Inovasi Kurikulum - p-ISSN 1829-6750 & e-ISSN 2798-1363
Volume 23 No 2 (2026) 409-426

Inovasi Kurikulum

https://ejournal-hipkin.or.id/index.php/jik

Psychometric properties of the formative assessment test for the statistics course

Helli Ihsan[1], Wardani Rahayu[2], Riyan Arthur[3]
[1,2,3] Universitas Negeri Jakarta, Jakarta, Indonesia
helli_psi@upi.edu[1]

ABSTRACT

This study is motivated by the importance of formative assessment in enhancing students' understanding and engagement in Psychological Statistics learning, as well as the limited availability of psychometrically sound instruments. The study aims to develop a structured formative assessment instrument that is valid and reliable, and to analyze item characteristics using Classical Test Theory (CTT) and Item Response Theory (IRT). This research employed an instrument development design involving 191 first-year psychology students. The initial instrument consisted of 40 multiple-choice items constructed based on learning outcomes and validated through expert judgment using Aiken's V. Data analysis was conducted in stages, including item-total correlation analysis, IRT 1PL for item difficulty, and IRT 2PL for both difficulty and discrimination parameters. The results showed that after the selection process, 29 items met the psychometric criteria, with difficulty levels predominantly in the easy-to-moderate range, good discrimination indices, and high reliability. These findings indicate that the developed instrument accurately and consistently measures students' understanding and can be effectively used as a formative assessment tool to support statistics learning in higher education.

ARTICLE INFO

Article History:
Received: 25 Sep 2025
Revised: 5 Apr 2026
Accepted: 11 Apr 2026
Publish online: 23 Apr 2026

Keywords:
classical test theory; formative assessment; item response theory; psychological statistics

Open access
Inovasi Kurikulum is a peer-reviewed open-access journal.

ABSTRAK

Penelitian ini dilatarbelakangi oleh pentingnya asesmen formatif dalam meningkatkan pemahaman dan keterlibatan mahasiswa pada pembelajaran Statistik Psikologi, serta keterbatasan ketersediaan instrumen yang memiliki kualitas psikometris yang teruji. Penelitian ini bertujuan untuk mengembangkan instrumen asesmen formatif terstruktur yang valid dan reliabel, serta menganalisis karakteristik butir menggunakan pendekatan Classical Test Theory (CTT) dan Item Response Theory (IRT). Metode yang digunakan adalah penelitian pengembangan instrumen dengan melibatkan 191 mahasiswa semester pertama Program Studi Psikologi. Instrumen awal terdiri dari 40 butir soal pilihan ganda yang disusun berdasarkan capaian pembelajaran dan divalidasi melalui expert judgment menggunakan Aiken's V. Analisis data dilakukan secara bertahap melalui uji korelasi item-total, analisis IRT 1PL untuk parameter kesulitan, serta IRT 2PL untuk parameter kesulitan dan daya diskriminasi. Hasil penelitian menunjukkan bahwa setelah proses seleksi, diperoleh 29 butir soal yang memenuhi kriteria psikometris dengan tingkat kesulitan yang dominan mudah hingga sedang, daya diskriminasi yang baik, serta reliabilitas yang tinggi. Temuan ini menunjukkan bahwa instrumen yang dikembangkan mampu mengukur pemahaman mahasiswa secara akurat dan konsisten, serta dapat digunakan sebagai alat asesmen formatif yang efektif untuk mendukung pembelajaran statistik di pendidikan tinggi.

Kata Kunci: *asesmen formatif; classical test theory; item response theory; statistik psikologi*

How to cite (APA 7)
Ihsan, H., Rahayu, W., & Arthur, R. (2026). Psychometric properties of the formative assessment test for the statistics course. *Inovasi Kurikulum, 23*(2), 409-426.

Peer review
This article has been peer-reviewed through the journal's standard double-blind peer review, where both the reviewers and authors are anonymised during review.

https://doi.org/10.64014/jik.v23i2.325

INTRODUCTION

Terdapat banyak alasan mengapa penggunaan asesmen formatif penting dalam pendidikan perguruan tinggi. Alasan utama adalah bahwa asesmen berkelanjutan membantu mahasiswa memahami proses belajar dan kinerja akademik mereka. Dalam bidang psikologi, penguasaan teori dan metode statistik merupakan kompetensi esensial. Asesmen formatif membantu mahasiswa meningkatkan kemampuan analisis serta pemahaman terhadap metode statistik (Menéndez *et al*., 2019; Munaroh, 2024). Kesadaran terhadap kebutuhan belajar memungkinkan mahasiswa mengembangkan strategi pembelajaran yang lebih efektif. Selain itu, asesmen formatif mendorong partisipasi aktif mahasiswa, yang penting dalam memahami konsep statistik dalam penelitian psikologi (Leenknecht *et al*., 2021; Stanja *et al*., 2023). Dengan demikian, asesmen formatif berperan penting dalam membantu mahasiswa memahami dan mengaplikasikan pengetahuan statistik. Penggunaan tes formatif dan tugas terkait mendukung keterlibatan aktif mahasiswa dalam pembelajaran. Asesmen formatif memberikan umpan balik cepat yang memungkinkan mahasiswa menyesuaikan strategi belajar mereka (Chen *et al*., 2021; Dayal, 2021). Melalui kuis rutin dan tugas berbasis data, pemahaman statistik mahasiswa dapat meningkat secara bertahap. Selain itu, asesmen formatif membantu mahasiswa mengembangkan keterampilan dalam mengidentifikasi dan menerapkan konsep statistik. Proses yang dilakukan secara berulang juga memperdalam pemahaman terhadap istilah dan penerapannya dalam penelitian psikologi. Hal ini menunjukkan bahwa asesmen formatif berkontribusi terhadap pengembangan kompetensi akademik mahasiswa (Pai, 2025).

Asesmen formatif juga meningkatkan keterlibatan mahasiswa dalam pembelajaran. Interaksi dalam perkuliahan menjadi lebih intensif melalui pemberian umpan balik. Dalam pembelajaran psikologi, keterampilan statistik diperlukan untuk memahami dan menganalisis data penelitian (Lu & Cutumisu, 2022). Penerapan asesmen formatif membuat pembelajaran lebih menarik dan berdampak pada peningkatan kinerja akademik. Asesmen ini memberikan informasi mengenai capaian mahasiswa dan target pembelajaran yang harus dicapai. Keterlibatan yang berkelanjutan membantu mahasiswa memproses informasi secara lebih efektif, khususnya dalam tugas analisis statistik. Secara keseluruhan, asesmen formatif dapat meningkatkan kinerja dan keterlibatan mahasiswa serta mendukung pembelajaran aktif (Morris *et al*., 2021; Saekoko *et al*., 2025). Selain itu, asesmen formatif membantu dosen dalam menyesuaikan strategi pembelajaran berdasarkan kebutuhan mahasiswa. Data yang diperoleh dari asesmen formatif dapat digunakan untuk meningkatkan kualitas pengajaran dan pemahaman mahasiswa terhadap materi (Hanefar *et al*., 2022). Dengan demikian, penerapan asesmen formatif secara sistematis mendukung keberhasilan pembelajaran statistik dalam pendidikan psikologi.

Penelitian terdahulu menunjukkan bahwa asesmen formatif berkontribusi terhadap peningkatan hasil belajar mahasiswa. Asesmen formatif terbukti efektif ketika disertai umpan balik yang bermakna (Wafubwa, 2020). Dalam pendidikan psikologi, penguasaan statistik merupakan kompetensi inti. Mahasiswa yang mengikuti asesmen formatif secara rutin menunjukkan peningkatan pemahaman konsep statistik (Menéndez *et al*., 2019). Umpan balik yang cepat juga membantu mahasiswa menyesuaikan strategi belajar dan meningkatkan kemampuan analisis data (Lubis & Setiawan, 2025). Keterlibatan mahasiswa menjadi aspek penting dalam asesmen formatif. Penerapan asesmen formatif meningkatkan partisipasi aktif dan memperkuat pemahaman konsep statistik (Nisrina *et al*., 2025). Selain itu, pembelajaran aktif yang didukung asesmen formatif meningkatkan motivasi mahasiswa dalam mempelajari materi kompleks (Lase, 2024). Dari perspektif pengajaran, asesmen formatif memberikan informasi bagi dosen untuk menyesuaikan strategi pembelajaran. Data asesmen dapat digunakan untuk meningkatkan efektivitas pengajaran dan pemahaman mahasiswa (Dianti *et al*., 2025).

https://doi.org/10.64014/jik.v23i2.325

Meskipun manfaat asesmen formatif telah banyak diteliti, masih terdapat beberapa kesenjangan. Sebagian besar penelitian dilakukan di negara Barat sehingga konteks Indonesia belum banyak dikaji. Selain itu, penelitian cenderung berfokus pada dampak jangka pendek, sedangkan pengaruh jangka panjang terhadap kompetensi statistik masih terbatas. Variasi jenis asesmen formatif juga belum banyak dibandingkan secara mendalam, khususnya terkait efektivitasnya dalam meningkatkan keterampilan statistik mahasiswa psikologi. Oleh karena itu, penelitian ini berupaya mengkaji efektivitas asesmen formatif terstruktur dalam konteks pendidikan tinggi di Indonesia. Penelitian ini memberikan kontribusi dalam tiga aspek. Pertama, penelitian ini mengkaji asesmen formatif dalam konteks pendidikan psikologi di Indonesia. Kedua, penelitian ini mengintegrasikan keterlibatan mahasiswa dan kompetensi statistik sebagai variabel utama. Ketiga, penelitian ini menggunakan pendekatan longitudinal untuk melihat perkembangan mahasiswa selama satu semester.

Berdasarkan latar belakang dan kesenjangan penelitian yang telah diuraikan, penelitian ini bertujuan untuk menjawab permasalahan utama mengenai apakah penerapan asesmen formatif terstruktur dapat meningkatkan kompetensi statistik dan keterlibatan mahasiswa psikologi dibandingkan dengan pendekatan evaluasi konvensional. Selain itu, penelitian ini juga mengkaji apakah terdapat perbedaan efektivitas di antara berbagai jenis asesmen formatif, seperti kuis mingguan, tugas analisis data, dan umpan balik teman sebaya, dalam meningkatkan penguasaan metode statistik. Sejalan dengan rumusan masalah tersebut, hipotesis yang diajukan dalam penelitian ini adalah bahwa penerapan asesmen formatif terstruktur secara signifikan meningkatkan kompetensi statistik mahasiswa, meningkatkan keterlibatan mahasiswa dalam proses pembelajaran, serta bahwa setiap jenis asesmen formatif memberikan kontribusi yang berbeda terhadap peningkatan kompetensi statistik mahasiswa.

Penelitian ini bertujuan untuk menganalisis dampak penerapan asesmen formatif terstruktur terhadap peningkatan kompetensi statistik mahasiswa psikologi, serta mengkaji hubungan antara frekuensi asesmen formatif dan tingkat keterlibatan mahasiswa dalam proses pembelajaran. Selain itu, penelitian ini juga bertujuan untuk membandingkan efektivitas berbagai jenis asesmen formatif dalam meningkatkan penguasaan konsep dan metode statistik, serta mengidentifikasi faktor-faktor yang memengaruhi efektivitas penerapannya. Melalui pencapaian tujuan tersebut, penelitian ini diharapkan dapat memberikan kontribusi empiris dan praktis dalam pengembangan strategi pembelajaran statistik di pendidikan tinggi, khususnya dalam bidang psikologi, serta menjadi referensi bagi dosen dalam merancang asesmen yang lebih efektif, adaptif, dan berorientasi pada peningkatan kompetensi mahasiswa.

LITERATURE REVIEW

Tes Formatif

Tes formatif adalah penilaian yang digunakan selama proses pendidikan untuk memantau pembelajaran siswa dan memberikan umpan balik berkelanjutan guna meningkatkan pengajaran dan pembelajaran. Berbeda dengan penilaian sumatif yang mengevaluasi pembelajaran siswa di akhir periode pengajaran, tes formatif terintegrasi dalam proses pembelajaran dan dilakukan secara berulang (Abdullah, 2025; Ismail *et al*., 2022). Tujuan utama tes formatif adalah mengidentifikasi kekuatan dan kelemahan siswa sehingga pendidik dapat menyesuaikan strategi pengajaran secara personal. Penilaian ini dapat berbentuk kuis, diskusi, atau catatan jurnal yang dirancang untuk mendorong keterlibatan siswa tanpa tekanan penilaian nilai (Prastikawati *et al*., 2024).

https://doi.org/10.64014/jik.v23i2.325

Tes formatif menekankan peningkatan pembelajaran melalui umpan balik berkelanjutan, bukan nilai akhir. Umpan balik menjadi komponen utama yang memberikan wawasan kepada siswa tentang pemahaman mereka saat ini dan area yang perlu ditingkatkan, sehingga mempromosikan pola pikir berkembang (Munaroh, 2024). Jenis-jenis tes formatif meliputi kuis yang memberikan umpan balik langsung, observasi yang memungkinkan penilaian keterlibatan secara *real-time*, serta tinjauan sejawat yang mendorong pembelajaran kolaboratif dan keterampilan berpikir kritis (Wong *et al*., 2025; Zainuddin *et al*., 2020). Melalui siklus umpan balik yang berkelanjutan, tes formatif memfasilitasi identifikasi kesenjangan pembelajaran secara tepat waktu dan intervensi yang disesuaikan, sehingga berkontribusi pada pemahaman yang lebih mendalam serta peningkatan keterlibatan dan pencapaian akademik siswa (Lu & Cutumisu, 2022; Menéndez *et al*., 2019).

Asesmen Formatif dalam Pembelajaran Statistik

Statistik memiliki peran penting dalam psikologi sebagai landasan untuk menganalisis dan menginterpretasi data penelitian. Konsep-konsep seperti probabilitas, pengambilan sampel acak, statistik inferensial, dan distribusi-t merupakan kompetensi inti yang harus dikuasai mahasiswa psikologi (Wasserstein *et al*., 2019). Integrasi asesmen formatif dalam pembelajaran statistik menjadi penting karena memberikan kesempatan berkelanjutan bagi mahasiswa untuk menerapkan konsep-konsep tersebut dalam konteks praktis, sehingga dapat meningkatkan keterampilan analitis sekaligus memperkuat pemahaman terhadap konsep yang kompleks (Ajid *et al*., 2025; Panadero *et al*., 2018). Umpan balik yang diperoleh melalui asesmen formatif juga berkontribusi terhadap peningkatan kepercayaan diri mahasiswa dalam kemampuan statistik serta kesiapan mereka dalam melakukan penelitian psikologi (Dimova Popovska *et al*., 2024).

Asesmen formatif juga membantu menegaskan bahwa tujuan utama pembelajaran statistik bukan sekadar menghasilkan angka, tetapi memahami makna di balik data dalam kaitannya dengan teori dan hipotesis psikologi (Wasserstein *et al*., 2019). Perkembangan metodologi statistik yang semakin mengarah pada pendekatan estimasi dan Bayesian menuntut adanya integrasi praktik evaluasi yang lebih intensif dalam pembelajaran. Dalam konteks ini, asesmen formatif memungkinkan dosen mengidentifikasi kesalahan konseptual mahasiswa sejak dini serta menyesuaikan strategi pembelajaran agar tetap relevan dengan perkembangan metode statistik terkini (Lichtenberger *et al*., 2025). Selain itu, penggunaan soal pilihan ganda dengan distraktor yang dirancang secara efektif terbukti dapat membantu mahasiswa memahami konsep statistik yang abstrak serta mengurangi miskonsepsi (Arhin, 2024).

Analisis Butir dan Reliabilitas

Analisis butir dalam kerangka *Classical Test Theory* (CTT) merupakan teknik fundamental untuk menilai kualitas instrumen dalam penilaian pendidikan dan psikologi. Pendekatan ini menekankan keterkaitan antara tingkat kesulitan dan daya diskriminasi dalam menghasilkan tes yang valid dan andal (Meguellati *et al*., 2024). Melalui analisis ini, setiap butir soal dapat dievaluasi secara sistematis untuk memastikan bahwa instrumen mampu mengukur kemampuan peserta secara akurat.

Tingkat kesulitan butir menunjukkan proporsi peserta tes yang menjawab suatu butir dengan benar, yang dihitung dengan membandingkan jumlah jawaban benar terhadap total peserta tes (Sharma, 2021). Nilai indeks yang mendekati 1,0 menunjukkan bahwa butir tergolong mudah, sedangkan nilai yang mendekati 0 menunjukkan bahwa butir tergolong sulit. Analisis tingkat kesulitan penting untuk mengidentifikasi butir yang terlalu mudah atau terlalu sulit sehingga dapat direvisi guna menjamin kualitas dan keadilan penilaian (Kumar *et al*., 2021).

https://doi.org/10.64014/jik.v23i2.325

Daya diskriminasi butir mengukur kemampuan suatu butir dalam membedakan peserta dengan kemampuan tinggi dan rendah. Indeks ini umumnya dihitung menggunakan koefisien *point biserial correlation*l atau dengan membandingkan kinerja kelompok atas dan kelompok bawah (Gamage *et al*., 2019). Nilai diskriminasi yang tinggi menunjukkan bahwa butir tersebut efektif dalam membedakan kemampuan peserta, sedangkan nilai yang rendah mengindikasikan perlunya revisi atau penghapusan butir (Jordan & Spiess, 2019). Selain itu, analisis daya diskriminasi yang dikombinasikan dengan tingkat kesulitan juga membantu mengevaluasi efektivitas distraktor dalam soal pilihan ganda sehingga meningkatkan akurasi pengukuran terhadap capaian pembelajaran (Rezigalla *et al*., 2024).

Reliabilitas merupakan aspek penting yang menunjukkan konsistensi skor tes dalam berbagai kondisi pengukuran. Reliabilitas dapat diukur melalui beberapa pendekatan, seperti *test-retest*, konsistensi internal, dan bentuk paralel (Himelfarb, 2019). Salah satu indikator yang paling umum digunakan adalah koefisien Cronbach's alpha, yang mengukur sejauh mana butir-butir dalam suatu tes memiliki konsistensi dalam mengukur konstruk yang sama (Kumar *et al*., 2021). Tingkat reliabilitas yang tinggi menunjukkan bahwa instrumen memiliki stabilitas dan keandalan yang baik dalam menghasilkan data yang dapat dipercaya.

Item Response Theory (IRT) *One Parameter Logistic* (1PL)

Model *One Parameter Logistic* (1PL) atau model Rasch adalah pendekatan dasar dalam IRT yang mengukur sifat laten seperti kemampuan atau kompetensi hanya berdasarkan parameter kesulitan butir, dengan asumsi semua butir memiliki daya diskriminasi yang seragam. Kesederhanaannya menjadikan model ini sangat berguna dalam penilaian pendidikan dan psikologis karena menghasilkan estimasi kemampuan yang andal tanpa memerlukan asumsi kompleks atau sampel besar (Iliya, 2024; Iwintolu *et al*., 2024). Model 1PL efektif dalam konteks tes standar dengan butir yang dirancang seragam (Robitzsch, 2023). Namun, asumsi diskriminasi seragam menjadi keterbatasan ketika butir bervariasi dalam kemampuan membedakan responden; dalam situasi tersebut, model 2PL atau 3PL yang memasukkan parameter tambahan seperti diskriminasi bervariasi dan peluang menebak lebih dianjurkan (Kalkan & Çuhadar, 2020).

Item Response Theory Two Parameter Logistic (2PL)

Model *Two Parameter Logistic* (2PL) dalam IRT merupakan kerangka yang diakui luas dalam psikometrik karena kemampuannya meningkatkan analisis dan interpretasi tes melalui dua parameter utama: kesulitan dan diskriminasi butir. Parameter kesulitan menunjukkan titik pada skala kemampuan di mana sebuah butir memiliki probabilitas 50% untuk dijawab benar, sementara parameter diskriminasi mengukur ketajaman butir dalam membedakan tingkat kemampuan peserta (Gyamfi & Acquaye, 2023; Zhang *et al*., 2023). Dengan menggabungkan kedua parameter ini, model 2PL memberikan analisis yang lebih komprehensif dibandingkan model 1PL maupun CTT (Stenhaug & Domingue, 2022).

Dalam pengujian pendidikan, model 2PL memungkinkan estimasi kemampuan laten siswa yang lebih presisi, sehingga memudahkan pendidik menyesuaikan pengajaran sesuai kebutuhan individu (Jumini & Retnawati, 2022). Dalam penilaian psikologis, model 2PL efektif mendeteksi perbedaan kecil pada sifat psikologis yang krusial untuk tujuan diagnostik. Secara keseluruhan, fleksibilitas dan presisi model 2PL menjadikannya alat yang berharga untuk mengembangkan dan menyempurnakan instrumen penilaian, meningkatkan keandalan dan validitas interpretasi tes, serta mendukung pengambilan keputusan yang lebih baik dalam konteks pendidikan dan psikologi (Kalkan & Çuhadar, 2020).

https://doi.org/10.64014/jik.v23i2.325

METHODS

Penelitian ini menggunakan pendekatan kuantitatif dengan desain pengembangan instrumen (*instrument development research*). Pengembangan instrumen tes formatif mengikuti prosedur sistematis yang diadaptasi dari DeVellis pada tahun 2021 dalam buku "*Scale Development: Theory and Applications*", yang meliputi penentuan tujuan pengukuran, penyusunan *blueprint*, penulisan butir soal, validasi isi oleh ahli, dan uji coba empiris. Instrumen dikembangkan berdasarkan capaian pembelajaran mata kuliah Statistik Psikologi yang mencakup empat topik, yaitu probabilitas, *random sampling*, pengantar statistik inferensial (pengujian hipotesis *single mean Z*), dan distribusi T (perbandingan rata-rata sampel independen). Seluruh butir dirancang pada level kognitif memahami (C2) sesuai dengan Anderson dan Krathwohl pada tahun 2014 dalam buku *A Taxonomy for Learning, Teaching, and Assessing*, sehingga menghasilkan 40 butir soal pilihan ganda dengan distribusi materi sebagaimana tersaji pada **Tabel 1**.

Tabel 1. Distribusi Soal

Materi	Level Kognitif	Jumlah Soal	Kode Soal
Probabilitas	Memahami	6	1 – 6
Random Sampling	Memahami	9	7 – 15
Pengantar Statistik Inferensial: Menguji Hipotesis Single Mean (Z)	Memahami	15	16 – 30
Distribusi T (Perbandingan Rata-Rata Sampel Independen)	Memahami	10	31 – 40
Total Soal		**40**	

Sumber: Anderson dan Krathwohl (2014) dalam buku "*A Taxonomy for Learning, Teaching, and Assessing*"

Validitas isi instrumen dikonfirmasi melalui *expert judgment* yang melibatkan tiga ahli menggunakan koefisien Aiken's V dengan kriteria nilai minimum V ≥ 0,75 sebagaimana dikemukakan oleh Aiken pada tahun 1980 dalam bukunya yang berjudul "*Content Validity and Reliability of Single Items or Questionnaires*". Butir yang tidak memenuhi kriteria direvisi sebelum uji coba empiris. Partisipan penelitian adalah 191 mahasiswa semester 1 Program Studi Psikologi Universitas Pendidikan Indonesia yang dipilih menggunakan teknik *purposive sampling*. Pemilihan teknik ini mengacu pada Creswell pada tahun 2014 dalam buku "*Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*'', yang menekankan kesesuaian sampel dengan tujuan penelitian. Ukuran sampel ini dinilai memadai untuk analisis IRT model 1PL dan 2PL sebagaimana dijelaskan oleh Hambleton *et al*. dalam buku "*Fundamentals of Item Response Theory*". Pengumpulan data dilakukan secara daring melalui platform *e-learning*, dengan prosedur *informed consent* sebelum pengerjaan tes.

Analisis data dalam penelitian ini dilakukan secara bertahap untuk memastikan bahwa instrumen yang digunakan memiliki kualitas psikometris yang baik dan mampu mengukur kemampuan mahasiswa secara akurat. Tahap pertama adalah analisis korelasi item-total yang bertujuan untuk mengevaluasi konsistensi internal setiap butir soal. Pada tahap ini, setiap butir dikorelasikan dengan skor total untuk mengetahui sejauh mana butir tersebut berkontribusi terhadap keseluruhan instrumen. Butir yang memiliki nilai korelasi item-total di bawah 0,30 dianggap kurang representatif dalam mengukur konstruk yang dimaksud, sehingga dieliminasi. Dengan demikian, hanya butir yang memiliki hubungan kuat dengan skor total yang dipertahankan untuk tahap selanjutnya.

Tahap kedua adalah analisis menggunakan model IRT 1PL. Pada tahap ini, fokus analisis terletak pada parameter kesulitan butir (*item difficulty*). Setiap butir dievaluasi apakah berada dalam rentang kesulitan yang ideal, yaitu antara -2 hingga +2. Butir yang berada di luar rentang tersebut dianggap terlalu mudah atau terlalu sulit, sehingga kurang efektif dalam memberikan informasi tentang kemampuan responden

https://doi.org/10.64014/jik.v23i2.325

dan kemudian dieliminasi. Tahap ini bertujuan untuk memastikan bahwa butir-butir yang dipilih memiliki tingkat kesulitan yang proporsional dan sesuai dengan asumsi model Rasch.

Tahap ketiga merupakan analisis lanjutan menggunakan model IRT 2PL. Berbeda dengan model sebelumnya, pada tahap ini analisis tidak hanya mempertimbangkan tingkat kesulitan butir, tetapi juga daya diskriminasi (*discrimination power*). Parameter diskriminasi menunjukkan kemampuan suatu butir dalam membedakan responden dengan tingkat kemampuan yang berbeda. Selain itu, kesesuaian model juga diuji menggunakan nilai *p-value* dari uji Chi-Square, di mana butir dinyatakan sesuai dengan model apabila memiliki nilai *p-value* lebih dari 0,05. Butir yang memenuhi kedua kriteria tersebut, yaitu memiliki parameter kesulitan dan diskriminasi yang baik serta sesuai dengan model, dipilih sebagai bagian dari instrumen akhir.

Melalui rangkaian tahapan analisis tersebut, diperoleh seperangkat butir soal yang telah teruji secara empiris dan memiliki karakteristik yang baik. Instrumen yang dihasilkan diharapkan mampu mengukur pemahaman mahasiswa secara valid dan reliabel, khususnya pada materi probabilitas, *random sampling*, statistik inferensial, dan distribusi t.

RESULTS AND DISCUSSION

Kemudahan Item dan Daya Diskriminasi Item

Tingkat kesulitan item diukur berdasarkan nilai P, yaitu proporsi peserta yang menjawab suatu item dengan benar. Nilai P diklasifikasikan ke dalam tiga kategori: mudah (>0,700), sedang (0,30–0,70), dan sulit (<0,30). Sebagaimana ditunjukkan pada **Tabel 2**, sebagian besar item termasuk dalam kategori mudah, yang berarti mayoritas peserta dapat menjawab item-item tersebut dengan benar. Variasi kategori tingkat kesulitan juga tampak pada sejumlah item berkategori sedang yang memberikan tantangan yang cukup bagi peserta, serta beberapa item berkategori sulit yang hanya dapat dijawab benar oleh sebagian kecil peserta. Dengan demikian, distribusi tingkat kesulitan instrumen ini mencakup rentang yang cukup luas meskipun didominasi oleh item-item yang relatif mudah. Distribusi tingkat kesulitan dan daya diskriminasi setiap butir soal secara rinci disajikan pada **Tabel 2** berikut.

Tabel 2. Tingkat Kesulitan dan Daya Diskriminasi Soal

No Item	P	Point Biserial	No Item	P	Point Biserial
1	0.3526	**-0.2661**	20	0.8684	**0.3772**
2	0.1368	**0.4147**	21	0.8579	0.3754
3	0.9000	**0.0756**	22	0.7842	0.4342
4	0.8737	**0.3784**	23	0.8947	**0.4293**
5	0.0579	**0.3010**	24	0.7579	**0.4123**
6	0.9368	**0.2538**	25	0.1421	**0.3563**
7	0.9737	**0.1420**	26	0.7842	**0.3619**
8	0.7211	**0.5144**	27	0.8737	**0.3232**
9	0.8895	**0.3773**	28	0.8105	**0.5615**
10	0.8895	**0.4605**	29	0.7632	**0.5077**
11	0.9053	**0.2296**	30	0.8579	**0.4129**
12	0.8211	**0.4532**	31	0.9316	**0.2718**
13	0.6842	**0.4208**	32	0.6316	**0.3470**
14	0.2000	**0.0795**	33	0.1684	**0.0293**

https://doi.org/10.64014/jik.v23i2.325

No Item	P	Point Biserial	No Item	P	Point Biserial
15	0.8211	0.2375	34	0.6684	**0.4038**
16	0.8632	0.4110	35	0.5316	**0.1227**
17	0.7895	**0.5015**	36	0.7105	**0.3923**
18	0.7474	**0.3072**	37	0.5632	**0.1801**
19	0.6947	**0.4418**	38	0.8053	**0.3314**
20	0.8684	**0.3772**	39	0.0263	**0.1901**
21	0.8579	0.3754	40	0.6789	**0.3666**

Sumber: Penelitian 2024
*Koefisien yang ditebalkan adalah kunci jawaban

Selain tingkat kesulitan, daya diskriminasi item juga dianalisis menggunakan *point biserial correlation*, yang menunjukkan sejauh mana sebuah butir dapat membedakan peserta berkemampuan tinggi dari peserta berkemampuan rendah. Berdasarkan **Tabel 2**, sebagian besar butir memiliki daya diskriminasi yang baik hingga sangat baik, menandakan bahwa item-item tersebut efektif dalam membedakan peserta berdasarkan tingkat kemampuan mereka.

Meskipun demikian, terdapat sejumlah butir yang menunjukkan daya diskriminasi rendah sehingga kurang efektif dalam membedakan peserta. Lebih jauh, ditemukan pula butir dengan daya diskriminasi negatif, yang berarti peserta berkemampuan tinggi justru cenderung menjawab salah sementara peserta berkemampuan rendah menjawab benar. Kondisi ini mengindikasikan adanya ambiguitas atau kelemahan dalam formulasi soal, sehingga butir-butir tersebut perlu dieliminasi. Setelah proses eliminasi butir yang tidak memenuhi standar daya diskriminasi minimum, instrumen yang tersisa diharapkan memiliki konsistensi internal yang lebih kuat. Berdasarkan hasil analisis tingkat kesulitan dan daya diskriminasi tersebut, dapat disimpulkan bahwa instrumen awal memiliki kualitas yang beragam. Butir-butir yang tidak memenuhi kriteria psikometris selanjutnya dieliminasi melalui proses seleksi item yang diuraikan pada bagian berikut.

Seleksi Item

Seleksi item dilakukan berdasarkan nilai korelasi *point biserial correlation* sebagaimana tersaji pada **Tabel 2**. Proses seleksi ini menghasilkan pengurangan jumlah butir dari 40 menjadi 32 item. Sebagaimana ditunjukkan pada **Tabel 3**, perubahan jumlah butir tersebut berdampak pada sejumlah properti statistik tes secara keseluruhan.

Tabel 3. Distribusi Soal

Kategori	Item Awal	Item Terseleksi	Perbandingan
Jumlah Butir Soal	40	32	Jumlah butir soal berkurang dari 40 menjadi 32 setelah seleksi item.
Jumlah Peserta	190	190	Tidak ada perubahan pada jumlah peserta (tetap 190).
Nilai Minimum	10.00	8.00	Nilai minimum turun dari 10 menjadi 8 setelah seleksi item.
Nilai Maksimum	37.00	32.00	Nilai maksimum juga menurun dari 37 menjadi 32 setelah seleksi item.
Rata-rata	29.0895	25.3737	Rata-rata skor menurun dari 29,09 menjadi 25,37 setelah seleksi item.

https://doi.org/10.64014/jik.v23i2.325

Kategori	Item Awal	Item Terseleksi	Perbandingan
Median	31.0000	27.00	Median menurun dari 31 menjadi 27, menunjukkan pergeseran nilai tengah.
Standar Deviasi	5.9283	5.7329	Standar deviasi sedikit menurun, menunjukkan variasi skor peserta sedikit berkurang.
Rentang Interkuartil	9.0000	10.00	Rentang interkuartil meningkat dari 9 menjadi 10, menunjukkan penyebaran nilai tengah yang lebih besar.
Skewness	-0.8662	-0.8193	Skewness mendekati nol, tetapi tetap negatif, menunjukkan distribusi skor cenderung ke arah nilai tinggi.
Kurtosis	0.1037	-0.1574	Kurtosis menjadi sedikit lebih rendah, menunjukkan distribusi skor pasca seleksi lebih datar.
KR21	0.7941	0.8672	KR21 meningkat dari 0,7941 menjadi 0,8672, menunjukkan reliabilitas tes yang lebih tinggi.

Sumber: Penelitian 2024

Sebagaimana tampak pada **Tabel 3**, seleksi item berdampak pada perubahan distribusi skor peserta. Penurunan skor minimum, maksimum, rata-rata, dan median merupakan konsekuensi logis dari berkurangnya jumlah butir, bukan indikasi penurunan kemampuan peserta. Penurunan standar deviasi yang relatif kecil menunjukkan bahwa variasi skor antar individu tetap terjaga, sehingga instrumen masih mampu membedakan peserta berdasarkan kemampuan mereka.

Peningkatan yang paling signifikan tampak pada koefisien reliabilitas KR21 yang meningkat setelah seleksi, menunjukkan bahwa instrumen menjadi lebih konsisten dalam mengukur konstruk yang sama. Distribusi skor yang sedikit condong ke nilai tinggi (*skewness negatif*) dan kurtosis yang mendekati nol mencerminkan bahwa sebagian besar peserta mampu menjawab item-item yang tersisa dengan baik. Dengan demikian, instrumen pasca seleksi memiliki kualitas psikometris yang lebih baik dibandingkan instrumen awal. Selanjutnya, butir-butir terseleksi ini dianalisis lebih lanjut menggunakan pendekatan IRT untuk memperoleh gambaran yang lebih presisi tentang karakteristik setiap butir.

Analisis Item Berdasarkan IRT 1PL

Pada tahap ini, analisis dilakukan menggunakan model IRT 1PL dengan fokus pada parameter kesulitan butir. Analisis ini bertujuan untuk mengevaluasi distribusi tingkat kesulitan item dalam instrumen. Hasil pengujian disajikan pada **Tabel 4** berikut.

Tabel 4. Pengujian Soal berdasarkan IRT 1PL

No Item	Parameter b (Tingkat Kesulitan)	Chi-Kuadrat	Nilai P
2	-1.79	15.8894	0.6646
4	-2.39	20.7527	0.3506
5	0.28	27.1290	0.0766
6	-3.34	19.5674	0.3577
8	-1.23	20.4997	0.3651
9	-2.57	21.5229	0.2539
10	-2.57	20.7573	0.2918

https://doi.org/10.64014/jik.v23i2.325

No Item	Parameter b (Tingkat Kesulitan)	Chi-Kuadrat	Nilai P
11	-2.76	29.4817	0.0588
12	-1.91	14.7763	0.7367
13	-0.99	15.7231	0.6757
15	-1.95	26.7779	0.1099
16	-2.33	12.2960	0.8726
17	-1.67	18.3649	0.4982
18	-1.40	27.2866	0.0738
19	-1.09	34.5451	0.0108
20	-2.34	12.3440	0.8704
21	-2.23	22.0390	0.2823
22	-1.63	23.2764	0.2254
23	-2.69	15.0039	0.7223
24	-1.47	9.1713	0.9705
25	-1.82	19.9631	0.3968
26	-1.63	17.1633	0.5788
27	-2.43	13.5319	0.8103
28	-1.83	17.4295	0.5608
29	-1.48	18.3656	0.4318
30	-2.23	21.2948	0.3207
31	-3.15	14.1251	0.7764
32	-0.69	23.0157	0.2366
34	-0.90	24.7147	0.1701
36	-1.16	16.7703	0.5389
38	-1.79	39.4132	0.0039
40	-0.96	37.2371	0.0049

Sumber: Penelitian 2024

Analisis IRT 1PL berfokus pada parameter kesulitan butir (*b*) untuk mengevaluasi kesesuaian setiap butir dengan model Rasch. Sebagaimana tersaji pada **Tabel 4**, sebagian besar butir menunjukkan kesesuaian yang baik dengan model, ditandai oleh nilai *p* Chi-Square yang lebih besar dari 0,05. Distribusi parameter *b* yang sebagian besar berada pada rentang negatif menggambarkan bahwa butir-butir tersebut tergolong mudah dalam skala IRT, konsisten dengan temuan analisis CTT sebelumnya. Namun demikian, terdapat tiga butir yang tidak memenuhi kriteria kesesuaian model ($p \leq 0{,}05$) sehingga harus dieliminasi. Ketidaksesuaian butir-butir tersebut kemungkinan disebabkan oleh karakteristik respons yang tidak dapat dijelaskan hanya oleh parameter kesulitan, sehingga memerlukan parameter tambahan untuk dimodelkan secara lebih tepat.

Berdasarkan hasil seleksi tersebut, butir-butir yang memenuhi kriteria selanjutnya dianalisis menggunakan IRT 2PL yang mempertimbangkan parameter kesulitan dan daya diskriminasi. Hasil analisis kesesuaian item dengan model 2PL disajikan pada **Tabel 5** berikut.

https://doi.org/10.64014/jik.v23i2.325

Tabel 5. Kesesuaian Item dengan Model 2PL (Parameter Diskriminasi dan Tingkat Kesulitan)

No Item	Parameter a (Daya Diskriminasi)	Parameter b (Tingkat Kesulitan)	Chi-Kuadrat	Nilai P
2	1.23	-1.52	6.8501	0.9617
4	1.27	-1.98	12.8872	0.6110
5	0.81	0.30	9.1405	0.9075
6	1.29	-2.71	22.5024	0.0953
8	1.61	-0.92	15.1036	0.5171
9	1.50	-1.93	16.6814	0.4065
10	1.78	-1.78	20.3733	0.2039
11	1.10	-2.49	16.7143	0.4043
12	1.72	-1.91	9.4811	0.8511
13	1.37	-0.81	13.4286	0.6412
15	0.62	-2.74	15.2686	0.4323
16	1.65	-2.33	6.4498	0.9825
17	1.70	-1.21	12.9529	0.6762
18	0.83	-1.55	11.4676	0.7797
20	1.32	-1.90	13.2752	0.6525
21	1.28	-1.84	23.3069	0.1058
22	1.22	-1.40	18.7888	0.2798
23	1.79	-1.85	6.6390	0.9796
24	1.16	-1.31	8.8636	0.9189
25	1.12	-1.63	5.2289	0.9900
26	1.20	-1.41	18.6748	0.2859
27	1.21	-2.08	30.3295	0.0164
28	1.92	-1.26	13.8429	0.5375
29	1.74	-1.08	24.6277	0.0767
30	1.34	-1.80	19.3764	0.1972
31	1.11	-2.83	25.4112	0.0447
32	0.80	-0.78	12.1828	0.6651
34	1.37	-0.74	22.0118	0.1428
36	1.14	-1.04	10.7066	0.7731

Sumber: Penelitian 2024

Berdasarkan **Tabel 5**, sebagian besar butir menunjukkan nilai *p* Chi-Kuadrat yang lebih besar dari 0,05, yang mengindikasikan bahwa butir-butir tersebut memiliki kesesuaian yang baik dengan model 2PL. Selain itu, nilai parameter daya diskriminasi (*a*) pada sebagian besar item berada pada kategori sedang hingga tinggi, yang menunjukkan bahwa butir mampu membedakan peserta dengan tingkat kemampuan yang berbeda secara efektif. Sementara itu, parameter tingkat kesulitan (*b*) yang dominan berada pada rentang negatif mengindikasikan bahwa sebagian besar item tergolong relatif mudah bagi responden.

https://doi.org/10.64014/jik.v23i2.325

Namun demikian, terdapat beberapa butir yang tidak memenuhi kriteria kesesuaian model, ditandai dengan nilai $p \leq 0{,}05$, sehingga butir tersebut dieliminasi dari instrumen. Dengan demikian, hanya butir-butir yang memenuhi kriteria parameter kesulitan, daya diskriminasi, serta kesesuaian model yang dipertahankan sebagai bagian dari instrumen akhir. Berdasarkan hasil seleksi tersebut, distribusi butir final yang memenuhi kriteria model IRT 2PL disajikan pada **Tabel 6** berikut.

Tabel 6. Kesesuaian Item dengan Model 2PL (Parameter Diskriminasi dan Tingkat Kesulitan)

Materi	Level Kognitif	Jumlah Soal	Kode Soal
Probabilitas	Memahami	4	2, 4, 5, 6
Random Sampling	Memahami	7	8, 9, 10, 11, 12, 13, 15
Pengantar Statistik Inferensial: Menguji Hipotesis Single Mean (Z)	Memahami	14	16 – 18, 20-30
Distribusi T (Perbandingan Rata-Rata Sampel Independen)	Memahami	4	31, 32, 34, 36
Total Soal		**29**	

Sumber: Penelitian 2024

Hasil pada **Tabel 6** menunjukkan bahwa distribusi butir telah mencakup berbagai materi utama, yaitu probabilitas, random sampling, pengantar statistik inferensial, serta distribusi t, dengan dominasi pada level kognitif memahami. Jumlah butir yang relatif seimbang pada setiap materi menunjukkan bahwa instrumen telah disusun secara proporsional dan representatif terhadap cakupan materi yang diukur. Dengan demikian, instrumen akhir yang dihasilkan tidak hanya memenuhi kriteria psikometris, tetapi juga memiliki kesesuaian konten yang baik untuk mengukur pemahaman mahasiswa secara komprehensif.

Diskusi

Interpretasi Hasil Analisis

Distribusi tingkat kesulitan yang didominasi oleh butir-butir mudah perlu dibaca sebagai kesesuaian instrumen dengan konteks penggunaannya, bukan sebagai kelemahan. Tes formatif memang tidak dirancang untuk menyortir kemampuan mahasiswa secara ekstrem, melainkan untuk memantau pemahaman terhadap materi yang baru saja dipelajari serta mendorong keterlibatan aktif tanpa tekanan nilai yang berlebihan (Munaroh, 2024; Prastikawati *et al*., 2024). Kehadiran sejumlah butir dengan kesulitan sedang memberikan gradasi yang diperlukan agar instrumen tidak kehilangan fungsi pembedanya sama sekali. Profil kesulitan ini mencerminkan keseimbangan antara aksesibilitas bagi mahasiswa dan kemampuan diagnostik instrumen dalam mendeteksi celah pemahaman.

Daya diskriminasi yang tinggi pada sebagian besar butir memiliki makna penting bagi tujuan pedagogis pembelajaran statistik psikologi. Butir yang mampu membedakan mahasiswa berkemampuan tinggi dari yang berkemampuan rendah secara efektif menjadi alat diagnosis yang andal bagi dosen untuk mengidentifikasi siapa saja yang memerlukan perhatian lebih, sehingga intervensi pembelajaran dapat dilakukan secara tepat sasaran (Lichtenberger *et al*., 2025). Hal ini sangat relevan mengingat penguasaan statistik bukan sekadar hafalan rumus, melainkan pemahaman konseptual yang mendalam tentang makna data dalam konteks penelitian psikologi (Wasserstein *et al*., 2019). Sebaliknya, butir-butir dengan diskriminasi rendah atau negatif yang telah dieliminasi mengindikasikan bahwa soal-soal tersebut tidak mengukur kemampuan yang relevan, kemungkinan karena rumusan yang ambigu atau tidak selaras dengan capaian pembelajaran yang ditetapkan.

https://doi.org/10.64014/jik.v23i2.325

Peningkatan reliabilitas setelah proses seleksi butir menunjukkan bahwa pengurangan jumlah soal tidak selalu berdampak negatif terhadap kualitas pengukuran. Sebaliknya, eliminasi butir-butir yang lemah justru menghasilkan instrumen yang lebih koheren dan konsisten dalam mengukur konstruk yang sama. Koefisien reliabilitas yang berada di atas ambang batas yang baik mengindikasikan bahwa skor yang dihasilkan instrumen dapat dipercaya sebagai representasi pemahaman aktual mahasiswa, bukan sekadar variasi acak dalam respons (Himelfarb, 2019; Kumar *et al*., 2021). Dalam konteks asesmen formatif, keandalan data menjadi sangat krusial karena berbagai keputusan pedagogis dosen, seperti penyesuaian strategi pembelajaran dan pemberian umpan balik individual, sangat bergantung pada akurasi informasi yang dihasilkan oleh instrumen tersebut.

Temuan dari analisis IRT memperkuat dan memperdalam gambaran yang diperoleh melalui CTT. Kesesuaian sebagian besar butir dengan model 1PL menunjukkan bahwa instrumen ini memiliki kerangka pengukuran yang solid, di mana setiap butir berkontribusi secara konsisten terhadap pengukuran kemampuan laten mahasiswa. Lebih jauh, analisis 2PL mengungkap bahwa butir-butir yang lolos seleksi tidak hanya memiliki tingkat kesulitan yang proporsional dengan kemampuan mahasiswa semester pertama, tetapi juga memiliki kekuatan diskriminasi yang memadai untuk membedakan tingkat penguasaan konsep statistik di antara mereka. Butir-butir yang tidak fit pada kedua model IRT tersebut mengindikasikan ketidakkonsistenan pola respons yang tidak dapat dijelaskan oleh parameter standar, sehingga keberadaannya justru akan menurunkan akurasi pengukuran apabila dipertahankan. Pada akhirnya, proses seleksi bertahap ini menghasilkan instrumen yang secara psikometris lebih terpercaya dan efisien dalam mengukur pemahaman mahasiswa terhadap konsep statistik psikologi.

Profil psikometris instrumen yang dihasilkan dalam penelitian ini konsisten dengan temuan dari studi-studi pengembangan instrumen statistik pada populasi mahasiswa serupa. Instrumen yang dikembangkan untuk kelompok mahasiswa pemula umumnya memiliki distribusi kesulitan yang cenderung ke arah mudah, karena tujuan utamanya adalah membangun kepercayaan diri serta memetakan pemahaman awal, bukan untuk tujuan seleksi (Dimova Popovska *et al*., 2024). Keselarasan ini menunjukkan bahwa instrumen yang dikembangkan dalam penelitian ini telah berfungsi sesuai dengan peruntukannya. Selain itu, keseimbangan antara tingkat kesulitan dan daya diskriminasi merupakan prasyarat penting bagi instrumen yang valid dan andal dalam kerangka CTT, dan kondisi tersebut telah terpenuhi dalam instrumen ini (Meguellati *et al*., 2024).

Dari sisi metodologi, pendekatan bertahap yang menggabungkan CTT dan IRT dalam penelitian ini sejalan dengan rekomendasi terkini dalam literatur pengembangan tes. Model 2PL dinilai memberikan analisis yang lebih komprehensif dibandingkan CTT maupun model 1PL, karena mampu menangkap dua dimensi kualitas butir secara simultan, yaitu tingkat kesulitan dan daya diskriminasi (Stenhaug & Domingue, 2022). Penerapan pendekatan ini terbukti menghasilkan instrumen yang lebih ketat dan terpercaya. Pendekatan tersebut juga melampaui penelitian yang hanya berfokus pada efektivitas distraktor, dengan memasukkan uji kesesuaian model sebagai kriteria utama dalam seleksi butir (Arhin, 2024). Selain itu, model 2PL dinilai sebagai pilihan yang seimbang antara kompleksitas dan akurasi, khususnya untuk ukuran sampel menengah (Kalkan & Çuhadar, 2020).

Kelebihan penelitian ini terletak pada transparansi dan sistematisitas prosedur pengembangan instrumen, mulai dari validasi isi hingga seleksi butir berbasis IRT, serta pada konteks spesifiknya yang menjawab keterbatasan literatur psikometrik pada mata kuliah Statistik Psikologi di Indonesia. Meskipun demikian, penelitian ini belum mengadopsi pendekatan *Computer Adaptive Testing* (CAT) yang memungkinkan penyesuaian tingkat kesulitan soal secara real-time sesuai kemampuan individu mahasiswa (Jumini & Retnawati, 2022). Selain itu, tidak digunakannya model 3PL menyebabkan faktor tebakan (*guessing*) pada soal pilihan ganda belum diperhitungkan secara eksplisit dalam estimasi kemampuan. Kondisi ini membuka peluang bagi penelitian selanjutnya untuk mengembangkan instrumen yang lebih adaptif dan memiliki tingkat presisi yang lebih tinggi.

https://doi.org/10.64014/jik.v23i2.325

Instrumen yang dihasilkan dari penelitian ini siap digunakan secara langsung sebagai alat asesmen formatif yang telah memenuhi standar psikometris. Artinya, dosen statistik psikologi kini memiliki instrumen yang dapat dipercaya untuk memantau perkembangan pemahaman mahasiswa secara berkelanjutan, bukan sekadar soal latihan tanpa dasar pengukuran yang kuat. Penelitian terdahulu secara konsisten menunjukkan bahwa tes formatif yang berkualitas, ketika disertai umpan balik yang bermakna, terbukti meningkatkan pemahaman konsep statistik mahasiswa secara signifikan (Menéndez *et al*., 2019; Wafubwa, 2020). Dengan demikian, ketersediaan instrumen yang terstandarisasi ini menjadi prasyarat penting bagi penerapan siklus asesmen-umpan balik yang efektif dalam perkuliahan.

Selain dapat digunakan sebagai tes lengkap, struktur instrumen yang terbagi per topik memberikan fleksibilitas yang bernilai praktis tinggi. Dosen dapat memilih subset butir yang relevan untuk digunakan sebagai tes formatif per pertemuan atau per topik, sehingga asesmen menjadi lebih terpadu dengan alur perkuliahan. Integrasi ke dalam *platform e-learning* yang sudah digunakan juga memungkinkan pemberian umpan balik secara otomatis dan segera setelah pengerjaan, yang terbukti membantu mahasiswa mendiagnosis celah pemahaman mereka dan menyesuaikan strategi belajar (Lubis & Setiawan, 2025). Penerapan asesmen formatif yang rutin dan berkelanjutan seperti ini diharapkan dapat meningkatkan partisipasi aktif serta motivasi mahasiswa dalam mendalami materi statistik yang secara tradisional dianggap kompleks (Lase, 2024; Nisrina *et al*., 2025).

Dari perspektif pengajaran, keberadaan instrumen formatif yang terstandarisasi memungkinkan dosen mengambil keputusan pedagogis berdasarkan data, bukan semata intuisi. Pola respons kolektif mahasiswa terhadap butir-butir tertentu dapat mengungkap topik atau konsep mana yang belum dipahami secara merata, sehingga dosen dapat menyesuaikan pendekatan mengajar secara responsif dan berbasis bukti (Dianti *et al*., 2025). Lebih dari itu, pengalaman berhasil menjawab soal-soal formatif yang dirancang dengan baik berkontribusi pada penguatan kepercayaan diri mahasiswa terhadap kemampuan statistik mereka, yang merupakan prediktor penting bagi kesiapan mereka dalam melakukan penelitian psikologi di jenjang selanjutnya (Dimova Popovska *et al*., 2024).

Penelitian ini memiliki beberapa keterbatasan yang perlu diakui. Pertama, sampel penelitian terbatas pada mahasiswa semester 1 Program Studi Psikologi Universitas Pendidikan Indonesia, sehingga generalisasi temuan ke populasi mahasiswa psikologi di institusi lain harus dilakukan dengan hati-hati. Perbedaan kurikulum, kualitas pengajaran, dan latar belakang akademik mahasiswa antar perguruan tinggi dapat memengaruhi karakteristik psikometris butir. Kedua, penelitian ini hanya menggunakan model IRT 1PL dan 2PL tanpa menyertakan model 3PL yang memperhitungkan parameter peluang menebak (*guessing*). Mengingat format soal pilihan ganda dengan empat opsi jawaban yang digunakan, parameter tebakan berpotensi memengaruhi estimasi kemampuan mahasiswa terutama pada butir yang sangat sulit. Ketiga, validitas tes hanya dievaluasi melalui validitas isi (*content validity*) menggunakan Aiken's V, tanpa melakukan validasi konstruk secara empiris menggunakan analisis faktor konfirmatori atau kriteria eksternal. Keempat, seluruh data dikumpulkan dalam satu waktu pengujian (*cross-sectional*), sehingga stabilitas temporal (*test-retest reliability*) instrumen belum dapat dikonfirmasi.

CONCLUSION

Berdasarkan hasil penelitian, dapat disimpulkan bahwa pengembangan instrumen asesmen formatif terstruktur pada mata kuliah Statistik Psikologi berhasil menghasilkan alat ukur yang memiliki kualitas psikometris yang baik, ditunjukkan melalui tingkat kesulitan yang proporsional, daya diskriminasi yang memadai, serta reliabilitas yang tinggi setelah melalui proses seleksi butir menggunakan pendekatan CTT dan IRT model 1PL dan 2PL. Instrumen yang dihasilkan mampu mengukur pemahaman mahasiswa secara valid dan konsisten, sekaligus berfungsi efektif sebagai alat asesmen formatif dalam memantau perkembangan kompetensi statistik mahasiswa. Selain itu, karakteristik butir yang didominasi tingkat

https://doi.org/10.64014/jik.v23i2.325

kesulitan mudah hingga sedang serta diskriminasi yang baik menunjukkan bahwa instrumen ini sesuai untuk meningkatkan keterlibatan mahasiswa dalam proses pembelajaran tanpa memberikan tekanan evaluatif yang berlebihan. Dengan demikian, penelitian ini menjawab permasalahan utama bahwa asesmen formatif terstruktur dapat mendukung peningkatan kompetensi statistik dan keterlibatan mahasiswa, serta menunjukkan bahwa penerapan prosedur pengembangan instrumen yang sistematis mampu menghasilkan alat asesmen yang lebih akurat, adaptif, dan dapat digunakan sebagai dasar pengambilan keputusan pedagogis dalam pembelajaran statistik di pendidikan tinggi.

Berdasarkan keterbatasan yang telah diidentifikasi, beberapa rekomendasi untuk penelitian lanjutan dapat dikemukakan. Pertama, replikasi dan validasi silang instrumen ini perlu dilakukan pada sampel mahasiswa psikologi dari institusi lain di Indonesia untuk menguji stabilitas properti psikometrisnya. Kedua, analisis menggunakan model IRT 3PL disarankan untuk dilakukan guna memperoleh estimasi yang lebih akurat pada butir-butir dengan tingkat kesulitan tinggi dengan memperhitungkan peluang tebakan acak. Ketiga, studi longitudinal yang melacak pengaruh penggunaan tes formatif ini terhadap hasil belajar akhir semester mahasiswa akan memberikan bukti empiris tentang efektivitas instrumen dalam meningkatkan pemahaman statistik. Keempat, pengembangan bank butir yang lebih luas berdasarkan *framework* IRT yang telah dibangun dalam penelitian ini akan memungkinkan implementasi CAT di masa depan, yang dapat mengoptimalkan pengukuran kemampuan mahasiswa secara individual dan efisien. Kelima, perlu dilakukan uji diferensial fungsi butir (*Differential Item Functioning/DIF*) untuk memastikan bahwa instrumen tidak bias terhadap kelompok mahasiswa tertentu berdasarkan jenis kelamin, latar belakang pendidikan sebelumnya, atau faktor demografis lainnya.

AUTHOR'S NOTE

Penulis menyatakan bahwa tidak ada konflik kepentingan terkait publikasi artikel ini. Penulis menegaskan bahwa data dan isi artikel bebas dari plagiarisme.

REFERENCES

Abdullah, M. S. (2025). Model evaluasi formatif dan sumatif: strategi untuk meningkatkan proses dan hasil pembelajaran di pendidikan dasar pada Kurikulum Merdeka. *Dewantara: Jurnal Pendidikan*, *2*(4), 30-35.

Ajid, S. N., Kusumaningtyas, D. A., Ratih, K., & Lava, S. (2025). Strategies for integrating problem-based learning, teaching modules, and formative assessments to enhance learning outcomes and critical thinking skills. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, 218-232.

Arhin, A. K. (2024). Developing distractors for mathematics multiple choice items: a literature review. *Acta Educationis Generalis*, *14*(3), 103-120.

Chen, Z., Jiao, J., & Hu, K. (2021). Formative assessment as an online instruction intervention: student engagement, outcomes, and perceptions. *International Journal of Distance Education Technologies (IJDET)*, *19*(1), 50-65.

Dayal, H. C. (2021). How teachers use formative assessment strategies during teaching: evidence from the classroom. *Australian Journal of Teacher Education (Online)*, *46*(7), 1-21.

Dianti, K., Ulfah, M., Salam, A., Gunawan, G., & Luthfiyah, L. (2025). Analisis asesmen diagnostik, formatif dan sumatif serta implikasinya terhadap efektivitas sistem evaluasi pendidikan. *Jurnal Pendidikan dan Pembelajaran Indonesia (JPPI)*, *5*(2), 555-565.

https://doi.org/10.64014/jik.v23i2.325

Dimova Popovska, H., Popovski, F., & Popovska Nalevska, G. (2024). Using formative assessment to foster confidence and motivation to learn. *International Journal of Research Studies in Education*, *13*(1), 113-121.

Gamage, S. H., Ayres, J. R., & Behrend, M. B. (2022). A systematic review on trends in using Moodle for teaching and learning. *International Journal of STEM Education*, *9*(1), 1-24.

Gyamfi, A., & Acquaye, R. (2023). Parameters and models of Item Response Theory (IRT): a review of literature. *Acta Educationis Generalis*, *13*(3), 68-78.

Hanefar, S. B. M., Anny, N., & Rahman, S. (2022). Enhancing teaching and learning in higher education through formative assessment: teachers' perceptions. *International Journal of Assessment Tools in Education*, *9*(1), 61-79.

Himelfarb, I. (2019). A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating. *Journal of Chiropractic Education*, *33*(2), 151–163.

Iliya, A. (2024). Item parameters, scoring models and ability estimates of distance education students: implications for the national open University of Nigeria. *Sokoto Educational Review*, *23*(1), 162-172.

Ismail, S. M., Rahul, D. R., Patra, I., & Rezvani, E. (2022). Formative vs. summative assessment: impacts on academic motivation, attitude toward learning, test anxiety, and self-regulation skill. *Language Testing in Asia*, *12*(1), 1-23.

Iwintolu, R. O., Opesemowo, O. A. G., & Adetutu, P. O. (2024). Effect of 2-PL and 3-PL models on the ability estimate in Mathematics binary items. *Journal on Efficiency and Responsibility in Education and Science*, *17*(3), 257-272.

Jordan, P., & Spiess, M. (2019). Rethinking the interpretation of item discrimination and factor loadings. *Educational and Psychological Measurement*, *79*(6), 1103–1132.

Jumini, J., & Retnawati, H. (2022). Estimating item parameters and student abilities: an IRT 2PL analysis of Mathematics examination. *Al-Ishlah: Jurnal Pendidikan*, *14*(1), 385-398.

Kalkan, Ö. K., & Çuhadar, İ. (2020). An evaluation of 4PL IRT and DINA models for estimating pseudo-guessing and slipping parameters. *Journal of Measurement and Evaluation in Education and Psychology*, *11*(2), 131-146.

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.

Lase, A. G. (2024). Penerapan asesmen formatif berbasis Quizizz untuk meningkatkan hasil belajar matematika siswa SMA Negeri 2 Medan. *Education Journal: Journal Educational Research and Development*, *8*(2), 466-475.

Leenknecht, M., Wijnia, L., Köhlen, M., Fryer, L., Rikers, R., & Loyens, S. (2021). Formative assessment as practice: the role of students' motivation. *Assessment & Evaluation in Higher Education*, *46*(2), 236-255.

Lichtenberger, A., Hofer, S. I., Stern, E., & Vaterlaus, A. (2025). Enhanced conceptual understanding through formative assessment: results of a randomized controlled intervention study in physics classes. *Educational Assessment, Evaluation and Accountability*, *37*(1), 5-33.

https://doi.org/10.64014/jik.v23i2.325

Lu, C., & Cutumisu, M. (2022). Online engagement and performance on formative assessments mediate the relationship between attendance and course performance. *International Journal of Educational Technology in Higher Education*, *19*(1), 1-23.

Lubis, A., & Setiawan, A. (2025). Adaptive learning dalam desain instruksional: pendekatan strategis meningkatkan keterlibatan mahasiswa di e-learning perguruan tinggi. *Pendas: Jurnal Ilmiah Pendidikan Dasar*, *10*(2), 780-792.

Meguellati, S., Samia, A., Ferhat, A., & Djelloul, A. (2024). A critical analysis of the use of Classical Test Theory (CTT) in psychological testing: a comparison with Item Response Theory (IRT). *Pakistan Journal of Life & Social Sciences*, *22*(2), 9442-9449.

Menéndez, I. Y. C., Napa, M. A. C., Moreira, M. L. M., & Zambrano, G. G. V. (2019). The importance of formative assessment in the learning teaching process. *International Journal of Social Sciences and Humanities*, *3*(2), 238-249.

Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: a systematic review. *Review of Education*, *9*(3), 1-26.

Munaroh, N. L. (2024). Asesmen dalam pendidikan: memahami konsep, fungsi dan penerapannya. *Dewantara: Jurnal Pendidikan Sosial Humaniora, 3*(3), 281-297.

Nisrina, P., Primawati, R. I., & Nahadi, N. (2025). Analysis of the implementation of formative assessment on students' conceptual understanding in chemistry learning. *Hydrogen: Jurnal Kependidikan Kimia*, *13*(1), 174-185.

Pai, G. (2025). Using formative assessment and feedback from Student Response Systems (SRS) to revise statistics instruction and promote student growth for all. *Journal of Statistics and Data Science Education*, *33*(1), 16-25.

Panadero, E., Andrade, H., & Brookhart, S. (2018). Fusing self-regulated learning and formative assessment: a roadmap of where we are, how we got here, and where we are going. *The Australian Educational Researcher*, *45*(1), 13-31.

Prastikawati, E. F., Adeoye, M. A., & Ryan, J. C. (2024). Fostering effective teaching practices: integrating formative assessment and mentorship in Indonesian preservice teacher education. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, 230-253.

Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaifi, J., ALGhamdi, M. A., Al Ameer, A. Y., ... & Adam, M. I. E. (2024). Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*, *24*(1), 1-7.

Robitzsch, A. (2023). Comparing robust linking and regularized estimation for linking two groups in the 1PL and 2PL models in the presence of sparse uniform differential item functioning. *Stats*, *6*(1), 192-208.

Saekoko, N., Benu, S., Oematan, I. W. A., & Pa, H. D. B. (2025). Peran evaluasi formatif dalam meningkatkan kualitas pembelajaran di era digital. *Jurnal Ilmiah Literasi Indonesia*, *1*(2), 336-350.

Sharma, L. R. (2021). Analysis of difficulty index, discrimination index and distractor efficiency of multiple choice questions of speech sounds of English. *International Research Journal of MMC*, *2*(1), 15-28.

Stanja, J., Gritz, W., Krugel, J., Hoppe, A., & Dannemann, S. (2023). Formative assessment strategies for students' conceptions—the potential of learning analytics. *British Journal of Educational Technology*, *54*(1), 58-75.

https://doi.org/10.64014/jik.v23i2.325

Stenhaug, B. A., & Domingue, B. W. (2022). Predictive fit metrics for item response models. *Applied Psychological Measurement*, *46*(2), 136-155.

Wafubwa, R. N. (2020). Role of formative assessment in improving students' motivation, engagement, and achievement: A systematic review of literature. *International Journal of Assessment and Evaluation*, *28*(1), 17-31.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p< 0.05". *The American Statistician*, *73*(1), 1-19.

Wong, J. T., Richland, L. E., & Hughes, B. S. (2025). Immediate versus delayed low-stakes questioning: Encouraging the testing effect through embedded video questions to support students' knowledge outcomes, self-regulation, and critical thinking. *Technology, Knowledge and Learning*, *30*(3), 1421-1456.

Zainuddin, Z., Shujahat, M., Haruna, H., & Chu, S. K. W. (2020). The role of gamified e-quizzes on student learning and engagement: an interactive gamification solution for a formative assessment system. *Computers and Education*, *145*(2020), 1-48.

Zhang, D., Wang, C., Yuan, T., Li, X., Yang, L., Huang, A., ... & Zhang, L. (2023). Psychometric properties of the Coronavirus Anxiety Scale based on Classical Test Theory (CTT) and Item Response Theory (IRT) models among Chinese front-line healthcare workers. *BMC Psychology*, *11*(1), 1-10.

https://doi.org/10.64014/jik.v23i2.325