2021). Item analysis of multiple choice questions: A quality assurance test for an assessment tool. *Medical Journal Armed Forces India*, *77*, S85–S89. https://doi.org/10.1016/j.mjafi.2020.11.007

Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: advancing the methodology. *Implementation Science*, *5*(1), 69. https://doi.org/10.1186/1748-5908-5-69

Luyben, A., Barger, M., Avery, M., Bharj, K. K., O'Connell, R., Fleming, V., Thompson, J., & Sherratt, D. (2017). Exploring global recognition of quality midwifery education: Vision or fiction? *Women and Birth*, *30*(3), 184–192. https://doi.org/10.1016/j.wombi.2017.03.001

Nojomi, M., & Mahmoudi, M. (2022). Assessment of multiple-choice questions by item analysis for medical students' examinations. *Research and Development in Medical Education*, *11*, 24. https://doi.org/10.34172/rdme.2022.024

Padamjeet, P., Bheem, P., & Sarita, K. (2018). Multiple choice questions role in assessment of competency of knowledge in Anatomy. *International Journal of Anatomy and Research*, *6*(2.1), 5156–5162. https://doi.org/10.16965/ijar.2018.143

Pan, Q., & Jiang, Z. (2022). Examining distractor qualities of pediatrics subject tests from a national assessment. *Frontiers in Medicine*, *9*(1). https://doi.org/10.3389/fmed.2022.921719

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1), 1301013. https://doi.org/10.1080/2331186X.2017.1301013

Sandeep Prakash, N., & Dattatraya Hanumantrao, N. (2020). Effect of surprise test and instruction for negative marking on item analysis in Pharmacology. *IP International Journal of Comprehensive and Advanced Pharmacology*, *5*(1), 19–21. https://doi.org/10.18231/j.ijcaap.2020.005

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2018). How to Design a Useful Test. In *Understanding Medical Education* (pp. 275–289). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119373780.ch20

Yudkowsky, R., Park, Y. S., & Downing, S. M. (2020). *Assessment in Health Professions Education*. Routledge TNF.

Zhang, L., & VanLehn, K. (2021). Evaluation of auto-generated distractors in multiple choice questions from a semantic network. *Interactive Learning Environments*, *29*(6), 1019–1036. https://doi.org/10.1080/10494820.2019.1619586