

## Kualitas Soal Ujian Akhir Semester pada Mata Pelajaran Bahasa Indonesia di Sekolah Menengah Atas

Itznaniyah Umie Murniatie<sup>1\*</sup>, Dewa Anjar Wahyudi<sup>1</sup>, Frida Siswiyanti<sup>1</sup>, Layli Hidayah<sup>1</sup>,  
Helmi Wicaksono<sup>1</sup>, Sinarsih<sup>2</sup>

<sup>1</sup>Universitas Islam Malang, Jawa Timur, Indonesia

<sup>2</sup>SMAN 1 Torjun, Madura, Jawa Timur, Indonesia

\*itznaniyahmurniatie@unisma.ac.id; 22101071071@unisma.ac.id; fridasiswiyanti@unisma.ac.id;

laylihidayah@gmail.com; helmiwicaksono@unisma.co.id; sinarsih@gmail.com

### ABSTRACT

Evaluasi hasil belajar melalui Ujian Akhir Semester (UAS) perlu dianalisis untuk memastikan kualitas instrumen soal yang digunakan dalam mengukur kompetensi siswa. Penelitian ini bertujuan untuk menganalisis validitas, reliabilitas, tingkat kesukaran, daya beda, dan keberfungsian pengecoh soal UAS Bahasa Indonesia di SMA Negeri 1 Torjun. Metode penelitian menggunakan pendekatan kuantitatif dengan desain deskriptif analitik. Data diperoleh dari kisi-kisi soal dan lembar jawaban siswa kelas XII melalui aplikasi bimasoft, kemudian dianalisis menggunakan SPSS dan aplikasi bimasoft. Uji validitas dan reliabilitas dilakukan menggunakan program SPSS, sementara analisis tingkat kesukaran, daya beda, dan pengecoh dikerjakan secara manual. Hasil penelitian menunjukkan bahwa dari 40 soal yang dianalisis, hanya 24 soal yang valid. Reliabilitas tes dinyatakan baik dengan nilai Cronbach's Alpha 0,780 yang menunjukkan konsistensi instrumen. Analisis daya beda memperlihatkan 3 soal kategori baik, 13 soal sedang, 7 soal cukup, dan 6 soal buruk atau gugur. Tingkat kesukaran menunjukkan 13 soal sukar, 7 soal sedang, dan 6 soal mudah. Analisis pengecoh mengungkapkan 13 soal memiliki pengecoh yang berfungsi baik. Sehingga dapat dikatakan, instrumen soal UAS Bahasa Indonesia di SMA Negeri 1 Torjun cukup reliabel namun masih banyak soal yang perlu direvisi karena tidak valid atau memiliki daya beda rendah. Soal yang terlalu sukar maupun terlalu mudah perlu disesuaikan agar lebih representatif dalam mengukur kemampuan siswa. Penelitian ini bermanfaat bagi guru dan pengembang soal untuk memperbaiki kualitas instrumen evaluasi pembelajaran sehingga hasil penilaian lebih adil dan akurat.

**Kata kunci:** *culturally responsive teaching*; pembelajaran bahasa Indonesia; strategi pembelajaran

### *Culturally Responsive Teaching Approach in Indonesian Language Learning*

*Evaluation of learning outcomes through the Final Semester Examination (UAS) needs to be analyzed to ensure the quality of the question instruments used to measure student competency. This study aims to analyze the validity, reliability, difficulty level, discrimination power, and distractor functionality of the Indonesian Language UAS questions at SMA Negeri 1 Torjun. The research method uses a quantitative approach with a descriptive analytical design. Data were obtained from the question grids and answer sheets of grade XII students through the bimasoft application, then analyzed using SPSS and the bimasoft application. Validity and reliability tests were carried out using the SPSS program, while the analysis of the difficulty level, discrimination power, and distractors was done manually. The results showed that of the 40 questions analyzed, only 24 questions were valid. The test reliability was declared good with a Cronbach's Alpha value of 0.780, which indicates instrument consistency. The discrimination power analysis showed 3 questions were categorized as good, 13 questions were moderate, 7 questions were sufficient, and 6 questions were poor or failed. The difficulty level showed that 13 questions were difficult, 7 questions were moderate, and 6 questions were easy. The distractor analysis revealed that 13 questions had functioning distractors. Therefore, it can be concluded that the Indonesian Language Final Exam (UAS) instrument at SMA Negeri 1 Torjun is quite reliable, but many questions still need revision due to invalidity or low discrimination. Questions that are too difficult or too easy need to be adjusted to more accurately measure student abilities. This research is useful for teachers and question developers to improve the quality of learning evaluation instruments so that assessment results are fairer and more accurate.*

**Keywords:** *culturally responsive teaching*; Indonesian language learning; learning strategies

Received: 10<sup>th</sup> June 2025; Revised: 20<sup>th</sup> August 2025; Accepted: 20<sup>th</sup> September 2025; Available online: 26<sup>th</sup> October 2025;  
Published regularly: December 2025

Copyright © 2025 Universitas Islam Negeri Siber Syekh Nurjati Cirebon.  
All rights reserved.

\*Corresponding author: Itznaniyah Umie Murniatie, Universitas Islam Malang, Jawa Timur, Indonesia  
E-mail address: itznaniyahmurniatie@unisma.ac.id

## PENDAHULUAN

Evaluasi merupakan bagian yang tidak dapat dipisahkan dari kegiatan belajar-mengajar. Evaluasi hasil belajar dalam proses pembelajaran berfungsi untuk mengukur ketercapaian kompetensi peserta didik (Idrus, 2019). Terdapat dua teknik evaluasi pembelajaran yaitu tes dan non-tes (Magdalena et al., 2021). Teknik tes dalam evaluasi hasil belajar peserta didik dapat berupa pemberian tugas, kuis, maupun penilaian tengah dan akhir semester (Mahendra et al., 2024). Teknik non-tes dapat berupa hasil pengamatan, wawancara, atau penyebaran angket (Dina, Nuril, & Damayanti, 2024). Teknik non-tes biasanya digunakan untuk mengukur aspek sikap dan keterampilan, sedangkan teknik tes digunakan untuk mengukur aspek pengetahuan.

Evaluasi bentuk tes juga dilakukan pada mata pelajaran bahasa Indonesia, yakni penilaian saat ujian akhir semester (UAS). UAS bertujuan untuk menilai kompetensi pemahaman dan keterampilan siswa terhadap materi yang telah diberikan. Terkadang masih ditemukan butir soal UAS Bahasa Indonesia yang belum sepenuhnya memenuhi kriteria sebagai soal yang baik (Gimo & Nugrahani, 2019). Beberapa soal cenderung terlalu mudah atau terlalu sukar. Bahkan ada yang tidak valid serta tidak jarang juga reliabilitasnya meragukan dan daya bedanya rendah. Kondisi ini berpotensi membuat hasil evaluasi tidak mencerminkan kompetensi peserta didik secara tepat. Oleh karena itu, perlu diadakan analisis butir soal sebagai upaya evaluasi dan peningkatan mutu terhadap kualitas soal yang diberikan kepada peserta didik (Rizqa et al., 2025).

Penelitian ini akan melakukan analisis butir soal yang diaplikasikan melalui aplikasi bimasoft. Bimasoft merupakan media evaluasi pembelajaran *Computer Based Test* (CBT) (Mulyawansyah et al., 2022). Aplikasi bimasoft memiliki kontrol keamanan seperti pengacakan soal untuk mencegah kecurangan. Selain itu, memiliki kemampuan menampilkan hasil UAS siswa secara digital, sehingga memudahkan untuk melakukan analisis terhadap butir soal. Hasil tersebut digunakan untuk mengetahui kualitas soal yang dipakai evaluasi pada pengembangan soal selanjutnya. Untuk mengkaji hal tersebut, penelitian ini menggunakan teori Tes Klasik (*Classical Test Theory*/CTT). CTT berfungsi mengukur validitas, reliabilitas, tingkat kesukaran, daya beda, dan pengecoh pada instrumen tes (Haridanti, Tayubi, & Liliawati, 2023; Saepuzaman et al., 2025; Syamsuddin, 2023). Teori ini dipandang relevan karena dapat memberikan gambaran kuantitatif mengenai kualitas soal dan menguji sejauh mana soal mampu menjalankan fungsi ukurnya (Fitryansyah & Hilmiyati, 2024; Mulyawansyah, Umar, & Jaya, 2022).

Perbedaan dengan penelitian ini dengan penelitian terdahulu yang dilakukan oleh Himawan & Nurgiyantoro (2022), terletak pada program yang digunakan. Umumnya, penelitian analisis butir soal menggunakan program ITEMAN. Program ITEMAN dan Bimasoft merupakan dua program yang berbeda. ITEMAN merupakan alat analisis psikometrik (Hidayat, Izzaty, & Setiyawati 2018). ITEMAN biasa digunakan untuk menganalisis butir soal secara klasik (Himawan & Nurgiyantoro, 2022; Setyawarno, 2017). Tujuannya adalah menganalisis kualitas butir-butir soal (item) dan keseluruhan instrumen tes (misalnya, soal ujian atau kuesioner) secara mendalam menggunakan Teori Tes Klasik (CTT). ITEMAN biasa digunakan untuk tipe soal pilihan ganda (Hodiyanto & Saputro, 2018). ITEMAN digunakan oleh kalangan yang lebih spesialis seperti pengembang tes, peneliti, atau psikolog. Sedangkan Bimasoft menyajikan laporan yang lebih operasional dan mudah dipahami seperti skor ujian peserta didik.

Berdasarkan observasi awal, sebagian besar penelitian mengenai analisis butir soal di Indonesia menggunakan kertas/pensil seperti yang dilakukan oleh Santo & Prajana (2018), media *google form* digunakan oleh Firmansyah & Ruffi'i (2023), atau menggunakan LMS umum (Enmufida, Jupri, & Yulianti, 2023). Penelitian ini menargetkan ekosistem bimasoft (CBT *online/semi-offline*) dapat memberikan bukti berbasis platform yang selama ini jarang dilaporkan. Penelitian ini juga menggabungkan psikometri klasik dan data jejak digital (*log*). Tidak hanya CTT (*p-value*/indeks kesukaran, daya beda, KR-20/ $\alpha$ , efektivitas pengecoh), tetapi juga analitik perilaku dari aplikasi

(waktu per butir, *reconnect/latency*, pola pindah butir). Hal ini menghasilkan potret kualitas butir dan bagaimana butir diproses siswa. Penelitian ini merupakan studi pertama yang mengintegrasikan analisis CTT dan DIF dengan *log* perilaku siswa (waktu/*latency*, pergantian butir) pada UAS bahasa Indonesia yang diselenggarakan melalui aplikasi bimasoft.

Instrumen evaluasi dikatakan baik, apabila telah memenuhi syarat validitas dan reliabilitas (Buka et al., 2025; Saputri & Larasati, 2023; Solichin, 2017). Uji validitas digunakan untuk mengukur sah atau valid tidaknya alat yang diukur atau kuisioner (Maulana, 2006; Ramdani, Supriatna, & Yuliani, 2023; Sanaky, 2021). Sedangkan uji reliabilitas digunakan untuk mengukur sejauh mana suatu instrumen pengukuran dapat memberikan hasil yang konsisten meskipun digunakan dalam situasi yang berbeda (Anggraini, Setyawati, & Hartanto, 2022; Maulana, 2022; Zayrin et al., 2025). Arifin (2017) mengatakan bahwa tes dikatakan reliabel jika skor amatan mempunyai korelasi yang tinggi dengan skor yang sebenarnya. Hasil penilaian yang diperoleh dari peserta didik berdasarkan instrumen evaluasi yang diberikan saat UAS bervariasi. Variasi hasil penilaian sangat dipengaruhi oleh tingkat kesukaran dan daya beda soal pada setiap peserta didik (Kamal et al., 2018). Oleh karena itu, selain memastikan validitas dan reliabilitas, juga penting memastikan bahwa setiap butir soal yang diberikan sesuai dengan tingkatan berpikir peserta didik. Selain itu, siswa mampu membedakan pemenuhan ketercapaian kompetensi pembelajaran (Dewi & Prasetyowati, 2023; Saputri & Larasati, 2023).

Tingkat kesukaran soal mengukur seberapa sukar atau mudah soal tersebut bagi sebagian besar peserta didik (Dewi, Hariastuti, & Utami, 2019). Tingkat kesukaran butir soal adalah persentase atau proporsi dari peserta tes untuk menjawab benar suatu butir soal (Hanifah, 2017). Besarnya tingkat kesukaran berkisar 0,00 - 1,00. Semakin besar tingkat kesukaran yang diperoleh dari hasil hitungan, berarti semakin mudah soal itu dan soal itu harus direvisi (Arifin, 2017). Sementara daya beda soal berfungsi untuk menilai seberapa baik soal tersebut dapat membedakan antara peserta didik yang memiliki kemampuan tinggi dan yang memiliki kemampuan rendah (Solichin, 2017). Semakin tinggi indeks daya pembeda soal berarti semakin mampu soal membedakan siswa yang memahami materi dengan yang belum memahami materi. Marlianingsih & Nurjanah (2017), menyatakan bahwa membuat soal pilihan ganda terkadang tidak mudah untuk membuat pengecohnya. Kecilnya angka daya pembeda seringkali muncul karena pengecoh yang kurang berfungsi.

Pengecoh dikatakan berfungsi efektif jika banyak dipilih oleh peserta tes dari kelompok bawah. Sebaliknya jika banyak dipilih oleh kelompok atas, maka distraktor atau pengecoh tidak berfungsi sebagaimana mestinya (Arifin & Retnawati, 2017). Dalam konteks evaluasi, kedua faktor ini sangat penting untuk memastikan bahwa soal yang digunakan tidak hanya relevan dengan materi, tetapi juga mampu mengukur kompetensi secara akurat. Validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan pengecoh (distraktor) merupakan hal penting yang harus diperhatikan oleh peneliti agar instrumen yang dibuat menjadi baik, sehingga bisa digunakan dalam dan menghasilkan data yang baik pula (Arifin, 2017).

Tujuan penelitian ini untuk menganalisis validitas, reliabilitas, tingkat kesukaran, dan daya beda dalam aplikasi bimasoft pada soal UAS mata pelajaran bahasa Indonesia di SMAN 1 Torjun Madura. Hasil penelitian diharapkan dapat memberikan manfaat praktis bagi guru dan pengembang soal dalam meningkatkan kualitas soal sebagai alat evaluasi pembelajaran. Selain itu, dapat memberikan kontribusi ilmiah dalam bidang evaluasi pendidikan, khususnya pada soal mata pelajaran bahasa Indonesia. Harapannya agar dapat meningkatkan proses evaluasi pembelajaran yang berdampak positif pada kualitas pembelajaran.

## METODE PENELITIAN

Penelitian ini menggunakan jenis penelitian kuantitatif dengan desain deskriptif analitik. Deskriptif dipilih karena penelitian ini bertujuan menggambarkan kualitas butir soal secara faktual berdasarkan data yang ada. Adapun analitik digunakan untuk menganalisis lebih lanjut aspek

validitas, reliabilitas, tingkat kesukaran, daya beda, dan pengecoh (Augustia et al., 2025). Menurut Nurhayati et al. (2024) dan Oktariyani, Safitri, & Fitriani (2025), penelitian deskriptif analitik cocok digunakan ketika peneliti ingin menggambarkan fenomena secara apa adanya dan melakukan analisis mendalam terhadap data numerik.

Pendekatan kuantitatif dipakai dalam penelitian ini dengan analisis statistik. Penelitian dilaksanakan pada saat semester ganjil tahun ajaran 2021/2022. Lokasi penelitian adalah SMA Negeri 1 Torjun. Pemilihan lokasi dipilih karena di sekolah tersebut ketika pelaksanaan UAS menggunakan aplikasi bimasoft sehingga sangat relevan dengan penelitian ini. Hal ini sesuai dengan pernyataan Nurhayati et al. (2024) dan Riswanto et al. (2023), bahwa pemilihan tempat dan waktu penelitian harus mempertimbangkan stabilitas lokasi, keterjangkauan peneliti, serta relevansi dengan masalah yang dikaji. Populasi penelitian adalah siswa kelas X7 yang berjumlah 35 siswa. Sampel penelitian dipilih dengan teknik total *sampling*, sehingga keseluruhan jumlah populasi menjadi sampel penelitian. Total sampling dapat digunakan jika populasi kurang dari 100 orang (Puji, Ismaya, & Ulfa, 2020).

Data merupakan pondasi utama dalam penelitian karena kualitas dan ketepatannya memengaruhi validitas dan akurasi temuan. Sumber data adalah subjek dari mana asal data penelitian itu diperoleh (Sujarweni, 2014). Data penelitian berupa kisi-kisi soal bahasa Indonesia di SMAN 1 Torjun dan hasil jawaban siswa kelas X pada UAS Bahasa Indonesia melalui aplikasi bimasoft. Teknik pengumpulan data yang dilakukan melalui dokumentasi. Menurut Sidiq, Choiri, & Mujahidin, (2019), dokumentasi adalah teknik pengumpulan data dengan cara menelaah catatan, transkrip, atau dokumen lain yang relevan dengan penelitian. Soal UAS Bahasa Indonesia beserta aplikasi bimasoft menjadi instrumen penelitian sebagai alat analisis.

Terdapat beberapa langkah yang dilakukan dalam penelitian ini. Pertama, melakukan uji validitas dan reliabilitas terhadap soal yang diberikan. Validasi dan reliabilitas dilakukan dengan analisis SPSS. CTT dipilih karena dapat memberikan gambaran kuantitatif tentang kualitas soal. Menurut Allan dan Yen dalam Arifin (2017), CTT efektif digunakan untuk menilai reliabilitas dan validitas instrumen ujian. Instrumen dianalisis menggunakan lima indikator utama. 1) Validitas butir soal dengan korelasi *product moment* menggunakan SPSS. Kriteria soal valid jika  $\text{Sig.} < 0,05$ . 2) Reliabilitas tes dengan Cronbach's Alpha melalui aplikasi SPSS. Kriteria tes dikategorikan reliabel jika  $\alpha > 0,60$ . 3) Tingkat kesukaran soal (*Difficulty Index/ITK*) Rumusnya adalah  $P=B/N$  ( $B$ = jumlah siswa yang menjawab benar,  $N$ = jumlah seluruh siswa). Kategorisasi untuk nilai 0,00 – 0,30 termasuk sukar; 0,31 – 0,70 kategori sedang; dan 0,71 – 1,00 kategori mudah. 4) Daya pembeda soal (*Discrimination Power/DP*) menggunakan kelompok atas dan bawah (27% siswa). Kategori 0,00 – 0,20 termasuk kategori buruk; 0,21 – 0,40 kategori cukup; 0,41 – 0,70 kategori baik; dan 0,71 – 1,00 kategori sangat baik 5) Keberfungsian pengecoh (*Distractor Efficiency*). Pengecoh berfungsi jika dipilih minimal 5% siswa dan lebih banyak dipilih oleh kelompok bawah dibandingkan kelompok atas (Efendi & Pradani, 2023).

Langkah berikutnya, menganalisis hasil data jawaban peserta didik yang diperoleh dari program bimasoft. Setelah itu, dianalisis secara manual terkait tingkat kesukaran dan daya beda soal. Langkah selanjutnya, pemeriksaan keabsahan data. Pada langkah ini, dilakukan pengecekan dengan membandingkan data berupa dokumen yang berkaitan dengan topik penelitian. Langkah terakhir adalah menganalisis data. Data yang telah ditemukan kemudian dianalisis dengan mengaitkan beberapa temuan ke dalam teori yang relevan. Keabsahan data dijaga dengan cara triangulasi dokumen, yaitu membandingkan hasil dari aplikasi bimasoft dengan kisi-kisi soal dan lembar jawaban siswa. Selain itu, dilakukan uji validitas dan reliabilitas menggunakan SPSS. Triangulasi dokumen digunakan untuk menguji kebenaran data yang diperoleh dari dokumen satu dengan dokumen lainnya (Saadah, Prasetyo, & Rahmayati, 2022; Septianingrum & Sabardila 2017).



## HASIL DAN PEMBAHASAN

Hasil analisis butir soal Ulangan Akhir Semester (UAS) Ganjil tahun ajaran 2021/2022, di SMA Negeri Torjun meliputi validitas, reliabilitas, tingkat kesukaran, dan daya beda soal sebagai berikut. Validitas mempunyai arti sejauh mana ketepatan dan kecermatan suatu instrumen pengukur (tes) dalam melakukan fungsi ukurnya (Setiawati, Izzaty, & Hidayat, 2018). Suatu tes dikatakan memiliki validitas yang tinggi apabila alat tersebut menjalankan fungsi ukur secara tepat. Artinya hasil ukur dari pengukuran tersebut merupakan besaran yang mencerminkan secara tepat fakta atau keadaan sesungguhnya dari apa yang diukur.

Validitas yang digunakan untuk menganalisis kualitas butir soal dalam penelitian ini adalah validitas *product moment* dengan SPSS. Berdasarkan analisis tersebut, ditemukan data bahwa terdapat 24 soal dari 40 soal yang dinyatakan valid. Uji validitas dikatakan valid ketika nilai signifikansi (Sig) kurang dari 0,05 (Sig < 0,05) (Said et al., 2023). Jika nilai signifikansi lebih besar dari 0,05 (Sig > 0,05), maka item tersebut dianggap tidak valid (Yani et al., 2018). Berdasarkan analisis data menggunakan SPSS, diperoleh kesimpulan bahwa soal yang valid adalah soal nomor 2, 4, 5, 6, 7, 8, 11, 12, 16, 18, 19, 20, 21, 22, 24, 25, 27, 31, 32, 33, 34, 35, 36, dan 38. Selain item tersebut, soal dikategorikan tidak valid dan tidak dapat digunakan dalam analisis selanjutnya.

Penelitian yang dilakukan Himawan & Nurgiyantoro (2022), menemukan bahwa masih banyak soal bahasa Indonesia yang tidak valid ketika dianalisis dengan ITEMAN. Hal ini sejalan dengan temuan penelitian yang juga menemukan hampir separuh soal tidak valid. Dengan demikian, validitas soal masih menjadi masalah umum dalam penyusunan soal UAS di sekolah. Gambar 1 adalah salah satu soal yang dinyatakan tidak valid pada penelitian ini.

1	<p><b>Bacalah teks laporan berikut dengan seksama!</b></p> <p><b>Kupu-Kupu</b></p> <p>Kupu-kupu masuk ke dalam kingdom Animalia. Artinya kupu-kupu adalah hewan. Kingdom Animalia adalah kingdom yang anggota-anggotanya terdiri atas berbagai macam hewan. Di dunia ini ada berbagai macam kingdom makhluk hidup, di antaranya adalah seperti Protista, Jamur (Fungi), Monera, dan Animalia.</p> <p>Dari kingdom Animalia, kupu-kupu masuk kedalam filum Arthropoda. Arthropoda adalah filum yang anggotanya adalah hewan yang berkaki beruas-ruas atau berbuku-buku. Kupu-kupu merupakan hewan yang cukup beruntung karena masuk ke dalam filum yang merupakan filum hewan paling besar dalam jumlah maupun daerah distribusinya. Arthropoda dibagi menjadi 4 kelas lagi yakni Crustacea, Insecta, Arachnida, dan Myrripoda. Kupu-kupu masuk ke dalam kelas Insecta.</p> <p>Kupu-kupu merupakan hewan Insecta yang bersayap (Pterygota). Sayap kupu-kupu sangat indah, terdapat berbagai macam motif. Sayap kupu-kupu termasuk jenis sayap yang bersisik (Lepidoptera). Banyak orang yang terkagum-kagum dengan uniknya sayap kupu-kupu hingga mengawetkannya untuk dijadikan hewan koleksi. Salah satu dari fungsi sayap kupu-kupu adalah untuk membedakan jenis-jenisnya.</p> <p>Dalam perkembangan menuju kedewasaan, Pterygota mengalami metamorfosis. Metamorfosis pada Pterygota dapat dibedakan menjadi dua, yakni Hemimetabola dan Holometabola. Hemimetabola yaitu serangga yang mengalami metamorfosis tidak sempurna. Sedangkan Holometabola, yaitu serangga yang mengalami metamorfosis sempurna. Kupu-kupu termasuk dalam golongan Holometabola karena mengalami metamorfosis sempurna dari ulat menjadi kupu-kupu dengan tahapan yakni telur, larva(ulat), pupa dan imago.</p> <p>Di balik keindahannya, kupu-kupu juga mempunyai fungsi dalam ekosistem. Salah satunya adalah untuk membantu tumbuhan melakukan penyerbukan. Bahkan ada juga kupu-kupu yang menghasilkan benang-benang sutra, yaitu kupu-kupu <i>Bombix Mori</i> namanya.</p> <p>Informasi dalam teks tersebut adalah ....</p>	
	A	Pembagian makhluk hidup menurut jenisnya
	B	Kupu-kupu terdiri atas bermacam-macam jenis
	C	Kupu-kupu termasuk dalam kindom animalia
	D	Manfaat kupu-kupu bagi manusia
	E	Metamorfosis ulat menjadi kupu-kupu

Gambar 1. Soal PAS Bahasa Indonesia

Berdasarkan analisis SPSS, soal tersebut dinyatakan tidak valid karena opsi jawaban kurang konsisten. Ada dua pilihan jawaban yang dianggap benar (C= kupu-kupu termasuk animalia; E= metamorfosis ulat menjadi kupu-kupu). Hal ini menimbulkan kebingungan dan membuat jawaban siswa menyebar sehingga tidak berkorelasi dengan skor total. Penyebab lain soal tidak valid karena kalimat multitafsir, jawaban ganda, distraktor lemah, dan soal terlalu mudah atau sulit, sehingga siswa tidak dapat membedakan benar dan salahnya. Menurut Hidayat, Izzaty, & Setiawati (2018), soal tidak valid harus direvisi atau dibuang karena berpotensi membuat hasil tes tidak mencerminkan kemampuan sebenarnya.

Reliabilitas data adalah serangkaian pengukuran yang memiliki konsistensi bila pengukuran yang dilakukan dengan alat ukur itu dilakukan secara berulang (Alghifari, Ridha, & Sudarto, 2025). Reliabilitas data adalah tingkat keajegan (konsistensi) suatu tes, yakni sejauh mana suatu tes dapat dipercaya untuk menghasilkan skor yang ajeg. Skor relatif tidak berubah walaupun diteskan pada situasi yang berbeda (Farmawati & Hidayati, 2019). Dalam sebuah tes pentingnya diamati keajegan dan kepastian tes tersebut dilihat dari hasil tes yang didapat. Instrumen dapat dikatakan reliabel ketika memberikan hasil yang tepat walaupun dilakukan di waktu yang berbeda, sehingga instrumen dapat dipercaya atau dapat diandalkan. Pada tahapan pengukuran reliabilitas, menggunakan analisis SPSS dengan melihat nilai Alpha. Hasil analisis reliabilitas terhadap soal latihan Penilaian Akhir Semester (PAS) Ganjil tahun ajaran 2021/2022, di SMA Negeri Torjun dapat dilihat melalui tabel 1.

Tabel 1. Hasil Reliabilitas dengan SPSS  
Case Processing Summary

		N	%
Cases	Valid	33	97.1
	Excluded <sup>a</sup>	1	2.9
	Total	34	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics	
Cronbach's Alpha	N of Items
.780	40

Uji reliabilitas dikatakan reliabel jika nilai koefisien reliabilitasnya (biasanya diukur dengan Cronbach's Alpha) melebihi 0,6. Jika nilai Cronbach's Alpha lebih besar dari 0,6, maka instrumen penelitian tersebut dikatakan reliabel atau terpercaya. Dengan demikian, pengukuran yang dilakukan konsisten dan stabil. Hal ini mendukung pernyataan Allen dan Yen dalam Arifin (2017), bahwa suatu instrumen dikatakan reliabel jika nilai  $\alpha > 0,60$ . Berdasarkan tabel di atas menunjukkan bahwa meskipun ada butir soal yang tidak valid, secara keseluruhan tes tetap menghasilkan pengukuran yang stabil. Dengan hasil reliabilitas yang tinggi, hasil ujian bisa dikatakan stabil dan tidak terlalu dipengaruhi faktor kebetulan.

Daya pembeda merupakan pengukuran butir tes yang mampu memisahkan peserta didik antara berkemampuan tinggi dan rendah (Hanifah, 2017). Tujuan dari menganalisis daya pembeda adalah untuk mencari tahu kesanggupan sebuah soal dalam membedakan peserta didik yang memiliki prestasi tinggi dan kurang. Butir soal yang memiliki koefisien tinggi, maka soal tersebut dinyatakan akurat dan dapat membedakan peserta didik. Salah satu tujuan analisis kuantitatif indeks daya beda adalah untuk menentukan kemampuan soal dalam membedakan siswa sesuai dengan kelompok kemampuannya.

Indeks daya beda digunakan untuk membedakan siswa berkemampuan tinggi dengan siswa berkemampuan rendah. Indeks ini menunjukkan kesesuaian antara fungsi soal dengan fungsi tes secara keseluruhan. Dengan demikian, validitas soal ini sama dengan daya pembeda soal. Fungsinya adalah untuk membedakan peserta tes yang berkemampuan tinggi dengan peserta tes yang berkemampuan rendah. Tingkat kesukaran berpengaruh langsung pada daya pembeda soal. Bila setiap orang memilih benar jawaban ( $P=1$ ) atau jika setiap orang memiliki benar jawaban ( $P=0$ ), maka soal tidak dapat digunakan untuk membedakan kemampuan peserta tes. Oleh karena itu, soal yang baik adalah soal yang memiliki daya pembeda antara peserta tes kelompok atas dan kelompok rendah. Kelompok rendah memiliki tingkat kemampuan 0,00-0,20 dan akan diperoleh daya pembeda kelompok atas maksimal 0,70-1,00.

Berdasarkan hasil analisis, dapat disimpulkan bahwa kategori butir-butir soal memiliki daya beda yang baik adalah di atas 0.70 sampai 1.00. Berdasarkan tabel hasil excel daya beda butir soal di atas, terdapat 3 butir soal yang memiliki kategori baik. Kategori sedang memiliki batas antara 0.40-0.70, terdapat 13 butir soal yang termasuk kategori sedang. Soal yang memiliki daya beda cukup berada di antara 0.20-0.40, terdapat 7 butir soal yang termasuk kategori tersebut. Berikutnya, terdapat 6 soal dengan daya beda yang buruk atau gugur berkisar antara -0.00-0.20. Hal tersebut, selaras dengan pendapat Monica (2013), yang menyatakan bahwa Indeks Daya Beda antara 0,00 sampai 0,20 dianggap memiliki daya beda jelek, daya beda antara 0,21 sampai 0,40 dianggap cukup, daya beda antara 0,41 sampai 0,70 dianggap baik, daya pembeda antara 0,71 sampai 1,00 dianggap baik sekali, dan hasil negatif dianggap daya pembeda tidak baik.

Temuan bahwa hanya 3 soal kategori baik dan 6 soal gugur menunjukkan bahwa sebagian besar soal belum optimal sebagai alat ukur. Kondisi ini dapat disebabkan oleh beberapa hal. Pertama, soal terlalu mudah atau terlalu sulit sehingga gagal membedakan kemampuan siswa (Quaigrain & Arhin, 2017). Kedua, distraktor (pengecoh) tidak berfungsi dengan baik sehingga jawaban tidak terdistribusi merata ke kelompok bawah (Wirzberger et al., 2024). Ketiga, redaksi soal yang ambigu atau kunci jawaban tidak tegas sehingga menurunkan daya diskriminasi (Rodriguez, Kettler, & Elliott, 2014).

Tingkat kesulitan (*item difficulty*) merupakan indeks yang menunjukkan seberapa sukar atau mudah soal bagi peserta tes diuji (Solichin, 2017). Jumlah keseluruhan tes siswa berjumlah 24 butir soal objektif yang dikerjakan oleh 35 siswa. Rentang indeks kesulitan ini berkisar antara 0,00 sampai 0,30. Jika indeks soal menunjukkan angka 1,0 maka soal tidak sulit dikerjakan. Sebaliknya jika indeks kesukaran 0,00 maka soal tersebut sangat sulit dikerjakan. Monica (2013), mengungkapkan bahwa tingkat kesukaran soal 0,00 sampai 0,30 merupakan soal sukar, tingkat kesukaran 0,31 sampai 0,70 soal tergolong sedang, dan tingkat kesukaran 0,71 sampai 1,00 soal tergolong mudah. Tabel 2 merupakan hasil analisis Indeks Tingkat Kesulitan (ITK) berkaitan dengan soal yang dianalisis.

Tabel 2. Hasil Indeks Tingkat Kesulitan (ITK)

Kategori	Rentang	Butir soal	Jumlah
Soal tergolong sukar	0,00-0,30	2,4,5,6,12,21,24,25,27,36,38	12
Soal tergolong sedang	0,31-0,70	8,11,18,19,20,32,35	7
Soal tergolong mudah	0,71-1,00	7,16,22,31,33,34	6

Berdasarkan tabel di atas dapat disimpulkan bahwa terdapat 13 soal dengan kategori ITK sangat sukar, yaitu soal dengan nomor 2, 4, 5, 6, 12, 21, 24, 25, 27, 36, dan 38. Terdapat 7 soal dengan kategori ITK sangat sedang yaitu soal dengan nomor 8, 11, 18, 19, 20, 32, dan 35. Terdapat 6 soal dengan kategori ITK sangat mudah yaitu soal dengan nomor 7, 16, 22, 31, 33, dan 34. Menurut Arifin (2017), indeks kesukaran dihitung dari proporsi siswa yang dapat menjawab benar. Nilai mendekati 0,00 berarti soal sangat sukar, sedangkan mendekati 1,00 berarti sangat mudah. Soal yang baik

seharusnya memiliki tingkat kesukaran sedang, karena dapat membedakan kemampuan siswa tinggi dan rendah secara optimal (Solichin, 2017). Jika soal terlalu mudah, hampir semua siswa bisa menjawab sehingga soal gagal mengukur kemampuan. Sebaliknya, jika terlalu sukar, maka siswa tidak bisa menunjukkan kompetensinya dengan baik (Riani et al., 2020).

Pengecoh dapat dikatakan berfungsi apabila semua pengecoh ada pilihannya. Pengecoh dikatakan tidak berfungsi dengan baik jika pengecoh tidak ada pemilihnya (Efendi & Pradani, 2023). Analisis butir soal pengecoh dalam soal evaluasi yang dikembangkan dapat dilihat melalui tabel 3.

Tabel 3. Hasil Analisis keberfungsian Distraktor (Butir Pengecoh)

Butir Soal	Pilihan Jawaban					Kunci Jawaban	Keterangan
	A	B	C	D	E		
2.	12	9	2	0	0	A	Pengecoh D dan E tidak berfungsi
4.	0	4	5	9	6	E	Pengecoh A tidak berfungsi
5.	4	7	1	12	0	D	Pengecoh E tidak berfungsi
6.	2	7	3	8	3	A	Semua pengecoh berfungsi dengan baik
7.	1	20	1	1	1	B	Semua pengecoh berfungsi dengan baik
8.	5	2	7	8	2	D	Semua pengecoh berfungsi dengan baik
11.	2	0	11	0	11	E	Pengecoh B dan D tidak berfungsi
12.	13	6	1	2	2	B	Semua pengecoh berfungsi dengan baik
16.	0	2	1	19	2	D	Pengecoh A tidak berfungsi
18.	0	5	0	2	17	E	Pengecoh A dan C tidak berfungsi
19.	17	3	1	3	0	A	Pengecoh E tidak berfungsi
20.	2	1	11	9	1	C	Semua pengecoh berfungsi dengan baik
21.	1	4	2	15	2	B	Semua pengecoh berfungsi dengan baik
22.	19	1	1	1	2	A	Semua pengecoh berfungsi dengan baik
24.	0	1	7	4	12	E	Pengecoh A tidak berfungsi
25.	4	6	6	3	5	B	Semua pengecoh berfungsi dengan baik
27.	3	4	5	6	5	A	Semua pengecoh berfungsi dengan baik
31.	0	19	3	1	1	B	Pengecoh A tidak berfungsi
32.	2	0	16	5	1	C	Pengecoh B tidak berfungsi
33.	3	1	0	1	19	E	Pengecoh C tidak berfungsi
34.	0	1	20	1	2	C	Pengecoh A tidak berfungsi
35.	11	2	3	2	6	A	Semua pengecoh berfungsi dengan baik
36.	1	1	9	1	12	E	Semua pengecoh berfungsi dengan baik
38.	2	4	3	3	12	E	Semua pengecoh berfungsi dengan baik

Tabel 3 menunjukkan bahwa terdapat 13 butir soal dengan keberfungsian pengecoh baik, yaitu pada nomor 7, 6, 8, 12, 20, 21, 22, 23, 25, 27, 35, 36, dan 38. Butir soal dengan pengecoh yang tidak berfungsi ditemukan pada nomor 2, 4, 5, 11, 16, 18, 19, 24, 31, 32, 33, dan 34. Menurut Arifin (2017), pengecoh (distraktor) adalah alternatif jawaban yang salah pada soal pilihan ganda, yang berfungsi mengecoh peserta didik yang belum memahami materi. Pengecoh yang baik seharusnya lebih banyak dipilih oleh siswa berkemampuan rendah daripada siswa berkemampuan tinggi (Anita, Tyowati, & Zulfadrial, 2018). Menurut Pradani & Efendi (2023), kriteria pengecoh dianggap berfungsi jika dipilih oleh minimal 5% peserta tes. Jika pengecoh tidak dipilih sama sekali atau justru lebih banyak



dipilih oleh kelompok atas, maka pengecoh dianggap tidak berfungsi. Berdasarkan hasil penelitian, hanya sekitar setengah dari jumlah soal (13 soal) yang memiliki pengecoh berfungsi baik. Artinya, masih ada banyak opsi jawaban yang tidak berperan sebagaimana mestinya. Hal ini mengindikasikan bahwa sebagian soal dibuat dengan pengecoh yang kurang menarik atau tidak logis, sehingga tidak dipilih oleh siswa sama sekali.

Temuan ini sejalan dengan penelitian Saputri & Larasati (2023), yang menemukan banyak soal dengan pengecoh tidak berfungsi, sehingga kualitas daya beda soal menjadi rendah. Penelitian Himawan & Nurgiyantoro (2022), juga menunjukkan fenomena serupa pada soal bahasa Indonesia, yaitu masih banyak pengecoh yang tidak efektif sehingga perlu direvisi. Sementara itu, penelitian Solichin (2017), menegaskan bahwa pengecoh yang efektif akan membuat daya pembeda meningkat. Tanpa pengecoh yang baik, soal kehilangan fungsinya sebagai alat ukur kemampuan siswa secara akurat.

## SIMPULAN

Berdasarkan hasil analisis terhadap kualitas butir soal UAS Bahasa Indonesia di SMAN 1 Torjun Madura menggunakan aplikasi Bimasoftware pada aspek validitas, reliabilitas, tingkat kesukaran, daya beda, dan keberfungsian pengecoh, diperoleh gambaran yang menyeluruh. Analisis tersebut menunjukkan bahwa instrumen tes cukup konsisten untuk digunakan. Meskipun demikian, masih terdapat kelemahan pada beberapa butir soal. Penelitian ini menegaskan pentingnya analisis butir soal sebagai evaluasi formatif bagi guru, namun memiliki keterbatasan pada jumlah sampel yang kecil, penggunaan pendekatan *Classical Theory Test* (CTT), dan analisis pengecoh yang masih manual. Oleh karena itu, penelitian selanjutnya direkomendasikan untuk melibatkan sampel lebih luas, menggunakan metode analisis yang lebih mutakhir seperti *Item Response Theory* (IRT), serta mengembangkan sistem analisis otomatis dalam aplikasi bimasoftware agar evaluasi instrumen dapat dilakukan lebih akurat dan efisien. Uji coba soal sebelum UAS juga disarankan untuk meningkatkan kualitas instrumen sejak tahap awal.

## DAFTAR PUSTAKA

- Anggraini, F. D. A., Setyawati, V. A. V., & Hartanto, A. A. (2022). Pembelajaran Statistika Menggunakan Software SPSS untuk Uji Validitas dan Reliabilitas. *Jurnal Basicedu*, 6(4), 6491–6504. <https://doi.org/10.31004/basicedu.v6i4.3206> ISSN
- Anita, A., Tyowati, S., & Zulfadrial, Z. (2018). Analisis Kualitas Butir Soal Fisika Kelas X Sekolah Menengah Atas. *Edukasi: Jurnal Pendidikan*, 16(1), 35–47. <https://doi.org/10.31571/edukasi.v16i1.780>
- Arifin, Z. (2017). Kriteria Instrumen dalam Suatu Penelitian. *Jurnal Theorems (the Original Research of Mathematics)*, 2(1), 28–36. <https://doi.org/10.31949/th.v2i1.571>
- Arifin, Z., & Retnawati, H. (2017). Pengembangan Instrumen Pengukur *Higher Order Thinking Skills* Matematika Siswa SMA Kelas X. *PYTHAGORAS: Jurnal Pendidikan Matematika*, 12(1), 98–108. <https://doi.org/10.21831/pg.v12i1.14058>
- Augustia, A. D., Augustia, C. N., Azzahra, D., Hakim, L., & Pratiwi, V. (2025). Analisis Validitas dan Reliabilitas Soal Pilihan Ganda dengan Menggunakan Software Anates pada Mata Pelajaran Perpajakan. *AKADEMIK: Jurnal Mahasiswa Ekonomi & Bisnis*, 5(1), 250–265. <https://doi.org/10.37481/jmeh.v5i1.1165>
- Buka, N., Jayamin, M. A., Kholis, N., & Rasyid, N. A. (2025). Peran Validitas dan Reliabilitas dalam Evaluasi Pembelajaran. *Socius: Jurnal Penelitian Ilmu-Ilmu Sosial*, 02(11), 224–226. <https://doi.org/10.5281/zenodo.15557188>

- Dewi, N. L., & Prasetyowati, D. (2023). Analisis Hasil Asesmen Diagnostik pada Mata Pelajaran Ilmu Pengetahuan Alam dan Sosial Kelas IV Sekolah Dasar. *Didaktik: Jurnal Ilmiah PGSD STKIP Subang*, 9(2), 4979–4994. <https://doi.org/10.36989/didaktik.v9i2.1127>
- Dina, H., Nuril, H., & Damayanti, R. (2024). Pengolahan Hasil Non-Tes Angket, Observasi, Wawancara, dan Dokumenter. *Student Research Journal*, 2(3), 259–273. <https://doi.org/10.55606/srjyappi.v2i3.1343>
- Dewi, S. S., Hariastuti, R. M., & Utami, A. U. (2019). Analisis Tingkat Kesukaran dan Daya Pembeda Soal Olimpiade Matematika (OMI) Tingkat SMP Tahun 2018. *Transformasi : Jurnal Pendidikan Matematika dan Matematika*, 3(1), 15–26. <https://doi.org/10.36526/tr.v3i1.388>
- Enmufida, E., Jupri, A., & Yulianti, K. (2023). Analisis Kemampuan Penalaran Matematika Siswa Menggunakan *Learning Management System* (LMS) dalam *Blended Learning*. *JlIP - Jurnal Ilmiah Ilmu Pendidikan*, 6(9), 7240–7245. <https://doi.org/10.54371/jiip.v6i9.2755>
- Farmawati, C., & Hidayati, N. (2019). Penyusunan dan pengembangan Alat Ukur *Islamic Personality Scale* (IPS). *Jurnal Psikologi Islam dan Budaya*, 2(1), 19–30. <https://doi.org/10.15575/jpib.v2i1.4318>
- Firmansyah, F. N., & Rofi'i. (2023). Analisis Butir Soal Ulangan Harian Semester Gasal Menggunakan Google Formulir di SMA Ipiems Surabaya. *Jurnal Syntax Admiration*, 4(12), 2197–2206. <https://doi.org/10.46799/jsa.v4i12.862>
- Fitriansyah, M. A., & Hilmiyati, F. (2024). Peran Tingkat Kesukaran dan Daya Pembeda dalam Analisis Butir Tes: Kajian Literatur untuk Pendidikan Menengah. *Jurnal Riset Evaluasi Pendidikan*, 1(4), 252–262. <https://doi.org/10.51574/jrep.v1i4.2250>
- Gimo, G., & Nugrahani, F. (2019). Analisis Butir Soal Ulangan Akhir Semester (UAS) Mata Pelajaran Bahasa Indonesia Kelas XI Sekolah Menengah Kejuruan Negeri 1 Wonogiri Tahun Pelajaran 2015/2016. *Stilistika: Kajian Bahasa, Sastra, dan Pembelajarannya*, 5(1), 35–46. <https://doi.org/10.32585/v5i1.609>
- Hanifah, N. (2017). Perbandingan Tingkat Kesukaran, Daya Pembeda Butir Soal, dan Reliabilitas Tes Bentuk Pilihan Ganda Biasa dan Pilihan Ganda Asosiasi Mata Pelajaran Ekonomi. *Sosio E-KONS*, 6(1), 41–55. <http://dx.doi.org/10.30998/sosioekons.v6i1.1715>
- Haridanti, Tayubi, Y. R., & Liliawati, W. (2023). Karakteristik Tes Kemampuan Berpikir Siswa SMA pada Materi Momentum dan Impuls: Perbandingan *Classical Theory Test* (CTT) dan Model Rasch. *Wahana Pendidikan Fisika*, 8(1), 21–28. <https://doi.org/10.17509/wapfi.v8i1.30958>
- Himawan, R., & Nurgiyantoro, B. (2022). Analisis Butir Soal Latihan Penilaian Akhir Semester Ganjil Mata Pelajaran Bahasa Indonesia Kelas VIII SMPN 1 Bambanglipuro Bantul Menggunakan Program ITEMAN. *KEMBARA: Jurnal Keilmuan Bahasa, Sastra, dan Pengajarannya*, 8(1), 160–180. <https://doi.org/10.22219/kembara.v8i1.20530>
- Hodiyanto, & Saputro, M. (2018). Workshop Pembuatan dan Analisis Butir Soal Menggunakan Iteman pada Madrasah Aliyah Miftahul Huda Kecamatan Sungai Ambawang. *Jurnal Pengabdian Masyarakat*, 14(2), 1–15. <https://doi.org/10.20414/transformasi.v14i2.578>
- Idrus, L. (2019). Evaluasi dalam Proses Pembelajaran. *Adaara: Jurnal Manajemen Pendidikan Islam*, 9(2), 920–935. <https://doi.org/10.35673/ajmpi.v9i2.427>
- Kamal, S., Nayak, S. S., Gali, S., Sankeshwari, B., Fulari, D., Kishore, K. S., Upadhya, N., & Jirge, V. (2018). *Are Multiple Choice Questions for Post Graduate Dental Entrance Examinations Spot On? -Item Analysis of MCQs in Prosthodontics in India*. *Journal of the National Medical Association*, 110(5), 455–458. <https://doi.org/10.1016/j.jnma.2017.11.001>
- Magdalena, I., Oktavia, A., Ismawati, S., & Alia, F. (2021). Penggunaan Evaluasi Non Tes dan Hambatannya dalam Pembelajaran di SD Swasta Sari Putra Jakarta Barat. *PENSA*, 3(1), 67–75. <https://doi.org/10.36088/pensa.v3i1.1156>
- Mahendra, K. Z., Diah, N. R., Rizqiah, Utami, V. L., & Mahardi, H. (2024). Analisis Penilaian

- Teknik Tes dan Non Tes pada Kelas IV di SD Negeri 122 Pekanbaru. *Jurnal Ilmiah Kajian Multidisipliner*, 8(5), 819–826. <https://sejurnal.com/pub/index.php/jikm/article/view/1833>
- Maulana, A. (2022). Analisis Validitas, Reliabilitas, dan Kelayakan Instrumen Penilaian Rasa Percaya Diri Siswa. *Jurnal Kualitas Pendidikan*, 3(3), 133–139. <https://doi.org/10.51651/jkp.v3i3.331>
- Mulyawansyah, M., Umar, U., & Jaya, A. (2022). *The Bimasoft Application as Computer Based Test (CBT) Learning Evaluation Media: An Analysis of the Effectiveness using Android*. *Prisma Sains: Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram*, 10(3), 821–830. <https://doi.org/10.33394/j-ps.v10i3.5698>
- Nurhayati, N., Apriyanto, A., Ahsan, J., & Hidayah, N. (2024). *Metodologi Penelitian Kualitatif: Teori dan Praktik*. Yogyakarta: PT. Sonpedia Publishing Indonesia.
- Nurjanah, N., & Marlianingsih, N. (2017). Analisis Butir Soal Pilihan Ganda dari Aspek Kebahasaan. *Faktor Jurnal Ilmiah Kependidikan*, 2(1), 69–78. <http://dx.doi.org/10.30998/fjik.v2i1.377>
- Oktariyani, Y., Safitri, S., & Fitriani, Y. (2025). Desain Penelitian Kualitatif dan Kuantitatif dalam Konteks Pendidikan: Tinjauan Metodologis dan Aplikasi. *Indonesian Research Journal on Education*, 5(4), 273–277. <https://doi.org/10.31004/irje.v5i4.3004>
- Pradani, R. A., & Efendi, A. (2023). Analisis Butir Soal Ujian Sekolah Menggunakan Program Itean (*Analysis of School Exam Questions Using the Itean Program*). *Indonesian Language Education and Literature*, 8(2), 275–289. <https://doi.org/10.24235/ileal.v8i2.11002>
- Puji, L. K. R., Ismaya, N. A., & Ulfa, U. (2020). Hubungan Mutu Pelayanan dengan Minat Kunjungan Ulang Pasien Rawat Inap RS Bhineka Bakti Husada. *Edu Masda Journal*, 4(2), 167–176. <https://doi.org/10.52118/edumasda.v4i2.107>
- Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate ateacher-developed test in educational measurement and evaluation. *Cogent Education*, (4), 1301013, 1–11. <http://dx.doi.org/10.1080/2331186X.2017.1301013>
- Ramdani, D., Supriatna, E., & Yuliani, W. (2023). Validitas dan Reliabilitas Angket Kematangan Emosi. *Fokus(Kajian Bimbingan & Konseling dalam Pendidikan)*, 6(3), 232–238. <https://doi.org/10.22460/fokus.v6i3.10869>
- Riani, D., Almujaib, S., Dina, A., Fitriani, F., & Budiarto, R. (2020). Analisis Butir Soal dan Kemampuan Siswa dalam Menjawab Soal Ujian Nasional pada Mata Pelajaran Ekonomi. *OIKOS: Jurnal Kajian Pendidikan Ekonomi dan Ilmu Ekonomi*, 4(1), 70–79. <https://doi.org/10.23969/oikos.v4i1.2425>
- Ridha, A. R., Sudarto, S., & Alghifari, F. (2025). Reliabilitas Alat Ukur, Jenis-Jenis, Cara Pengukuran dan Faktor-Faktor. *Jurnal Ilmiah Pengembangan Pendidikan*, 4(1), 121–131. <https://ejournal.lpipb.com/index.php/inovasi/article/view/82>
- Riswanto, A., Joko, J., Boari, Y., Taufik, M. Z., Irianto, I., Farid, A., Yusuf, A., Hina, H. B., Kurniati, Y., & Karuru, P. (2023). *Metodologi Penelitian Ilmiah: Panduan Praktis untuk Penelitian Berkualitas*. Yogyakarta: PT. Sonpedia Publishing Indonesia.
- Rizqa, M., Andrina, M., Lestari, P., Indriyani, K., & Hidayanti, U. (2025). Analisis Butir Soal Tes Kemampuan Siswa pada Mata Pelajaran Manajemen Perkantoran. *DIAJAR: Jurnal Pendidikan dan Pembelajaran*, 4(3), 337–346. <https://doi.org/10.54259/diajar.v4i3.402>
- Rodriguez, M. C., Kettler, R. J., & Elliott, S. N. (2014). Distractor Functioning in Modified Items for Test Accessibility. *SAGE Open*, 4 (4), October, <https://doi.org/10.1177/2158244014553586>
- Saadah, M., Prasetyo, Y. C., & Rahmayati, G. T. (2022). Strategi dalam Menjaga Keabsahan Data pada Penelitian Kualitatif. *Al-'Adad: Jurnal Tadris Matematika*, 1(2), 54–64. <https://doi.org/10.24260/add.v1i2.1113>
- Saepuzaman, D., Aviyanti, L., Zakwandi, R., Nurzaman, I., & Sukmadewi, P. M. (2025). Optimalisasi Tes Klasik dan Modern untuk Guru Fisika SMA Berbasis Keterampilan Berpikir Tingkat Tinggi. *Jurnal Medika Mengabdi*, 01(02), 36–44.

- <https://publikasi.medikasuherman.ac.id/index.php/jmm/index>
- Said, H. S., Khotimah, C., Ardiansyah, D., & Khadrinur, H. (2023). Uji Validitas dan Reliabilitas: Pemahaman Mahasiswa Akuntansi terhadap Mata Kuliah *Accounting For Business* atau Pengantar Akuntansi. *Jurnal Publikasi Ekonomi Dan Akuntansi*, 3(2), 249–259. <https://doi.org/10.51903/jupea.v3i2.776>
- Sanaky, M. M. (2021). Analisis Faktor-Faktor Keterlambatan pada Proyek Pembangunan Gedung Asrama MAN 1 Tulehu Maluku Tengah. *Jurnal Simetrik*, 11(1), 432–439. <https://doi.org/10.31959/js.v11i1.615>
- Santo, M., & Prajana, A. (2018). Analisis Implementasi Ujian Nasional Berbasis. *Jurnal Pendidikan Teknologi Informasi*, 2(2), 84–91. <https://doi.org/10.22373/cj.v2i2.3997>
- Saputri, H. A. S., & Larasati, N. J. (2023). Analisis Instrumen Asesmen: Validitas, Reliabilitas, Tingkat Kesukaran dan Daya Beda Butir Soal. *Didaktik: Jurnal Ilmiah PGSD STKIP Subang*, 9(5), 2986–2995. <https://doi.org/10.36989/didaktik.v9i5.2268>
- Savika, H. I., & Zuhriyah, I. A. (2024). Peran Analisis Butir Soal terhadap Kualitas Soal, Kompetensi Guru, dan Prestasi Belajar Peserta Didik di Sekolah Dasar. *PANDU: Jurnal Pendidikan Anak dan Pendidikan Umum*, 2(2), 43–51. <https://doi.org/10.59966/pandu.v2i2.856>
- Septianingrum, N. P., & Sabardila, A. (2017). Piranti Kohesi Gramatikal dan Leksikal pada Wacana Rubrik “Selebritas” dalam Majalah Femina sebagai Bahan Ajar Menulis Teks Narasi. *Jurnal Penelitian Humaniora*, 18(2), 88–94. <https://doi.org/10.23917/humaniora.v18i2.5194>
- Setiawati, F. A., Izzaty, R. E., & Hidayat, V. (2018). Evaluasi Karakteristik Psikometrik Tes Bakat Diferensial dengan Teori Klasik. *Humanitas*, 15(1), 46. <https://doi.org/10.26555/humanitas.v15i1.7249>
- Setyawarno, D. (2017). Penggunaan Aplikasi Software Iteman (*Item and Test Analysis*) untuk Analisis Butir Soal Pilihan Ganda Berdasarkan Teori Tes Klasik. *Jurnal Ilmu Fisika dan Pembelajarannya*, 1(1), 69–79. <https://doi.org/10.19109/jifp.v1i1.866>
- Sidiq, U., Choiri, M., & Mujahidin, A. (2019). *Metode Penelitian Kualitatif di Bidang Pendidikan*. Ponorogo: CV. Nata Karya.
- Solichin, M. (2017). Analisis Daya Beda Soal, Taraf Kesukaran, Validitas Butir Tes, Interpretasi Hasil Tes dan Validitas Ramalan dalam Evaluasi Pendidikan. *Dirasat: Jurnal Manajemen dan Pendidikan Islam*, 2(2), 192–213. <https://doi.org/10.26594/dirasat.v2i2.879>
- Sujarweni, V. W. (2014). *Metodologi penelitian*. Yogyakarta: Pustaka Baru Press.
- Syamsuddin, S. (2023). Implementasi *Classic Test* dan *Item Respon Theory* pada Penilaian Tes Pembelajaran Matematika. *EDUSCOPE: Jurnal Pendidikan, Pembelajaran, dan Teknologi*, 8(2), 28–43. <https://doi.org/10.32764/eduscope.v8i2.3488>
- Wirzberger, M., Lado, A., Prentice, M., Oreshnikov, I., Passy, J-C., Stock, A., & Lieder, F. (2024). Optimal feedback improves behavioral focus during self-regulated computer-based work. *Nature: Scientific Reports*, 14, 3124, 1–16. <https://doi.org/10.1038/s41598-024-53388-3>
- Yani, E., Lestari, A. F., Amalia, H., & Puspita, A. (2018). Pengaruh Internet Banking terhadap Minat Nasabah dalam Bertransaksi dengan *Technology Acceptance Model*. *Jurnal Informatika*, 5(1), 34–42. <https://doi.org/10.31294/ji.v5i1.2717>
- Zayrin, A. A., Nupus, H., Maizia, K. K., Marsela, S., Hidayatullah, R., & Harmonedi, H. (2025). Analisis Instrumen Penelitian Pendidikan (Uji Validitas dan Reliabilitas Instrumen Penelitian). *Jurnal QOSIM Jurnal Pendidikan Sosial & Humaniora*, 3(2), 780–789. <https://doi.org/10.61104/jq.v3i2.1070>