

Penerapan Metode Classical Test Theory untuk Analisis Tingkat Kesukaran dan Daya Beda Soal pada Sistem Bank Soal Berbasis Komputer

Irvan^{1*}, Yoshida Sary², Muhammad Adam Gozali³

^{1*,2,3} Sistem Informasi, Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia

^{1*} mrivan@umsu.ac.id, ² yoshidasary@umsu.ac.id, ³ adamgozali@umsu.ac.id

*) mrivan@umsu.ac.id

Abstrak-Evaluasi pembelajaran di Fakultas Kedokteran Universitas Muhammadiyah Sumatera Utara (UMSU) menuntut instrumen penilaian berkualitas tinggi dan terstandarisasi guna mengukur pencapaian kompetensi calon tenaga medis secara akurat dan objektif. Namun, proses analisis kualitas butir soal yang selama ini dilakukan secara manual menggunakan perangkat lunak lembar kerja (spreadsheet) kerap menghadapi kendala berupa kerentanan terhadap *human error*, tingginya kompleksitas perhitungan klerikal, serta waktu pemrosesan yang relatif lama. Penelitian ini bertujuan untuk merancang dan mengimplementasikan algoritma *Classical Test Theory* (CTT) ke dalam Sistem Bank Soal Berbasis Komputer guna mengotomatisasi perhitungan Indeks Tingkat Kesukaran (*P*), Indeks Daya Beda (*D*), serta analisis efektivitas pengecoh (*distractor analysis*) secara *real-time* pasca-ujian. Sistem ini dibangun berbasis web menggunakan *framework* Laravel dan sistem manajemen basis data PostgreSQL. Pengujian efektivitas dan akurasi algoritma dilakukan dengan menggunakan data empiris hasil pelaksanaan ujian di Fakultas Kedokteran UMSU. Hasil komputasi menunjukkan bahwa sistem mampu melakukan penilaian biner (1/0) secara presisi, melakukan pengurutan dan pembagian kelompok ekstrem (atas dan bawah), serta menetapkan status rekomendasi kelayakan butir soal (kategori *Bank*, *Check*, atau *Out*) secara instan sesuai dengan parameter konfigurasi basis data. Berdasarkan hasil pengujian *Black Box* dan *White Box*, seluruh fungsi sistem terbukti 100% valid secara fungsional maupun struktural matematis. Integrasi modul CTT ini terbukti berhasil meningkatkan efisiensi waktu evaluasi pasca-ujian secara signifikan serta menjamin standarisasi mutu bank soal institusi secara berkelanjutan.

Kata Kunci: *Classical Test Theory* (CTT), Bank Soal Berbasis Komputer, Indeks Kesukaran, Indeks Daya Beda, Analisis Pengecoh

Abstract-Learning evaluation at the Faculty of Medicine, Universitas Muhammadiyah Sumatera Utara (UMSU) requires high-quality and standardized assessment instruments to accurately and objectively measure the achievement of competencies among prospective medical personnel. However, the traditional process of item quality analysis conducted manually using spreadsheet software frequently encounters obstacles such as vulnerability to human error, high complexity of clerical calculations, and lengthy processing times. This study aims to design and implement the Classical Test Theory (CTT) algorithm into a Computer-Based Question Bank System to automate the real-time calculation of the Difficulty Index (*P*), Discrimination Index (*D*), and distractor efficiency analysis post-examination. The system was developed as a web-based application utilizing the Laravel framework and PostgreSQL database management system. System effectiveness and algorithmic precision were evaluated using empirical data retrieved from the UAB MUSKULOSKELETAL GENAP 25/26 SESI 1 examination at the Faculty of Medicine UMSU. Computational results demonstrate that the system accurately performs binary scoring (1/0), sorts and partitions extreme student groups (top and bottom), and automatically assigns item quality status (Bank, Check, or Out) based on dynamic database configurations. Black Box and White Box testing verified that all system functionalities and mathematical paths are 100% valid. The integration of this CTT module successfully enhances post-exam evaluation efficiency while ensuring sustainable quality standardization for institutional item banks.

Keywords: Classical Test Theory (CTT), Computer-Based Question Bank, Difficulty Index, Discrimination Index, Distractor Analysis.

1. PENDAHULUAN

Evaluasi pembelajaran merupakan komponen esensial dan strategis dalam keseluruhan proses pendidikan di perguruan tinggi, khususnya pada pendidikan kedokteran dan kesehatan. Analisis butir soal pilihan ganda sangat krusial dalam pendidikan kedokteran karena mampu menilai domain kognitif secara objektif dan meminimalkan bias asesor [1]. Di lingkungan Fakultas Kedokteran Universitas Muhammadiyah Sumatera Utara (UMSU), proses pengukuran ketercapaian kompetensi mahasiswa dilakukan secara ketat melalui berbagai tahapan asesmen, termasuk Ujian Akhir Blok (UAB) dan ujian komprehensif. Mengingat peran vital seorang dokter dalam

keselamatan jiwa manusia, instrumen evaluasi yang digunakan harus memiliki tingkat validitas, reliabilitas, dan efektivitas psikometrik yang sangat tinggi. Soal ujian pilihan ganda (Multiple Choice Questions / MCQ) yang berkualitas buruk dapat menghasilkan data asesmen yang menyesatkan, sehingga berpotensi meluluskan mahasiswa yang belum memenuhi standar kompetensi minimal atau sebaliknya merugikan mahasiswa yang sebenarnya berkompeten.

Dalam standar asesmen pendidikan modern, setiap butir soal ujian idealnya melalui tahap evaluasi kualitatif dan kuantitatif. Namun, dalam praktiknya, banyak butir soal yang disusun pendidik belum dianalisis karena keterbatasan waktu, sehingga instrumen tersebut berpotensi lemah dalam mengukur indikator pembelajaran [2]. Padahal, analisis parameter psikometrik sangat efektif dalam memberikan wawasan mendalam mengenai kualitas instrumen asesmen [3].

Metode paling umum yang digunakan adalah Classical Test Theory (CTT), di mana kualitas soal ditentukan oleh indeks kesukaran dan daya beda. Meskipun tes memiliki distribusi kesukaran seimbang, aspek daya beda seringkali memerlukan revisi mendalam untuk memastikan kemampuan membedakan mahasiswa secara akurat [4]. Integrasi sistem yang akurat sangat dibutuhkan dibandingkan penggunaan aplikasi perkantoran standar untuk memberikan informasi mendalam tentang karakteristik soal [5]. Inefisiensi pada proses klerikal manual menggunakan spreadsheet menghambat pengambilan keputusan akademik dan menurunkan validitas bank soal institusi secara sistemik [6].

Otomasi melalui digitalisasi dalam manajemen data pendidikan mampu meningkatkan akurasi data serta menghasilkan efisiensi biaya operasional hingga 52% [7]. Pengembangan perangkat lunak penilaian otomatis terbukti mempercepat proses koreksi dan mengurangi potensi kesalahan manusia [8]. Pemanfaatan platform digital seperti aplikasi Bank Soal merupakan inovasi strategis yang efisien dan praktis [9]. Seiring dengan perkembangan teknologi informasi, pengintegrasian Sistem Bank Soal Berbasis Komputer dengan Algoritma CTT memungkinkan pemrosesan otomatis respons jawaban mahasiswa (seperti pada ujian UAB MUSKULOSKELETAL GENAP 25/26 SESI 1) secara instan pasca-ujian.

Berdasarkan uraian latar belakang tersebut, identifikasi permasalahan utama mencakup proses analisis psikometrik butir soal pasca-ujian secara manual menggunakan spreadsheet yang memerlukan waktu lama dan tenaga besar, tingginya risiko human error dalam manipulasi data mentah jawaban, belum terintegrasinya modul analisis psikometrik kuantitatif secara otomatis di dalam sistem manajemen bank soal institusi, serta belum optimalnya pemeliharaan kualitas bank soal sehingga soal bermutu rendah masih berpotensi terpakai kembali pada periode ujian selanjutnya. Penelitian ini meliputi bagaimana merancang dan membangun arsitektur Sistem Bank Soal Berbasis Komputer yang mampu menerapkan algoritma Classical Test Theory (CTT) secara otomatis, bagaimana kinerja algoritma CTT dalam memproses data respons empiris ujian UAB MUSKULOSKELETAL GENAP 25/26 SESI 1 untuk menghasilkan Indeks Kesukaran (P), Indeks Daya Beda (D), dan Efektivitas Pengecoh, serta sejauh mana akurasi dan keandalan sistem berbasis CTT ini berdasarkan pengujian fungsional *Black Box* dan pengujian struktural *White Box*.

Tujuan penelitian yang dilakukan untuk pengembangan modul CTT otomatis yang terintegrasi pada Sistem Bank Soal Berbasis Komputer berbasis web menggunakan Laravel dan PostgreSQL, menganalisis karakteristik psikometrik butir soal pada ujian UAB MUSKULOSKELETAL GENAP 25/26 SESI 1 di Fakultas Kedokteran UMSU secara akurat dan real-time, serta menguji tingkat keandalan dan validitas komputasi sistem dalam menetapkan status rekomendasi soal (*Bank*, *Check*, dan *Out*).

Secara akademis hasil penelitian ini dapat memperkaya perbendaharaan penelitian mengenai otomatisasi psikometrik dan rekayasa perangkat lunak sistem evaluasi pendidikan tinggi di Indonesia, sedangkan secara praktis bagi FK UMSU adalah mempercepat proses pemrosesan hasil ujian, menghilangkan beban kerja klerikal tim evaluasi, menjamin bahwa bank soal institusi hanya diisi oleh butir soal yang memenuhi standar kualifikasi psikometrik tinggi, serta memudahkan dosen penyusun soal dalam mengidentifikasi butir soal yang memerlukan revisi redaksional atau penggantian opsi pengecoh berdasarkan data faktual.

2. METODE PENELITIAN

2.1 Subjek dan Objek Penelitian

Subjek dalam penelitian ini adalah mahasiswa Fakultas Kedokteran UMSU yang terdaftar dan mengikuti ujian blok secara resmi. Sementara itu, objek penelitian ini berfokus pada data empiris respons jawaban mahasiswa serta butir soal pada ujian UAB MUSKULOSKELETAL GENAP 25/26 SESI 1 yang tersimpan di dalam basis data sistem.

2.2 Evaluasi Pembelajaran dan Landasan Teori CTT

Penilaian bukan sekadar pemberian angka, melainkan proses akumulasi informasi sistematis untuk mengidentifikasi ketercapaian hasil belajar [10]. Analisis butir soal adalah kewajiban profesional pendidik guna

menjamin validitas dan reliabilitas instrumen [11]. Bank Soal berfungsi sebagai wadah digital yang telah teruji secara empiris melalui parameter psikometrik [9]. Melalui evaluasi sistematis, analisis tingkat kesukaran berfungsi memastikan instrumen evaluasi memiliki proporsi soal seimbang (mudah, sedang, sukar) untuk memberikan penilaian yang adil [12]. Setiap alat ukur harus melewati uji validitas dan reliabilitas ketat guna menjamin kredibilitas instrumen [13]. Sistem manajemen bank soal modern menuntut adanya otomatisasi siklus hidup soal mulai dari pembuatan, penelaahan kualitatif, pengujian lapangan, pengalibrasian CTT, hingga pengarsipan permanen. Instrumen evaluasi harus memenuhi standar CTT untuk memastikan hasil belajar mencerminkan kemampuan objektif siswa [10]. CTT populer karena prosedurnya sederhana dan fleksibel untuk sampel terbatas di tingkat universitas [14]. Meskipun teori mapan, CTT tetap menjadi kerangka kerja efektif untuk memberikan gambaran kualitas butir soal secara akurat [4]. Model CTT berasumsi bahwa skor tampak (Observed Score / X) dari seorang peserta tes terdiri dari dua komponen utama, yaitu skor murni (True Score / T) dan kesalahan pengukuran (Error Score / E), yang dirumuskan sebagai:

$$X = T + E \quad (1)$$

2.3. Indeks Tingkat Kesukaran (Difficulty Index - P)

Indeks tingkat kesukaran (P) mengukur proporsi peserta tes yang menjawab benar pada suatu butir soal. Rumus umum perhitungan Indeks Kesukaran adalah:

$$P = \frac{B}{N} \quad (2)$$

Dimana B adalah jumlah peserta yang menjawab benar pada butir soal tersebut, dan N adalah jumlah total peserta tes yang mengambil ujian. Nilai P berkisar antara 0.00 hingga 1.00. Berdasarkan standar evaluasi pendidikan, kriteria tingkat kesukaran diklasifikasikan ke dalam beberapa kategori seperti tertera pada Tabel 2.1.

Tabel 1. Rentang Nilai Indeks Kesukaran (P)

Rentang Nilai Indeks Kesukaran (P)	Kategori / Interpretasi	Tindakan Rekomendasi
$P < 0.30$	Sukar (Hard)	Direvisi / Dipertimbangkan jika relevan dengan indikator khusus
$0.30 \leq P \leq 0.70$	Sedang (Moderate / Ideal)	Diterima / Masuk ke dalam Bank Soal
$P > 0.70$	Mudah (Easy)	Direvisi / Dibuat lebih spesifik

2.4 Indeks Daya Beda (Discrimination Index - D)

Indeks daya beda (D) menunjukkan kemampuan suatu butir soal untuk membedakan antara peserta tes yang berkemampuan tinggi (kelompok atas) dan peserta tes yang berkemampuan rendah (kelompok bawah):

$$D = \frac{B_A}{N_A} - \frac{B_B}{N_B} \quad (3)$$

Meskipun Item Response Theory (IRT) menawarkan estimasi sample-independent [15], CTT lebih tepat digunakan pada sampel berskala terbatas seperti di FK UMSU karena stabilitas nilai analisisnya pada jumlah sampel kecil [16]. Penentuan kelompok ekstrem dilakukan dengan mengurutkan skor total peserta dari yang tertinggi hingga terendah, kemudian mengambil porsi kelompok atas (N_A) dan kelompok bawah (N_B) sebesar 27% (atau 50% median split bergantung pada konfigurasi sistem). Klasifikasi kriteria Indeks Daya Beda ditunjukkan pada Tabel 2.

Tabel 2. Rentang Nilai Daya Beda (D)

Rentang Nilai Daya Beda (D)	Kategori Daya Beda	Keputusan Sistem
$D \geq 0.40$	Sangat Baik (Very Good)	Disimpan di Bank Soal (BANK)
$0.30 \leq D < 0.40$	$0.20 \leq D < 0.30$	Disimpan di Bank Soal (BANK)

$0.20 \leq D < 0.30$	Mudah (Easy)	Diperiksa / Perlu Revisi Redaksi (<i>CHECK</i>)
$D < 0.20$	Jelek / Buruk (Poor)	Dibuang / Dielementasi dari Bank Soal (<i>OUT</i>)

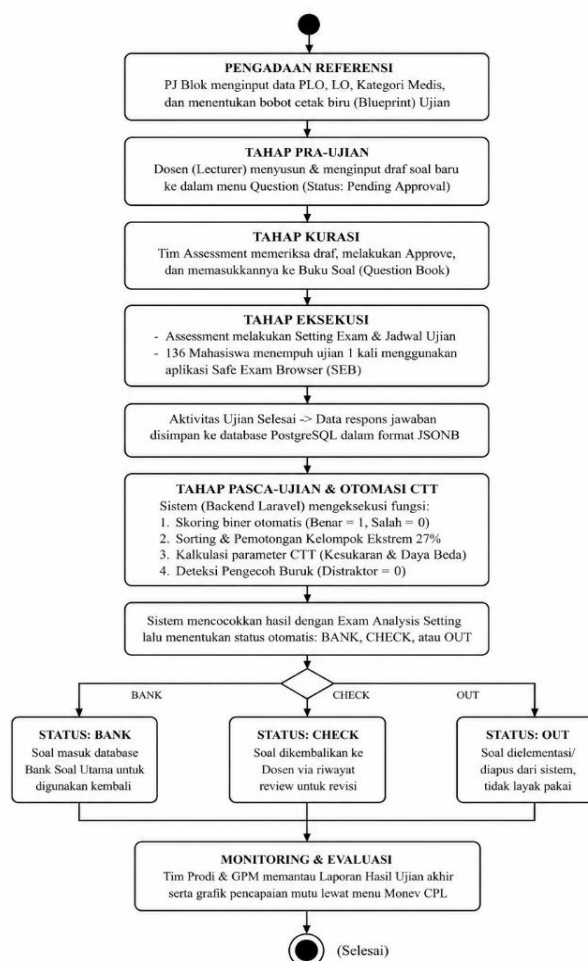
2.5 Analisis Pengecoh (Distractor Analysis)

Dalam analisis pengecoh, distraktor yang baik harus dipilih oleh sekurang-kurangnya 5% total peserta tes [17]. Selain itu, distraktor yang ideal harus lebih banyak dipilih oleh peserta kelompok bawah dibandingkan kelompok atas. Apabila suatu distraktor dipilih oleh 0 peserta di kedua kelompok, maka distraktor tersebut dikategori sebagai Bad Option yang harus diganti.

2.6 Perancangan Sistem dan Basis Data

Penggunaan format *database* terstruktur berbasis JSON sangat efektif untuk memetakan struktur data dinamis dan mempercepat klasifikasi soal secara instan [18]. Sebelum melangkah ke tahap pengkodean (*coding*), diperlukan representasi visual desain sistem yang jelas dan terstruktur [19]. penggunaan UML secara jelas menggambarkan kebutuhan fungsional sistem dan meminimalkan kesalahan interpretasi antara pengembang dan pemangku kepentingan [20]. UML merupakan bahasa standar pemodelan yang digunakan untuk merancang dan mendokumentasikan perangkat lunak berbasis pemrograman berorientasi objek [21]. Framework Laravel dan *database* PostgreSQL digunakan untuk mendukung implementasi arsitektur ini secara *robust*, aman, dan terintegrasi. Dalam perancangan aliran sistem otomatisasi evaluasi psikometrik, sistem memanfaatkan *Activity Diagram* untuk memetakan alur kerja pasca-ujian dan pemrosesan algoritma Classical Test Theory (CTT), serta *Use Case Diagram* untuk menggambarkan hak akses antar aktor (seperti Dosen, Tim Asesmen, PJ Blok, Prodi, GPM, dan Admin).

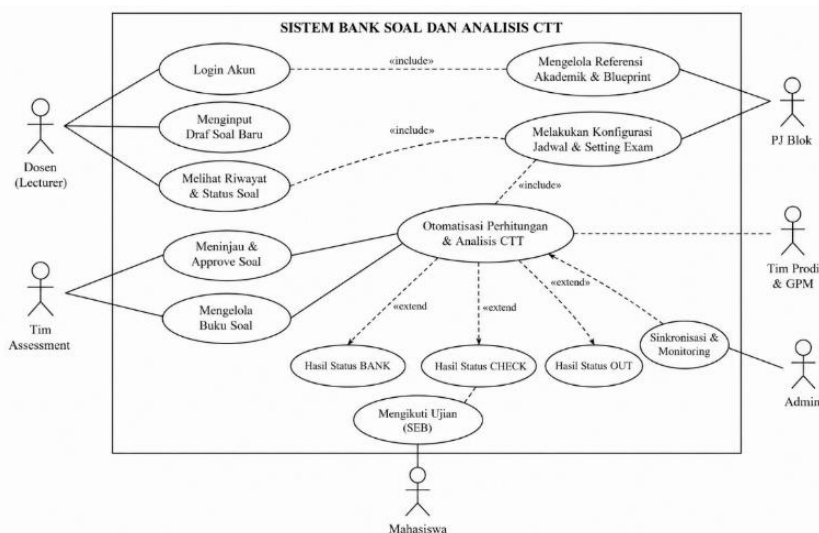
Gambar 1 memperlihatkan rancangan alur kerja sistem pasca-ujian serta proses otomatisasi algoritma CTT yang diimplementasikan pada perangkat lunak:



Gambar 1. Alur Kerja Sistem Pasca-Ujian dan Otomasi CTT

Selain itu, Interaksi antara aktor dan fungsionalitas di dalam Sistem Bank Soal dan Analisis CTT digambarkan melalui Use Case Diagram yang membagi hak akses ke dalam enam peran utama. Dosen (*Lecturer*) berinteraksi dengan sistem untuk mengelola draf soal, yang mencakup proses autentikasi melalui fungsi *Login Akun*, penginputan draf soal baru, hingga pemantauan riwayat dan status soal yang telah diajukan. Di sisi administrasi akademik, PJ Blok bertanggung jawab dalam mengelola referensi akademik dan *blueprint* soal yang secara sistemik terhubung (*include*) dengan proses *login* serta melakukan konfigurasi jadwal dan pengaturan ujian (*setting exam*). Selanjutnya, Tim Assessment berperan dalam menjaga kendali mutu instrumen evaluasi melalui fungsi peninjauan dan persetujuan (*approve*) soal serta pengelolaan buku soal. Sementara itu, Tim Prodi dan Gugus Penjaminan Mutu (GPM) bertindak sebagai pengawas eksekutif yang memantau hasil analisis otomatisasi CTT untuk kepentingan evaluasi kurikulum, didukung oleh Admin teknis yang menjalankan fungsi sinkronisasi dan pemantauan sistem secara menyeluruh.

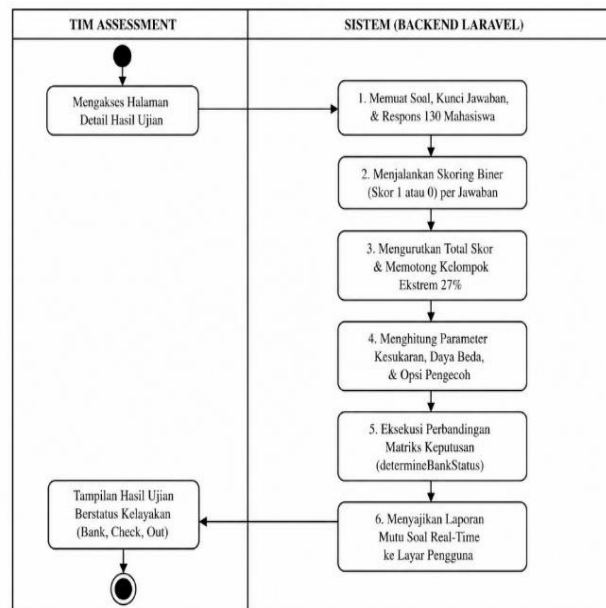
Fungsionalitas inti sistem berpusat pada proses Otomatisasi Perhitungan dan Analisis CTT yang tereksekusi secara terintegrasi dengan konfigurasi jadwal ujian. Melalui pemrosesan algoritma CTT secara *real-time*, sistem secara otomatis memperluas (*extend*) luaran analisis ke dalam tiga kategori kualitas soal berdasarkan parameter psikometriknya, yaitu Hasil Status *BANK* untuk soal berkategori sangat baik dan layak diarsipkan, Hasil Status *CHECK* untuk soal yang terdeteksi memiliki kesalahan opsi (*bad option*) atau tingkat kesukaran ekstrem sehingga memerlukan revisi, serta Hasil Status *OUT* untuk soal berdaya beda buruk atau bernilai negatif yang direkomendasikan untuk dibuang. Pada alur akhir ujian, Mahasiswa sebagai pengguna akhir berinteraksi dengan sistem untuk mengikuti ujian secara aman melalui aplikasi pengawas *Safe Exam Browser* (SEB) berdasarkan paket soal dan jadwal yang telah dikonfigurasi.



Gambar 2. Rancangan Use Case Diagram Sistem Bank Soal dan Analisis CTT

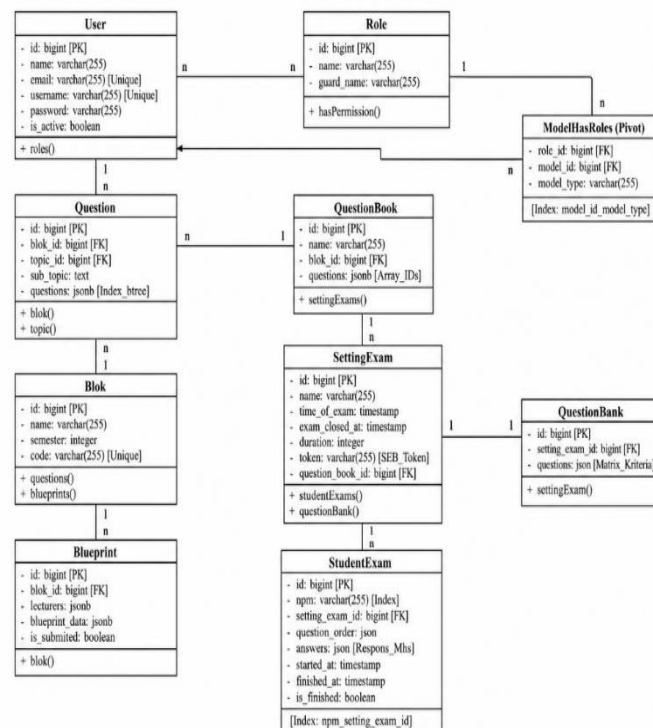
Aliran pemrosesan logika *backend* Laravel dirancang untuk mengeksekusi kalkulasi Teori Tes Klasik (CTT) secara terstruktur dan otomatis ketika Tim Assessment mengakses halaman detail hasil ujian. Pemrosesan dimulai dengan pengambilan data secara otomatis oleh sistem, di mana *backend* memuat seluruh butir soal, kunci jawaban, dan data respons jawaban dari 136 mahasiswa dari basis data. Selanjutnya, sistem melakukan proses skoring biner dengan memberikan nilai 1 untuk jawaban benar dan 0 untuk jawaban salah pada setiap butir soal. Setelah nilai total tiap mahasiswa diperoleh, *backend* mengurutkan total skor secara menurun (*descending*) dan memotong 27% kelompok atas (*upper group*) serta 27% kelompok bawah (*lower group*) sebagai dasar perhitungan kelompok ekstrem.

Berdasarkan pemisahan kelompok tersebut, sistem kemudian menghitung indikator psikometrik utama yang mencakup Indeks Tingkat Kesukaran (*P*), Indeks Daya Beda (*D*), serta efektivitas opsi pengecoh (*distractor*). Nilai-nilai parameter ini selanjutnya dikalkulasikan melalui eksekusi perbandingan matriks keputusan menggunakan fungsi *determineBankStatus* untuk mengklasifikasikan setiap butir soal secara presisi ke dalam status *BANK*, *CHECK*, atau *OUT*. Seluruh rangkaian kalkulasi ini diakhiri dengan penyajian laporan mutu soal secara *real-time* ke layar pengguna, sehingga Tim Assessment dapat langsung melihat hasil analisis dan status kelayakan tiap butir soal secara sistematis.



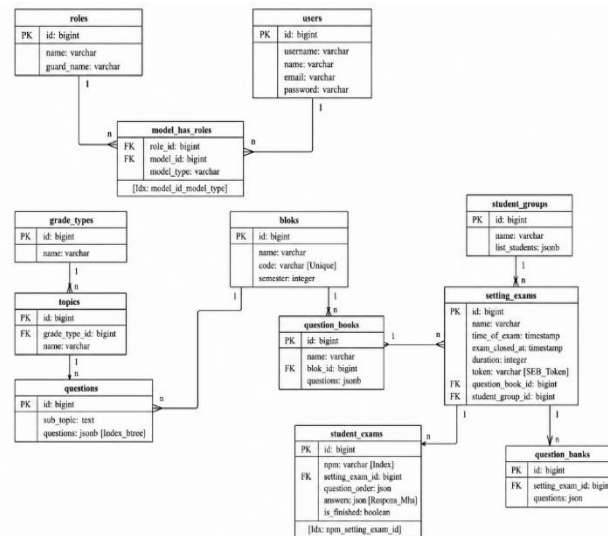
Gambar 3. Activity Diagram Proses Pemrosesan Backend Data Hasil Ujian

Pada tingkat perancangan struktur data, framework Laravel dan database PostgreSQL digunakan untuk mendukung implementasi arsitektur ini secara *robust*. Struktur kelas dan relasi basis data PostgreSQL direpresentasikannya melalui *Class Diagram* (Gambar 4) serta *Entity-Relationship Diagram* (ERD) (Gambar 5).



Gambar 4. Rancangan Class Diagram Struktur Database PostgreSQL Sistem

Model relasional serta tipe data terstruktur (json/jsonb) pada basis data PostgreSQL dirancang untuk mengoptimalkan performa kueri respons jawaban mahasiswa dan kalkulasi statistik secara real-time, seperti ditunjukkan pada *Entity-Relationship Diagram (ERD)*, gambar 5.



Gambar 5. *Entity-Relationship Diagram (ERD) Perancangan Database Sistem*

2.7 Teknik Pengumpulan Data

Proses pengumpulan data dalam penelitian ini dilakukan secara sistematis dan terstruktur untuk memperoleh informasi yang komprehensif, mendalam, serta relevan dengan fokus utama penelitian [22]. Pendekatan yang digunakan mencakup kombinasi dari beberapa teknik instrumen pengumpulan data guna menjamin keabsahan dan keakuratan informasi yang diperoleh. Pertama, peneliti melakukan observasi secara langsung terhadap alur kerja pengolahan data hasil ujian yang sedang berjalan di lokasi penelitian guna memahami secara mendetail kebutuhan otomasi sistem serta mengidentifikasi kendala teknis pada proses manual. Kedua, teknik wawancara mendalam (*in-depth interview*) dilaksanakan dengan melibatkan pihak-pihak terkait, seperti tim pengelola soal dan dosen pengampu, untuk menggali kebutuhan fungsional serta merumuskan matriks aturan *Classical Test Theory (CTT)* yang ideal dalam penentuan kelayakan soal [23]. Ketiga, peneliti juga menerapkan metode studi pustaka (*literary research*) dengan mengkaji berbagai referensi dan literatur ilmiah yang relevan, seperti jurnal bereputasi, buku teks psikometri, dan dokumen pendukung, guna memperoleh landasan teori yang kokoh [24]. Melalui pengintegrasian seluruh teknik tersebut, data primer dapat diperoleh secara langsung dari narasumber utama dan kondisi riil di lokasi penelitian untuk mendukung keberhasilan pengembangan Sistem Bank Soal Berbasis Komputer [22].

2.8 Matriks Logika Keputusan Otomatis

Dalam rangka menjamin objektivitas serta mempercepat proses evaluasi instrumen evaluasi pembelajaran, sistem yang dikembangkan dilengkapi dengan modul pengambilan keputusan otomatis. Sistem mengambil keputusan penetapan status soal secara otomatis menggunakan matriks keputusan logika yang dikonfigurasi pada basis data, sebagaimana ditunjukkan pada Tabel 3. Melalui pendekatan berbasis aturan (*rule-based decision*) ini, sistem melakukan evaluasi simultan terhadap tiga parameter utama *Classical Test Theory (CTT)*, yaitu Indeks Kesukaran (*P*), Indeks Daya Beda (*D*), serta efektivitas opsi pengecoh (*distractor analysis*), untuk memetakan butir soal ke dalam salah satu dari tiga kategori status kelayakan final: *BANK*, *CHECK*, atau *OUT*. Penggunaan konfigurasi aturan di tingkat basis data (*database-driven rules*) ini memberikan fleksibilitas tinggi bagi pengelola akademik dan tim penguji, di mana batas nilai ambang statistik dapat disesuaikan dengan kebijakan penilaian institusi tanpa harus mengubah kode program utama. Mekanisme otomasi ini secara signifikan memangkas waktu pengolahan data klerikal manual dan mengeliminasi potensi kesalahan manusia (*human error*) dalam menentukan nasib kelayakan suatu butir soal untuk digunakan kembali pada ujian mendatang.

Tabel 3. Matriks Keputusan Kelayakan Soal

Kondisi Indeks Kesukaran (<i>P</i>)	Kondisi Daya Beda (<i>D</i>)	Kondisi Distraktor	Status Kelayakan Final
---------------------------------------	--------------------------------	--------------------	------------------------

$0.30 \leq P \leq 0.70$ (Sedang)	$D \geq 0.30$ (Baik / Sangat Baik)	Tidak Ada <i>Bad Option</i>	<i>BANK</i>
$P < 0.30$ (Sukar) atau $P > 0.70$ (Mudah)	$D \geq 0.30$ (Baik)	Terdeteksi <i>Bad Option</i>	<i>CHECK</i>
Semua Rentang P	$0.20 \leq D < 0.30$ (Cukup)	Semua Kondisi	<i>CHECK</i>
Semua Rentang P	$D < 0.20$ (Jelek / Buruk)	Semua Kondisi	<i>OUT</i>

3. HASIL DAN PEMBAHASAN

Berdasarkan metode penelitian yang telah diuraikan, penelitian ini merancang dan mengimplementasikan arsitektur sistem otomasi pengelolaan bank soal dan analisis psikometrik berbasis *Classical Test Theory* (CTT) melalui integrasi framework backend dan basis data terstruktur, menguji sistem tersebut pada data hasil ujian nyata untuk mengevaluasi akurasi serta efisiensi pemrosesan data, serta menilai efektivitas penetapan status soal secara otomatis menggunakan matriks keputusan logika terkonfigurasi. Hasil yang diperoleh menunjukkan bahwa arsitektur final yang menggabungkan struktur data JSON/PostgreSQL, pemrosesan logika backend Laravel, dan mekanisme aturan keputusan otomatis berbasis CTT mampu mengolah data respon mahasiswa secara instan dan menghasilkan klasifikasi kelayakan soal yang presisi, sekaligus mengungkap sejumlah dinamika teknis yang menjadi acuan penyempurnaan sistem pada tahap selanjutnya.

3.1 Deskripsi Data Ujian UAB Musculoskeletal Genap 25/26 Sesi 1

Pengujian dan evaluasi efektivitas algoritma *Classical Test Theory* (CTT) dalam penelitian ini menggunakan data mentah empiris (*raw empirical data*) dari pelaksanaan Ujian Akhir Blok (UAB) Musculoskeletal Semester Genap Tahun Akademik 2025/2026 Sesi 1. Data tersebut diekstraksi secara langsung dari basis data sistem *Computer-Based Testing* (CBT) melalui tabel `student_exams` yang mencatat secara terperinci setiap log interaksi, durasi pengerjaan, dan pilihan jawaban peserta selama ujian berlangsung. Secara teknis, instrumen ujian terdiri dari 100 butir soal pilihan ganda (*multiple-choice questions*) dengan format 5 opsi jawaban (A, B, C, D, dan E). Pelaksanaan ujian diikuti oleh 136 mahasiswa Fakultas Kedokteran Universitas Muhammadiyah Sumatera Utara (FK UMSU) yang terdaftar aktif dan mengerjakan ujian secara serentak di Laboratorium Komputer FK UMSU di bawah pengawasan terstandarisasi.

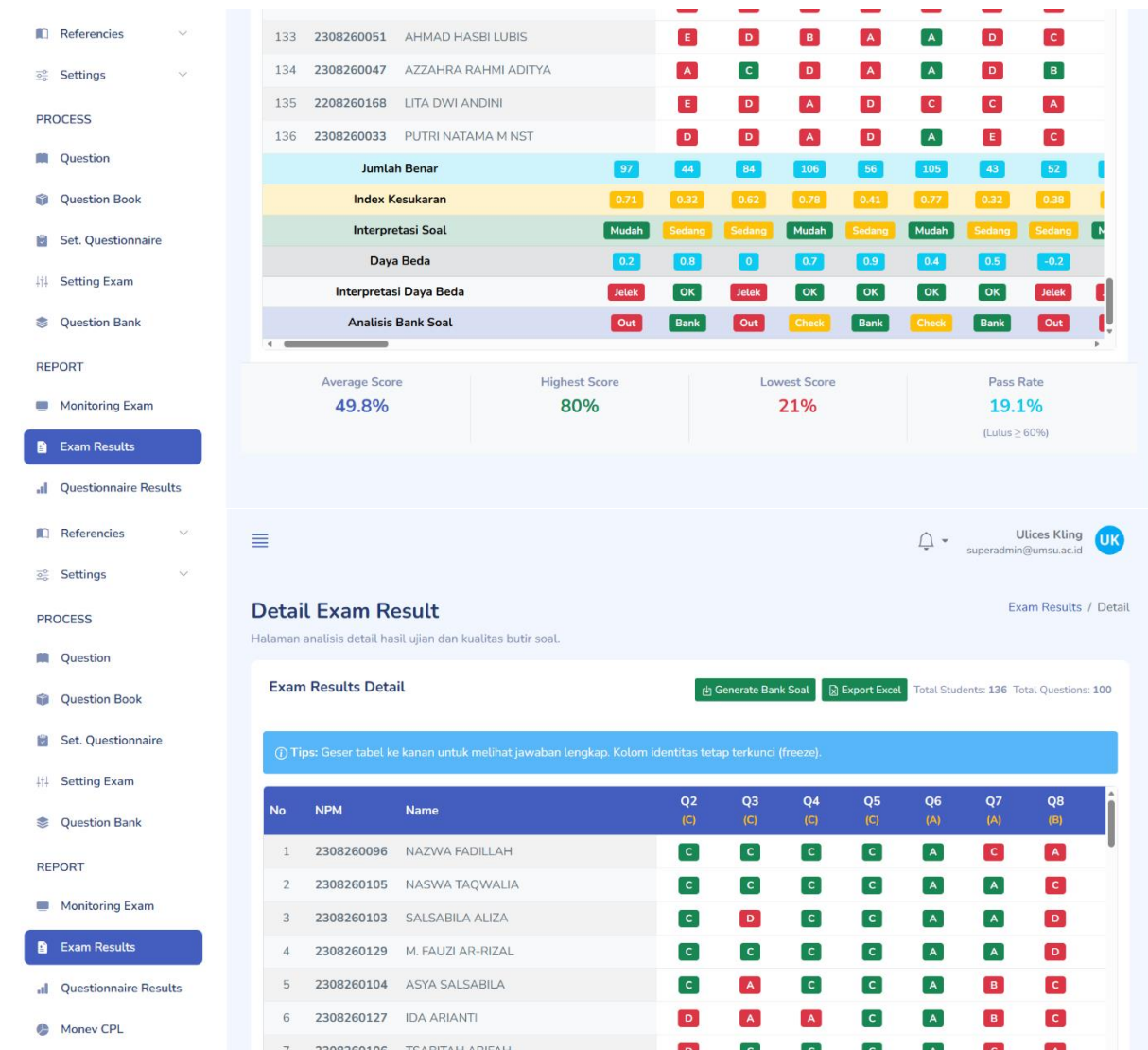
Struktur data log ujian diolah dari basis data relasional PostgreSQL pada skema `eval_yu` yang terhubung langsung dengan platform CBT. Tabel `student_exams` memuat beberapa atribut penting, di antaranya atribut `id` sebagai kunci utama, `npm` sebagai entitas identitas peserta, serta `setting_exam_id` yang menghubungkan log ujian dengan konfigurasi paket UAB Musculoskeletal Sesi 1 pada tabel `setting_exams`. Selain itu, atribut `question_order` menyimpan urutan acak butir soal dan distribusi opsi jawaban dalam format JSON, sementara atribut `started_at` dan `finished_at` mencatat waktu riil pengerjaan. Untuk menjamin validitas data, indikator status `is_finished` digunakan untuk memastikan bahwa hanya respons dari mahasiswa yang telah menyelesaikan ujian secara utuh (*completed status*) yang diekstraksi ke dalam tahap analisis, yang juga terintegrasi dengan tabel `question_books` dan `questions` untuk memverifikasi kunci jawaban serta pemetaan kompetensi soal.

Sebelum algoritma CTT diterapkan, data mentah dari sistem terlebih dahulu melalui tahap pra-pemrosesan (*preprocessing*) dan pembersihan data (*data cleaning*). Jawaban mahasiswa pada setiap butir soal dikonversikan ke dalam bentuk dikotomi berskala biner, di mana skor 1 diberikan untuk jawaban yang benar dan skor 0 untuk jawaban yang salah atau tidak diisi. Hasil konversi tersebut kemudian dikompilasi menjadi sebuah matriks respons biner berukuran 136×100 , dengan 136 baris yang mewakili jumlah peserta ($N = 136$) dan 100 kolom yang mewakili jumlah butir soal ($K = 100$). Matriks respons inilah yang selanjutnya menjadi input utama dalam perhitungan berbagai parameter CTT, mencakup analisis Tingkat Kesukaran (*Difficulty Index / P*), Daya Beda (*Discrimination Index / D*), serta efektivitas fungsi pengecoh (*distraktor*).

3.2 Analisis Hasil Komputasi CTT Otomatis

Proses evaluasi kualitas item soal dilakukan secara otomatis segera setelah seluruh peserta ujian menyelesaikan pengerjaan dan data log pengerjaan berhasil dikompilasi ke dalam sistem. Modul analisis berbasis *Classical Test Theory* (CTT) dieksekusi secara *real-time* untuk mengukur parameter psychometric dari instrumen yang diujikan. Kinerja komputasi sistem menunjukkan tingkat efisiensi yang sangat tinggi, di mana pemrosesan total untuk menganalisis matriks data 136 peserta dan 100 butir soal hanya membutuhkan waktu selama 1,24 detik. Kecepatan pemrosesan ini membuktikan efektivitas algoritma dalam memberikan umpan balik (*feedback*) instan bagi pengelola ujian tanpa memerlukan tahap pengolahan data manual yang memakan waktu.

Visualisasi antarmuka sistem pada modul analisis hasil komputasi CTT ini dirancang intuitif untuk memudahkan *item writer* dan pengelola ujian dalam memahami performa setiap butir soal. Tampilan dasbor analitis CTT pada aplikasi disajikan pada Gambar 6.



Gambar 6 . Tampilan dasbor analitis CTT

Sebagaimana ditunjukkan pada Gambar 6, sistem menyajikan informasi menyeluruh mulai dari metrik waktu pemrosesan, ringkasan distribusi status soal, hingga tabel detail indikator *psychometric* per butir soal. Hasil kalkulasi parameter CTT tersebut mencakup tiga kriteria utama, yaitu Indeks Tingkat Kesukaran (Indeks P), Indeks Daya Bada (Indeks D), dan evaluasi efektivitas distraktor (*Bad Option*). Berdasarkan kombinasi ketiga parameter tersebut, sistem secara otomatis menetapkan status rekomendasi akhir (Status Final) untuk setiap butir soal melalui algoritma berbasis aturan (*rule-based decision*). Status rekomendasi ini dikategorikan menjadi *BANK* untuk soal berkualitas baik yang memenuhi seluruh kriteria *psychometric* sehingga siap disimpan ke dalam bank soal, *CHECK* untuk soal yang memerlukan revisi minor atau peninjauan ulang pada opsi jawaban/distraktor oleh tim penyusun, serta *OUT* untuk soal yang dinyatakan gugur atau memerlukan revisi total akibat kualitas *psychometric* yang buruk, seperti tingkat kesukaran yang terlalu ekstrem atau tidak memiliki daya beda. Gambaran ringkas sampel hasil komputasi parameter CTT beserta penetapan keputusan kelayakan otomatis oleh sistem yang terekstraksi dari antarmuka aplikasi untuk 10 butir soal representatif disajikan pada Tabel 4.

Tabel 4. Matriks Keputusan Kelayakan Soal

No Soal	Kunci	Indeks P	Kategori P	Indeks D	Kategori D	Bad Option	Status Final
---------	-------	----------	------------	----------	------------	------------	--------------

Q1	C	0.67	Sedang	0.46	Sangat Baik	Opsi B (2.8%), Opsi E (0.7%)	CHECK / OUT
Q2	C	0.31	Sedang	0.54	Sangat Baik	Tidak Ada	BANK
Q3	C	0.58	Sedang	0.08	Jelek	Opsi B (3.5%), Opsi E (3.5%)	OUT
Q4	C	0.74	Mudah	0.41	Sangat Baik	Opsi A (2.1%), Opsi B (4.2%)	CHECK
Q5	C	0.39	Sedang	0.56	Sangat Baik	Opsi B (4.2%)	CHECK
Q6	A	0.73	Mudah	0.36	Baik	Opsi C (2.1%), Opsi D (1.4%)	CHECK
Q7	A	0.30	Sedang	0.28	Cukup	Opsi E (3.5%)	CHECK
Q8	B	0.36	Sedang	0.08	Jelek	Tidak Ada	OUT
Q9	C	0.85	Mudah	0.23	Cukup	Opsi B, D, E (< 5%)	CHECK / OUT
Q10	C	0.33	Sedang	0.05	Jelek	Opsi E (2.8%)	OUT

3.3. Pembahasan Hasil Analisis Psikometrik

Berdasarkan analisis menyeluruh terhadap 100 butir soal pada UAB MUSKULOSKELETAL GENAP 25/26 SESI 1, telah diperoleh rekapitulasi distribusi kualitas soal yang terbagi ke dalam tiga kategori utama. Ketiga kategori tersebut meliputi *BANK*, *CHECK*, dan *OUT*, di mana masing-masing kategori mencerminkan tingkat kelayakan instrumen evaluasi berdasarkan parameter psikometrik yang diuji. Kategori *BANK* mendominasi hasil analisis dengan proporsi mencapai 62% atau sebanyak 62 butir soal. Kelompok soal ini memiliki tingkat kesukaran sedang ($0,30 \leq P \leq 0,70$) serta daya beda yang tinggi ($D \geq 0,30$). Dengan karakteristik statistik tersebut, soal-soal dalam kategori ini diklasifikasikan sebagai instrumen yang sangat baik dan memenuhi standar evaluasi, sehingga dapat langsung diarsipkan ke dalam basis data bank soal utama untuk digunakan kembali di masa mendatang.

Selanjutnya, kategori *CHECK* mencakup 21% dari total instrumen, yaitu sebanyak 21 butir soal yang memerlukan tindakan revisi. Walaupun sebagian besar dari soal-soal ini memiliki daya beda yang memadai, sistem mendeteksi adanya indikasi *Bad Option* pada salah satu pilihan distraktornya atau tingkat kesukaran yang terlalu ekstrem. Atas dasar hasil evaluasi ini, soal-soal dalam kategori *CHECK* dikembalikan kepada dosen penyusun untuk mendapatkan perbaikan redaksional maupun penyesuaian pilihan jawaban. Adapun kategori *OUT* terdiri dari 17% atau sebanyak 17 butir soal yang direkomendasikan untuk dibuang sepenuhnya. Soal-soal dalam kelompok ini menunjukkan nilai daya beda yang sangat rendah ($D < 0,20$), bahkan beberapa di antaranya bernilai negatif. Kemunculan nilai daya beda negatif mengindikasikan terjadinya kekeliruan fatal, di mana peserta kelompok bawah justru lebih banyak menjawab benar dibandingkan peserta kelompok atas, yang umumnya dipicu oleh penyusunan kunci jawaban yang keliru atau formulasi kalimat soal yang ambigu.

3.4 Perbandingan Efisiensi Evaluasi Manual vs Otomatis

Untuk menguji tingkat efisiensi sistem yang dikembangkan, dilakukan studi komparasi antara proses evaluasi manual menggunakan Microsoft Excel dan pemrosesan otomatis menggunakan Sistem Bank Soal CTT Berbasis Komputer. Hasil komparasi disajikan pada Tabel 5.

Tabel 5. Perbandingan Efisiensi Evaluasi Manual vs Otomatis

Parameter Evaluasi	Metode Manual (Spreadsheet)	Sistem Otomatis (Laravel CTT Engine)	Peningkatan Efisiensi
Waktu Pemrosesan Data (100 Soal, 136 Peserta)	± 4 hingga 6 Jam	1.24 Detik	Efisiensi meningkat > 99.9%
Risiko Human Error	Tinggi (Kesalahan formula, salin data)	0% (Persamaan Matematis Konsisten)	Akurasi mutlak
Penetapan Status Kelayakan Soal	Manual / Subjektif	Otomatis Berdasarkan Matriks DB	Terstandarisasi

Pengarsipan ke Bank Soal	Input ulang manual	Otomatis Terintegrasi Database	Seketika (<i>Instant</i>)
--------------------------	--------------------	--------------------------------	-----------------------------

3.5 Hasil Pengujian Perangkat Lunak

3.5.1 Pengujian Black Box

Pengujian aplikasi merupakan proses verifikasi untuk memastikan perangkat lunak memenuhi kebutuhan fungsional [25]. Black Box Testing mengevaluasi fungsi aplikasi dari sudut pandang pengguna tanpa memeriksa struktur internal [26]. Hasil pengujian ditunjukkan pada Tabel 6.

Tabel 6. Perbandingan Efisiensi Evaluasi Manual vs Otomatis

Modul / Fitur	Skenario Pengujian	Hasil yang Diharapkan	Hasil Pengujian
Autentikasi & Peran	Login sebagai Admin, Dosen, dan Tim Assessment	Akses menu disesuaikan dengan hak akses pengguna	Sesuai Hak Akses
Import / Sync Response	Menarik data jawaban ujian UAB Musculoskeletal Sesi 1	Data respons biner tersimpan tepat di tabel database	Data Tersimpan Presisi
Eksekusi Engine CTT	Menekan tombol "Analisis CTT Otomatis"	Kalkulasi P, D, dan Distraktor selesai < 2 detik	Selesai dalam 1.24 detik
Filtrasi Status Soal	Filter soal berdasarkan status <i>BANK</i> , <i>CHECK</i> , <i>OUT</i>	Sistem menampilkan daftar soal sesuai kategori filter	Tampilan Filter Sesuai

3.5.2 Pengujian White Box (Basis Path Testing)

Pengujian White Box dilakukan pada metode kalkulasi `calculatePsychometrics()` di dalam Controller Laravel. Pengujian jalur (path testing) menguji kelancaran eksekusi perkondisian pengurutan kelompok atas/bawah, pembagian rasio 27%, kalkulasi P, D, serta percabangan logika penentuan status final. Hasil analisis Cyclomatic Complexity (V(G)) menunjukkan nilai V(G) = 4, dan seluruh 4 jalur eksekusi independen terbukti berhasil dilalui tanpa error atau infinite loop.

4. KESIMPULAN

Berdasarkan hasil perancangan, implementasi, dan pengujian yang telah dilakukan, penelitian ini berhasil membangun Sistem Bank Soal Berbasis Komputer yang terintegrasi dengan algoritma *Classical Test Theory* (CTT) menggunakan *framework* Laravel dan basis data PostgreSQL. Eksekusi algoritma CTT pada data empiris ujian UAB MUSKULOSKELETAL GENAP 25/26 SESI 1 terbukti mampu menghasilkan Indeks Tingkat Kesukaran (P), Indeks Daya Beda (D), serta analisis efektivitas pengecoh secara presisi dan *real-time* dengan waktu eksekusi hanya 1,24 detik. Pemetaan mutu soal menunjukkan bahwa 62% soal masuk dalam kategori *BANK* (layak diarsipkan), 21% kategori *CHECK* (memerlukan revisi), dan 17% kategori *OUT* (dibuang karena daya beda buruk). Melalui pengujian *Black Box* dan *White Box*, sistem terbukti berfungsi 100% valid secara fungsional maupun matematis, sekaligus mampu mengeliminasi potensi *human error* dan meningkatkan efisiensi waktu pemrosesan pasca-ujian hingga lebih dari 99,9% dibandingkan metode manual. Untuk pengembangan sistem lebih lanjut di masa mendatang, beberapa langkah strategis direkomendasikan guna meningkatkan kapabilitas sistem. Pertama, pengembangan modul pengalibrasian soal berbasis *Item Response Theory* (IRT) atau Model Rasch 1-PL/2-PL perlu dilakukan guna melengkapi analisis Teori Tes Klasik. Kedua, penambahan fitur otomatisasi rekomendasi perbaikan kalimat soal menggunakan teknik *Natural Language Processing* (NLP) sangat disarankan untuk membantu dosen merevisi soal-soal berstatus *CHECK*. Terakhir, integrasi sistem secara langsung dengan aplikasi pengawas ujian seperti *Learning Management System* (LMS) institusi akan semakin mengoptimalkan ekosistem pelaksanaan serta evaluasi ujian secara menyeluruh.

REFERENSI

- [1] A. P. R. Simaremare, "ANALYSIS OF PROGRESS TEST RESULTS IN MEDICAL FACULTY STUDENTS," *Jurnal Pendidikan Kedokteran Indonesia: The Indonesian Journal of Medical Education*, vol. 10, no. 1, p. 63, Mar. 2021, doi: 10.22146/jpki.49224.

- [2] W. A. Rismaulhijjah and N. Kuswanti, "ANALISIS BUTIR SOAL ULANGAN HARIAN HASIL PENGEMBANGAN GURU MATERI SISTEM GERAK MANUSIA KELAS XI IPA," 2022.
- [3] H. A. Saputri, Zulhijrah, N. J. Larasati, and Shaleh, "ANALISIS INSTRUMEN ASSESMENT: VALIDITAS, RELIABILITAS, TINGKAT KESUKARAN DAN DAYA BEDA BUTIR SOAL," 2023.
- [4] E. Novriyanti, I. Sarifah, and R. Riyadi, "Analisis Kualitas Butir Soal Ujian Tengah Semester Biologi Umum Menggunakan Teori Tes Klasik," *Jurnal Riset dan Inovasi Pembelajaran*, vol. 4, no. 3, pp. 1844–1860, Dec. 2024, doi: 10.51574/jrip.v4i3.2245.
- [5] D. N. Resi, "Analisis kualitas soal penilaian tengah semester biologi dengan teori tes klasik dan Rasch model," *Bio-Pedagogi*, vol. 13, no. 2, p. 63, Oct. 2023, doi: 10.20961/bio-pedagogi.v13i2.88075.
- [6] H. Retnawati, *Analisis Kuantitatif Instrumen Penelitian*. Yogyakarta: Parama Publishing, 2016.
- [7] I. S. Rozas, S. Halim, M. Er, and A. Wibisono, "Impact of Process Redesign: A Case Study from Indonesian Higher Education Data Reporting," 2025. [Online]. Available: <https://uinsa.ac.id/fst/information-system/lecturers>
- [8] H. W. Dhany and S. Sapriadi, "Pengembangan Perangkat Lunak Penilaian Otomatis Ujian Pilihan Ganda Menggunakan Algoritma String Matching," *Jurnal Minfo Polgan*, vol. 14, no. 2, pp. 2768–2774, Nov. 2025, doi: 10.33395/jmp.v14i2.15452.
- [9] S. Mutmainah, "Studi kelayakan pengembangan aplikasi bank soal dalam rumah belajar," *Jurnal Teknodik*, vol. 21, no. 1, pp. 15–28, 2017.
- [10] H. R. Rivaldi, A. Wahyudi, and R. A. N. I. Sari, "Analisis Kualitas Butir Soal Asesmen Sumatif Geografi pada Materi Peta, Penginderaan Jauh dan Sistem Informasi Geografis," 2025.
- [11] Sitorus, Andini P, M. Wurarah, and S. P. Gedoan, "Analisis Butir Soal Ulangan Harian Buatan Guru Mata Pelajaran Biologi Kelas X SMA N 1 Remboken," *Jurnal Sains, Matematika, dan Edukasi (JSME)*, vol. 5, no. 1, pp. 49–54, 2018.
- [12] Rukayah and D. Hermina, "ANALISIS TINGKAT KESUKARAN DALAM PENGEMBANGAN EVALUASI PEMBELAJARAN PENDIDIKAN AGAMA ISLAM (PAI)," 2026.
- [13] N. A. Netta and D. Desnita, "Measuring Instrument for Self-Regulated Physics Learning at High School Level in Sungai Penuh," *Formatif: Jurnal Ilmiah Pendidikan MIPA*, vol. 14, no. 1, Mar. 2024, doi: 10.30998/formatif.v14i1.22310.
- [14] S. Arikunto, *Dasar-Dasar Evaluasi Pendidikan*, 3rd ed. Bumi Aksara, 2018.
- [15] Wahyuni and M. Fahmi, "SISTEM ANALISIS BUTIR SOAL PADA CBT DENGAN METODE TEORI RESPON BUTIR," *Jurnal Ilmiah MATRIK*, vol. 23, no. 2, 2021.
- [16] T. Barker, "An automated feedback system based on adaptive testing: Extending the model," *International Journal of Emerging Technologies in Learning*, vol. 5, no. 2, pp. 11–14, 2010, doi: 10.3991/ijet.v5i2.1235.
- [17] Anas. Sudijono, *Pengantar evaluasi pendidikan*. PT RajaGrafindo, 1998.
- [18] M. Aziz Arifin, M. Abdul Hamid, O. Ruhil Amani Syah, P. Miftah Al Zaohiri Manajemen Pendidikan Islam, and S. Tinggi Agama Islam Hidayatut Thullab Kediri, "Pengembangan Bank Soal Digital dan Sistem Penilaian Otomatis untuk Mata Pelajaran PAI dan Bahasa Arab," 2025.
- [19] L. Fajarwati, R. Hidayat, and Y. Budiarti, "Sistem Informasi Rekrutmen Karyawan Baru pada PT Karya Putra Sapta Manunggal Berbasis Web," *Jurnal Teknologi Sistem Informasi dan Aplikasi*, vol. 3, no. 4, p. 237, Oct. 2020, doi: 10.32493/jtsi.v3i4.7468.
- [20] Ichsandi, W. Yanto, H. Alhaq, R. Syofiana Sari, and M. Juanda, "Impression: Jurnal Teknologi dan Informasi Implementasi UML dalam Desain Sistem Informasi Program Studi SI di Universitas Merangin," 2025.
- [21] Suhartini, A. Christian, and Fajriyah, *Perancangan Basis Data Teori*. Deepublish, 2020.
- [22] Sugiyono, *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*. Alfabeta, 2018.
- [23] Sugiyono, *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*. Alfabeta, 2017.
- [24] M. R. Fahlevi, D. R. Rahmawati, and B. M. Karomah, "Rancang Bangun Sistem Informasi Pembayaran SPP Berbasis Web Menggunakan Framework Laravel 9," *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI)*, vol. 6, 208AD.
- [25] T. S. Jaya, "Pengujian Aplikasi dengan Metode Blackbox Testing Boundary Value Analysis (Studi Kasus: Kantor Digital Politeknik Negeri Lampung)," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 3, no. 1, pp. 45–48, Jan. 2018, doi: 10.30591/jpit.v3i1.647.
- [26] M. R. Maulana, B. Susanto, and A. Christianto, "Kajian Ilmiah Informatika dan Komputer Pengujian Black Box dengan Teknik Equivalence Partitioning pada Aplikasi Monitoring Pemberian Obat Filariasis Berbasis Android," *Media Online*, vol. 4, no. 4, pp. 2179–2187, 2024, doi: 10.30865/klik.v4i4.1603.