Jurnal Pendidikan dan Teknologi Indonesia (JPTI)
Vol. 5, No. 11, November 2025, Hal. 3269-3278
DOI: https://doi.org/10.52436/1.jpti.1024
p-ISSN: 2775-4227
e-ISSN: 2775-4219

Evaluasi Kualitas Butir Soal Matematika Materi Matriks Kelas XI dengan Pendekatan Analisis Kuantitatif Menggunakan ANATES 4.1.0 dan SPSS 26

Rivani Adistia Dewi[*1], Fauzan Syahbudin[2], Muhammad Dwiky Febriansyah[3], Jarnawi Afgani Dahlan[4]

[1,2,3,4]Pendidikan Matematika, Universitas Pendidikan Indonesia, Bandung, Indonesia
Email: [1]*rivaniadistiadewi@upi.edu*, [2]*fauzan.syah@upi.edu*, [3]*mdwiky64@gmail.com*, [4]*jarnawi@upi.edu*

Abstrak

Analisis kualitas butir soal evaluasi pembelajaran matematika pada materi matriks kelas XI dilakukan dengan pendekatan kuantitatif deskriptif menggunakan *software* ANATES versi 4.1.0 dan IBM SPSS Statistics 26. Penelitian ini dilakukan untuk menilai sejauh mana butir soal yang dikembangkan telah memenuhi kriteria sebagai instrumen evaluasi pembelajaran yang baik. Instrumen terdiri dari 25 butir soal yang mencakup soal pilihan ganda, pilihan ganda kompleks, dan uraian, yang disusun berdasarkan indikator Kurikulum Merdeka. Subjek penelitian terdiri dari 18 siswa kelas XI SMA AQL Islamic School 2 Purwakarta. Proses analisis dilakukan dengan menilai lima aspek utama, yaitu validitas, reliabilitas, daya pembeda, tingkat kesukaran, dan efektivitas pengecoh. Hasil analisis menunjukkan bahwa sebagian besar butir soal memiliki validitas sedang hingga tinggi dan reliabilitas yang baik. Tingkat kesukaran soal bervariasi dari mudah hingga sukar, dengan daya pembeda dan efektivitas pengecoh yang beragam. Beberapa soal menunjukkan kelemahan, seperti pengecoh yang tidak berfungsi (non-fungsional) serta daya pembeda yang rendah, sehingga memerlukan revisi untuk meningkatkan kualitasnya. Temuan ini mengindikasikan bahwa sebagian besar soal telah memenuhi kriteria instrumen evaluasi yang baik, namun tetap diperlukan revisi pada beberapa butir untuk mencapai kualitas optimal. Disimpulkan bahwa analisis butir soal berbasis teknologi sangat penting dalam menjamin kualitas, keadilan, dan efektivitas asesmen, serta mendukung praktik evaluasi yang akuntabel dalam pembelajaran matematika di sekolah.

Kata kunci: *analisis butir soal, ANATES, matriks, reliabilitas, SPSS, validitas.*

Evaluation of the Quality of Grade XI Mathematics Test Items on Matrix Material Using a Quantitative Analysis Approach with ANATES 4.1.0 and SPSS 26

Abstract

The analyze of the quality of mathematics learning evaluation items on matrix material for grade XI was conducted using a descriptive quantitative approach with ANATES version 4.1.0 and IBM SPSS Statistics 26 software. This research aimed to assess the extent to which the developed test items met the criteria of a good learning evaluation instrument. The instrument consisted of 25 items, including multiple-choice, complex multiple-choice, and essay questions, constructed based on indicators of the Merdeka Curriculum. The research subjects were 18 students from grade XI at SMA AQL Islamic School 2 Purwakarta. The analysis process focused on five main aspects: validity, reliability, discriminating power, difficulty level, and distractor efficiency. The results showed that most items had moderate to high validity and good reliability. The item difficulty levels ranged from easy to difficult, with varying discriminating power and distractor efficiency. Some items exhibited weaknesses, such as non-functional distractors and low discriminating power, requiring revision to improve their quality. These findings indicate that most of the items met the criteria of a good evaluation instrument, although improvements are still needed in some items to achieve optimal quality. It is concluded that technology-based item analysis is crucial in ensuring the quality, fairness, and effectiveness of assessment, as well as supporting accountable evaluation practices in mathematics learning at school.

Keywords: *ANATES, item analysis, matrix, reliability, SPSS, validity.*

p-ISSN: 2775-4227
e-ISSN: 2775-4219

1. PENDAHULUAN

Salah satu tantangan utama dalam dunia pendidikan saat ini adalah memastikan bahwa proses evaluasi benar-benar mencerminkan pencapaian belajar siswa secara adil dan objektif. Penilaian bukan sekadar alat ukur hasil belajar, tetapi juga merupakan sarana penting untuk memberikan umpan balik dan merancang strategi pembelajaran yang lebih efektif [1]. Dalam konteks ini, peran guru sebagai perancang dan pelaksana asesmen yang berkualitas menjadi sangat krusial. Kualitas asesmen sangat bergantung pada kompetensi guru dalam mengembangkan dan menganalisis instrumen yang valid, reliabel, serta mampu membedakan tingkat penguasaan siswa secara akurat [2], [3].

Kompetensi guru dalam merancang dan menganalisis penilaian, yang dikenal sebagai *assessment literacy*, masih menjadi tantangan di berbagai negara, termasuk Indonesia, karena keterbatasan pelatihan menganalisis butir soal secara empirik [1], [4]. Dalam konteks digital, *assessment literacy* mencakup kemampuan menilai secara profesional dan memanfaatkan teknologi untuk evaluasi yang akurat [3]. Hasil penelitian menunjukkan bahwa *assessment literacy* berkorelasi positif dengan peningkatan praktik instruksional di kelas, sehingga kompetensi ini merupakan fondasi penting dalam pengambilan keputusan pembelajaran [2].

Penilaian yang efektif bergantung pada kualitas instrumen yang digunakan dalam proses evaluasi. Instrumen evaluasi yang baik harus memenuhi prinsip validitas, reliabilitas, daya pembeda, serta tingkat kesukaran yang sesuai [5]. Prinsip-prinsip ini ditegaskan dalam dokumen *Standards for Educational and Psychological Testing* yang dikeluarkan oleh AERA, APA, dan NCME [6]. Instrumen yang memenuhi prinsip-prinsip tersebut akan menghasilkan data yang akurat dan dapat digunakan sebagai dasar pengambilan keputusan pembelajaran yang tepat.

Dalam praktiknya, analisis butir soal digunakan untuk menilai sejauh mana setiap item mengukur kompetensi yang dimaksud melalui indikator daya pembeda, tingkat kesukaran, dan efektivitas pengecoh [7]. Beberapa penelitian menunjukkan bahwa pengecoh yang tidak fungsional, yaitu dipilih oleh kurang dari 5% peserta, dapat menurunkan daya pembeda dan validitas soal [8]–[11]. Selain itu, soal dengan dua hingga tiga pengecoh yang berfungsi cenderung memiliki kualitas yang optimal, sehingga penting untuk merancang pengecoh yang efektif dan tidak mudah dieliminasi oleh peserta didik [12].

Penggunaan *software*, seperti ANATES versi 4.1.0 dan IBM SPSS Statistics 26, mendukung analisis butir soal secara efisien dan akurat. Keduanya memudahkan evaluasi validitas, reliabilitas, daya pembeda, tingkat kesukaran, dan efektivitas pengecoh, serta menunjang analisis statistik lanjutan untuk memperkuat interpretasi data. Teknologi digital tentu bermanfaat untuk mempercepat dan meningkatkan akurasi analisis soal dalam berbagai konteks asesmen [13]. Selain itu, analisis pengecoh berbasis statistik mampu mengidentifikasi soal yang berkualitas dengan lebih efektif [8]. Dengan demikian, penggunaan teknologi dalam analisis butir soal tidak hanya meningkatkan efisiensi kerja guru, tetapi juga kualitas evaluasi pembelajaran secara keseluruhan.

Namun, sebagian besar penelitian di Indonesia belum mengintegrasikan pendekatan teknologi secara menyeluruh dalam analisis butir soal matematika, khususnya pada materi matriks. Penelitian [14] menganalisis soal matematika tingkat SMP, tanpa fokus pada materi matriks dan belum mengoptimalkan penggunaan *software* untuk analisis. Sedangkan [15] menyoroti validitas soal matematika di madrasah tanpa membahas indikator daya pembeda dan efektivitas pengecoh. Semantara itu, [16] menggunakan ANATES versi 4.1.0 untuk menganalisis soal PAI, sehingga berbeda dari fokus kajian matematika. Adapun [17] menganalisis soal PTS matematika dengan teori klasik, yang tidak secara eksplisit mengkaji materi matriks.

Berdasarkan paparan tersebut, terdapat celah penelitian yang belum banyak dieksplorasi, yaitu analisis butir soal matematika pada materi matriks kelas XI dengan menggunakan *software* ANATES versi 4.1.0 dan IBM SPSS Statistics 26, sehingga dapat diperoleh gambaran yang lebih objektif, akurat, dan relevan bagi pengembangan instrumen evaluasi pembelajaran. Oleh karena itu, penelitian ini secara spesifik bertujuan untuk mengevaluasi kualitas butir soal matematika materi matriks kelas XI berdasarkan lima indikator utama, yaitu validitas, reliabilitas, daya pembeda, tingkat kesukaran, dan efektivitas pengecoh, dengan menggunakan *software* ANATES versi 4.1.0 dan IBM SPSS Statistics 26. Hasil penelitian diharapkan dapat memberikan kontribusi terhadap pengembangan instrumen evaluasi pembelajaran matematika yang lebih objektif, adil, dan sesuai dengan karakteristik materi.

2. METODE PENELITIAN

2.1 Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan desain deskriptif. Tujuannya adalah untuk mengevaluasi kualitas butir soal matematika pada materi matriks kelas XI melalui analisis statistik menggunakan *software* ANATES versi 4.1.0 dan IBM SPSS Statistics 26. Penelitian ini berfokus pada lima indikator utama kualitas soal, yaitu validitas, reliabilitas, daya pembeda, tingkat kesukaran, dan efektivitas pengecoh.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

2.2 Subjek Penelitian

Subjek penelitian ini adalah seluruh siswa kelas XI di SMA AQL Islamic School 2 Purwakarta, yang berjumlah 18 orang. Penentuan subjek dilakukan dengan teknik total sampling karena seluruh populasi dijadikan sampel. Para siswa yang dipilih telah menyelesaikan pembelajaran materi matriks dan memiliki latar belakang akademik yang beragam, sehingga dianggap representatif dalam konteks evaluasi kualitas butir soal. Pemilihan sekolah didasarkan pada kesesuaian antara kurikulum yang digunakan dan tujuan penelitian, yakni implementasi Kurikulum Merdeka dan penerapan asesmen berbasis kompetensi.

2.3 Penyusunan Instrumen

Instrumen penelitian ini berupa tes evaluasi yang terdiri atas 25 butir soal, yang terdiri dari 15 soal pilihan ganda, 5 soal pilihan ganda kompleks, dan 5 soal uraian. Instrumen dikembangkan berdasarkan kisi-kisi yang disusun sesuai kompetensi dasar dan indikator capaian hasil belajar dalam Kurikulum Merdeka. Soal-soal dirancang dengan memperhatikan variasi bentuk dan tingkat kognitif guna mengakomodasi prinsip asesmen berbasis kompetensi.

Sebelum digunakan, instrumen divalidasi melalui teknik *expert judgment* yang melibatkan dua ahli: seorang guru matematika dari sekolah yang diteliti dan seorang dosen ahli di bidang evaluasi pembelajaran. Proses validasi ini bertujuan untuk menilai kesesuaian isi soal, indikator, dan tujuan pembelajaran. Masukan dari para ahli digunakan untuk merevisi instrumen sebelum pelaksanaan uji coba kepada siswa.

2.4 Teknik Analisis Data

Data yang diperoleh dianalisis secara kuantitatif menggunakan *software* ANATES versi 4.1.0 dan IBM SPSS Statistics 26. Analisis dilakukan terhadap lima aspek utama kualitas butir soal, yaitu validitas, reliabilitas, daya pembeda, tingkat kesukaran, dan efektivitas pengecoh. Validitas butir dihitung dengan korelasi point biserial untuk mengetahui kontribusi butir terhadap total skor. Reliabilitas instrumen diukur menggunakan koefisien KR-20 untuk menilai konsistensi internal tes. Daya pembeda dihitung untuk mengetahui kemampuan soal membedakan siswa berkemampuan tinggi dan rendah. Tingkat kesukaran dihitung berdasarkan persentase siswa yang menjawab benar tiap soal. Efektivitas pengecoh ditentukan dari persentase pilihan yang dipilih oleh minimal 5% siswa. Kelima parameter tersebut dianalisis dan diinterpretasikan berdasarkan kaidah dalam *Classical Test Theory* (CTT) [18].

3. HASIL DAN PEMBAHASAN

3.1 Soal Pilihan Ganda

Analisis terhadap 15 butir soal pilihan ganda pada tes evaluasi materi matriks kelas XI dilakukan berdasarkan lima parameter utama, yaitu validitas, reliabilitas, daya pembeda, tingkat kesukaran, dan efektivitas pengecoh. Proses analisis dibantu oleh *software* ANATES versi 4.1.0. Analisis ini bertujuan untuk mengevaluasi sejauh mana butir soal mampu mengukur kompetensi siswa secara objektif dan berkualitas.

Tabel 1. Validitas soal pilihan ganda

No	Korelasi	Signifikansi	Validitas
1	0,879	Sangat signifikan	Valid
2	0,457	-	Tidak valid
3	0,136	-	Tidak valid
4	0,785	Sangat signifikan	Valid
5	0,533	Signifikan	Valid
6	0,387	-	Tidak valid
7	0,656	Sangat signifikan	Valid
8	0,665	Sangat signifikan	Valid
9	0,215	-	Tidak valid
10	0,332	-	Tidak valid
11	0,623	Sangat signifikan	Valid
12	0,630	Sangat signifikan	Valid
13	0,053	-	Tidak valid
14	0,465	-	Tidak valid
15	0,457	-	Tidak valid

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Validitas butir soal dihitung menggunakan korelasi point biserial. Hasil analisis yang disajikan pada Tabel 1 menunjukkan bahwa dari 15 butir soal pilihan ganda, terdapat 7 soal yang tergolong valid karena memiliki nilai korelasi yang melebihi batas signifikansi (r>0,4702 untuk N=18). Butir-butir valid tersebut adalah soal nomor 1, 4, 5, 7, 8, 11, dan 12. Soal-soal ini dianggap mampu merepresentasikan kompetensi yang diukur dan berkontribusi secara signifikan terhadap skor total tes. Sebaliknya, 8 butir soal lainnya memiliki nilai korelasi yang rendah dan tidak signifikan, sehingga dinyatakan tidak valid. Soal-soal tersebut disarankan untuk direvisi atau dikaji ulang guna meningkatkan akurasi pengukuran. Hasil ini menunjukkan pentingnya uji validitas empirik sebagai bagian dari proses penyusunan instrumen [5].

Selain validitas, reliabilitas instrumen juga dianalisis untuk mengetahui sejauh mana konsistensi internal antar-butir soal dalam mengukur kemampuan siswa. Hasil perhitungan dengan *software* ANATES versi 4.1.0 menunjukkan bahwa instrumen tes memiliki koefisien reliabilitas sebesar 0,88, yang tergolong dalam kategori tinggi. Nilai ini mengindikasikan bahwa butir-butir soal memiliki tingkat konsistensi yang baik dalam menghasilkan skor yang stabil dan dapat dipercaya, sesuai dengan prinsip-prinsip pengujian dalam CTT [18]. Dengan demikian, secara umum instrumen ini mampu memberikan hasil evaluasi yang reliabel, sehingga dapat digunakan sebagai dasar dalam pengambilan keputusan pembelajaran.

Daya pembeda menunjukkan kemampuan soal dalam membedakan siswa berkemampuan tinggi dan rendah. Berdasarkan hasil analisis pada Tabel 2, terdapat 6 soal yang memiliki daya pembeda sangat baik, yaitu soal nomor 1, 4, 5, 7, 11, dan 12. Sementara 3 soal lainnya berkategori baik, 2 soal tergolong cukup, dan 3 soal termasuk kategori jelek. Adapun 1 soal, yakni soal nomor 13, dikategorikan sangat jelek. Hasil ini menunjukkan bahwa soal nomor 13 gagal menjalankan fungsi diskriminatifnya. Soal dengan daya pembeda sangat jelek seperti ini sebaiknya direvisi secara menyeluruh atau dipertimbangkan untuk dieliminasi dari instrumen evaluasi agar tidak mengganggu keakuratan pengukuran. Kategori ini menegaskan pentingnya daya pembeda dalam menghindari soal yang hanya mengukur hafalan atau tidak membedakan level pemahaman siswa [7].

Tabel 2. Daya pembeda soal pilihan ganda

No	Kelompok Atas	Kelompok Bawah	Beda	Indeks Daya Pembeda (%)	Kategori
1	5	0	5	100	Sangat baik
2	5	2	3	60	Baik
3	2	1	1	20	Jelek
4	5	1	4	80	Sangat baik
5	5	1	4	80	Sangat baik
6	4	2	2	40	Cukup
7	4	0	4	80	Sangat baik
8	5	2	3	60	Baik
9	2	1	1	20	Jelek
10	3	2	1	20	Jelek
11	4	0	4	80	Sangat baik
12	5	0	5	100	Sangat baik
13	2	2	0	0	Sangat jelek
14	5	2	3	60	Baik
15	2	0	2	40	Cukup

Tingkat kesukaran soal dianalisis berdasarkan proporsi siswa yang menjawab benar. Hasil analisis yang disajikan pada Tabel 3 menunjukkan bahwa dari 15 soal pilihan ganda, sebanyak 10 soal berada pada kategori sedang, 2 soal termasuk mudah (soal nomor 4 dan 8), 2 soal dikategorikan sukar (soal nomor 3 dan 13), dan 1 soal diklasifikasikan sangat sukar (soal nomor 15), karena hanya dijawab benar oleh 2 dari 18 siswa (11,11%). Distribusi ini menunjukkan bahwa sebagian besar soal sudah berada dalam rentang kesukaran yang optimal. Namun, soal yang sangat mudah atau sangat sukar perlu ditinjau ulang agar tidak bias dalam mengukur kemampuan siswa. Soal yang sangat mudah cenderung tidak menantang dan tidak mampu mengukur variasi kemampuan siswa secara maksimal, sedangkan soal yang sangat sukar berpotensi menimbulkan frustrasi atau hilangnya motivasi, sehingga dapat mengganggu akurasi pengukuran kemampuan siswa yang sebenarnya.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Tabel 3. Tingkat kesukaran soal pilihan ganda

No	Jumlah Benar	Tingkat Kesukaran (%)	Tafsiran
1	12	66,67	Sedang
2	12	66,67	Sedang
3	5	27,78	Sukar
4	14	77,78	Mudah
5	11	61,11	Sedang
6	12	66,67	Sedang
7	8	44,44	Sedang
8	14	77,78	Mudah
9	7	38,89	Sedang
10	9	50,00	Sedang
11	8	44,44	Sedang
12	9	50,00	Sedang
13	4	22,22	Sukar
14	11	61,11	Sedang
15	2	11,11	Sangat sukar

Efektivitas pengecoh dianalisis berdasarkan proporsi siswa yang memilih opsi selain jawaban benar. Pengecoh dianggap fungsional jika dipilih oleh minimal 5% responden. Hasil pada Tabel 4 menunjukkan bahwa beberapa soal seperti nomor 1 dan 8 memiliki pengecoh yang tidak fungsional dengan kode "--" hingga "---". Hal ini menunjukkan bahwa pengecoh tersebut terlalu lemah dan tidak efektif sebagai pengalih perhatian. Sebaliknya, soal nomor 7 dan 14 memiliki tiga pengecoh fungsional dengan kategori "+" atau "++". Pengecoh yang fungsional dapat meningkatkan daya pembeda soal [10], [11]. Oleh karena itu, kualitas pengecoh perlu diperhatikan agar fungsi evaluatif soal tetap optimal. Temuan ini sejalan dengan penelitian sebelumnya yang menekankan pentingnya identifikasi dan revisi pengecoh yang tidak fungsional karena terjadi korelasi negatif antara jumlah pengecoh yang tidak fungsional dengan kualitas soal secara keseluruhan [19], [20].

Tabel 4. Kualitas pengecoh soal pilihan ganda

No.	A	B	C	D	E
1	3--	1+	2+	12**	0--
2	12**	4---	0--	2+	0--
3	8---	3++	5**	2+	0--
4	2--	14**	0--	2--	0--
5	11**	2++	4---	0--	1+
6	2+	4---	12**	0--	0--
7	2++	2++	4-	2++	8**
8	0--	14**	0--	3---	1++
9	5--	7**	0--	1-	5--
10	9**	4--	3+	1-	1-
11	4-	0--	8**	3++	3++
12	1-	3+	4--	1-	9**
13	2+	4**	5+	3++	4++
14	1+	2++	2++	2++	11**
15	2**	4++	3+	6+	3+

Keterangan:
** : Kunci jawaban
++ : Sangat baik
+ : Baik
- : Kurang baik
-- : Buruk
--- : Sangat Buruk

Berdasarkan hasil analisis terhadap 15 butir soal pilihan ganda, ditemukan bahwa sebanyak 7 soal memenuhi kriteria validitas dan memiliki kualitas teknis yang memadai, yaitu soal nomor 1, 4, 5, 7, 8, 11, dan 12. Soal-soal tersebut dinilai layak untuk digunakan kembali dalam evaluasi pembelajaran karena mampu mengukur kompetensi siswa secara efektif dan membedakan tingkat penguasaan secara jelas. Sementara itu, 8 soal lainnya memerlukan

p-ISSN: 2775-4227
e-ISSN: 2775-4219

revisi, baik karena validitasnya tidak signifikan, daya pembeda yang rendah, maupun karena adanya pengecoh yang tidak fungsional. Dengan demikian, dapat disimpulkan bahwa sebagian butir soal pilihan ganda telah dinilai layak, namun sebagian lainnya memerlukan revisi, khususnya pada aspek pengecoh dan daya pembeda. Temuan ini menegaskan pentingnya proses revisi dan uji empiris dalam pengembangan instrumen evaluasi, agar setiap butir soal benar-benar mencerminkan kompetensi yang diukur secara adil, objektif, dan sesuai dengan tujuan asesmen.

Kecenderungan tersebut selaras dengan temuan yang menunjukkan bahwa penggunaan ANATES versi 4.1.0 dapat memudahkan identifikasi soal yang berkualitas dan mendeteksi kelemahan teknis seperti pengecoh yang tidak fungsional [16]. Selain itu, statistik juga berperan penting dalam proses revisi soal untuk memastikan bahwa soal benar-benar mampu mengukur kompetensi siswa secara tepat [8]. Dengan demikian, analisis butir soal berbasis *software* terbukti mendukung pengembangan instrumen evaluasi yang lebih akurat, efisien, dan bermakna dalam konteks pembelajaran matematika.

3.2 Soal Pilihan Ganda Kompleks

Analisis terhadap 5 butir soal pilihan ganda kompleks dilakukan menggunakan *software* IBM SPSS Statistics 26 untuk mengevaluasi kualitas instrumen. Seperti halnya soal pilihan ganda, evaluasi ini bertujuan untuk memastikan bahwa setiap butir soal mampu mengukur kompetensi siswa secara objektif, adil, dan berkualitas.

Validitas soal dianalisis menggunakan korelasi antara skor masing-masing soal dengan skor total tes. Hasil analisis yang ditampilkan pada Tabel 5 menunjukkan bahwa soal nomor 1, 3, dan 4 dinyatakan valid secara statistik. Sebaliknya, soal nomor 2 dan 5 dinyatakan tidak valid dan perlu direvisi. Hal ini memperlihatkan bahwa meskipun kompleks secara bentuk, tidak semua soal kompleks memiliki kualitas yang bagus. Temuan ini mendukung pernyataan bahwa soal matematika yang berbentuk kompleks justru rentan terhadap kesalahan validitas jika tidak diuji secara empirik [17].

Tabel 5. Validitas soal pilihan ganda kompleks

No.	Korelasi	Validitas
1	0,540	Valid
2	0,438	Tidak valid
3	0,661	Valid
4	0,531	Valid
5	0,418	Tidak valid

Daya pembeda dihitung berdasarkan selisih proporsi jawaban benar antara kelompok atas dan kelompok bawah. Hasil analisis pada Tabel 6 menunjukkan bahwa terdapat satu soal (soal nomor 3) dengan kategori cukup, sedangkan tiga soal lainnya berada pada kategori jelek (soal nomor 1 dan 4) dan sangat jelek (soal nomor 2). Hanya satu soal (soal nomor 5) yang tergolong baik, meskipun nilainya masih relatif rendah untuk diandalkan sebagai pembeda yang efektif. Temuan ini menunjukkan bahwa sebagian besar soal belum memiliki kemampuan diskriminatif yang memadai. Oleh karena itu, diperlukan revisi terhadap soal-soal tersebut untuk meningkatkan kualitas instrumen secara keseluruhan. Rendahnya daya pembeda berpotensi menghambat identifikasi tingkat penguasaan siswa secara akurat, yang pada akhirnya dapat menurunkan validitas hasil evaluasi [7].

Tabel 6. Daya pembeda soal pilihan ganda kompleks

No	Indeks Daya Pembeda (%)	Kategori
1	18,9	Jelek
2	-1,9	Sangat jelek
3	35,4	Cukup
4	16,7	Jelek
5	45,5	Baik

Analisis tingkat kesukaran dilakukan berdasarkan persentase siswa yang menjawab benar pada setiap butir soal. Hasil analisis menunjukkan bahwa dua soal, yaitu soal nomor 2 dan 4, berada dalam kategori sedang karena memiliki persentase jawaban benar antara 30% hingga 70%. Sementara itu, soal nomor 1, 3, dan 5 berada pada kategori mudah dengan tingkat kesukaran sebesar 78%. Komposisi ini mencerminkan adanya variasi tingkat kesukaran yang cukup baik, namun tetap perlu perhatian terhadap keseimbangan distribusi soal agar tingkat kesukaran lebih merata dan sesuai dengan prinsip penyusunan tes yang proporsional.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Tabel 7. Tingkat kesukaran soal pilihan ganda kompleks

No	Tingkat Kesukaran (%)	Tafsiran
1	78	Mudah
2	41,5	Sedang
3	78	Mudah
4	47	Sedang
5	78	Mudah

Berdasarkan hasil analisis soal pilihan ganda kompleks, dapat disimpulkan bahwa masih terdapat butir soal yang perlu direvisi sebelum digunakan dalam evaluasi pembelajaran. Meskipun tiga soal telah memenuhi kriteria validitas dan satu soal menunjukkan daya pembeda yang baik, sebagian besar soal masih menunjukkan kelemahan, baik dari segi validitas yang tidak signifikan, maupun daya pembeda yang rendah. Selain itu, distribusi tingkat kesukaran belum sepenuhnya merata, sehingga dapat berdampak pada keakuratan hasil pengukuran. Oleh karena itu, diperlukan revisi terhadap soal-soal tertentu, baik dari segi konstruksi maupun kesesuaian dengan indikator pembelajaran, guna meningkatkan validitas, daya pembeda, serta kualitas instrumen secara keseluruhan agar lebih representatif dan mampu mengukur kompetensi siswa secara objektif, adil, dan efektif.

Temuan ini mempertegas bahwa kompleksitas format soal tidak secara otomatis menjamin kualitas instrumen, terlebih jika tidak didukung dengan konstruksi yang matang dan pengujian statistik yang tepat. Pemanfaatan teknologi dalam analisis soal memiliki peran penting untuk mengidentifikasi kekurangan dengan lebih cepat dan akurat [13]. Oleh karena itu, pengembangan soal pilihan ganda kompleks harus disertai dengan validasi sistematis dan pemantauan ketat terhadap struktur dan fungsi setiap pilihan jawaban, agar instrumen yang dihasilkan benar-benar mampu menjadi alat ukur yang sahih, efektif, dan adil.

3.3 Soal Uraian

Untuk menilai kualitas 5 butir soal uraian dalam penelitian ini, dilakukan analisis kuantitatif berdasarkan data empiris dengan menggunakan *software* IBM SPSS Statistics 26. Analisis dilakukan guna memastikan bahwa setiap soal mampu mengukur kompetensi siswa secara akurat dan bermakna.

Validitas diuji menggunakan korelasi Pearson antara skor butir soal dan skor total. Berdasarkan hasil analisis pada Tabel 8, sebanyak 4 soal (soal nomor 1, 2, 3, dan 4) termasuk dalam kategori valid dan dapat digunakan dalam evaluasi pembelajaran tanpa revisi. Sedangkan soal nomor 5 tegolong tidak valid dan masih memerlukan revisi. Hal ini mengindikasikan bahwa sebagian besar soal uraian telah dirancang secara selaras dengan indikator pembelajaran. Temuan ini menunjukkan bahwa mayoritas soal uraian telah disusun secara selaras dengan indikator pembelajaran yang ditetapkan. Hal ini memperkuat pandangan bahwa soal uraian yang dikembangkan dengan baik dapat memberikan informasi mengenai pencapaian kompetensi siswa secara utuh dan kontekstual [5].

Tabel 8. Validitas soal uraian

No.	Korelasi	Validitas
1	0,847	Valid
2	0,853	Valid
3	0,872	Valid
4	0,550	Valid
5	0,462	Tidak valid

Tabel 9. Daya pembeda soal uraian

No	Indeks Daya Pembeda (%)	Kategori
1	67,6	Baik
2	71,8	Sangat baik
3	75,2	Sangat baik
4	43,8	Baik
5	32,2	Cukup

Daya pembeda menunjukkan kemampuan soal dalam membedakan siswa yang telah menguasai materi dengan yang belum. Berdasarkan hasil analisis pada Tabel 9, diketahui bahwa 2 soal (soal nomor 1 dan 4) memiliki daya pembeda sangat baik, 2 soal (soal nomor 2 dan 3) berada pada kategori baik, dan 1 soal (soal nomor 5) berada dalam kategori cukup. Hasil ini menunjukkan bahwa sebagian besar soal uraian efektif dalam membedakan tingkat penguasaan siswa. Efektivitas ini juga mendukung pernyataan bahwa format soal uraian memiliki potensi lebih besar untuk mengungkap kemampuan analisis dan pemahaman mendalam siswa [7].

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Tingkat kesukaran dihitung berdasarkan persentase siswa yang menjawab benar untuk setiap soal. Dari hasil analisis pada Tabel 10, terdapat dua soal dengan kategori sukar (soal nomor 4 dan 5) dan tiga soal berada dalam kategori sedang. Distribusi ini tergolong proporsional, namun perlu diwaspadai bahwa soal yang terlalu mudah atau terlalu sukar dapat mengurangi efektivitas evaluasi jika tidak ditangani dengan tepat [1].

Tabel 10. Tingkat kesukaran soal uraian

No	Tingkat Kesukaran (%)	Tafsiran
1	51,5	Sedang
2	61	Sedang
3	32	Sedang
4	16,75	Sukar
5	11	Sukar

Berdasarkan hasil analisis terhadap 5 butir soal uraian, diperoleh bahwa soal uraian memiliki kualitas yang cukup baik dari segi validitas dan daya pembeda, serta distribusi tingkat kesukaran yang relatif proporsional. Temuan ini mendukung pernyataan bahwa soal uraian yang disusun dan diuji dengan cermat, dapat menjadi alat asesmen yang efektif dalam mengungkap kemampuan berpikir siswa. Dalam konteks ini, penggunaan *software* statistik seperti IBM SPSS Statistics 26 turut memperkuat proses validasi secara objektif dan sistematis [13].

3.4 Pembahasan dan Sintesis Temuan Berdasarkan Jenis Soal

Hasil penelitian menunjukkan bahwa kualitas butir soal evaluasi pembelajaran matematika pada materi matriks bervariasi tergantung pada bentuk soal dan indikator analisis yang digunakan. Pendekatan kuantitatif berbasis *software* seperti ANATES dan SPSS terbukti efektif dalam mengevaluasi kualitas instrumen evaluasi secara objektif, sebagaimana ditunjukkan oleh hasil validitas dan reliabilitas yang tinggi pada sebagian besar soal. Temuan ini memperkuat pentingnya pemanfaatan teknologi dalam asesmen untuk menghasilkan instrumen yang valid, reliabel, dan akuntabel.

Secara khusus, efektivitas pengecoh menjadi perhatian penting dalam evaluasi instrumen. Sebagaimana ditegaskan dalam beberapa penelitian sebelumnya, pengecoh yang tidak fungsional dapat menurunkan daya pembeda dan kualitas soal secara keseluruhan [8], [12] Oleh karena itu, identifikasi dan revisi pengecoh menjadi langkah penting dalam pengembangan soal pilihan ganda. Selain itu, temuan bahwa soal pilihan ganda kompleks cenderung memiliki validitas dan daya pembeda yang rendah memperkuat pandangan bahwa kompleksitas format soal tidak selalu menjamin kualitas, terutama jika tidak didukung oleh proses konstruksi dan validasi yang tepat [17]. Hal ini menunjukkan bahwa penggunaan format soal yang lebih kompleks membutuhkan perhatian khusus dalam penyusunannya.

Penelitian ini juga memberikan dukungan terhadap pentingnya kompetensi guru dalam *assessment literacy* berbasis teknologi. Kemampuan untuk menggunakan alat analisis data seperti ANATES dan SPSS secara efektif dapat meningkatkan akurasi evaluasi serta kualitas pembelajaran [1], [2]. Dalam konteks ini, hasil penelitian berkontribusi nyata terhadap penguatan praktik evaluasi berbasis data yang akuntabel dan professional di tingkat sekolah.

Jika ditinjau berdasarkan jenis soal, ditemukan perbedaan yang signifikan dalam hal kualitas. Soal pilihan ganda secara umum menunjukkan kualitas yang cukup seimbang, dengan sejumlah butir soal memenuhi kriteria validitas, daya pembeda, dan tingkat kesukaran yang sesuai. Sebaliknya, soal pilihan ganda kompleks menunjukkan kelemahan paling menonjol, terutama pada aspek validitas dan daya pembeda, yang mengindikasikan bahwa jenis soal ini belum optimal dalam membedakan siswa berkemampuan tinggi dan rendah secara efektif. Adapun soal uraian cenderung memiliki validitas dan daya pembeda yang tinggi, meskipun terdapat satu soal yang perlu direvisi. Temuan ini mengindikasikan bahwa meskipun soal pilihan ganda dan uraian memiliki potensi besar sebagai instrumen evaluasi yang efektif, penyusunan soal pilihan ganda kompleks masih membutuhkan perhatian dan validasi yang lebih cermat agar berfungsi sebagai alat ukur yang akurat, representatif, dan adil terhadap kompetensi siswa.

4. KESIMPULAN

Hasil penelitian ini menunjukkan bahwa kualitas butir soal evaluasi pembelajaran materi matriks kelas XI bervariasi, tergantung pada jenis dan karakteristik soal. Dari 15 soal pilihan ganda, sebanyak 7 soal memenuhi kriteria validitas, daya pembeda, dan tingkat kesukaran yang optimal, sementara 8 soal lainnya menunjukkan kelemahan pada aspek validitas atau efektivitas pengecoh. Sementara itu, soal pilihan ganda kompleks cenderung memiliki kualitas yang lebih rendah, terutama dari segi daya pembeda dan validitas, meskipun beberapa di

p-ISSN: 2775-4227
e-ISSN: 2775-4219

antaranya memiliki tingkat kesukaran yang memadai. Adapun soal uraian menunjukkan performa yang lebih baik, dengan validitas dan daya pembeda yang tinggi serta distribusi tingkat kesukaran yang proporsional. Temuan ini menegaskan pentingnya proses validasi empirik dan analisis kuantitatif dalam menjamin kualitas instrumen evaluasi.

Berdasarkan temuan tersebut, guru dan pengembang asesmen disarankan untuk merevisi soal-soal yang tidak memenuhi kriteria validitas atau memiliki daya pembeda dan pengecoh yang lemah sebelum digunakan dalam evaluasi pembelajaran. Pemanfaatan *software* untuk analisis butir soal, seperti ANATES versi 4.1.0 dan IBM SPSS Statistics 26, terbukti dapat meningkatkan efektivitas evaluasi dengan mempercepat proses analisis dan memperkuat akurasi data. Oleh karena itu, pelatihan *assessment literacy* berbasis teknologi perlu diperluas agar guru memiliki kompetensi dalam merancang, menganalisis, dan merevisi instrumen evaluasi sesuai dengan prinsip-prinsip pengukuran yang baik. Untuk selanjutnya, penelitian ini dapat diperluas pada materi atau jenjang pendidikan yang berbeda untuk meningkatkan generalisasi temuan dan memperkaya pemahaman mengenai kualitas instrumen evaluasi dalam pembelajaran matematika.

DAFTAR PUSTAKA

[1] S. Pastore, "Teacher assessment literacy: a systematic review," *Front. Educ.*, vol. 8, 2023, doi: 10.3389/feduc.2023.1217167.

[2] S. Ahmadi, S. Ghaffary, and M. Shafaghi, "Examining Teacher Assessment Literacy and Instructional Improvement of Iranian High School Teachers on Various Fields of Study," *Int. J. Lang. Test.*, vol. 12, no. 1, pp. 1–25, 2022, doi: 10.22034/IJLT.2022.146981.

[3] Z. Banitalebi, M. Estaji, and G. T. L. Brown, "Measuring teacher assessment literacy in digital environments: Development and validation of a scenario-based instrument," *Educ. Technol. Soc.*, vol. 28, no. 2, pp. 169–215, 2025, doi: 10.30191/ETS.202504_28(2).RP10.

[4] C. DeLuca and D. A. Klinger, "Assessment literacy development: Identifying gaps in teacher education," *Assess. Educ. Princ. Policy Pract.*, vol. 27, no. 2, pp. 136–155, 2020, doi: 10.1080/0969594X.2019.1659431.

[5] A. J. Nitko and S. M. Brookhart, *Educational assessment of students*, 7th ed. Pearson Education, 2014.

[6] A. E. R. Association, A. P. Association, and N. C. on M. in Education, *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, 2014.

[7] T. M. Haladyna and M. C. Rodriguez, *Developing and validating test items*, 3rd ed. Routledge, 2013.

[8] D. P. Sahoo and R. Singh, "Item and distracter analysis of multiple choice questions (MCQs) from a preliminary examination of undergraduate medical students," *Int. J. Res. Med. Sci.*, vol. 5, no. 12, pp. 5351–5355, 2017, doi: 10.18203/2320-6012.ijrms20175453.

[9] I. Burud, K. Nagandla, and P. Agarwal, "Impact of distractors in item analysis of multiple choice questions," *Int. J. Res. Med. Sci.*, vol. 7, no. 4, pp. 1136–1139, 2019, doi: 10.18203/2320-6012.ijrms20191313.

[10] T. Puthiaparampil and M. Rahman, "How important is distractor efficiency for grading Best Answer Questions?," *BMC Med. Educ.*, vol. 21, no. 1, p. 29, 2021, doi: 10.1186/s12909-020-02463-0.

[11] A. A. Rezigalla *et al.*, "Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items," *BMC Med. Educ.*, vol. 24, 2024, doi: 10.1186/s12909-024-05433-y.

[12] G. R. Chauhan, B. R. Chauhan, J. V Vaza, and P. R. Chauhan, "Relations of the Number of Functioning Distractors With the Item Difficulty Index and the Item Discrimination Power in the Multiple Choice Questions," *Cureus*, vol. 15, no. 7, 2023, doi: 10.7759/cureus.42492.

[13] H. Retnawati *et al.*, "A Systematic Review of the Use of Technology in Educational Assessment Practices: Lesson Learned and Direction for Future Studies," *Int. J. Robot. Control Syst.*, vol. 4, no. 4, pp. 1656–1693, 2024, doi: 10.31763/ijrcs.v4i3.1572.

[14] A. L. F. Tilaar and Hasriyanti, "Analisis Butir Soal Semester Ganjil Mata Pelajaran Matematika pada Sekolah Menengah Pertama," *J. Pengukuran Psikol. dan Pendidik. Indones.*, vol. 8, no. 1, pp. 57–68, 2019, doi: 10.15408/jp3i.v8i1.13068.

[15] B. Utomo, "Analisis Validitas Isi Butir Soal Sebagai Salah Satu Upaya," *J. Pendidik. Mat.*, vol. 2, no. 1, pp. 145–159, 2018.

[16] Elviana, "Analisis Butir Soal Evaluasi Pembelajaran Pendidikan Agama Islam Menggunakan Program

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Anates," *J. MUDARRISUNA*, vol. 10, no. 2, pp. 58–74, 2020, [Online]. Available: https://jurnal.ar-raniry.ac.id/index.php/mudarrisuna/article/view/7839.

[17] Gunawan and L. Asria, "Analisis Butir Soal Penilaian Tengah Semester (PTS) Matematika Kelas XI Berdasarkan Teori Klasik," *MATH LOCUS J. Ris. dan Inov. Pendidik. Mat.*, vol. 4, no. 1, pp. 1–11, 2023, doi: 10.31002/mathlocus.v4i1.3177.

[18] S. C. Ohiri and R. O. Okoye, "Application of Classical Test Theory As Linear Modeling To Test Item Development and Analysis," *Int. Res. J. Mod. Eng. Technol. Sci.*, vol. 5, no. 10, 2023, doi: 10.56726/irjmets45379.

[19] M. Ansari, R. Sadaf, A. Akbar, S. Rehman, Z. R. Chaudhry, and S. Shakir, "Assessment of distractor efficiency of MCQS in item analysis.," *Prof. Med. J.*, vol. 29, no. 05, pp. 730–734, 2022, doi: 10.29309/tpmj/2022.29.05.6955.

[20] M. Sajjad, S. Iltaf, and R. A. Khan, "Nonfunctional distractor analysis: An indicator for quality of multiple choice questions," *Pakistan J. Med. Sci.*, vol. 36, no. 5, pp. 982–986, 2020, doi: 10.12669/pjms.36.5.2439.