IJCEE Vol.9 No.2 December 2023, Page 44-57 **ISSN 2598-2931**

Analisis Kualitas Tes dan Butir Soal Integral pada Evaluasi Formatif Matematika Teknik

Lilis Trianingsih[1]
Email: lilistrianingsih@staff.uns.ac.id

Diterima : 12 Desember 2023
Disetujui : 30 Desember 2023
Terbit : 31 Desember 2023

Abstrak: Analisis kualitas butir soal pilihan ganda merupakan alat penting untuk mengidentifikasi *item* yang dapat dipertahankan, direvisi, atau dikeluarkan. Tujuan penelitian adalah menguji kualitas *item* berdasarkan validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efek pengecoh. Metode penelitian berfokus pada *item analysis* dari 35 soal pilihan ganda yang dilakukan untuk 83 mahasiswa Pendidikan Teknik Bangunan. Sampel penelitian adalah *total sampling*. Analisis statistik menggunakan Microsoft Excel dan IBM SPSS. Teknik pengumpulan data melalui *LMS* spada.uns.ac.id. Analisis data menggunakan analisis deskriptif kuantitatif. Hasil analisis menghasilkan validitas isi 0,89, konstruk 97,14%, valid. Reliabilitas *Kuder-Richardson* 0,876 dan *Intraclass Correlation Coefficient* 0,880, reliabilitas tinggi. *Difficulty index* (DIF I) 28 (80%) soal memiliki tingkat kesukaran baik, tiga (8,57%) terlalu sulit, empat (11,43%) terlalu mudah, dan Mean ± SD 61,83% ± 16,61%. *Discrimination index* (DI) 35 (100%) soal memiliki daya pembeda yang dapat diterima hingga sangat baik dengan Mean±SD 53,90% ± 17,26%. *Distractor effectiveness* (DE) Mean ± SD 92,86% ± 17,75% dengan 92,86% distraktor fungsional secara keseluruhan. Dari keseluruhan hasil analisis disimpulkan butir soal tes integral pada evaluasi formatif Matematika Teknik memiliki kualitas soal yang baik untuk penilaian kognitif mahasiswa. Penelitian selanjutnya dapat dilakukan investigasi tentang korelasi DIF I, DI, dan DE untuk meningkatkan kualitas *item* pada bank soal.
Kata Kunci: analisis kualitas tes; daya pembeda; evaluasi formatif; efektifitas pengecoh; tingkat kesukaran

Abstract: *The item analysis of multiple choice questions (MCQ) is essential for identifying items that can be retained, revised, or removed. The research analyzes items' quality based on validity, reliability, difficulty index, discrimination index, and distractor effectiveness. The study focuses on item analysis of 35 MCQ administered to 83 Building Engineering Education students. The research sample is total sampling—statistical analysis using Microsoft Excel and IBM SPSS—data analysis quantitative descriptive analysis. The study results had a content validity of 0.89 and a construct of 97.14%, which is valid—Kuder-Richardson reliability of 0.876 and Intreclass Correlation Coefficient of 0.880, high reliability. Difficulty index (DIF I) 28 (80%) items had a good difficulty, three (8.57%) were too difficult, four (11.43%) were too easy, Mean ± SD 61.83% ± 16.61%. Discrimination index (DI) 35 (100%) items are acceptable to excellent with Mean ± SD 53.90% ± 17.26%. Distractor effectiveness (DE) Mean ± SD 92.86% ± 17.75% with 92.86% functional distractors. The study concluded that the integral test in the formative evaluation of Engineering Mathematics had good-quality questions for students' cognitive assessment. Further research investigates the correlation of DIF I, DI, and DE to enhance the quality of items in the question bank.*
Keywords: *difficulty index; discrimination index; distractor effectiveness; formative evaluation; item analysis*

[1] Pendidikan Teknik Bangunan, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Sebelas Maret

PENDAHULUAN

Menyiapkan mahasiswa menjadi sumber daya manusia yang potensial untuk meningkatkan pertumbuhan ekonomi, menjadi prioritas yang perlu diperhatikan demi mewujudkan negara yang sejahtera dan stabil. Pembangunan ekonomi yang maju suatu bangsa ditentukan dari sumber daya manusia (SDM) yang berkualitas, karenanya harus diberi prioritas untuk penciptaan negara yang makmur dan kompetitif (Bates, 2014; Pavlova, 2014; Setiawan, 2015, 2022; Setiawan et al., 2020, 2021; Towip et al., 2021; Triyono et al., 2018). Sejalan dengan peta jalan *Sustainable Development Goals* (SDGs) Indonesia menuju 2030 pada tujuan nomor 4 meningkatkan kualitas pendidikan dengan memastikan kualitas pendidikan yang inklusif, merata, dan mempromosikan kesempatan belajar seumur hidup untuk semua, sebagai target nomor tiga hingga enam tujuan pembangunan berkelanjutan (Badan Pusat Statistik, 2016; Salim, 2018). Pada tahun 2030, memastikan akses yang sama bagi setiap individu untuk mendapatkan pendidikan teknis, kejuruan dan tersier yang terjangkau dan berkualitas, meningkatkan keterampilan yang relevan, menghilangkan disparitas gender termasuk penyandang disabilitas, mencapai literasi, dan numerasi (Salim, 2018; Setiawan et al., 2020). Oleh karenanya, keberhasilan proses pembelajaran berkontribusi dalam pencapaian pendidikan yang berkualitas dan inklusif. Efektivitas semua sistem pendidikan sangat bergantung pada kualitas pengajaran dan pembelajaran di ruang kelas, bengkel, laboratorium, dan ruang lain di mana pendidikan berlangsung (Bruri Triyono et al., 2018; Setiawan & Takaoka, 2020).

Asesmen berperan penting dalam membantu menginterpretasikan besarnya kemampuan mahasiswa dan kemajuan belajarnya sendiri (Kumar et al., 2021). Asesmen berbasis kompetensi berpusat pada kinerja mahasiswa, yaitu kemajuan mahasiswa akan bergantung pada demonstrasi kompetensi (Davis & Harden, 2003; Wagner et al., 2019). Penilaian yang digambarkan sebagai "formatif", berlangsung secara terus-menerus, membimbing peserta didik di sepanjang jalan menuju kompetensi melalui umpan balik langsung (Sadler, 1989). Umpan balik formatif telah diidentifikasi sebagai salah satu pengaruh yang paling kuat dalam pembelajaran (Hattie & Timperley, 2007). Sejalan dengan (Mardapi, 2017) yang menjelaskan bahwa tes formatif bertujuan untuk memperoleh masukan tentang tingkat keberhasilan pelaksanaan proses pembelajaran dan memperbaiki strategi pembelajaran. Melalui umpan balik formatif, dapat meningkatkan pembelajaran dengan membantu mahasiswa melacak kemajuan belajarnya, menargetkan pengajaran dan sumber daya untuk kebutuhan mahasiswa, serta mengidentifikasi mahasiswa yang membutuhkan dukungan ekstra sejak dini (Sonnadara et al., 2013).

Fitur utama dari setiap penilaian adalah validitas isi dan konstruk, reliabilitas, serta objektivitas (Kumar et al., 2021). Validitas berkaitan dengan derajat sejauh mana tes menilai apa yang ingin dinilai (Bennett et al., 1984; Popham, 1999). Reliabilitas adalah tingkat atau derajat konsistensi dari suatu instrumen, apakah suatu tes teliti dan dapat dipercaya sesuai dengan kriteria yang telah ditetapkan (Zainal, 2009). Reliabilitas dapat diukur dari tiga kriteria, yaitu *stability*, *dependability*, dan *predictability* (Schneider & Kerlinger, 1979). Lebih lanjut Kumar et al., menjelaskan bahwa reliabilitas berkaitan dengan seberapa baik skor mewakili kemampuan individu sedangkan objektivitas merupakan penilaian dengan satu jawaban yang benar (Kumar et al., 2021). Keputusan penilaian hasil belajar dapat dilakukan dengan tepat hanya apabila didasarkan pada data hasil belajar yang baik (Purwanto, 2014). Dalam penilaian hasil belajar, tes diharapkan dapat menggambarkan sampel perilaku dan menghasilkan nilai yang objektif serta akurat (Zainal, 2009). Oleh sebab itu, tes yang digunakan harus memiliki kualitas yang baik untuk mengukur dan menjadi dasar untuk menilai hasil belajar secara adil dan objektif mahasiswa dalam kemampuannya pada materi integral.

Integral merupakan konsep yang penting

dalam kalkulus selain turunan. Prinsip-prinsip integral diformulasikan oleh Isaac Newton dan Gottfried Leibniz pada abad 17 dengan memanfaatkan hubungan erat antara anti turunan dan integral tentu (Monariska, 2019). Penerapan integral dalam bidang teknik sipil digunakan untuk menghitung momen inersia, volume, luas, titik berat yang semuanya digunakan sebagai alat bantu dalam merancang kekuatan/ketahanan suatu bangunan. Dengan semakin berkembangnya infrastruktur khususnya Indonesia maka perlu perhitungan konstruksi yang aman dan nyaman salah satunya perhitungan lendutan yang terjadi pada konstruksi. Dalam menghitung lendutan yang terjadi maka digunakan fungsi matematika yaitu pengintegralan dari pada fungsi momen sehingga didapatkan nilai dari pada lendutan (Zega, 2019). Matematika Teknik menjadi salah satu mata kuliah wajib program studi Pendidikan Teknik Bangunan yang menjadi dasar untuk mata kuliah pada konsentrasi struktur. Oleh karenanya untuk dapat mengetahui karakteristik butir soal yang baik, maka perlu dilakukan analisis terhadap butir soal tes yang diberikan kepada mahasiswa sehingga akan memberikan kriteria soal yang layak/baik dan mendapatkan data hasil belajar yang akurat dan objektif (Saputra et al., 2022).

Pertanyaan pilihan ganda biasanya digunakan dalam penilaian karena efisien, andal, dan dapat distandarisasi dengan mudah (Kumar et al., 2021). Soal tes bentuk pilihan ganda dapat digunakan untuk mengukur hasil belajar yang lebih kompleks dan berkenaan dengan aspek ingatan, pengertian, aplikasi, analisis, sintesis, dan evaluasi (Zainal, 2009). Kualitas soal pilihan ganda penting karena berpengaruh pada hasil tingkat kompetensi mahasiswa secara keseluruhan selama penilaian. Kesimpulan berdasarkan penilaian hasil tes pilihan ganda memiliki konsekuensi yang berisiko tinggi, maka perlu untuk memastikan bahwa tes adalah penilaian pembelajaran siswa yang valid dan dapat diandalkan (Hicks, 2014; Karim et al., 2021). Soal pilihan ganda yang dibangun dengan baik memungkinkan penilaian keterampilan kognitif tingkat tinggi seperti interpretasi, pemikiran analitis dan kritis, aplikasi atau sintesis dalam kerangka taksonomi Bloom dan piramida Miller (Kumar et al., 2021; McCoubrie, 2004). Membuat dan merevisi soal pilihan ganda adalah tugas yang menantang. Bank soal pilihan ganda yang dibangun dengan cermat setelah analisis butir soal menyeluruh adalah alat yang berguna untuk lembaga akademik manapun untuk melakukan penilaian.

Karakteristik atau kualitas tes berhubungan dengan validitas, reliabilitas, dan kepraktisan (Zainal, 2009). Analisis kualitas butir soal dari alat penilaian memberikan masukan tentang validitas dan reliabilitas butir soal tersebut. Analisis soal pilihan ganda terdiri dari indeks kesukaran (persentase siswa yang menjawab soal dengan benar), indeks diskriminasi (membedakan antara yang berprestasi tinggi dan tidak berprestasi), efektivitas pengecoh (baik tidaknya butir soal dibangun dengan baik) dan reliabilitas konsistensi internal (seberapa baik butir soal berkorelasi satu sama lain) (Kiat et al., 2018; Kumar et al., 2021). Setiap butir soal dievaluasi untuk indeks ini karena jika *item* cacat, maka dapat menjadi distraktor dan penilaian dapat gagal. Oleh karenanya analisis kualitas butir soal ini akan membantu mempertahankan soal pilihan ganda yang berkualitas, membuang atau merevisi ulang *item* yang belum baik. Sehingga nantinya akan menghasilkan soal tes yang layak digunakan untuk pencapaian hasil belajar mahasiswa pada aspek kognitif dan meningkatkan kualitas keputusan yang dibuat berdasarkan penilaian ini.

Ada banyak penelitian yang menemukan bahwa analisis soal sangat penting dalam memperbaiki soal dan menghilangkan *item* yang menyesatkan dalam tes (Botterman et al., 2022; Charles Secolsky, 2017; Hansen & Dexter, 1997; Kumar et al., 2021; Quaigrain & Arhin, 2017). Namun tidak banyak penelitian yang menganalisis soal dengan melakukan uji validitas dan reliabilitas secara internal. Dalam konteks ini, penelitian ini bertujuan untuk menganalisis butir soal tes dengan karakteristik atau kualitas tes berhubungan dengan validitas, reliabilitas

internal eksternal, tingkat kesukaran, daya beda, dan efektivitas pengecoh untuk mendapatkan soal pilihan ganda yang baik dan ideal yang dapat menjadi bagian dari bank soal untuk masa depan.

METODE PENELITIAN

Metode pada penelitian ini berfokus pada *item analysis* dari 35 soal pilihan ganda yang dilakukan untuk 83 mahasiswa Pendidikan Teknik Bangunan semester gasal Agustus 2022-Januari 2023 tahun pertama. Sampel penelitian menggunakan teknik *total sampling* artinya keseluruhan populasi menjadi sampel penelitian. Analisis statistik dilakukan dengan menggunakan Microsoft Excel dan IBM SPSS, versi 22. Teknik pengumpulan data dilakukan dengan dokumentasi melalui *Learning Management System* spada.uns.ac.id yang berupa pemberian soal tes pilihan ganda berupa *quiz*; kunci jawaban; dan lembar jawaban soal tes. Analisis data menggunakan teknik analisis deskriptif kuantitatif. Setiap *item* dianalisis untuk lima indeks, yaitu validitas, reliabilitas, tingkat kesukaran (*difficulty index* (DIF I)), daya pembeda (*discrimination index* (DI)), dan efek pengecoh soal (*distractor effectiveness*) (DE).

Validitas

Pengujian instrumen penelitian dilakukan melalui uji validitas konstruk menggunakan korelasi *point-biserial* dan validitas isi. Pengujian validitas isi instrumen terkait dengan keterwakilan pertanyaan terhadap kemampuan khusus yang harus diukur (Rudner, 1995). Validitas isi ditentukan menggunakan kesepakatan rater, instrumen pengukuran berupa tes dibuktikan valid jika ahli (*expert*) meyakini bahwa instrumen tersebut mengukur penguasaan kemampuan yang didefinisikan dalam domain ataupun juga konstruk psikologi yang diukur (Aiken, 1980). Pengujian validitas isi instrumen berupa soal tes matematika teknik materi persamaan integral dilakukan dengan menggunakan indeks kesepakatan rater mengenai validitas butir yang diusulkan oleh Aiken V. Asumsinya adalah suatu butir instrumen dianggap valid jika memiliki nilai indeks V ≥ 0,40 (Aiken, 1980, 1985; Retnawati, 2016; Setiawan et al., 2021). Indeks Aiken V dirumuskan sebagai berikut:

$$V = \frac{\sum s}{n\,(c-1)} \quad (1)$$

$$V_{untuk\ 35\ item} = \frac{\sum s_{untuk\ 35\ item}}{n(c-1)} = \frac{5{,}31}{2(4-1)} = 0{,}89$$

dengan V adalah indeks kesepakatan rater mengenai validitas butir, s skor yang ditetapkan setiap rater dikurangi skor terendah dalam kategori yang dipakai (s=r-I_o dengan r = skor kategori pilihan rater dan I_o skor terendah dalam kategori penskoran); n banyaknya rater, dan c banyaknya kategori yang dapat dipilih rater.

Pengujian validitas selanjutnya yaitu validitas konstruk menggunakan korelasi *point-biserial*, rumusnya sebagai berikut (Arikunto, 2019).

$$r_{pbis} = \frac{M_p - M_t}{SD_t}\sqrt{\frac{p}{q}} \quad (2)$$

dengan r_{pbis} adalah koefisien biserial, M_p mean skor total yang menjawab benar, M_t mean skor total, SD_t standar deviasi; p proporsi mahasiswa yang menjawab benar (p = banyaknya mahasiswa yang menjawab benar dibagi dengan jumlah seluruh mahasiswa), q proporsi mahasiswa yang menjawab salah (q = 1 – p). Butir soal tes dinyatakan valid jika nilai r_{pbis} > r_{tabel} *product moment* dengan taraf signifikansi 5%.

Reliabilitas

Reliabilitas pada penelitian ini menggunakan dua metode pengujian yang pertama menggunakan rumus koefisien korelasi antar kelas (*Interclass Correlation Coefficient*) dan rumus *Kuder-Richardson* (KR21). Rumus *Kuder-Richardson* digunakan untuk menghitung reliabilitas tes yang terdiri atas *item* dikotomi, dengan menggunakan rata-rata proporsi subjek yang mendapat skor 1 (KR21) (Cronbach, 1951; Mardapi, 2017). Formula KR21 sebagai berikut (Allen & Yen, 1979; Arikunto, 2019; Cronbach, 1951; Kuder & Richardson, 1937; Retnawati, 2017).

$$r_{ii} = \frac{n}{n-1}\left(1 - \frac{\bar{X}(n-\bar{X})}{n.{\sigma_t}^2}\right) \quad (3)$$

dengan r_{ii} adalah koefisien reliabilitas

skor instrumen, n banyaknya butir pertanyaan, $\sigma_t{}^2$ varians total, $\underline{X}$= skor rata-rata (Allen & Yen, 1979; Arikunto, 2019; Cronbach, 1951; Kuder & Richardson, 1937; Retnawati, 2017). Kriteria untuk menentukan tingkat reliabilitas disajikan pada Tabel 1.

Tabel 1. Kategori Tingkat Reliabilitas

Nilai r_{ii}	Kategori
0,00-0,20	Sangat Rendah
0,21-0,40	Rendah
0,41-0,70	Cukup
0,71-0,90	Tinggi
0,91-1,00	Sangat Tinggi

Reliabilitas antar rater (*Interclass Correlation Coefficient* (ICC)) digunakan untuk menilai reliabilitas alat ukur yang telah disusun melalui instrumen rating yang menghasilkan data ordinal (Widhiarso, 2007). Kriteria yang digunakan adalah "…..*for the minimum acceptable value for the reliability coefficient* ≥ *0,75*…..(Shrout & Fleiss, 1979)". Rumus reliabilitas ICC sebagai berikut.

$$r = \frac{MS_{people} - MS_{residual}}{MS_{people} + (df_{people}\ x\ MS_{residual})} \quad (4)$$

$$= \frac{0{,}434 - 0{,}028}{0{,}434 + (1\ x\ 0{,}028)} = 0{,}880$$

dimana r adalah koefisien reliabilitas ICC, MS_{people} mengacu pada *mean square between people*, $MS_{residual}$ adalah *mean square within people residual*, dan df_{people} mengacu pada *the degree of freedom within the people.*

DIF I

Indeks kesukaran (*difficulty index*) (DIF I) soal dihitung sebagai persentase dari jumlah total jawaban yang benar terhadap butir-butir tes (Crocker & Algina, 2008; Kumar et al., 2021; Purwanto, 2014; Quaigrain & Arhin, 2017; Widoyoko, 2014; Zainal, 2009). Persentase berkisar antara 0% dan 100%. Semakin tinggi indeks kesukaran maka semakin mudah *item* dipahami. Kriteria klasifikasi DIF I adalah sebagai berikut: DIF I <30% (terlalu sulit), DIF I antara 30% dan 70% (baik/dapat diterima/rata-rata), DIF I >70% (terlalu mudah) dan DIF I antara 50% dan 60% (sangat baik/ideal) (Christian et al., 2017; Date et al., 2019; Kumar et al., 2021; Quaigrain & Arhin, 2017; Rao et al., 2016; Zainal, 2009). DIF I dihitung dengan menggunakan rumus:

$$p = \frac{R}{T} \quad (5)$$

di mana p adalah indeks kesukaran soal, R adalah jumlah jawaban yang benar, dan T adalah jumlah total jawaban (yang mencakup jawaban benar dan salah (Quaigrain & Arhin, 2017). Lebih lanjut (Aiken, 1979; Hotiu, 2016; Quaigrain & Arhin, 2017) menjelaskan p-value antara 20% dan 90% (baik dan dapat diterima). Diantaranya, *item* dengan nilai p antara 40% dan 60% dianggap sangat baik. Soal dengan nilai p (indeks kesulitan) kurang dari 20% (terlalu sulit) dan lebih dari 90% (terlalu mudah) tidak dapat diterima dan memerlukan modifikasi. Perlu dikonseptualisasikan bahwa nilai p pada dasarnya adalah ukuran perilaku. Kesulitan didefinisikan dalam hal frekuensi relatif dimana peserta tes memilih jawaban yang tepat, bukan dalam hal beberapa fitur intrinsik dari *item* (Thorndike et al., 1991).

DI

Item DI adalah pengukuran sejauh mana suatu tes mampu membedakan individu yang memiliki tingkat kemampuan atau pengetahuan yang berbeda. Semakin tinggi koefisien DI suatu *item* tes semakin mampu membedakan kompetensi individu (Kumar et al., 2021; Quaigrain & Arhin, 2017; Zainal, 2009). Lebih lanjut, menentukan 27% sebagai kelompok atas yang memperoleh jawaban benar dan menentukan 27% sebagai kelompok bawah yang memperoleh jawaban benar. DI dihitung dengan menggunakan rumus:

$$DI = \frac{UG - LG}{n} \quad (6)$$

dimana UG adalah banyaknya mahasiswa kelompok atas yang menjawab soal benar. LG adalah banyaknya mahasiswa kelompok bawah yang menjawab soal benar. n adalah 27% dari jumlah seluruh mahasiswa (Quaigrain & Arhin, 2017).

Kriteria indeks daya pembeda yang dikembangkan oleh Ebel & Garvin (1980) (Garvin & Ebel, 1980; Rao et al., 2016) dijelaskan sebagai berikut: (1) *If DI ≥ 0,40, then the item is functioning satisfactorily. (2) If 0,30 ≤ DI ≤ 0,39, then little or no revision is required. (3) If 0,20 ≤ DI ≤ 0,29, then the item*

is marginal and needs revision. (4) If DI ≤ 0,19, then the item should be eliminated or completely revised. Lebih lanjut (Botterman et al., 2022; Christian et al., 2017; Date et al., 2019; Kelley, 1939; Kumar Namdeo & Dev Rout, 2016; Masters, 1996; Shakurnia et al., 2022) mengklasifikasikan DI dalam 4 kriteria antara lain*: DI ≤ 0,20 (poor), DI 0,21-0,24 (acceptable), DI 0,25-0,34 (good) and DI ≥ 0,35 (excellent)*.

Item dengan indeks negatif harus diperiksa untuk menentukan mengapa nilai negatif diperoleh. Menurut Mehrens dan Lehman (1991), ada berbagai alasan *item* mungkin memiliki daya pembeda yang rendah. Pertama semakin sulit atau mudah *item*, semakin rendah daya pembedanya namun *item* tersebut sering dibutuhkan untuk sampel yang memadai dan representatif dari isi dan tujuan mata kuliah. Kedua tujuan butir dalam kaitannya dengan total tes akan mempengaruhi besarnya daya pembedanya (Rae, 1978).

DE

Distractor Effectiveness (DE) adalah kemampuan jawaban salah untuk mengalihkan perhatian siswa (Date et al., 2019). Distraktor yang tidak berfungsi didefinisikan sebagai opsi dengan frekuensi jawaban <5% (Christian et al., 2017; Kumar et al., 2021). Butir soal yang baik, pengecohnya akan dipilih secara merata oleh individu yang menjawab salah. Sebaliknya *item* yang kurang baik, pengecohnya akan dipilih secara tidak merata (Zainal, 2009). Menulis *Functioning Distractors* (FD) adalah aspek penting dalam menyusun pertanyaan pilihan ganda yang berkualitas. Butir soal pilihan ganda dengan distraktor yang efektif sangat penting untuk memperoleh tes yang valid (Shakurnia et al., 2022).

DE ditentukan berdasarkan jumlah NFD dalam suatu *item* dan berkisar antara 0-100%. Efisiensi pengecoh *item* dinilai sebagai rendah (memiliki 3-4 NFD), sedang (memiliki 1-2 NFD) dan tinggi (memiliki 0 NFD). Setelah mengumpulkan informasi dasar tentang *Non Functioning Distractors* (NFD) dan distraktor yang berfungsi (*Functioning Distractors* (FD), *items* tersebut dikategorikan berdasarkan jumlah NFD; yaitu 3 NFD, 2 NFD, 1 NFD, dan 0 NFD. DE akan menjadi 0% (buruk), 33,3% (sedang), 66,6% (baik) dan 100% (sangat baik) (Christian et al., 2017; Date et al., 2019; Kumar et al., 2021). Lebih lanjut jika jumlah pilihan jawaban memiliki 5 opsi, (Sajjad et al., 2020) mengklasifikasikan DE berdasarkan jumlah NFD sebagai berikut: rendah (memiliki 3-4 NFD) (<50%), sedang (memiliki 1-2 NFD) (50%-75%) dan tinggi (memiliki 0 NFD) (100%).

HASIL DAN PEMBAHASAN

Hasil

Hasil analisis kualitas tes dan butir soal yang diterapkan pada evaluasi formatif mata kuliah Matematika Teknik ditentukan berdasarkan validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efek pengecoh soal. Hasil analisis data menghasilkan uji validitas isi dari 35 butir pertanyaan sebesar 0,89 > nilai indeks V yaitu 0,40 maka butir instrumen dinyatakan valid (Aiken, 1980, 1985; Retnawati, 2016; Setiawan et al., 2021). Lebih lanjut hasil uji validitas konstruk menggunakan analisis korelasi *point-biserial* dari 35 butir pertanyaan sebanyak 34 (97,14%) soal dengan nilai r_{pbis} > r_{tabel} yaitu 0,216 maka butir instrumen dinyatakan valid, 1 (2,86%) soal dengan nilai r_{pbis} < r_{tabel} yaitu 0,216 maka butir instrumen dinyatakan tidak valid (Arikunto, 2019). Mean r_{pbis} sebesar 0,469 dan standar deviasi 0,139. Hasil analisis validitas konstruk disajikan pada Tabel 2 sebagai berikut.

Tabel 2. Ringkasan Hasil Analisis Validitas Konstruk

N (35 Soal)		r_{pbis}			
Valid	Tidak Valid	*Minimum*	*Maximum*	*Mean*	*Std. Deviation*
34	1	0,179	0,686	0,469	0,139

Uji reliabilitas dilakukan menggunakan dua metode pengujian yang pertama menggunakan rumus koefisien korelasi antar kelas (*Interclass Correlation Coefficient*) dan rumus *Kuder-Richardson* (KR21). Hasil uji

reliabilitas soal tes antar rater (2 rater) menunjukkan bahwa nilai koefisien reliabilitas ICC sebesar 0,880 telah memenuhi syarat minimal nilai koefisien reliabilitas ICC yaitu ≥ 0,75 sehingga dapat dimaknai bahwa instrumen reliabel. Lebih lanjut hasil uji reliabilitas menggunakan formula KR21 sebesar 0,876 termasuk pada kategori tingkat reliabilitas tinggi.

Tes terdiri dari 35 *item*, range skor nilai tes 83 mahasiswa antara 7 hingga 33 (dari 35). Nilai rata-rata tes adalah 21,639 dan standar deviasi adalah 7,488. Mediannya adalah 23, nilai skewness sebesar -0,353 dan kurtosis sebesar -1,040. Skewness bernilai negatif dan kurtosis lebih kecil dari nilai referensi (<3). Hal ini menunjukkan kemencengan yang negatif (*negative skew*) dan ekor kurva yang runcing *(platy kurtosis)* (Kothandaraman & Pachaiyappan, 2013).

Kemencengan yang negatif berarti ada sangat banyak mahasiswa yang memiliki skor nilai yang sangat tinggi, sementara hanya sedikit mahasiswa yang memiliki skor nilai rendah. Ekor kurva yang runcing berarti ada banyak jumlah mahasiswa yang memiliki skor nilai yang hampir sama. Hasil analisis deskriptif nilai tes mahasiswa disajikan pada Tabel 3.

Tabel 3. Hasil Analisis Deskriptif Nilai Tes

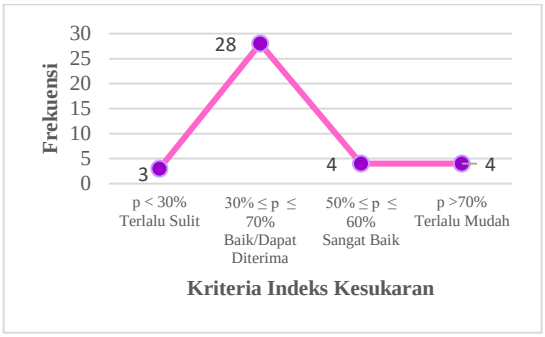
Parameter	Hasil
Jumlah *item*	35
Mean ± SD	21,639 ± 7,488
Median	23
Range skor nilai tes	7-33
Skewness	-0,353
Kurtosis	-1,040

Hasil DIF I menunjukkan bahwa dari 35 soal pilihan ganda, 28 (80%) soal memiliki tingkat kesukaran baik/dapat diterima (DIF I 30-70%), sedangkan tiga (8,57%) soal terlalu sulit (DIF I <30%) dan empat (11,43%) soal terlalu mudah (DIF I >70%). Di antara semua soal pilihan ganda, empat (11,43%) soal pilihan ganda memiliki tingkat kesulitan sangat baik/ideal (DIF I 50-60%). Rata-rata dan standar deviasi (SD) untuk indeks kesulitan (%) 61,83 ± 16,61. Ringkasan hasil analisis DIF I disajikan pada Tabel 4.

Tabel 4. Klasifikasi Soal Menurut Indeks Kesukaran (DIF I)

DIF (p)	*Item* (%)	*Difficulty Index* (Mean±SD)
<30% (Terlalu Sulit)	3 (8,57%)	27,70±1,20
30-70% (Baik/Dapat Diterima)	28 (80%)	61,89±11,12
50-60% (Sangat Baik/Ideal)	4 (11,43%)	56.60±4,04
>70% (Terlalu Mudah)	4 (11,43%)	87,03±0,65

Distribusi frekuensi indeks kesukaran butir tes dari 35 *item* disajikan pada Gambar 1 berikut.



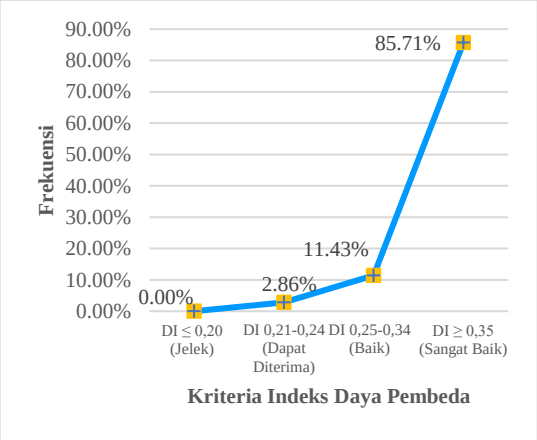
Gambar 1. Distribusi Frekuensi Indeks Kesukaran Butir Tes

Dari total 35 soal pilihan ganda, nol (0%) soal memiliki DI jelek (DI≤0,20), satu (2,86%) soal memiliki DI dapat diterima (0,21≤DI≤0,24), empat (11,43%) soal menunjukkan DI baik (0,25≤DI≤0,34) dan 30 (85,71%) soal menunjukkan DI yang sangat baik (DI ≥0,35). Rata-rata dan standar deviasi (SD) untuk indeks daya beda (%) 53,90 ± 17,26. Ringkasan hasil analisis DI disajikan pada Tabel 5.

Tabel 5. Klasifikasi Soal Menurut Indeks Daya Beda (DI)

DI	*Item* (%)	*Discrimination index* (Mean±SD)
≤0,20 (Jelek)	0 (0%)	0±
0,21-0,24 (Dapat Diterima)	1 (2,86%)	22,73±
0,25-0,34 (Baik)	4 (11,43%)	30,68±2,27
≥0,35 (Sangat Baik)	30 (85,71%)	58,03±14,94

Distribusi frekuensi indeks daya beda tes dari 35 *item* disajikan pada Gambar 2 berikut.



Gambar 2. Distribusi Frekuensi Indeks Daya Beda (DI)

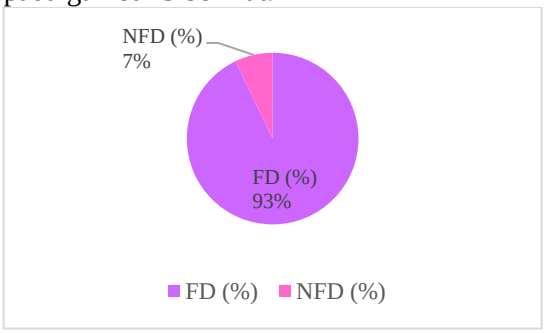
Secara keseluruhan 140 pengecoh untuk 35 *item*, yang memiliki frekuensi pilihan jawaban <5% yaitu 10 (7,14%) adalah NFD dan 130 (92,86%) merupakan pengecoh fungsional (FD). Dari seluruh *item*, satu *item* (2,86%) memiliki DE rendah (<50%), lima *item* (14,29%) memiliki DE sedang (50%-75%), dan 29 *item* (82,86%) memiliki DE tinggi (100%). Rata-rata dan standar deviasi (SD) efektifitas pengecoh (%) 92,86 ± 17,75. Analisis pengecoh dan efektivitas pengecoh (DE) disajikan pada Tabel 6.

Tabel 6. Analisis pengecoh dan efektivitas pengecoh (DE)

Parameter	Jumlah (%)
Jumlah *item*	35
Jumlah pengecoh	140
FD (DE > 5%)	130 (92,86%)
NFD (DE < 5%)	10 (7,14%)
Jumlah *item* dengan 3-4 NFD/0 FD (DE = <50%, rendah)	1 (2,86%)
Jumlah *item* dengan 1-2 NFD/3-2 FD (DE= 50%-75%, sedang)	5 (14,29%)
Jumlah *item* dengan 0 NFD/4 FD (DE=100%, tinggi)	29 (82,86%)
Mean ± SD	92,86 ± 17,75

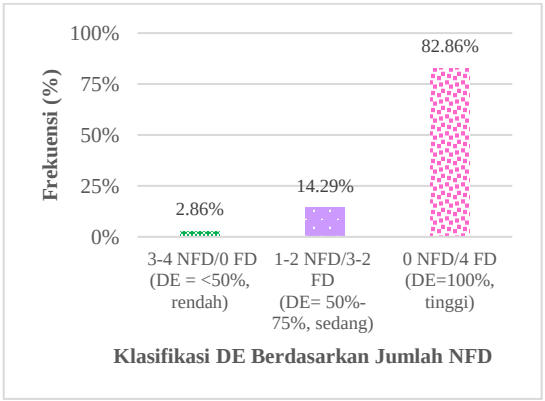
DE, *distractor effectiveness*

Distractor performance (n=140) disajikan pada gambar 3 berikut.



Gambar 3. *Distractor performance* (n=140)

Klasifikasi efisiensi pengecoh DE berdasarkan jumlah NFD berkisar antara 0-100% disajikan pada Gambar 4 berikut.



Gambar 4. Klasifikasi DE Berdasarkan Jumlah NFD (n=140)

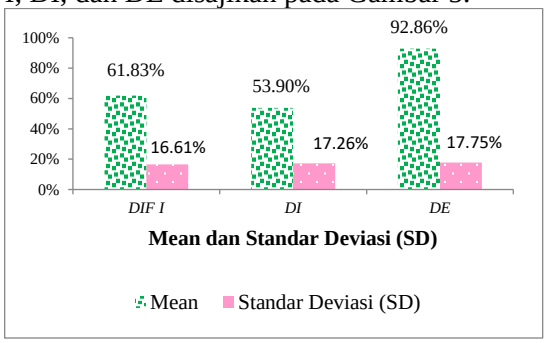
Berdasarkan keseluruhan analisis

menunjukkan bahwa 23 (65,71%) soal pilihan ganda memenuhi ketiga kriteria (tingkat DIF I yang baik/dapat diterima (30-70%) dengan DI baik (0,25-0,34), dan sangat baik (>0,35), serta DE 100%). Nilai rata-rata DIF I, DI, dan DE diberikan pada Tabel 7.

Tabel 7. Nilai rata-rata DIF I, DI dan DE

Parameter Analisis *Item*	DIF I	DI	DE
Mean ± SD (%)	61,83 ± 16,61	53,90 ± 17,26	92,86 ± 17,75

Gambar nilai rata-rata dan standar deviasi DIF I, DI, dan DE disajikan pada Gambar 5.



Gambar 5. Mean dan Standar Deviasi DIF I, DI, serta DE

Mean DIF I 62,83% termasuk pada kriteria indeks kesukaran baik/dapat diterima (30-70%). Mean DI 53,90% termasuk pada kriteria sangat baik (≥0,35). Mean DE 92,86% termasuk pada kriteria tinggi.

Pembahasan

Pengukuran efektif atas pengetahuan dan penerapan integral dalam bidang teknik sipil yang diperoleh mahasiswa merupakan komponen penting dalam Pendidikan Teknik Bangunan. Bentuk butir soal pilihan ganda pada evaluasi formatif Matematika Teknik merupakan alat penilaian yang berguna dalam mengukur ingatan faktual dan jika dibangun dengan cermat dapat menguji keterampilan berpikir tingkat tinggi yang sangat penting bagi lulusan Pendidikan Teknik Bangunan. Metode penilaian harus dievaluasi secara berkala. Mengembangkan strategi penilaian yang tepat adalah bagian penting dalam pengembangan kurikulum (Rao et al., 2016). Sejalan dengan (Pande et al., 2013) yang menjelaskan bahwa penting untuk mengevaluasi soal-soal pilihan ganda untuk melihat seberapa efektif soal-soal tersebut dalam menilai pengetahuan mahasiswa.

Butir soal pilihan ganda tipe jawaban tunggal yang benar adalah alat yang efisien untuk menilai mahasiswa. Desain soal pilihan ganda yang sistematis dan penggunaan soal pilihan ganda yang valid dan reliabel sangat penting jika hasil penilaian dianggap valid dan reliabel. Validitas isi dan konstruk harus ditetapkan oleh tinjauan para ahli dan validitas konstruk harus ditetapkan, berdasarkan DIF I, DI, dan DE (Date et al., 2019). Lebih lanjut reliabilitas ditentukan berdasarkan reliabilitas internal (ICC) dan eksternal (KR21).

Dalam penelitian ini, hasil uji validitas isi dari 35 butir pertanyaan sebesar 0,89 > nilai indeks V yaitu 0,40 maka butir instrumen dinyatakan valid (Aiken, 1980, 1985; Retnawati, 2016; Setiawan et al., 2021). Hasil uji validitas konstruk sebanyak 34 (97,14%) soal dengan nilai r_{pbis} > r_{tabel} yaitu 0,216 maka butir instrumen dinyatakan valid (Arikunto, 2019). Tingkat validitas isi dan konstruk sudah menunjukkan tinggi sehingga soal dapat digunakan. Penelitian serupa yang mendukung temuan penelitian dilakukan oleh (Muaja et al., 2013; Saputra et al., 2022) hasil uji validitas konstruk sebanyak 22 (40%) dari 55 butir soal valid dan sebanyak 20 (80%) dari 25 butir soal valid.

Hasil uji reliabilitas menggunakan formula KR21 sebesar 0,876 termasuk pada kategori tingkat reliabilitas tinggi (Allen & Yen, 1979; Arikunto, 2019; Cronbach, 1951; Kuder & Richardson, 1937; Retnawati, 2017). Hasil uji reliabilitas intrumen antar rater (2 rater) menggunakan aplikasi IBM SPSS version 22 menunjukkan bahwa nilai koefisien reliabilitas ICC sebesar 0,880 telah memenuhi syarat minimal nilai koefisien reliabilitas ICC yaitu ≥ 0,75 sehingga dapat dimaknai bahwa instrumen reliabel (Shrout & Fleiss, 1979). Tingkat reliabilitas yang tinggi pada semua hasil uji menunjukkan bahwa soal memiliki tingkat atau derajat konsistensi yang dapat dipercaya sesuai dengan kriteria yang telah ditetapkan. Penelitian serupa yang mendukung temuan penelitian dilakukan oleh (Christian et al., 2017; Kumar et al., 2021;

Quaigrain & Arhin, 2017; Saputra et al., 2022) hasil uji reliabilitas Kuder-Richardson berturut-turut 0,51; 0,71; 0,77; dan 0,89.

Hasil DIF I menunjukkan bahwa dari 35 soal pilihan ganda, 28 (80%) soal memiliki tingkat kesukaran baik/dapat diterima (DIF I 30-70%), sehingga dapat disimpan sebagai bank soal untuk digunakan di masa depan. sedangkan tiga (8,57%) soal terlalu sulit (DIF I <30%) dan empat (11,43%) soal terlalu mudah (DIF I >70%). Di antara semua soal pilihan ganda, empat (11,43%) soal pilihan ganda memiliki tingkat kesulitan sangat baik/ideal (DIF I 50-60%). Rata-rata dan standar deviasi (SD) untuk indeks kesulitan (%) 61,83 ± 16,61.

Empat soal mudah sedikit direvisi dan disimpan untuk digunakan di masa mendatang guna meningkatkan kepercayaan diri mahasiswa. Lebih lanjut pertanyaan sulit disimpan dan digunakan untuk memilih mahasiswa dengan kemampuan atas. Tiga *item* sulit diperiksa untuk kemungkinan kontroversi bahasa yang membingungkan, untuk setiap kunci yang salah dan setelah revisi disimpan untuk mengembangkan bank soal pilihan ganda (Date et al., 2019; Suryadevara & Bano, 2018).

Temuan DIF I ini lebih mirip dengan penelitian yang dilakukan oleh (Ingale et al., 2017) tentang analisis soal pilihan ganda menunjukkan bahwa 80%, 7%, dan 13% soal pilihan ganda masing-masing dapat diterima, terlalu mudah, dan terlalu sulit. Studi lain yang dilakukan oleh (Kumar et al., 2021) bahwa 82%, 10%, dan 8% soal pilihan ganda masing-masing dapat diterima, terlalu mudah, terlalu sulit. Lebih lanjut (Rao et al., 2016) melaporkan bahwa 85% yang dapat diterima 5% soal mudah dan 10% soal sulit. Penelitian yang dilakukan oleh (Date et al., 2019) bahwa 70% dapat diterima, 20% terlalu mudah dan 10% terlalu sulit.

Dari total 35 soal pilihan ganda, tidak ada (0%) soal memiliki DI jelek (DI≤0,20), satu (2,86%) soal memiliki DI dapat diterima (0,21≤DI≤0,24), empat (11,43%) soal menunjukkan DI baik (0,25≤DI≤0,34) dan 30 (85,71%) soal menunjukkan DI yang sangat baik (DI ≥0,35). Rata-rata dan standar deviasi (SD) untuk indeks daya beda (%) 53,90 ± 17,26. Nilai DI pada penelitian ini sebanding dengan penelitian pada analisis *item* oleh (Kumar et al., 2021) bahwa 80% *item* memiliki daya pembeda yang dapat diterima hingga sangat baik dan 20% memiliki daya pembeda yang buruk. Lebih lanjut penelitian oleh (Date et al., 2019) 77,5% *item* memiliki daya pembeda yang dapat diterima hingga sangat baik dan 22,5% memiliki daya pembeda yang buruk. Nilai DI cenderung lebih rendah untuk tes pengukuran berbagai bidang konten dibandingkan tes yang lebih homogen. *Item* dengan DI rendah sering kali memiliki kata-kata yang ambigu.

Secara keseluruhan, 35 *item* dan 140 pengecoh, sebesar 10 pengecoh (7,14%) NFD tidak dipilih oleh siapa pun dan 130 (92,86%) dari semua pengecoh diklasifikasikan sebagai pengecoh yang berfungsi (FD), artinya pengecohnya dipilih secara merata oleh individu yang menjawab soal. Dari seluruh *item*, satu *item* (2,86%) memiliki DE rendah (<50%), lima *item* (14,29%) memiliki DE sedang (50%-75%), dan 29 *item* (82,86%) memiliki DE tinggi (100%). Rata-rata dan standar deviasi (SD) efektifitas pengecoh (%) 92,86 ± 17,75. Penelitian yang sejalan dilakukan oleh (Kumar et al., 2021) yang melaporkan bahwa 73% FD dan 27% NFD. Penelitian (Date et al., 2019; Quaigrain & Arhin, 2017; Rao et al., 2016; Sajjad et al., 2020; Shakurnia et al., 2022) yang melaporkan bahwa 70%; 63,71%; 89,99%; 74,7%; dan 63,9% *item* adalah FD, 30%; 33,29%; 10,01%; 25,3%; dan 31,6% *item* adalah NFD.

Saat menyusun soal pilihan ganda yang berkualitas baik, aturan utamanya adalah bahwa pengecoh harus masuk akal, yaitu ditempatkan dekat dengan jawaban yang benar yang akan meningkatkan kemungkinan memilih pengecoh ini dibandingkan jawaban yang benar oleh mahasiswa (Date et al., 2019). Sejalan dengan hasil penelitian (Shakurnia et al., 2022) yang menyimpulkan bahwa NFD berbanding terbalik dengan kualitas tes. Namun, soal dengan FD yang lebih banyak akan lebih sulit dan memiliki kekuatan diskriminatif yang lebih besar.

Analisis fungsi pengecoh dan revisi NFD berfungsi sebagai metode penting untuk meningkatkan kualitas butir soal.

Kriteria soal pilihan ganda yang ideal, tingkat DIF I harus memiliki tingkat DIF I yang baik/dapat diterima, dengan DI tinggi dan DE tinggi. Penelitian ini menunjukkan bahwa total 65,71% soal pilihan ganda memenuhi semua tiga kriteria soal pilihan ganda yang ideal. Penelitian lain (Kumar Namdeo & Dev Rout, 2016) melaporkan bahwa 20% soal pilihan ganda memenuhi ketiga kriteria soal pilihan ganda yang ideal. Lebih lanjut nilai rata-rata dan standar deviasi untuk DIF I (baik), DI (sangat baik) dan DE (tinggi) yang menunjukkan bahwa sebagian besar soal pilihan ganda adalah dalam kategori soal pilihan ganda yang baik. Hasil ini sebanding dengan penelitian yang dilakukan (Ingale et al., 2017; Kumar et al., 2021) yang menganalisis 30 dan 90 soal pilihan ganda.

KESIMPULAN

Hasil Penelitian menunjukkan bahwa tingkat validitas isi dan validitas konstruk tinggi. Lebih lanjut tingkat reliabilitas yang tinggi pada semua hasil uji menunjukkan bahwa soal memiliki tingkat atau derajat konsistensi yang dapat dipercaya. Sejumlah besar soal pilihan ganda memiliki tingkat DIF I yang dapat diterima (80%) dan DI sangat baik dalam membedakan siswa berkemampuan tinggi dan rendah (85,71%). Efisiensi pengecoh (DE) (92,86%) dari semua pengecoh yang diklasifikasikan sebagai pengecoh yang berfungsi. Analisis soal bila digabungkan secara teratur dapat membantu mengembangkan bank soal yang sangat berguna, valid, dan reliabel dengan soal pilihan ganda yang dikategorikan ke dalam soal mudah, sulit, dan ideal. Dari keseluruhan hasil analisis maka dapat disimpulkan bahwa butir soal tes integral pada evaluasi formatif Matematika Teknik memiliki kualitas soal yang baik untuk digunakan sebagai alat penilaian kognitif mahasiswa.

SARAN

Penelitian lebih lanjut dapat dilakukan investigasi untuk mengetahui korelasi *item* yang memiliki DIF I, DI, dan DE untuk dapat meningkatkan kualitas butir soal tes sebagai kegiatan penjaminan mutu yang penting dalam mempersiapkan bank soal di setiap departemen. Distraktor yang efektif sangat penting untuk memperoleh tes yang valid, maka penelitian tentang hubungan antara NFD dan parameter analisis *item* lainnya perlu dilakukan di masa depan untuk menemukan jumlah distraktor yang optimal. Di samping itu tidak banyak penelitian yang dilakukan pada aspek kualitatif *item* individual dengan efisiensi pengecoh yang rendah. Maka menganalisis kekurangan penulisan *item* pada *item* efisiensi pengecoh yang rendah hingga sedang dapat dilakukan di masa depan, untuk mendapatkan wawasan tentang kekurangan struktural dalam *item* yang berdampak negatif pada efisiensi pengecoh.

DAFTAR PUSTAKA

Aiken, L. R. (1979). Relationships between the item difficulty and discrimination indexes. *Educational and Psychological Measurement*, *39*(4), 821–824. https://doi.org/10.1177/001316447903900415

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, *40*(4), 955–959. https://doi.org/10.1177/001316448004000419

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, *45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory, Monterey, CA: Brooks/Cole, 1979. *Google Scholar*.

Arikunto, S. (2019). Prosedur penelitian suatu pendekatan praktik. In *Jakarta: Rineka Cipta*. PT Rineka Cipta.

Badan Pusat Statistik. (2016). Potret awal tujuan pembangunan berkelanjutan

(sustainable development goals) di Indonesia. In *Katalog BPS*.

Bates, R. (2014). Improving human resources for health planning in developing economies. *Human Resource Development International*, *17*(1), 88–97. https://doi.org/10.1080/13678868.2013.857509

Bennett, N., Borg, W. R., & Gall, M. D. (1984). Educational Research: An Introduction. *British Journal of Educational Studies*, *32*(3), 274–274. https://doi.org/10.2307/3121583

Botterman, L., De Cock, I., Blot, S. I., & Labeau, S. O. (2022). A knowledge test on pressure injury in adult intensive care patients: Development, validation, and item analysis. *Journal of Tissue Viability*, *31*(4), 718–725. https://doi.org/10.1016/j.jtv.2022.08.007

Bruri Triyono, M., Köhler, T., & Trianingsih, L. (2018). Technical working skills of vocational high school students at the interface between digital workplaces and school. An empirical study about construction engineering drawings in Indonesia. *Communities in New Media: Research on Knowledge Communities in Science, Business, Education and Public Administration - Proceedings of 21th Conference GeNeMe*, 191–200. https://d-nb.info/1233869000/34

Charles Secolsky, D. B. D. (2017). Handbook on Measurement, Assessment, and Evaluation in Higher Education. In C. Secolsky & D. B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (2nd Edition). Routledge. https://doi.org/10.4324/9781315709307

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

Crocker, L., & Algina, James. (2008). Introduction to classical and modern test theory- Procedures for Estimating Reliability. In *Harcourt Brace Jovanovich College*.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

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), 1999–2003. https://doi.org/10.18203/2319-2003.ijbcp20194106

Davis, M. H., & Harden, R. M. (2003). Competency-based assessment: Making it a reality. In *Medical Teacher* (pp. 565–568). https://doi.org/10.1080/0142159032000153842

Garvin, A. D., & Ebel, R. L. (1980). Essentials of Educational Measurement. *Educational Researcher*, *9*(9), 21. https://doi.org/10.2307/1175572

Hansen, J. D., & Dexter, L. (1997). Quality Multiple-Choice Test Questions: Item-Writing Guidelines and an Analysis of Auditing Testbanks. *Journal of Education for Business*, *73*(2), 94–97. https://doi.org/10.1080/08832329709601623

Hattie, J., & Timperley, H. (2007). The power of feedback. In *Review of Educational Research* (Vol. 77, Issue 1). Sage. https://doi.org/10.3102/003465430298487

Hicks, N. A. (2014). Establishing the validity and reliability of the Fairness of Items Tool. *ProQuest Dissertations and Theses*, 282. https://hybridlogin.monash.edu/

Hotiu, A. (2016). *The relationship between item difficulty and discrimination indices in multiple-choice tests in a Physical science course* . Florida

Atlantic University.

Ingale, A. S., A. Giri, P., & Doibale, M. K. (2017). Study on item and test analysis of multiple choice questions amongst undergraduate medical students. *International Journal Of Community Medicine And Public Health*, *4*(5), 1562–1565. https://doi.org/10.18203/2394-6040.ijcmph20171764

Karim, S. A., Sudiro, S., & Sakinah, S. (2021). Utilizing test items analysis to examine the level of difficulty and discriminating power in a teacher-made test. *EduLite: Journal of English Education, Literature and Culture*, *6*(2), 256–269. https://doi.org/10.30659/e.6.2.256-269

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, *30*(1), 17–24. https://doi.org/10.1037/h0057123

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

Kothandaraman, M., & Pachaiyappan, A. (2013). Comparison of Independent Component Analysis techniques for Acoustic Echo Cancellation during Double Talk scenario. *Australian Journal of Basic and Applied Sciences*, *7*(4), 108–113.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160. https://doi.org/10.1007/BF02288391

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*(1), S85–S89. https://doi.org/10.1016/j.mjafi.2020.11.007

Kumar Namdeo, S., & Dev Rout, S. (2016). Assessment of Functional and Nonfunctional Distracter in an Item Analysis. *International Journal of Contemporary Medical Research ISSN*, *3*(1), 1891–1893.

Mardapi, D. (2017). Pengukuran, Penilaian, dan Evaluasi Pendidikan. *Academia Edu*, *7*(2), 107–115.

Masters, K. (1996). *Designing and Managing Multiple Choice Questions*. University of Cape Town, South Africa.

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical Teacher*, *26*(8), 709–712. https://doi.org/10.1080/01421590400013495

Monariska, E.-. (2019). Analisis kesulitan belajar mahasiswa pada materi integral. *Jurnal Analisa*, *5*(1), 9–19. https://doi.org/10.15575/ja.v5i1.4181

Muaja, J. R. T., Setiawan, A., & Mahatma, T. (2013). Uji validitas dan uji reliabilitas menggunakan metode bootstrap. *Prosiding Seminar Nasional Penelitian, Pendidikan Dan Penerapan MIPA, Fakultas MIPA, Universitas Negeri Yogyakarta*, 513–519. https://www.researchgate.net/publication/301558948

Pande, S. S., Pande, S. R., Parate, V. R., Nikam, A. P., & Agrekar, S. H. (2013). Correlation between difficulty & discrimination indices of MCQs in formative exam in Physiology. *South-East Asian Journal of Medical Education*, *7*(1), 45–50. https://doi.org/10.4038/seajme.v7i1.149

Pavlova, M. (2014). TVET as an important factor in country's economic development. *SpringerPlus*. https://doi.org/10.1186/2193-1801-3-S1-K3

Popham, W. J. (1999). Modern educational measurement. *Practical Guidelines for the Education Leader. Michigan: Pearson*, 35–90.

Purwanto. (2014). *Evaluasi Hasil Belajar*. Pustaka Pelajar.

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), 1–11. https://doi.org/10.1080/2331186X.2017.1301013

Rae, G. (1978). Measurement and Evaluation in Psychology and Education (4th Ed.), R. L. Thorndike and E. P. Hagen (Wiley, 1977) pp. viii plus 693, £11.75. *Scottish Educational Review*, *10*(2), 69–71. https://doi.org/10.1163/27730840-01002012

Rao, C., Kishan Prasad, H., Sajitha, K., Permi, H., & Shetty, J. (2016). Item analysis of multiple choice questions: Assessing an assessment tool in medical students. *International Journal of Educational and Psychological Researches*, *2*(4), 201–204. https://doi.org/10.4103/2395-2296.189670

Retnawati, H. (2016). Proving content validity of self-regulated learning scale (The comparison of Aiken index and expanded Gregory index). *Research and Evaluation in Education*, *2*(2), 155. https://doi.org/10.21831/reid.v2i2.11029

Retnawati, H. (2017). Reliabilitas Instrumen Penelitian. *Jurnal Pendidikan Teknik Mesin Unnes*, *12*(1). http://staffnew.uny.ac.id/upload/132255129/pengabdian/8 Reliabilitas3 alhamdulillah.pdf

Rudner, L. M. (1995). Questions to ask when evaluating tests. *Practical Assessment, Research and Evaluation*, *4*(2).

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144. https://doi.org/10.1007/BF00117714

Sajjad, M., Iltaf, S., & Khan, R. A. (2020). Nonfunctional distractor analysis: An indicator for quality of multiple choice questions. *Pakistan Journal of Medical Sciences*, *36*(5), 982–986. https://doi.org/10.12669/pjms.36.5.2439

Salim, E. (2018). Tujuan Pembangunan Berkelanjutan di Indonesia: Konsep, Target dan Strategi Implementasi. 感染症誌, *91*, 399–404.

Saputra, H. D., Purwanto, W., Setiawan, D., Fernandez, D., & Putra, R. (2022). Hasil Belajar Mahasiswa: Analisis Butir Soal Tes. *Edukasi: Jurnal Pendidikan*, *20*(1), 15–27. https://doi.org/10.31571/edukasi.v20i1.3432

Schneider, K. C., & Kerlinger, F. N. (1979). Behavioral Research: A Conceptual Approach. *Journal of Marketing Research*, *16*(4), 599–600. https://doi.org/10.2307/3150838

Setiawan, A. H. (2015). The Contribution of the Vocational Teachers Professional Competence toward Vocational High Schools Performance. *Proceedings of the 3rd UPI International Conference on Technical and Vocational Education and Training*, 1–6. https://doi.org/10.2991/ictvet-14.2015.1

Setiawan, A. H. (2022). Enhancing collaborative mindset by blended online learning platform in a civil engineering education course. *Journal of East Asian Studies*, *20*(3), 1–35. http://petit.lib.yamaguchi-u.ac.jp/28854/files/165507

Setiawan, A. H., & Takaoka, R. (2020). Designing PBL steps in vocational course based on students' readiness and teachers' discussion. In Mashoedah, I. Hidayatulloh, N. Hidayat, & I. W. Djatmiko (Eds.), *Journal of Physics: Conference Series*. IOP. https://doi.org/10.1088/1742-6596/1456/1/012045

Setiawan, A. H., Takaoka, R., Tamrin, A., Roemintoyo, Murtiono, E. S., & Trianingsih, L. (2021). Contribution of collaborative skill toward construction drawing skill for developing vocational course. *Open Engineering*, *11*, 755–771. https://doi.org/10.1515/eng-2021-0073

Setiawan, A. H., Takaoka, R., & Trianingsih, L. (2020). Investigation of Vocational Students' Skills for Determining Learning Experiences on CAD Construction Drawing Course. In H. Mitsuhara, Y. Goda, Y. Ohashi, Ma. M.

T. Rodrigo, J. Shen, N. Venkatarayalu, G. Wong, M. Yamada, & C.-U. Le (Eds.), *IEEE International Conference on Engineering, Technology and Education, TALE* (pp. 748–753). IEEE. https://doi.org/10.1109/TALE48869.2020.9368338

Shakurnia, A., Ghafourian, M., Khodadadi, A., Ghadiri, A., Amari, A., & Shariffat, M. (2022). Evaluating Functional and Non-Functional Distractors and Their Relationship with Difficulty and Discrimination Indices in Four-Option Multiple-Choice Questions. *Education in Medicine Journal*, *14*(4), 55–62. https://doi.org/10.21315/eimj2022.14.4.5

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Sonnadara, R., McQueen, S., Mironova, P., Safir, O., Nousiainen, M., Ferguson, P., Alman, B., Kraemer, W., & Reznick, R. (2013). Reflections on current methods for evaluating skills during joint replacement surgery. *The Bone & Joint Journal*, *95-B*(11), 1445–1449. https://doi.org/10.1302/0301-620X.95B11.30732

Suryadevara, V. K., & Bano, Z. (2018). Item analysis to identify quality multiple choice questions/items in an assessment in Pharmacology of II MBBS students in Guntur Medical College of Andhra Pradesh, India. *International Journal of Basic & Clinical Pharmacology*, *7*(8), 1517–1521. https://doi.org/10.18203/2319-2003.ijbcp20183004

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). Measurement and evaluation in psychology and education, 5th ed. In *Measurement and evaluation in psychology and education, 5th ed.*

Towip, Widiastuti, I., Saputra, T. W., Noviansyah, W., & Trianingsih, L. (2021). TVET Institutions' Perspective on Implementation of Public-Private Partnerships Model in the Southeast Asia Countries. *IOP Conference Series: Earth and Environmental Science*, *1808*(1), 1–9. https://doi.org/10.1088/1742-6596/1808/1/012007

Triyono, M. B., Trianingsih, L., & Nurhadi, D. (2018). Students' employability skills for construction drawing engineering in Indonesia. *World Transactions on Engineering and Technology Education*, *16*(1), 29–35.

Wagner, N., Acai, A., McQueen, S. A., McCarthy, C., McGuire, A., Petrisor, B., & Sonnadara, R. R. (2019). Enhancing Formative Feedback in Orthopaedic Training: Development and Implementation of a Competency-Based Assessment Framework. *Journal of Surgical Education*, *76*(5), 1376–1401. https://doi.org/10.1016/j.jsurg.2019.03.015

Widhiarso, W. (2007). *Mengestimasi Reliabilitas*. https://repository.ugm.ac.id/

Widoyoko, E. P. (2014). *Penilaian Hasil Pembelajaran di Sekolah* (Ratih, Ed.; 1st ed.). Pustaka Pelajar.

Zainal, A. (2009). *Evaluasi pembelajaran prinsip, teknik, prosedur* (P. Latifah, Ed.; 1st ed.). PT. Remaja Rosdakarya.

Zega. (2019). *Penerapan integral dan diferensial pada Mekanika Struktur*. http://repository.uhn.ac.id/handle/123456789/3345