EDUTECH

Jurnal Teknologi Pendidikan

Journal homepage https://ejournal.upi.edu/index.php/edutech

Ujian Melalui Gform, Mudahkah Ditebak? Analisis Item Response Theory 3 Parameter Logistik

Helli Ihsan, Awaluddin Tjalla, Soeprijanto
Penelitian dan Evaluasi Pendidikan, Pascasarjana, Universitas Negeri Jakarta, Jakarta, Indonesia
Email: helli_psi@upi.edu

A B S T R A C T

Quality assessment is essential for measuring students' abilities, but challenges such as guessing in multiple-choice questions can reduce test validity. This study aims to evaluate the quality of test items using the 3 Parameter Logistic (IRT 3PL) Item Response Theory model. A total of 191 students answered 40 multiple-choice questions. The analysis included item-total correlation, point biserial correlation, and IRT 3PL parameters (discrimination, difficulty, guessing). Results showed that most items had good quality (A > 1; p > 0.05), although some items, such as item 6 and 31, required revision. Conclusion: The test instrument is adequate, but items with high guessing probabilities or low discrimination need improvement to enhance the test's validity and reliability.

A B S T R A K

Penilaian berkualitas penting untuk mengukur kemampuan peserta didik, tetapi tantangan seperti menebak dalam soal pilihan ganda dapat mengurangi validitas tes. Penelitian ini bertujuan mengevaluasi kualitas butir soal menggunakan Model Item Response Theory 3 Parameter Logistic (IRT 3PL). Sebanyak 191 mahasiswa mengerjakan 40 soal pilihan ganda. Analisis meliputi korelasi item-total, point biserial, dan parameter IRT 3PL (daya pembeda, tingkat kesulitan, peluang menebak). Hasilnya sebagian besar butir memiliki kualitas baik (A > 1; p > 0,05), meskipun beberapa butir seperti item 6 dan 31 memerlukan revisi. Kesimpulan: Instrumen tes memadai, tetapi butir dengan peluang menebak tinggi atau daya pembeda

A R T I C L E I N F O

Article History:
Submitted/Received 8 Jan 2025
First Revised 5 Feb 2024
Accepted 16 Feb 2025
First Available online 20 Feb 2025
Publication Date 20 Feb 2025

Keyword:
Gform, tes formatif, daya tebak, item response theory, mahasiswa

rendah perlu diperbaiki untuk meningkatkan validitas dan reliabilitas tes. © 2023 Teknologi Pendidikan UPI		

DOI: http://dx.doi.org/10.xxxxx/ijost.v2i2
p- ISSN 2528-1410 e- ISSN 2527-8045

1. PENDAHULUAN

Meningkatkan pembelajaran melalui penilaian berkualitas merupakan aspek penting dalam praktik pendidikan modern. Penilaian memainkan peran integral dalam proses pendidikan dengan memberikan wawasan kritis tentang pemahaman dan kemajuan siswa. Penilaian yang dirancang dengan baik tidak hanya mengukur pengetahuan peserta didik, tetapi juga membantu pendidik mengevaluasi efektivitas strategi pengajaran mereka. Dengan memahami kekuatan dan kelemahan siswa, pendidik dapat menyesuaikan metode pengajaran untuk memenuhi kebutuhan belajar yang beragam, sehingga meningkatkan hasil pendidikan. Pada akhirnya, penilaian berkualitas menjadi jembatan antara pengajaran dan pembelajaran, menciptakan lingkungan di mana siswa termotivasi untuk terlibat secara mendalam dengan kurikulum dan mencapai potensi akademik mereka.

Penilaian memiliki peran penting dalam proses pembelajaran dengan membentuk strategi pengajaran dan pembelajaran untuk meningkatkan hasil pendidikan. Penilaian berfungsi sebagai instrumen utama bagi pendidik untuk mengidentifikasi area di mana siswa unggul atau membutuhkan dukungan tambahan, sehingga memungkinkan intervensi pengajaran yang lebih terarah. Menurut Stevens dan Levi, penggunaan rubrik dalam penilaian membantu menyampaikan umpan balik yang efektif, yang sangat penting untuk mendorong pembelajaran siswa (Stevens & Levi, 2023). Selain itu, penilaian formatif dan umpan balik berperan dalam menciptakan lingkungan yang mendorong siswa untuk terlibat aktif dengan materi pembelajaran, yang pada akhirnya meningkatkan kinerja akademik mereka (Irons & Elkington, 2021). Dengan mengintegrasikan penilaian yang selaras dengan tujuan pembelajaran, pendidik dapat memastikan metode pengajaran mereka responsif terhadap kebutuhan beragam siswa, sehingga menciptakan pengalaman pendidikan yang lebih dinamis dan efektif.

Penilaian berkualitas dalam pendidikan ditandai oleh validitas, reliabilitas, dan keadilan, yang masing-masing memiliki peran penting dalam memastikan bahwa penilaian mencerminkan pemahaman dan kinerja siswa secara akurat. Validitas mengacu pada sejauh mana penilaian mengukur apa yang dimaksudkan untuk diukur, sehingga memberikan informasi yang bermakna tentang hasil belajar siswa (Beerkens, 2020). Reliabilitas, di sisi lain, berkaitan dengan konsistensi hasil penilaian dari waktu ke waktu dan di antara penilai yang berbeda, yang sangat penting untuk menjaga kepercayaan dalam proses penilaian (Beerkens, 2020). Keadilan memastikan bahwa penilaian bebas dari bias dan adil bagi semua siswa, sehingga memberikan representasi kemampuan mereka secara akurat, terlepas dari latar belakang atau gaya belajar mereka (Stevens & Levi, 2023). Karakteristik-karakteristik ini mendasari terciptanya penilaian yang tidak hanya mendukung tujuan pendidikan tetapi juga berkontribusi pada lingkungan belajar yang lebih inklusif dan efektif.

Menerapkan penilaian yang efektif membutuhkan kombinasi praktik berbasis bukti dan pendekatan inovatif yang memenuhi kebutuhan belajar yang beragam. Salah satu strateginya adalah integrasi penilaian formatif, yang terbukti secara signifikan meningkatkan keterlibatan siswa dan hasil belajar melalui pemberian umpan balik berkelanjutan (Irons & Elkington, 2021). Penggunaan tes formatif dalam format pilihan ganda memiliki peran penting dalam lingkungan pendidikan karena memberikan wawasan berharga bagi guru tentang pembelajaran dan pemahaman siswa. Penilaian ini membantu memantau kemajuan siswa secara berkelanjutan, memungkinkan pendidik menyesuaikan pengajaran dan memberikan umpan balik tepat waktu untuk

DOI: https://doi.org/10.17509/e.v24i1.77831
P-ISSN 0852-1190 E-ISSN 2502-0781

meningkatkan hasil pembelajaran. Dengan menggunakan format pilihan ganda, guru dapat mengevaluasi berbagai bidang materi secara efisien, memastikan siswa memahami konsep dasar sekaligus mengembangkan keterampilan berpikir kritis. Selain itu, sifat terstruktur dari tes ini memungkinkan penilaian yang konsisten dan objektif, menciptakan lingkungan belajar yang adil.

Tes formatif pilihan ganda menawarkan berbagai manfaat dalam pendidikan. Salah satu manfaat utama adalah penyediaan umpan balik langsung, memungkinkan siswa untuk segera mengetahui kekuatan dan kelemahan mereka, sehingga mendukung pembelajaran berkelanjutan (Ryan et al., 2020). Tes ini juga mudah dinilai, mengurangi beban administratif guru dan memungkinkan mereka lebih fokus pada kegiatan pengajaran (Subramaniam et al., 2019). Selain itu, tes ini memungkinkan cakupan materi yang luas secara efisien, memastikan evaluasi pemahaman siswa secara komprehensif (Kusairi, 2020). Tes pilihan ganda juga mendukung pembelajaran dan retensi siswa dengan menyediakan peluang latihan yang terstruktur dan berulang. Dengan format ini, siswa dapat secara sistematis memahami materi, memperkuat pengetahuan, dan meningkatkan keterampilan berpikir kritis melalui aplikasi dalam berbagai konteks (Say et al., 2022).

Meskipun bermanfaat, tes pilihan ganda juga memiliki tantangan, seperti desain soal yang buruk yang dapat menyebabkan ambiguitas dan mengurangi validitas penilaian (Yılmaz et al., 2020). Tes ini juga cenderung mendorong pembelajaran permukaan, di mana siswa lebih fokus pada hafalan daripada pemahaman mendalam. Selain itu, format ini sering kali kurang efektif dalam menilai keterampilan berpikir tingkat tinggi (Bulut, 2021). Untuk mengatasi tantangan ini, pendidik dapat merancang soal yang mendorong keterampilan berpikir tingkat tinggi, seperti analisis dan evaluasi (Çekiç & Bakla, 2021). Menggunakan variasi soal, seperti soal berbasis skenario, juga dapat meningkatkan validitas tes dan mencerminkan tujuan pembelajaran dengan lebih baik (Kusairi, 2020). Selain itu soal pilihan ganda juga ada peluang siswa hanya mengandalkan tebakan untuk menjawab. Implementasi tes formatif pilihan ganda yang efektif memerlukan penyesuaian dan harus selaras dengan tujuan instruksional untuk memastikan relevansi dan memberikan umpan balik yang bermakna (Cosi et al., 2020). Salah satu teknik dalam pelaksanaan ujian adalah menggunakan gform sebagai lembar soal sekaligus melakukan penyekoran.

Penggunaan Google Forms untuk ujian semakin populer di dunia pendidikan seiring meningkatnya integrasi teknologi dalam penilaian. Alat ini menawarkan fleksibilitas dan kemudahan dalam format ujian daring, menjadikannya pilihan menarik bagi pendidik. Namun, meskipun memiliki banyak keunggulan, Google Forms juga menimbulkan pertanyaan terkait efektivitas dan keterbatasannya bagi siswa dan guru. Penting untuk mengkaji manfaat dan tantangannya guna memahami dampaknya dalam pendidikan.

Google Forms memiliki beberapa keunggulan dalam pelaksanaan ujian, terutama karena aksesibilitas, kemampuan penilaian otomatis, dan efisiensinya. Siswa dapat mengakses ujian dari mana saja dengan koneksi internet, memungkinkan partisipasi di berbagai lingkungan (Elsalem et al., 2020). Fitur penilaian otomatis secara signifikan mengurangi beban kerja pendidik, memungkinkan mereka untuk lebih fokus pada tugas pengajaran daripada tugas administratif. Selain itu, efisiensi waktu yang ditawarkan Google Forms memungkinkan guru untuk dengan cepat mendapatkan hasil dan memberikan umpan balik tepat waktu kepada siswa, yang penting untuk menjaga momentum akademik. Sebagai contoh, selama pandemi COVID-19, institusi berhasil menggunakan Google

DOI: https://doi.org/10.17509/e.v24i1.77831
P-ISSN 0852-1190 E-ISSN 2502-0781

Forms untuk mengelola ujian, menunjukkan kemampuannya beradaptasi dalam berbagai konteks pendidikan (Alea et al., 2020).

Namun, penggunaan Google Forms untuk ujian juga memiliki kelemahan, terutama terkait masalah teknis, integritas akademik, dan keterbatasan dalam pemberian umpan balik. Tantangan teknis seperti masalah koneksi internet dapat menghambat siswa dalam mengakses atau menyelesaikan ujian, yang dapat meningkatkan stres dan menyebabkan ketidaksetaraan akademik (Mahyoob, 2020). Selain itu, kemungkinan kecurangan menjadi perhatian utama, karena kurangnya pengawasan dalam ujian daring membuka peluang terjadinya ketidakjujuran akademik yang dapat merusak kredibilitas penilaian (Elsalem et al., 2021). Format Google Forms juga sering membatasi pemberian umpan balik yang dipersonalisasi, yang esensial untuk pembelajaran dan pengembangan siswa. Tantangan-tantangan ini menekankan perlunya pendidik dan institusi untuk mempertimbangkan implikasi penggunaan Google Forms dalam ujian, dengan memastikan langkah-langkah yang tepat diambil untuk menjaga integritas dan efektivitas proses penilaian.

Menebak dalam ujian adalah fenomena umum yang memainkan peran penting dalam penilaian pendidikan. Hal ini terutama terjadi dalam ujian pilihan ganda, di mana siswa sering menggunakan strategi ini ketika tidak yakin akan jawaban yang benar. Signifikansi menebak tidak hanya terletak pada potensinya untuk memengaruhi hasil ujian, tetapi juga pada implikasinya terhadap pemahaman perilaku siswa dan desain penilaian. Dalam konteks penilaian akademik, menebak sering digunakan sebagai pendekatan strategis oleh siswa ketika mereka tidak yakin akan jawaban yang benar, terutama dalam format pilihan ganda. Strategi ini menjadi relevan ketika struktur tes memungkinkan tebakan terdidik, di mana siswa dapat mengeliminasi opsi yang tidak mungkin untuk meningkatkan peluang memilih jawaban yang benar (Photopoulos & Triantis, 2022). Praktik menebak bukan hanya sekadar memilih secara acak, melainkan sering melibatkan proses pengambilan keputusan yang terencana berdasarkan pengetahuan parsial dan pengalaman mengerjakan ujian (Lee, 2019). Selain itu, desain soal pilihan ganda dapat memengaruhi efektivitas menebak, karena format tertentu mungkin secara tidak sengaja memberikan petunjuk halus pada jawaban yang benar (Lions et al., 2022). Dengan demikian, menebak mencakup lebih dari sekadar peluang, melainkan merupakan interaksi kompleks antara pengetahuan siswa, desain ujian, dan faktor psikologis yang muncul selama penilaian.

Menebak dalam ujian dapat secara signifikan mengubah hasil, dengan potensi untuk membantu atau merugikan kinerja siswa. Di satu sisi, menebak terdidik dapat memberikan hasil positif dengan memungkinkan siswa menjawab pertanyaan yang mungkin akan dibiarkan kosong, terutama jika mereka dapat mengeliminasi opsi yang tidak mungkin (Lions et al., 2022). Di sisi lain, sifat acak dari menebak dapat menyebabkan kesalahan, sehingga menurunkan reliabilitas skor ujian dan mungkin tidak merepresentasikan pengetahuan siswa yang sebenarnya (Soland & Kuhfeld, 2019). Faktor-faktor seperti desain ujian, kompleksitas soal, dan karakteristik individu siswa, termasuk pengalaman mereka dalam mengerjakan tes dan kondisi psikologisnya, dapat memengaruhi keberhasilan menebak (Lee, 2019). Pemahaman terhadap dinamika ini penting bagi pendidik dan perancang tes, karena menyoroti perlunya mempertimbangkan implikasi menebak terhadap validitas dan keadilan penilaian.

DOI: https://doi.org/10.17509/e.v24i1.77831
P-ISSN 0852-1190 E-ISSN 2502-0781

Tindakan menebak dalam ujian membawa banyak implikasi psikologis dan pendidikan, yang memengaruhi pola pikir siswa dan proses pendidikan secara keseluruhan. Dari perspektif psikologis, tekanan untuk menebak dapat meningkatkan tingkat stres, karena siswa mungkin merasa cemas saat harus membuat pilihan acak dalam batasan waktu (Soland & Kuhfeld, 2019). Stres ini dapat memengaruhi kemampuan pengambilan keputusan, yang berpotensi menghasilkan tebakan yang terburu-buru atau salah dan tidak mencerminkan pengetahuan atau kemampuan siswa secara benar (Lee, 2019). Secara pendidikan, menebak dapat menjadi alat pembelajaran yang mendorong keterampilan seperti berpikir kritis dan penilaian risiko, karena siswa harus mempertimbangkan opsi mereka dan membuat tebakan yang terinformasi berdasarkan informasi parsial (St. Hilaire et al., 2024). Dengan demikian, meskipun menebak menghadirkan tantangan, ia juga menawarkan peluang untuk mengembangkan keterampilan kognitif yang berharga di luar lingkungan ujian.

Model Item Response Theory Tiga Parameter Logistik: Teknik menganalisis soal beradasarkan tebakan

Model Item Response Theory Tiga Parameter Logistik (IRT 3PL) memiliki peran penting dalam penilaian pendidikan dengan secara efektif mengatasi masalah menebak dalam skenario ujian. Model ini menawarkan kerangka kerja yang kuat untuk memahami dan menginterpretasi respons siswa melalui tiga parameter utama: kesulitan, diskriminasi, dan menebak. Parameter menebak yang dimasukkan dalam model ini membedakannya dari model respons butir lainnya, memungkinkan untuk mempertimbangkan kemungkinan jawaban benar akibat menebak secara acak. Dalam penilaian pendidikan, pengukuran kemampuan siswa sering kali dipersulit oleh potensi menebak, yang dapat mengaburkan hasil tes. Model IRT 3PL memberikan metode canggih untuk mengurangi distorsi ini, meningkatkan akurasi dan reliabilitas skor tes, yang sangat penting untuk pengambilan keputusan dan penilaian pendidikan yang tepat.

Model IRT 3PL adalah kerangka statistik yang digunakan dalam penilaian pendidikan untuk mengevaluasi kemungkinan jawaban benar terhadap butir tes berdasarkan tiga parameter: kesulitan, diskriminasi, dan menebak. Parameter kesulitan (b) merepresentasikan tingkat kemampuan yang dibutuhkan untuk memiliki peluang 50% menjawab benar. Parameter diskriminasi (a) menunjukkan sejauh mana sebuah butir dapat membedakan individu dengan tingkat kemampuan yang berbeda. Sementara itu, parameter menebak (c) secara khusus memperhitungkan kemungkinan jawaban benar akibat menebak secara acak, skenario yang tidak ditangani oleh model yang lebih sederhana seperti model Rasch (Almaleki & Alomrany, 2021). Dengan menggabungkan ketiga parameter ini, model IRT 3PL meningkatkan ketepatan estimasi kemampuan, memastikan penilaian pendidikan lebih mencerminkan kemampuan siswa secara akurat (Rios, 2022).

Model IRT 3PL mencakup tiga parameter utama yang memiliki peran penting dalam proses penilaian. Parameter pertama, kesulitan (b), berfungsi untuk mengidentifikasi tingkat kemampuan yang diperlukan agar seseorang memiliki peluang 50% untuk menjawab suatu butir soal dengan benar. Parameter ini membantu memastikan bahwa setiap butir soal dapat disesuaikan dengan tingkat keterampilan yang berbeda, sehingga memberikan evaluasi yang lebih tepat terhadap kemampuan peserta tes. Selanjutnya, parameter diskriminasi (a) berfungsi untuk mengukur sejauh mana sebuah butir mampu membedakan individu dengan tingkat kemampuan yang berbeda. Hal ini penting untuk memastikan bahwa soal-soal dalam tes dapat secara akurat mencerminkan perbedaan kompetensi di antara peserta. Terakhir, parameter menebak (c) secara khusus dirancang untuk memperhitungkan kemungkinan jawaban benar yang berasal dari menebak secara acak. Dengan memasukkan parameter ini, model IRT 3PL dapat mengurangi distorsi yang

DOI: https://doi.org/10.17509/e.v24i1.77831
P-ISSN 0852-1190 E-ISSN 2502-0781

disebabkan oleh menebak, sehingga menghasilkan estimasi kemampuan yang lebih akurat. Keseluruhan parameter ini bekerja secara sinergis untuk memberikan analisis yang lebih mendalam dan andal terhadap performa peserta tes, seperti yang dijelaskan oleh Robitzsch (2022). Parameter menebak (c) adalah fitur unik model ini, yang secara khusus mengatasi dampak menebak dengan menyesuaikan estimasi kemampuan siswa berdasarkan kemungkinan jawaban benar yang tidak berasal dari pengetahuan sebenarnya.

Efektivitas model IRT 3PL dalam mengelola menebak terbukti melalui berbagai penelitian dan aplikasi. Penelitian menunjukkan bahwa parameter menebak (c) meningkatkan akurasi estimasi kemampuan, terutama dalam lingkungan tes berisiko tinggi di mana menebak dapat mendistorsi hasil (Rios, 2022). Model ini telah diterapkan dalam tes standar, di mana mempertimbangkan menebak mengurangi kemungkinan salah klasifikasi kemampuan siswa akibat respons acak yang benar (Almaleki & Alomrany, 2021).

Dibandingkan dengan model lain seperti model Rasch satu parameter atau model logistik dua parameter, model IRT 3PL unggul karena memasukkan parameter menebak (c). Model Rasch, misalnya, tidak dapat menangani menebak, yang dapat mengurangi akurasi estimasi kemampuan (Robitzsch, 2022). Sementara itu, model dua parameter mencakup diskriminasi tetapi tidak mempertimbangkan menebak. Meskipun model 3PL lebih kompleks dalam estimasi parameter, kemampuannya untuk mengatasi menebak menjadikannya alat yang berharga dalam meningkatkan keakuratan penilaian pendidikan.

Pertanyaan penelitian:

(i) Bagaimana tingkat kesulitan soal?
(ii) Bagaimana daya diskriminasi soal?
(iii) Bagaimana kemudahan soal ditebak?

2. METODE

2.1. Struktur Soal

Tes formatif disusun sesuai tujuan pembelajaran dari materi Probabilitas, Random Sampling, Pengenalan Statistik Inferensial, dan Distribusi T. Tes terdiri dari 40 butir soal dengan format soal pilihan ganda dengan 4 opsi jawaban. Setiap butir soal dibuat untuk mengukur kemampuan mahasiswa dalam memahami konsep teoretis sekaligus penerapan praktis dari setiap topik.

Table 1. Kisi-Kisi Soal

Materi Kuliah **Probabilitas**	Kemampuan Kognitif	Jumlah soal
Mengetahui bahwa probabilitas suatu peristiwa berada di antara 0 dan 1.	Remember	1
Mengetahui bahwa probabilitas tidak dapat memberi tahu kita apa yang akan terjadi dalam jangka pendek	Remember	1
Melakukan estimasi probabilitas	Understanding	1
Menjelaskan teorema penjumlahan dan teorema perkalian probabilitas	Understanding	1
Memahami apa yang dimaksud dengan distribusi probabilitas (frekuensi relatif teoretis)	Understanding	1

DOI: https://doi.org/10.17509/e.v24i1.77831
P-ISSN 0852-1190 E-ISSN 2502-0781

Menjelaskan mengapa distribusi binomial adalah contoh dari distribusi probabilitas	Understanding	1
	Jumlah Soal	**6**
Random Sampling		
Mendefinisikan dan menjelaskan Random Sampling	Understanding	2
Menggunakan Tabel Angka Acak	Understanding	3
Memahami Distribusi Sampling Acak	Understanding	1
Menjelaskan Central Limit Theorem	Understanding	1
Gunakan distribusi sampling dari rata-rata X untuk menjawab pertanyaan tentang probabilitas memperoleh rata-rata X dalam rentang nilai tertentu	Understanding	2
	Jumlah soal	**9**
Pengantar Statistik Inferensial:		
Menguji Hipotesis Single Mean (Z)	Understanding	2
Merumuskan hipotesis nol dan hipotesis alternatif;	Understanding	2
Memahami apa yang dimaksud dengan tingkat signifikansi dan bagaimana cara menggunakannya untuk menguji hipotesis;	Understanding	1
Mengetahui kapan harus menggunakan uji satu sisi (one-tailed test) atau uji dua sisi (two-tailed test).	Understanding	3
Mengetahui cara menguji hipotesis tentang rata-rata populasi ketika simpangan baku populasi diketahui (z).	Understanding	3
Memahami asumsi-asumsi dalam pengujian hipotesis tentang rata-rata suatu populasi.	Understanding	4
	Jumlah soal	**15**
Distribusi T		
Memahami karakteristik distribusi t	Understanding	1
Memahami konsep derajat kebebasan (degrees of freedom);	Understanding	1
Mengetahui cara menguji hipotesis tentang rata-rata populasi ketika simpangan baku populasi tidak diketahui (t)	Understanding	2
Memahami perbedaan antara tingkat signifikansi dan nilai p (p-value)	Understanding	1
Memahami sifat distribusi sampling acak dari $(X - Y)$	Understanding	2
Menghitung $s_{\bar{x}-\bar{y}}$ (perkiraan standar error dari selisih dua rata-rata) dan t untuk menguji hipotesis tentang perbedaan antara dua rata-rata independen		
Memahami sifat distribusi sampling acak dari (X − Y) dan D untuk sampel yang tergantung;	Understanding	2
Menghitung $s_{\bar{x}-\bar{y}}$ untuk sampel dependen dan menjelaskan bagaimana hal ini berbeda dari simpangan baku estimasi perbedaan untuk sampel independen.	Understanding	1
		10

2.2. Peserta tes

Partisipan atau testee terdiri dari mahasiswa semester 1 Program Studi Psikologi Universitas Pendidikan Indonesia yang mengambil mata kuliah Statistik. Sebanyak 191 mahasiswa berpartisipasi dalam tes ini, testee tidak dipilih tetapi mereka mengikuti karena tes ini adalah bagian dari evaluasi pembelajaran.

2.3. Analisis Data

Tahap 1: Analisis Korelasi Item-Total

Langkah awal dalam proses analisis melibatkan pengujian koherensi internal setiap butir soal dengan menggunakan korelasi item-total. Metode ini bertujuan untuk menilai sejauh mana setiap butir memiliki hubungan yang kuat dengan skor total, sebagai

DOI: https://doi.org/10.17509/e.v24i1.77831
P-ISSN 0852-1190 E-ISSN 2502-0781

indikator konsistensi internal keseluruhan instrumen. Butir soal dengan nilai korelasi item-total kurang dari 0,3 dianggap tidak mendukung keandalan tes dan harus dihapus. Langkah ini menghasilkan sekumpulan butir soal yang memiliki kontribusi nyata terhadap pengukuran yang dituju.

Tahap 2: Analisis Item IRT 3PL

Proses ini bertujuan untuk menghasilkan butir-butir soal yang tidak hanya sesuai dengan asumsi model IRT 3PL tetapi juga memberikan informasi yang akurat dan bermakna dalam pengukuran kemampuan mahasiswa.

a. Parameter item 3PL

Setelah melewati seleksi awal, butir-butir soal dianalisis menggunakan model logistik tiga parameter (3PL). Dalam model ini, tiga parameter utama diperhitungkan: kesulitan butir (difficulty), daya pembeda (discrimination), dan peluang menjawab benar secara acak (guessing).

1) Kesulitan butir dievaluasi untuk memastikan bahwa setiap butir berada dalam rentang ideal (-2 hingga +2). Butir dengan kesulitan terlalu ekstrem dieliminasi karena dianggap kurang informatif.
2) Daya pembeda digunakan untuk menentukan sejauh mana butir dapat membedakan peserta dengan kemampuan tinggi dan rendah. Butir dengan nilai daya pembeda rendah (misalnya, kurang dari 0,2) dianggap tidak efektif dan dihapus.
3) Peluang menjawab benar secara acak memastikan butir tidak mudah ditebak. Butir dengan nilai parameter guessing terlalu tinggi (misalnya, lebih dari 0,3) juga dieliminasi.

b. Analisis Pengujian Kesesuaian (Fit Test) Item

1) Hitung Indeks Kesesuaian Item

Kesesuaian suatu item terhadap model biasanya diuji menggunakan statistik Chi-Square Item Fit: Mengukur perbedaan antara respons yang diharapkan (prediksi model) dan respons yang diamati. Nilai $p > 0,05$ menunjukkan bahwa item fit dengan model.

2) Evaluasi Grafik ICC (Item Characteristic Curve)

Grafik ICC menunjukkan hubungan antara tingkat kemampuan peserta dan probabilitas menjawab benar untuk suatu item. Item yang fit akan menunjukkan kurva yang sesuai dengan prediksi model. Item dengan deviasi signifikan (misalnya, titik data nyata terlalu jauh dari kurva prediksi) dianggap misfit.

c. Eliminasi atau Revisi Item Misfit

Butir-butir yang terbukti misfit dieliminasi jika tidak dapat diperbaiki atau direvisi untuk mengatasi masalah, seperti perbaikan kalimat soal atau pengurangan ambiguitas. Item yang fit tetapi masih memiliki tingkat tebakan yang tinggi juga akan dieliminasi untuk benar-benar memberikan item yang tidak mudah ditebak oleh testee yang kemampuannya rendah.

Tahap 3: Hasil Akhir

Hasil dari pengujian ini adalah sekumpulan butir yang memenuhi syarat fit model IRT 3PL, sehingga memberikan informasi yang lebih akurat dan dapat diandalkan dalam mengukur kemampuan peserta.

3. HASIL DAN PEMBAHASAN

DOI: https://doi.org/10.17509/e.v24i1.77831
P-ISSN 0852-1190 E-ISSN 2502-0781

3.1. Hasil

3.1.1 Korelasi item-total

Tabel 2 menggambarkan hasil analisis statistik terhadap butir-butir soal berdasarkan nilai korelasi item-total dan korelasi point biserial. Kedua indikator tersebut digunakan untuk mengevaluasi kualitas butir dalam mengukur konsistensi internal dan daya pembeda masing-masing soal.

Korelasi item-total menunjukkan sejauh mana setiap butir soal berkontribusi terhadap skor total. Nilai yang lebih tinggi menunjukkan bahwa butir tersebut memiliki hubungan kuat dengan keseluruhan instrumen. Dalam tabel ini, butir 8 memiliki nilai korelasi item-total tertinggi sebesar 0,5144, mengindikasikan kontribusi yang sangat baik terhadap konsistensi internal. Sebaliknya, butir 6 memiliki nilai korelasi item-total terendah sebesar 0,2538, yang menunjukkan hubungan lemah dengan skor total.

Sementara itu, korelasi point biserial mengukur kemampuan butir dalam membedakan peserta dengan kemampuan tinggi dan rendah. Nilai yang lebih tinggi menunjukkan daya pembeda yang lebih baik. Butir 28 memiliki nilai korelasi point biserial tertinggi sebesar 0,5615, menunjukkan kemampuan membedakan yang sangat baik. Sebaliknya, butir 31 memiliki nilai terendah sebesar 0,2718, mengindikasikan daya pembeda yang lemah.

Secara keseluruhan, sebagian besar butir dalam tabel memiliki kualitas yang memadai dengan nilai korelasi di atas 0,3, yang merupakan batas minimal untuk mendukung konsistensi dan daya pembeda. Namun, beberapa butir seperti butir 6 dan 31 memerlukan evaluasi lebih lanjut untuk menentukan apakah perlu diperbaiki atau dieliminasi guna meningkatkan kualitas instrumen.

Table 1. Item dengan Korelasi item-total yang baik.

Item	Korelasi Item-Total	Item	Point Biserial
2	0.4147	22	0.4342
4	0.3784	23	0.4293
5	0.3010	24	0.4123
8	0.5144	25	0.3563
9	0.3773	26	0.3619
10	0.4605	27	0.3232
12	0.4532	28	0.5615
13	0.4208	29	0.5077
16	0.4110	30	0.4129
17	0.5015	32	0.3470
18	0.3072	34	0.4038
19	0.4418	36	0.3923
20	0.3772	38	0.3314
21	0.3754	40	0.3666

3.1.2 Parameter 3PL dan item fit

Analisis 3PL dilakukan beberapa kali dimana setiap kali analisis item-item yang tidak fit dihapus dari perangkat item. Item yang tidak fit adalah item yang memiliki p=value <0,05. Analisis pertama item 27 dihapus, sedangkan dalam analisis kedua ada dua item yang dihapus yaitu item 8 dan item 38.

Table 1. 3 Parameter Item dan Item Fit.

DOI: https://doi.org/10.17509/e.v24i1.77831
P-ISSN 0852-1190 E-ISSN 2502-0781

Kode item	A (SE)	B (SE)	C (SE)	Chi-Kuadrat	df	p-value
2	1.17 (0.27)	-1.20 (0.35)	0.22 (0.10)	8.9436	9	0.4425
4	1.20 (0.34)	-1.21 (0.74)	0.50 (0.23)	8.3294	9	0.5013
9	1.91 (0.50)	-1.07 (0.44)	0.50 (0.19)	9.4559	10	0.4895
10	1.72 (0.39)	-1.13 (0.48)	0.50 (0.21)	4.6339	9	0.8650
12	1.60 (0.32)	-1.15 (0.25)	0.21 (0.09)	11.9937	12	0.4462
13	1.55 (0.31)	-0.40 (0.23)	0.20 (0.08)	7.8331	10	0.6451
17	1.73 (0.32)	-0.95 (0.21)	0.19 (0.09	8.0142	9	0.5327
19	2.14 (0.38)	-0.31 (0.17)	0.25 (0.07)	20.9234	12	0.0515
20	1.50 (0.39)	-1.01 (0.50)	0.50 (0.18	6.3939	9	0.6999
23	1.78 (0.41)	-1.18 (0.46)	0.50 (0.21)	15.3484	10	0.1199
24	1.20 (0.25)	-0.93 (0.30)	0.20 (0.09)	9.6984	10	0.4673
25	1.23 (0.28)	-1.19 (0.33)	0.22 (0.10)	9.7909	11	0.5493
26	1.62 (0.39)	-0.77 (0.29)	0.27 (0.10)	20.1174	12	0.0649
28	2.02 (0.34)	-1.05 (0.17)	0.16 (0.08)	8.7444	10	0.5565
29	2.13 (0.35)	-0.75 (0.17)	0.18 (0.08)	8.1484	9	0.5193
30	1.24 (0.28)	-1.59 (0.36)	0.21 (0.10)	6.0645	10	0.8098
32	1.00 (0.24)	-0.18 (0.34)	0.20 (0.09)	5.0398	10	0.8885
34	2.08 (0.39)	-0.18 (0.17)	0.24 (0.07)	9.9264	10	0.4470
36	1.60 (0.40)	-0.38 (0.27)	0.27 (0.09)	10.0733	9	0.3446
40	1.40 (0.32)	-0.34 (0.27)	0.22 (0.09)	11.3832	11	0.4117

Tabel ini menyajikan hasil analisis butir soal menggunakan model IRT 3 Parameter Logistic (3PL), yang melibatkan tiga parameter utama: daya pembeda (A), tingkat kesulitan (B), dan peluang menebak benar (C). Setiap parameter dilengkapi dengan standar error (SE) untuk menunjukkan tingkat ketidakpastian estimasi. Selain itu, tabel juga mencantumkan nilai Chi-kuadrat, derajat kebebasan (df), dan p-value yang digunakan untuk mengevaluasi sejauh mana model IRT cocok dengan data empiris.

Parameter A (daya pembeda) menunjukkan kemampuan butir soal dalam membedakan peserta dengan kemampuan tinggi dan rendah. Nilai A yang lebih tinggi menunjukkan daya pembeda yang lebih baik. Dalam tabel ini, butir seperti item 19 dan 29 memiliki nilai A tertinggi, masing-masing sebesar 2.14 dan 2.13, yang menunjukkan kemampuan membedakan yang sangat baik. Sebaliknya, item 18 dan 32 memiliki nilai A yang lebih rendah, masing-masing 0.97 dan 1.00, menunjukkan daya pembeda yang lebih lemah dibandingkan butir lainnya.

Parameter B (tingkat kesulitan) menunjukkan posisi butir soal di sepanjang skala kemampuan peserta. Nilai negatif B menunjukkan butir yang lebih mudah, sedangkan nilai positif menunjukkan butir yang lebih sulit. Pada tabel ini, sebagian besar butir memiliki nilai B negatif, seperti item 4 dan 30 dengan nilai B masing-masing -1.21 dan -1.59, mengindikasikan butir tersebut relatif mudah. Sementara itu, butir seperti item 19 dan 34 dengan nilai B mendekati nol menunjukkan tingkat kesulitan sedang.

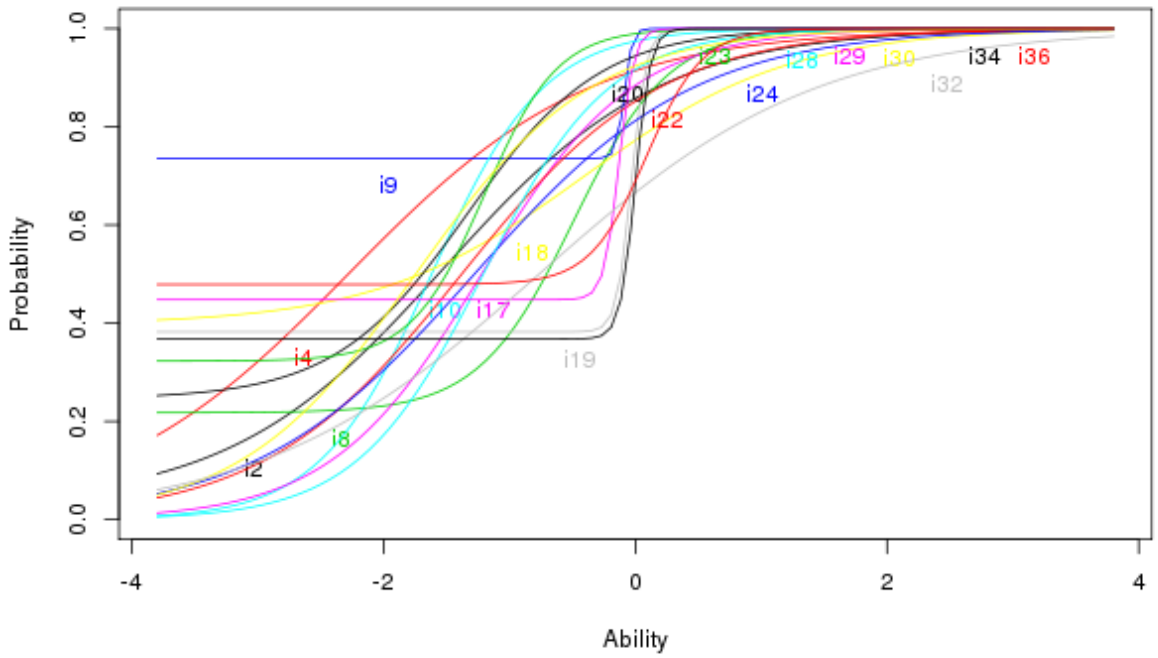
Parameter C (guessing) merepresentasikan peluang peserta dengan kemampuan rendah menjawab benar karena faktor tebak-menebak. Nilai C bervariasi dalam tabel, dengan beberapa butir seperti item 28 dan 29 memiliki nilai C rendah di sekitar 0.16-0.18, menunjukkan peluang menebak yang kecil. Sebaliknya, butir seperti item 4, 9, 10, dan 20 memiliki nilai C sebesar 0.50, menunjukkan potensi tebak-menebak yang lebih tinggi.

DOI: https://doi.org/10.17509/e.v24i1.77831
P-ISSN 0852-1190 E-ISSN 2502-0781

Nilai Chi-kuadrat dan p-value digunakan untuk menguji sejauh mana butir soal cocok dengan asumsi model 3PL. Nilai p-value yang lebih besar dari 0.05 menunjukkan bahwa model cocok dengan data empiris. Mayoritas butir dalam tabel memiliki p-value yang signifikan di atas 0.05, seperti item 10 dengan p-value 0.8650 dan item 32 dengan p-value 0.8885, menunjukkan kecocokan yang baik dengan model. Namun, beberapa butir seperti item 18 dan 19 memiliki p-value mendekati batas signifikan, yang menunjukkan perlunya evaluasi lebih lanjut terhadap butir tersebut.

Secara keseluruhan, tabel ini memberikan gambaran tentang kualitas butir soal dari segi daya pembeda, tingkat kesulitan, peluang tebak-menebak, dan kesesuaian dengan model IRT. Sebagian besar butir memiliki karakteristik yang baik, meskipun beberapa butir dengan daya pembeda rendah atau peluang menebak tinggi memerlukan perhatian lebih lanjut.

3.1.3 Item Characteristic Curves



Gambar 1. Item Characteristic Curves

Grafik ini menggambarkan Item Characteristic Curve (ICC) dari beberapa butir soal menggunakan model IRT 3 Parameter Logistic (3PL). Model ini mengevaluasi tiga parameter utama, yaitu tingkat kesulitan butir, daya pembeda, dan peluang menjawab benar secara acak. Sumbu horizontal (X) mewakili tingkat kemampuan peserta (theta) yang berkisar dari -4 hingga +4, sementara sumbu vertikal (Y) menunjukkan probabilitas peserta menjawab benar terhadap suatu butir soal, yang berkisar antara 0 hingga 1.

Setiap kurva pada grafik ini mewakili satu butir soal yang dilabeli, seperti i2, i4, i9, i20, dan lainnya. Bentuk dan posisi kurva memberikan informasi penting mengenai karakteristik butir tersebut. Posisi kurva sepanjang sumbu X menunjukkan tingkat kesulitan suatu butir; semakin ke kanan letaknya, semakin tinggi tingkat kesulitannya,

DOI: https://doi.org/10.17509/e.v24i1.77831
P-ISSN 0852-1190 E-ISSN 2502-0781

yang berarti hanya peserta dengan kemampuan tinggi yang memiliki probabilitas lebih besar untuk menjawab benar. Sebaliknya, kurva yang berada di sisi kiri grafik menunjukkan butir yang lebih mudah. Misalnya, butir seperti i22 dan i32 tampak bergeser jauh ke kanan, mengindikasikan bahwa soal ini cukup sulit, sementara butir seperti i4 dan i8 berada di sisi kiri, menunjukkan soal yang lebih mudah.

Kemiringan kurva mencerminkan daya pembeda butir. Kurva yang curam, seperti pada butir i20 dan i24, menunjukkan bahwa butir tersebut memiliki daya pembeda yang baik karena mampu membedakan peserta dengan kemampuan tinggi dan rendah secara efektif. Di sisi lain, kurva yang lebih landai, seperti pada butir i9, mengindikasikan daya pembeda yang rendah karena perbedaannya tidak signifikan antara peserta dengan kemampuan rendah dan tinggi. Selain itu, elevasi di bagian bawah kurva, terutama di sisi kiri grafik, menunjukkan adanya peluang peserta menjawab benar secara acak (guessing). Misalnya, pada butir i9 dan i12, terlihat bahwa probabilitas menjawab benar sudah cukup tinggi bahkan untuk peserta dengan kemampuan rendah, mengindikasikan adanya guessing parameter yang besar.

Dari grafik ini dapat disimpulkan bahwa sebagian besar butir memiliki karakteristik yang berbeda-beda. Butir dengan kurva yang tajam di tengah grafik menunjukkan kualitas yang baik karena mampu memberikan informasi yang akurat untuk membedakan peserta berdasarkan tingkat kemampuan mereka. Namun, terdapat butir-butir tertentu yang perlu dievaluasi lebih lanjut, seperti butir dengan guessing tinggi, daya pembeda rendah, atau tingkat kesulitan yang ekstrem, agar instrumen tes dapat memberikan hasil yang lebih akurat dan andal.

3.2. Pembahasan

Hasil analisis korelasi item-total menunjukkan bahwa sebagian besar butir soal dalam instrumen memiliki kualitas yang memadai dalam hal konsistensi internal dan daya pembeda. Nilai korelasi item-total tertinggi dicapai oleh butir 8 (0,5144), sedangkan nilai terendah terdapat pada butir 6 (0,2538). Nilai korelasi item-total di atas 0,3, sebagaimana diusulkan oleh Ebel dan Frisbie (1991), menunjukkan bahwa butir tersebut dapat dianggap memiliki kontribusi positif terhadap reliabilitas instrumen. Namun, nilai korelasi item-total yang lebih rendah pada butir 6 mengindikasikan bahwa butir ini memerlukan evaluasi lebih lanjut untuk menentukan relevansinya.

Korelasi point biserial juga digunakan untuk menilai daya pembeda butir. Nilai tertinggi ditemukan pada butir 28 (0,5615), yang menunjukkan bahwa butir ini sangat baik dalam membedakan peserta dengan kemampuan tinggi dan rendah. Sebaliknya, butir 31 dengan korelasi point biserial terendah (0,2718) perlu ditinjau ulang, karena nilai di bawah 0,3 menunjukkan kemampuan membedakan yang kurang optimal (Mueller, 1986). Butir 6 dan 31 dapat dipertimbangkan untuk direvisi atau dihapus untuk meningkatkan konsistensi internal dan daya pembeda instrumen.

Analisis 3 Parameter Logistic (3PL) memberikan wawasan mendalam mengenai kualitas butir berdasarkan daya pembeda (A), tingkat kesulitan (B), dan peluang menebak benar (C). Sebagian besar butir menunjukkan daya pembeda yang baik, seperti item 19 (A = 2,14) dan item 29 (A = 2,13), yang sesuai dengan kriteria Baker (2001), di mana nilai A > 1 dianggap sangat baik. Sebaliknya, butir seperti item 32 dengan nilai A =

DOI: https://doi.org/10.17509/e.v24i1.77831
P-ISSN 0852-1190 E-ISSN 2502-0781

1,00 menunjukkan daya pembeda yang lebih rendah dan memerlukan evaluasi lebih lanjut.

Dari sisi tingkat kesulitan (B), sebagian besar butir memiliki nilai negatif, seperti item 4 (B = -1,21), yang menunjukkan bahwa butir ini relatif mudah. Namun, butir seperti item 34 dengan nilai B mendekati nol (B = -0,18) menunjukkan tingkat kesulitan sedang. Keseimbangan dalam tingkat kesulitan instrumen penting untuk memastikan bahwa tes dapat mencakup rentang kemampuan peserta (Hambleton, Swaminathan, & Rogers, 1991).

Nilai peluang menebak benar (C) berkisar antara 0,16 hingga 0,50, dengan beberapa butir seperti item 28 dan 29 memiliki nilai rendah (C = 0,16 dan 0,18), menunjukkan bahwa peluang tebak-menebak pada butir ini cukup kecil. Sebaliknya, nilai C sebesar 0,50 pada beberapa butir (misalnya, item 9 dan 10) mengindikasikan bahwa butir tersebut lebih rentan terhadap tebak-menebak, yang dapat mengurangi validitas tes. Secara keseluruhan, mayoritas butir memiliki nilai p-value > 0,05, seperti item 10 (p = 0,8650) dan item 32 (p = 0,8885), yang menunjukkan kecocokan dengan model IRT. Namun, item seperti 19 (p = 0,0515) memerlukan perhatian khusus karena mendekati batas signifikan. Butir dengan daya pembeda rendah (A < 1,5), tingkat kesulitan ekstrem, atau nilai guessing tinggi (C ≥ 0,50) perlu direvisi untuk memastikan validitas dan reliabilitas instrumen.

Item Characteristic Curve (ICC) memberikan gambaran visual tentang bagaimana tiap butir soal berfungsi berdasarkan kemampuan peserta. Kurva dengan kemiringan curam, seperti pada butir i20 dan i24, menunjukkan daya pembeda yang baik, sementara kurva yang landai, seperti pada butir i9, mengindikasikan daya pembeda yang rendah. Kurva yang bergeser ke kanan, seperti pada i22 dan i32, menunjukkan tingkat kesulitan yang tinggi, sedangkan kurva di sisi kiri, seperti i4 dan i8, mencerminkan soal yang lebih mudah.

Adanya elevasi di sisi kiri kurva pada butir seperti i9 dan i12 menunjukkan peluang guessing yang tinggi, yang sesuai dengan nilai parameter C pada butir tersebut. Hal ini mengindikasikan bahwa peserta dengan kemampuan rendah memiliki peluang yang tidak seharusnya untuk menjawab benar, sehingga perlu dilakukan modifikasi atau eliminasi terhadap butir dengan karakteristik ini. Tes harus mempertahankan butir dengan kurva tajam dan proporsi kesulitan yang beragam. Butir dengan peluang guessing tinggi atau daya pembeda rendah sebaiknya diperbaiki untuk meningkatkan akurasi tes dalam mengukur kemampuan peserta.

4. SIMPULAN

Sebagian besar butir soal dalam instrumen memiliki kualitas yang baik berdasarkan analisis konsistensi internal, daya pembeda, tingkat kesulitan, dan kecocokan dengan model IRT 3 Parameter Logistic (3PL). Nilai korelasi item-total dan point biserial menunjukkan kontribusi yang signifikan dari mayoritas butir, meskipun beberapa butir, seperti item 6 dan 31, memerlukan revisi.

Hasil analisis parameter 3PL mengindikasikan bahwa mayoritas butir memiliki daya pembeda baik, tingkat kesulitan beragam, dan peluang guessing rendah, dengan nilai p-value menunjukkan kecocokan yang baik terhadap model. Analisis Item Characteristic Curves (ICC) juga mendukung bahwa instrumen ini efektif dalam membedakan peserta berdasarkan tingkat kemampuan.

DOI: https://doi.org/10.17509/e.v24i1.77831
P-ISSN 0852-1190 E-ISSN 2502-0781

Namun, revisi diperlukan untuk beberapa butir dengan daya pembeda rendah atau peluang guessing tinggi guna meningkatkan validitas dan reliabilitas instrumen secara keseluruhan.

5. PERNYATAAN PENULIS

Penulis menyatakan bahwa tidak terdapat konflik kepentingan terkait penerbitan artikel ini. Penulis menegaskan bahwa naskah artikel bebas dari plagiarisme.

6. REFERENSI

Almaleki, D. A., & Alomrany, A. G. (2021). The effect of methods of estimating the ability on the accuracy and items parameters according to 3PL model. International Journal of Computer Science & Network Security, 21(7), 93–102. https://koreascience.kr/article/JAKO202123563866608.page

Baker, F. B. (2001). The Basics of Item Response Theory. ERIC Clearinghouse.

Beerkens, M. (2020). Evidence-based policy and higher education quality assurance: progress, pitfalls and promise. In *Impact Evaluation of Quality Management in Higher Education* (pp. 38–53). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9780429293276-4/evidence-based-policy-higher-education-quality-assurance-progress-pitfalls-promise-maarja-beerkens

Bulut, O. (2021). Beyond multiple-choice with digital assessments. *ELearn*, *2021*(Special Issue), 1–10. https://dl.acm.org/doi/abs/10.1145/3472394

Çekiç, A., & Bakla, A. (2021). A review of digital formative assessment tools: Features and future directions. *International Online Journal of Education and Teaching*, *8*(3), 1459–1485. https://eric.ed.gov/?id=EJ1308016

Cosi, A., Voltas, N., Lázaro-Cantabrana, J. L., Morales, P., Calvo, M., Molina, S., & Quiroga, M. Á. (2020). Formative assessment at university through digital technology tools. *Profesorado, Revista de Currículum y Formación Del Profesorado*, *24*(1), 164–183. https://revistaseug.ugr.es/index.php/profesorado/article/view/9314

Ebel, R. L., & Frisbie, D. A. (1991). Essentials of Educational Measurement (5th ed.). Prentice Hall.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory. Sage Publications.

Hill, J., & West, H. (2020). Improving the student learning experience through dialogic feed-forward assessment. *Assessment & Evaluation in Higher Education*. https://www.tandfonline.com/doi/shareview/10.1080/02602938.2019.1608908

Irons, A., & Elkington, S. (2021). Enhancing learning through formative assessment and feedback. In *taylorfrancis.com*. Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9781138610514/enhancing-learning-formative-assessment-feedback-alastair-irons-sam-elkington

Kusairi, S. (2020). A web-based formative feedback system development by utilizing isomorphic multiple choice items to support physics teaching and learning. *Journal of Technology and Science Education*, *10*(1), 117–126. https://eric.ed.gov/?id=EJ1247013

Mueller, D. J. (1986). Measuring Social Attitudes: A Handbook for Researchers and Practitioners. Teachers College Press.

DOI: https://doi.org/10.17509/e.v24i1.77831
P-ISSN 0852-1190 E-ISSN 2502-0781

Rios, J. A. (2022). Assessing the accuracy of parameter estimates in the presence of rapid guessing misclassifications. *Educational and Psychological Measurement*, *82*(1), 122–150. https://journals.sagepub.com/doi/abs/10.1177/00131644211003640

Robitzsch, A. (2022). Four-parameter guessing model and related item response models. *Mathematical and Computational Applications*, *27*(6), 95. https://www.mdpi.com/2297-8747/27/6/95

Ryan, A., Judd, T., Swanson, D., Larsen, D. P., Elliott, S., Tzanetos, K., & Kulasegaram, K. (2020). Beyond right or wrong: More effective feedback for formative multiple-choice tests. *Perspectives on Medical Education*, *9*, 307–313. https://link.springer.com/article/10.1007/S40037-020-00606-Z

Say, R., Visentin, D., Cummings, E., Carr, A., & King, C. (2022). Formative online multiple-choice tests in nurse education: An integrative review. *Nurse Education in Practice*, *58*, 103262. https://www.sciencedirect.com/science/article/pii/ S147159532 1002 985

Stevens, D. D., & Levi, A. J. (2023). Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning. In *taylorfrancis.com*. Routledge. https://www.taylorfrancis.com/books/mono/10.432 4/97810034 45432/introduction-rubrics-dannelle-stevens-antonia-levi

Subramaniam, A. V. V., Gupta, R., Singh, N., & Ravishankar, M. (2019). Usefulness of Multiple Choice Question-Based Online Formative Assessments for Determination of Item Statistics. *Journal of Research in Medical Education & Ethics*, *9*(2), 119–125. https://www.indianjournals.com/ijor.aspx?target=ijor:jrmee&volume=9&issue=2 &article=007

Yılmaz, F. G. K., Ustun, A. B., & Yılmaz, R. (2020). Investigation of pre-service teachers' opinions on advantages and disadvantages of online formative assessment: an example of online multiple-choice exam. *Journal of Teacher Education and Lifelong Learning*, *2*(1), 1–8. https://dergipark.org.tr/en/pub/tell/issue/52517/718396

Alea, L. A., Fabrea, M. F., Roldan, R. D. A., & Farooqi, A. Z. (2020). Teachers' Covid-19 awareness, distance learning education experiences and perceptions towards institutional readiness and challenges. *International Journal of Learning, Teaching and Educational Research*, *19*(6), 127–144. http://ijlter.net/index.php/ijlter/ article/view/308

Elsalem, L., Al-Azzam, N., Jum'ah, A. A., & Obeidat, N. (2021). Remote E-exams during Covid-19 pandemic: A cross-sectional study of students' preferences and academic dishonesty in faculties of medical sciences. *Annals of Medicine and Surgery*, *62*, 326–333. https://journals.lww.com/annals-of-medicine-and-surgery/fulltext/2021/02 000/Remote_E_exams_during_Covid_19_pandemic__A.69.aspx

Elsalem, L., Al-Azzam, N., Jum'ah, A. A., Obeidat, N., Sindiani, A. M., & Kheirallah, K. A. (2020). Stress and behavioral changes with remote E-exams during the Covid-19 pandemic: A cross-sectional study among undergraduates of medical sciences. *Annals of Medicine and Surgery*, *60*, 271–279. https://www.sciencedirect.com/ science/article/pii/S2049080120304131

Guangul, F. M., Suhail, A. H., Khalit, M. I., & Khidhir, B. A. (2020). Challenges of remote assessment in higher education in the context of COVID-19: a case study of Middle East College. *Educational Assessment, Evaluation and Accountability*, *32*, 519–535. https://link.springer.com/article/10.1007/s11092-020-09340-w

Mahyoob, M. (2020). Challenges of e-Learning during the COVID-19 Pandemic Experienced by EFL Learners. *Arab World English Journal (AWEJ)*, *11*(4). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3652757

DOI: https://doi.org/10.17509/e.v24i1.77831
P-ISSN 0852-1190 E-ISSN 2502-0781

Selvaraj, A., Radhin, V., Nithin, K. A., Benson, N., & Mathew, A. J. (2021). Effect of pandemic based online education on teaching and learning system. *International Journal of Educational Development*, *85*, 102444. https://www.sciencedirect.com/science/article/pii/S0738059321000973

DOI: https://doi.org/10.17509/e.v24i1.77831
P-ISSN 0852-1190 E-ISSN 2502-0781