UNIVERSALIA
JOURNAL OF SCIENTIFIC RESEARCH
Vol. 1, No. 1, April 2026, | E-ISSN: XXXX-XXXX

Analisis Kualitas Butir Soal Ujian Tengah Semester Menggunakan Pendekatan Teori Respons Butir (Item Response Theory)

Norina Liranti Pellokila[1], Agnes Demarci Nuban[2], Frengki Lado[3]
[1,2,3]Institut Agama Kristen Negeri Kupang, Indonesia
norinalirantipellokila@gmail.com

Abstract

This article presents a conceptual and theoretical review of the application of Item Response Theory (IRT), specifically the three-parameter logistic model (3PL), as a framework for analyzing the quality of Mid-Semester Examination (MSE) items at the senior high school level. Unlike conventional empirical articles, this study does not rely on directly collected examinee response data; instead, it explores the theoretical foundations, mathematical structure, and practical implications of the 3PL model through a comprehensive synthesis of psychometric literature. The review examines the three core parameters of the 3PL model discrimination (a), difficulty (b), and pseudo-guessing (c) along with their interpretive criteria, assumption verification procedures, and a contextually grounded IRT implementation framework for educational institutions in Indonesia, particularly in East Nusa Tenggara (NTT). The article also critically compares the IRT framework with Classical Test Theory (CTT) to elucidate the conceptual advantages and practical limitations of each approach. The findings of this review are intended to serve as a conceptual foundation for teachers, evaluators, and education policymakers in adopting IRT systematically to enhance the validity and reliability of assessment instruments at the secondary school level.

Keywords: *Conceptual Review, Item Response Theory, Three-Parameter Logistic Model, Item Quality, Educational Assessment.*

Abstrak

Penelitian ini menyajikan kajian konseptual dan teoretis tentang penerapan Teori Respons Butir (Item Response Theory/IRT), khususnya model logistik tiga parameter (3PL), sebagai kerangka analisis kualitas butir soal Ujian Tengah Semester (UTS) di jenjang sekolah menengah atas. Berbeda dari artikel empiris konvensional, kajian ini tidak bertumpu pada data respons peserta tes yang dikumpulkan secara langsung, melainkan mengeksplorasi fondasi teoretis, struktur matematis, dan implikasi praktis model 3PL melalui sintesis literatur psikometrik yang komprehensif. Kajian ini menelaah tiga parameter inti model 3PL daya pembeda (a), tingkat kesukaran (b), dan pseudo-guessing (c) beserta kriteria interpretasinya, prosedur verifikasi asumsi, serta kerangka kerja implementasi IRT yang kontekstual untuk satuan pendidikan di Indonesia, khususnya di wilayah Nusa Tenggara Timur (NTT). Artikel ini juga mengkomparasi secara kritis kerangka IRT dengan Teori Tes Klasik (CTT) untuk menjelaskan keunggulan konseptual dan keterbatasan praktis masing-masing pendekatan. Temuan kajian ini diharapkan

dapat menjadi landasan konseptual bagi guru, evaluator, dan pengambil kebijakan pendidikan dalam mengadopsi pendekatan IRT secara sistemik guna meningkatkan validitas dan reliabilitas instrumen asesmen di tingkat sekolah menengah.

Keywords: Kajian Konseptual, Teori Respons Butir, Model Logistik Tiga Parameter, Kualitas Butir Soal, Asesmen Pendidikan.

1. PENDAHULUAN

Kualitas instrumen evaluasi merupakan prasyarat fundamental bagi validitas keputusan pedagogis yang dihasilkan dari proses asesmen pendidikan. Instrumen yang dibangun secara ceroboh atau dianalisis secara superfisial tidak hanya menghasilkan data pengukuran yang tidak dapat dipercaya, tetapi juga berpotensi menciptakan distorsi sistematis dalam pencitraan kemampuan peserta didik, yang pada gilirannya dapat mengarah pada keputusan yang tidak adil dan tidak akurat di tingkat individual maupun institusional (Hambleton & Swaminathan, 1985). Ujian Tengah Semester (UTS) merupakan salah satu titik evaluasi paling krusial dalam siklus pembelajaran, karena hasilnya digunakan sebagai cerminan pencapaian kompetensi sementara yang memandu penyesuaian strategi pembelajaran sebelum tahap akhir semester.

Dalam konteks praktik evaluasi pendidikan di Indonesia, analisis kualitas butir soal umumnya masih dilakukan menggunakan kerangka Teori Tes Klasik (Classical Test Theory/CTT) yang mengandalkan indeks sederhana seperti tingkat kesukaran (p) dan daya beda (D) berbasis proporsi. Meskipun pendekatan ini memiliki keunggulan dalam hal kemudahan komputasi dan interpretasi, ia mengandung keterbatasan fundamental yang tidak dapat diabaikan dalam konteks pengukuran yang menuntut presisi tinggi: parameter butir bersifat sampel-dependen sehingga tidak dapat dibandingkan lintas kelompok peserta, dan estimasi kemampuan peserta bergantung penuh pada perangkat tes yang digunakan (Baker & Kim, 2004). Keterbatasan-keterbatasan ini secara logis mengarahkan perhatian pada pendekatan pengukuran yang lebih mutakhir, yaitu Teori Respons Butir (Item Response Theory/IRT).

IRT menawarkan kerangka pengukuran yang secara konseptual lebih kokoh karena memisahkan karakteristik butir soal dari kemampuan peserta secara matematis, sehingga menghasilkan parameter yang bersifat invarian tidak berubah secara sistematis meskipun diestimasi dari kelompok peserta yang berbeda (Lord, 1980). Di antara berbagai model IRT yang tersedia untuk data dikotomi, Model Logistik Tiga Parameter (3PL) dipandang paling relevan untuk analisis soal pilihan ganda karena secara eksplisit memodelkan efek tebakan acak melalui parameter pseudo-guessing (Birnbaum, 1968). Namun demikian, adopsi IRT dalam praktik

evaluasi di sekolah menengah terutama di luar Pulau Jawa masih sangat terbatas, sebagian besar karena minimnya literatur konseptual yang menjelaskan kerangka ini dalam bahasa yang aksesibel bagi praktisi pendidikan.

Penelitian ini hadir untuk menjawab kesenjangan tersebut melalui tiga kontribusi utama. Pertama, menyajikan kajian teoretis yang mendalam tentang fondasi matematis dan konseptual model IRT 3PL dalam konteks analisis butir soal UTS. Kedua, mengkomparasi secara kritis kerangka IRT dengan CTT untuk membantu pembaca memahami posisi epistemologis masing-masing pendekatan. Ketiga, merumuskan sebuah kerangka kerja implementasi IRT yang kontekstual dan operasional bagi satuan pendidikan menengah di Indonesia, dengan merujuk secara khusus pada kondisi dan tantangan yang dihadapi sekolah-sekolah di wilayah Nusa Tenggara Timur (NTT). Dengan demikian, artikel ini diharapkan dapat berfungsi sebagai jembatan konseptual antara teori psikometrik yang kompleks dan praktik evaluasi yang dapat diimplementasikan di lapangan.

Tujuan spesifik artikel ini adalah: (1) memaparkan fondasi teoretis dan asumsi dasar IRT model 3PL; (2) menguraikan interpretasi parameter a, b, dan c beserta kriteria kualitasnya; (3) membandingkan keunggulan dan keterbatasan IRT versus CTT secara konseptual; (4) menyajikan prosedur operasional analisis butir berbasis IRT; dan (5) merumuskan implikasi kebijakan dan rekomendasi implementasi yang relevan untuk konteks pendidikan di NTT.

2. KAJIAN PUSTAKA

EVOLUSI PARADIGMA PENGUKURAN: DARI CTT KE IRT

Teori Tes Klasik (CTT) yang telah mendominasi praktik pengukuran pendidikan selama hampir satu abad bertumpu pada model linier sederhana yang memandang skor observasi sebagai penjumlahan skor benar (true score) dan galat pengukuran (measurement error): $X = T + E$. Model ini, meskipun intuitif dan mudah diaplikasikan, menyimpan keterbatasan konseptual yang mendasar: skor benar hanya terdefinisi dalam konteks tes tertentu dan sampel tertentu, bukan sebagai representasi kemampuan yang bersifat universal (Lord & Novick, 1968). Akibatnya, perbandingan skor lintas perangkat tes yang berbeda tidak dapat dilakukan secara langsung tanpa prosedur ekuating yang rumit, dan interpretasi kualitas butir soal secara inheren terikat pada karakteristik kelompok peserta yang mengerjakan tes tersebut.

Perkembangan IRT yang dimulai secara sistematis sejak karya-karya Lord (1952, 1980) dan Birnbaum (1968) merepresentasikan pergeseran paradigma yang fundamental dalam pengukuran psikologis dan pendidikan. Alih-alih bekerja pada tingkat tes secara keseluruhan,

IRT memodelkan hubungan antara kemampuan laten individual (θ) dan probabilitas memberikan respons yang benar terhadap setiap butir secara terpisah melalui sebuah fungsi matematis yang dikenal sebagai Item Response Function (IRF) atau Item Characteristic Curve (ICC). Pergeseran dari skor tes ke kemampuan laten ini memungkinkan pemisahan yang bersih antara properti butir soal dan kemampuan peserta, menghasilkan parameter yang bersifat invarian secara teoretis terhadap komposisi sampel (Embretson & Reise, 2000).

PERBANDINGAN KONSEPTUAL CTT DAN IRT

Tabel 1. Perbandingan Konseptual Teori Tes Klasik (CTT) dan Teori Respons Butir (IRT)

Aspek Perbandingan	Teori Tes Klasik (CTT)	Teori Respons Butir (IRT)
Unit analisis	Skor tes secara keseluruhan	Butir soal secara individual
Parameter butir	Bergantung pada sampel (sample-dependent)	Invarian terhadap kelompok peserta (sample-free)
Estimasi kemampuan	Bergantung pada karakteristik tes	Invarian terhadap perangkat tes yang digunakan
Informasi pengukuran	Satu indeks untuk keseluruhan tes (alpha Cronbach)	Fungsi informasi butir dan tes di setiap titik θ
Deteksi bias butir	Tidak tersedia secara langsung	Tersedia melalui analisis DIF
Penanganan tebakan	Tidak dimodelkan secara eksplisit	Dimodelkan melalui parameter pseudo-guessing (c)
Kompleksitas estimasi	Relatif sederhana, dapat dihitung manual	Memerlukan perangkat lunak khusus (BILOG-MG, R)
Asumsi utama	Tau equivalence, error homogen	Unidimensionalitas, independensi lokal, fit model

Sumber: Diadaptasi dari Embretson & Reise (2000), Baker & Kim (2004), dan de Ayala (2009)

Tabel 1 merekam perbedaan mendasar antara kedua paradigma pengukuran tersebut secara sistematis. Perbedaan yang paling krusial dari perspektif analisis butir soal terletak pada sifat parameter yang dihasilkan: CTT menghasilkan indeks butir yang secara intrinsik terikat pada sampel pengukuran, sementara IRT menghasilkan parameter yang dalam kondisi asumsi terpenuhi bersifat invarian lintas kelompok peserta yang berbeda. Invariansi ini merupakan properti yang sangat berharga dalam konteks pengembangan bank soal dan pelaksanaan tes adaptif berbasis komputer (Computer Adaptive Testing/CAT), di mana butir-butir soal dari kalibrasi yang berbeda perlu dibandingkan dan dirakit secara bersama.

MODEL-MODEL IRT UNTUK DATA DIKOTOMI

Tabel 2. Perbandingan Model IRT untuk Data Respons Dikotomi

Model	Param. B	Param. a	Param. c	Karakteristik Utama
1PL (Rasch)	Ada	Tidak	Tidak	Mengasumsikan semua butir memiliki daya pembeda yang sama; paling parsimonious dan ketat secara teoretis
2PL	Ada	Ada	Tidak	Memperhitungkan variasi daya pembeda antarbutir; cocok bila tebakan acak tidak signifikan
3PL	Ada	Ada	Ada	Model paling komprehensif; direkomendasikan untuk soal pilihan ganda di mana efek tebakan relevan

Sumber: Hambleton, Swaminathan & Rogers (1991) dan de Ayala (2009)

Dalam ekosistem model IRT untuk data dikotomi, pemilihan antara model 1PL, 2PL, dan 3PL seyogyanya didasarkan pada pertimbangan teoretis tentang konstruk yang diukur dan karakteristik instrumen, bukan semata-mata pada pertimbangan parsimoni komputasional. Untuk instrumen pilihan ganda di jenjang sekolah menengah di mana respons peserta berkemampuan rendah rentan terhadap pengaruh tebakan acak model 3PL secara konseptual lebih tepat digunakan. Sebaliknya, apabila instrumen yang dianalisis berbentuk tes essay, wawancara terstruktur, atau rubrik dengan pensekoran dikotomi yang tegas, model 1PL (Rasch) atau 2PL dapat menjadi pilihan yang lebih parsimonious dan lebih mudah diestimasi dengan sampel yang lebih kecil.

LANDASAN TEORETIS MODEL LOGISTIK TIGA PARAMETER (3PL)

Formulasi Matematis Model 3PL

Model Logistik Tiga Parameter (3PL) yang dirumuskan oleh Birnbaum (1968) mendefinisikan probabilitas peserta dengan kemampuan laten θ memberikan respons benar terhadap butir ke-i melalui fungsi:

$$P(X_i = 1 \mid \theta) = c_i + (1 - c_i) \cdot [e^{(Da_i(\theta - b_i))} / (1 + e^{(Da_i(\theta - b_i))})]$$

di mana D = 1,702 adalah konstanta skala yang menyamakan logit dengan unit normal deviate

Kurva yang dihasilkan oleh fungsi ini yang dikenal sebagai Item Characteristic Curve (ICC) memiliki bentuk huruf S (sigmoid) yang asimtotis: batas bawah kurva tidak menyentuh nol melainkan berkonvergensi pada nilai c_i, sementara batas atas mendekati 1,0 seiring meningkatnya θ. Interpretasi geometris ketiga parameter dapat dipahami sebagai berikut: parameter b_i menentukan posisi horizontal ICC pada sumbu θ, parameter a_i menentukan kecuraman (slope) ICC pada titik infleksinya, dan parameter c_i mendefinisikan nilai asimtot

bawah ICC. Pemahaman geometris ini penting bagi praktisi yang ingin menggunakan informasi visual ICC dalam proses review dan revisi butir soal.

Interpretasi Parameter a: Daya Pembeda

Parameter daya pembeda (a_i) secara matematis merepresentasikan slope kurva ICC pada titik infleksinya, yang terjadi ketika $\theta = b_i$. Nilai slope ini sebanding dengan a_i semakin besar a_i, semakin curam ICC di sekitar b_i, yang berarti tes dapat membedakan peserta dengan sangat presisi di sekitar tingkat kesukaran butir tersebut. Secara psikometrik, butir dengan a_i tinggi memberikan kontribusi informasi yang besar terhadap pengukuran, sementara butir dengan a_i mendekati nol hampir tidak memberikan informasi diferensial yang bermakna dan dapat dihilangkan dari tes tanpa kehilangan akurasi pengukuran yang signifikan (Baker, 2001).

Dalam konteks konstruksi soal, nilai a_i yang rendah sering kali merupakan sinyal diagnostik tentang permasalahan konten: butir mungkin mengukur konstruk yang berbeda dari yang dimaksudkan (construct-irrelevant variance), memiliki kunci jawaban yang ambigu, atau berisi clue inadvertent yang memudahkan peserta berkemampuan rendah untuk menjawab benar (Haladyna & Rodriguez, 2013). Oleh karena itu, investigasi kualitatif terhadap butir-butir dengan a_i rendah selalu diperlukan sebelum keputusan revisi atau penggantian dibuat.

Interpretasi Parameter b: Tingkat Kesukaran

Parameter tingkat kesukaran (b_i) dinyatakan dalam unit skala kemampuan θ, sehingga interpretasinya harus selalu dikaitkan dengan distribusi kemampuan populasi sasaran. Secara operasional, b_i merepresentasikan nilai θ di mana probabilitas menjawab benar adalah tepat $(1 + c_i)/2$ yakni titik tengah antara asimtot bawah (c_i) dan asimtot atas (1,0). Butir dengan $b_i = 0$ adalah butir yang tepat berada di tengah distribusi kemampuan populasi (apabila diasumsikan $\theta \sim N(0,1)$), dan memberikan informasi pengukuran terbanyak pada kemampuan rata-rata. Butir dengan $b_i < -1{,}0$ terlalu mudah untuk memberikan diskriminasi yang bermakna di antara peserta berkemampuan rata-rata hingga rendah, sementara butir dengan $b_i > 1{,}0$ hanya efektif membedakan peserta di segmen atas distribusi kemampuan (de Ayala, 2009).

Implikasi penting dari perspektif perakitan tes adalah bahwa distribusi nilai b_i dalam suatu perangkat tes seyogyanya mencerminkan distribusi kemampuan populasi yang menjadi target pengukuran. Apabila tes dimaksudkan untuk mengukur seluruh spektrum kemampuan secara proporsional, distribusi b_i yang ideal adalah yang menyebar merata di sekitar nol. Apabila tes dirancang untuk seleksi kelompok berkemampuan tinggi, perangkat dengan mayoritas butir $b_i > 0{,}5$ lebih sesuai.

Interpretasi Parameter c: Pseudo-Guessing

Parameter pseudo-guessing (c_i) merupakan aspek model 3PL yang paling sering disalahpahami. Istilah 'pseudo' menekankan bahwa parameter ini tidak semata-mata merepresentasikan probabilitas tebakan acak murni (yang untuk soal dengan k opsi jawaban akan sama dengan 1/k), melainkan mencakup semua mekanisme yang memungkinkan peserta berkemampuan sangat rendah menjawab benar: tebakan acak, eliminasi sebagian distraktor, atau respons bias terhadap opsi tertentu (Lord, 1980). Nilai c_i yang jauh di atas 1/k misalnya, $c_i > 0{,}25$ untuk soal dengan 5 opsi di mana nilai 1/k = 0,20 mengindikasikan bahwa distraktor tidak berfungsi efektif: sebagian besar opsi pengecoh terlalu lemah sehingga bahkan peserta yang sama sekali tidak memahami materi pun dapat mengeliminasi sebagian besar distraktor dan meningkatkan peluang tebakannya secara signifikan.

Tabel 3. Kriteria Kualitas Parameter Butir Soal dalam Model IRT 3PL beserta Implikasi Praktisnya

Parameter	Kategori	Rentang	Implikasi Praktis
A	Rendah	0,00 – 0,34	Butir tidak mampu membedakan peserta; perlu diganti atau direkonstruksi secara menyeluruh
A	Sedang	0,35 – 0,64	Butir cukup membedakan; dapat dipertahankan dengan revisi pada distractor
A	Tinggi	0,65 – 1,34	Butir baik; direkomendasikan masuk bank soal
A	Sangat Tinggi	≥ 1,35	Butir sangat baik; prioritas utama dalam perakitan tes
B	Mudah	$b < -1{,}0$	Soal terlalu mudah; kurang efektif untuk diferensiasi peserta kemampuan rendah–sedang
B	Sedang	$-1{,}0 \leq b \leq 1{,}0$	Tingkat kesukaran ideal; memberikan informasi pengukuran terbanyak
B	Sukar	$b > 1{,}0$	Soal terlalu sukar; efektif hanya untuk peserta kemampuan tinggi
C	Dapat diterima	$0{,}00 \leq c \leq 0{,}25$	Peluang tebakan rendah; distraktor berfungsi efektif
C	Bermasalah	$c > 0{,}25$	Peluang tebakan tinggi; distraktor perlu direkonstruksi

Sumber: Baker (2001), de Ayala (2009), dan Haladyna & Rodriguez (2013)

ASUMSI IRT DAN PROSEDUR VERIFIKASINYA

Asumsi Unidimensionalitas

Unidimensionalitas adalah asumsi bahwa satu dan hanya satu konstruk kemampuan laten yang secara dominan menentukan pola respons terhadap seluruh set butir soal dalam tes. Asumsi ini tidak mensyaratkan bahwa hanya satu faktor yang sama sekali tidak ada kontribusi faktor lain hal itu hampir mustahil terpenuhi dalam praktik karena setiap respons tes dipengaruhi oleh berbagai faktor seperti kecemasan tes, motivasi, dan kecepatan kerja. Yang disyaratkan adalah bahwa satu faktor dominan menghasilkan apa yang oleh Stout (1987) disebut sebagai essential unidimensionality dimensionalitas yang secara praktis cukup untuk memvalidasi penerapan model IRT.

Prosedur verifikasi unidimensionalitas yang paling direkomendasikan dalam literatur psikometrik kontemporer mencakup dua pendekatan komplementer: (1) analisis faktor konfirmatori (CFA) dengan model satu faktor, menggunakan indeks kecocokan CFI $\geq 0{,}95$ dan RMSEA $\leq 0{,}08$ sebagai kriteria penerimaan; dan (2) pemeriksaan rasio eigenvalue faktor pertama terhadap faktor kedua dalam analisis komponen utama matriks korelasi tetrakori (*polychoric correlation*), di mana rasio $\lambda_1/\lambda_2 \geq 3{,}0$ umumnya dianggap menunjukkan dominasi faktor pertama yang memadai (Hambleton, Swaminathan & Rogers, 1991).

Asumsi Independensi Lokal

Independensi lokal mensyaratkan bahwa setelah kemampuan laten θ dikontrol secara statistik, tidak terdapat korelasi residual yang bermakna antara respons terhadap pasangan butir soal manapun. Pelanggaran asumsi ini paling sering terjadi karena dua sebab: (1) tumpang tindih konten, di mana jawaban butir satu secara langsung mengungkap jawaban butir lainnya; atau (2) dependensi testlet, di mana sekelompok butir merujuk pada satu stimulus bersama (seperti paragraf bacaan atau diagram) sehingga performa pada seluruh butir dalam kelompok tersebut dipengaruhi oleh seberapa baik peserta memahami stimulus tersebut.

Alat diagnostik yang paling umum untuk mendeteksi pelanggaran independensi lokal adalah statistik Q3 yang dikembangkan oleh Yen (1984). Q3 dihitung sebagai korelasi Pearson antara residual (selisih antara respons observasi dan probabilitas yang diprediksi model) untuk setiap pasangan butir. Nilai Q3 yang melebihi 0,20 secara umum dianggap mengindikasikan dependensi lokal yang bermakna yang memerlukan penanganan, baik melalui penghapusan salah satu butir dalam pasangan tersebut maupun melalui penggunaan model testlet yang mengakomodasi struktur dependensi.

Uji Kecocokan Model (*Goodness-of-Fit*)

Bahkan apabila asumsi unidimensionalitas dan independensi lokal terpenuhi secara keseluruhan, setiap butir secara individual masih perlu dievaluasi kesesuaiannya dengan model 3PL. Butir yang tidak fit (*misfit*) adalah butir yang pola responsnya secara sistematis menyimpang dari prediksi model baik karena butir tersebut sebenarnya mengukur konstruk yang berbeda, memiliki konten yang ambigu, atau memiliki distribusi respons yang anomali karena alasan non-psikometrik (misalnya, kesalahan cetak atau kunci jawaban yang diperdebatkan).

Statistik chi-square (χ^2) berbasis kelompok kemampuan merupakan uji fit yang paling luas digunakan dalam paket estimasi IRT konvensional seperti BILOG-MG, namun ia diketahui sangat sensitif terhadap ukuran sampel sehingga dengan sampel yang sangat besar, hampir semua butir akan menunjukkan misfit yang signifikan secara statistik (Van der Linden & Hambleton, 1997). Alternatif yang lebih direkomendasikan dalam literatur kontemporer adalah statistik S-X^2 yang dikembangkan oleh Orlando dan Thissen (2000), yang berbasis pada skor tes total dan kurang sensitif terhadap ukuran sampel, serta indeks kecocokan berbasis informasi seperti nilai RMSEA butir yang dapat diinterpretasikan secara substantif terlepas dari ukuran sampel.

FUNGSI INFORMASI BUTIR DAN TES

Item Information Function (IIF)

Salah satu keunggulan paling signifikan IRT atas CTT adalah ketersediaan fungsi informasi butir (Item Information Function/IIF), yang secara matematis mendefinisikan kontribusi setiap butir terhadap ketepatan pengukuran pada setiap titik sepanjang kontinum kemampuan θ. Untuk model 3PL, IIF didefinisikan sebagai:

$$\mathbf{I_i(\theta) = D^2 a_i^2 \cdot [(P(\theta) - c_i)^2 / ((1 - c_i)^2 \cdot P(\theta) \cdot Q(\theta))]}$$

di mana $Q(\theta) = 1 - P(\theta)$

Interpretasi IIF sangat intuitif: nilai $I_i(\theta)$ yang tinggi pada suatu nilai θ berarti butir tersebut memberikan pengukuran yang presisi untuk peserta dengan kemampuan di sekitar θ tersebut. Puncak IIF butir 3PL selalu terjadi di suatu nilai θ yang sedikit di atas b_i (akibat adanya parameter c_i yang menggeser asimtot bawah), dan nilai puncaknya sebanding dengan a_i^2. Implikasi praktis: butir dengan a_i tinggi memberikan kontribusi informasi yang jauh lebih besar informasi butir berbanding lurus dengan kuadrat daya pembeda. Inilah mengapa peningkatan daya pembeda butir adalah strategi paling efektif untuk meningkatkan ketepatan pengukuran keseluruhan tes.

Test Information Function (TIF) dan Reliabilitas IRT

Fungsi informasi tes (Test Information Function/TIF) merupakan penjumlahan aditif IIF seluruh butir pada setiap titik θ: $I(\theta) = \Sigma_i I_i(\theta)$. Ini adalah salah satu properti paling elegant dari

kerangka IRT: karena informasi bersifat aditif, penambahan butir baru ke dalam tes secara langsung meningkatkan TIF di sekitar tingkat kesukaran butir baru tersebut. Hubungan antara TIF dan standard error pengukuran (SE) juga bersifat langsung dan elegan: $SE(\theta) = 1/\sqrt{I(\theta)}$. Dengan demikian, TIF memberikan gambaran komprehensif tentang profil presisi pengukuran tes sesuatu yang sama sekali tidak tersedia dalam CTT, di mana reliabilitas dinyatakan sebagai satu angka skalar tunggal yang mengasumsikan ketepatan pengukuran yang sama untuk seluruh rentang kemampuan.

Indeks reliabilitas marginal berbasis IRT dapat dihitung dari TIF melalui berbagai formula, yang paling umum adalah $\rho = 1 - [1/\bar{A}I(\theta)]$, di mana $\bar{A}I(\theta)$ adalah rerata berbobot TIF atas distribusi kemampuan peserta. Indeks ini secara konsisten menghasilkan estimasi yang lebih presisi dibandingkan Cronbach Alpha, khususnya ketika distribusi kemampuan peserta tidak homogen atau ketika tes dirancang untuk mengukur rentang kemampuan yang luas. Untuk tes UTS yang diterapkan kepada populasi peserta dengan keragaman kemampuan yang substansial seperti yang lazim dijumpai di sekolah menengah umum — perbedaan antara indeks reliabilitas IRT dan Cronbach Alpha dapat mencapai 0,05–0,10 poin, yang bukan merupakan perbedaan yang dapat diabaikan dalam konteks pengambilan keputusan psikometrik.

KERANGKA KERJA IMPLEMENTASI IRT UNTUK SEKOLAH MENENGAH

Tahapan Operasional Analisis Butir Berbasis IRT

Tabel 4. Kerangka Kerja Implementasi IRT untuk Analisis Butir Soal UTS di Sekolah Menengah

Tahap	Aktivitas	Prosedur	Kriteria Kelayakan
1	Persiapan instrument	Menyusun soal pilihan ganda sesuai kisi-kisi; memastikan kejelasan stem dan efektivitas distractor	Soal pilihan ganda dengan 4–5 opsi; minimal 30–40 butir per perangkat tes
2	Pengumpulan data respons	Administrasi tes kepada sampel kalibrasi; entri data ke format matriks butir-peserta	Minimal 200 peserta untuk model 3PL; 500+ untuk estimasi yang stabil
3	Verifikasi asumsi	Uji unidimensionalitas (CFA/PCA) dan independensi lokal (statistik Q3)	CFI ≥ 0,95; RMSEA ≤ 0,08; nilai Q3 < 0,20
4	Estimasi parameter	Menjalankan estimasi MML menggunakan BILOG-MG atau paket R (mirt/ltm)	Konvergensi algoritma EM; standard error parameter yang kecil
5	Uji kecocokan model	Mengevaluasi fit setiap butir menggunakan statistik χ^2 atau indeks RMSEA butir	Nilai χ^2 tidak signifikan pada $\alpha = 0,05$; atau S-X^2 dari Orlando & Thissen

Tahap	Aktivitas	Prosedur	Kriteria Kelayakan
6	Interpretasi dan keputusan	Mengklasifikasikan butir berdasarkan kategori parameter; merekomendasikan revisi atau retensi	Butir dengan a rendah, c tinggi, atau misfit → revisi/ganti; butir baik → bank soal

Sumber: Diadaptasi dari Hambleton, Swaminathan & Rogers (1991) dan Retnawati (2014)

Kerangka kerja enam tahap yang disajikan dalam Tabel 4 dirancang sebagai panduan operasional yang dapat diadopsi secara bertahap oleh institusi pendidikan dengan kapasitas sumber daya yang bervariasi. Tidak semua tahap perlu diimplementasikan secara bersamaan dalam adopsi awal: institusi dapat memulai dengan tahapan 1–3 (persiapan, pengumpulan data, dan verifikasi asumsi) sambil membangun kapasitas teknis untuk tahapan estimasi parameter yang lebih kompleks. Pendekatan implementasi bertahap semacam ini lebih berkelanjutan dibandingkan adopsi menyeluruh yang membutuhkan investasi kapasitas yang terlalu besar sekaligus.

Tantangan Implementasi IRT di Konteks NTT

Penerapan IRT di sekolah-sekolah menengah di Nusa Tenggara Timur dan lebih luas di kawasan Indonesia Timur menghadapi sejumlah tantangan struktural yang perlu diakui dan diantisipasi secara eksplisit dalam perencanaan implementasi. Pertama, tantangan kapasitas teknis: literasi statistik di kalangan guru dan evaluator pendidikan di NTT masih sangat bervariasi, dan akses terhadap pelatihan analisis psikometrik tingkat lanjut sangat terbatas. Solusi yang paling realistis dalam jangka pendek adalah pengembangan materi pelatihan berbasis software open-source seperti paket R (khususnya paket mirt dan ltm) yang dapat diakses secara gratis, dilengkapi dengan panduan operasional dalam bahasa Indonesia yang komprehensif.

Kedua, tantangan ukuran sampel: model 3PL secara teoretis memerlukan sampel minimal 200–500 peserta untuk menghasilkan estimasi parameter yang stabil, sementara banyak sekolah di NTT terutama di daerah terpencil memiliki populasi siswa per angkatan yang jauh di bawah angka tersebut. Solusi yang dapat ditempuh adalah pooling data lintas sekolah dalam satu rayon atau kabupaten, yang secara bersamaan juga meningkatkan representativitas sampel kalibrasi. Pendekatan kolaboratif ini memerlukan koordinasi institusional yang lebih kuat antara sekolah, dinas pendidikan, dan perguruan tinggi setempat.

Ketiga, tantangan infrastruktur teknologi: keandalan akses komputer dan internet yang masih tidak merata di sebagian wilayah NTT dapat menghambat penggunaan perangkat lunak estimasi IRT. Untuk mengatasi hal ini, paket R dengan kemampuan estimasi offline merupakan

alternatif yang paling praktis, karena hanya memerlukan satu kali unduhan instalasi dan selanjutnya dapat dijalankan tanpa koneksi internet.

Implikasi bagi Pengembangan Bank Soal

Salah satu manfaat jangka panjang paling strategis dari implementasi IRT di satuan pendidikan adalah kemampuannya untuk mendukung pengembangan bank soal yang terkalibrasi secara psikometrik. Bank soal berbasis IRT memungkinkan perakitan tes paralel perangkat tes berbeda yang mengukur konstruk yang sama dengan distribusi informasi pengukuran yang ekuivalen tanpa memerlukan prosedur ekuating yang kompleks, karena semua butir telah dikalibrasi dalam skala θ yang sama. Kemampuan ini sangat relevan dalam konteks UTS di mana soal berbeda dapat diberikan kepada kelas-kelas paralel untuk mencegah kebocoran soal, sambil tetap memastikan komparabilitas skor antar-kelas.

Lebih jauh, bank soal berbasis IRT membuka jalan menuju implementasi Computer Adaptive Testing (CAT) yang adaptif secara psikometrik mode pengujian di mana soal yang diberikan kepada setiap peserta dipilih secara algoritmik berdasarkan estimasi kemampuan sementara, sehingga setiap peserta mengerjakan soal pada tingkat kesukaran yang paling informatif untuk kemampuannya. Meskipun CAT skala penuh masih merupakan aspirasi jangka panjang bagi sebagian besar sekolah di Indonesia, pengembangan bank soal tervalidasi merupakan langkah fondasi yang tidak dapat dilewati dalam trajektori menuju mode pengujian yang lebih canggih tersebut.

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kajian konseptual (conceptual review) dengan metode studi kepustakaan (library research) yang bertujuan untuk menganalisis secara mendalam penerapan Teori Respons Butir (Item Response Theory/IRT), khususnya Model Logistik Tiga Parameter (Three-Parameter Logistic Model/3PL), dalam evaluasi kualitas butir soal Ujian Tengah Semester (UTS) pada jenjang sekolah menengah atas. Sumber data penelitian berasal dari berbagai literatur ilmiah yang relevan, meliputi buku-buku psikometri, artikel jurnal nasional dan internasional bereputasi, laporan penelitian, serta dokumen akademik yang membahas teori pengukuran pendidikan, analisis butir soal, dan pengembangan instrumen evaluasi pembelajaran. Literatur yang digunakan dipilih berdasarkan relevansi substansi, kredibilitas sumber, serta kontribusinya terhadap pengembangan konsep IRT dan implementasinya dalam konteks asesmen pendidikan. Pengumpulan data dilakukan melalui penelusuran sistematis terhadap berbagai basis data akademik dan sumber referensi yang memiliki keterkaitan dengan topik penelitian.

Analisis data dilakukan menggunakan teknik analisis isi (content analysis) dan sintesis literatur secara kritis. Tahapan analisis mencakup identifikasi konsep-konsep utama dalam IRT, pengkajian asumsi dasar model 3PL, penelaahan karakteristik parameter daya pembeda (a), tingkat kesukaran (b), dan pseudo-guessing (c), serta perbandingan konseptual antara IRT dan Teori Tes Klasik (Classical Test Theory/CTT). Selanjutnya, berbagai temuan teoretis dari literatur disintesiskan untuk merumuskan kerangka implementasi analisis butir berbasis IRT yang relevan dengan kondisi sekolah menengah di Indonesia, khususnya di Nusa Tenggara Timur (NTT). Untuk menjaga validitas kajian, dilakukan triangulasi sumber melalui perbandingan berbagai referensi yang memiliki perspektif berbeda sehingga diperoleh pemahaman yang komprehensif mengenai keunggulan, keterbatasan, serta implikasi praktis penggunaan IRT dalam peningkatan kualitas instrumen asesmen pendidikan.

4. HASIL DAN PEMBAHASAN

Sebagai artikel kajian konseptual, bagian hasil dan pembahasan ini tidak menyajikan data empiris dari pengumpulan respons peserta tes secara langsung. Sebaliknya, bagian ini menyintesis temuan-temuan konseptual dan teoretis yang diperoleh melalui analisis mendalam terhadap literatur psikometrik, kemudian membahas implikasinya bagi praktik evaluasi di sekolah menengah, khususnya di konteks Nusa Tenggara Timur (NTT). Struktur pembahasan mengikuti lima tema sentral yang sesuai dengan tujuan spesifik kajian ini: fondasi matematis dan diagnostik model IRT 3PL, perbandingan kapasitas diagnostik IRT versus CTT, verifikasi asumsi sebagai prasyarat implementasi, fungsi informasi sebagai keunggulan unik IRT, dan implikasi implementasi untuk konteks pendidikan NTT.

Fondasi Matematis Model 3PL sebagai Kerangka Diagnostik Butir Soal

Kajian terhadap formulasi matematis model logistik tiga parameter (3PL) mengungkapkan bahwa kerangka ini menawarkan kapasitas diagnostik yang secara kualitatif melampaui indeks konvensional dalam Teori Tes Klasik (CTT). Probabilitas respons benar dalam model 3PL, yang diformulasikan sebagai $P(X_i = 1 \mid \theta) = c_i + (1 - c_i) \cdot [e^{(Da_i(\theta - b_i))} / (1 + e^{(Da_i(\theta - b_i))})]$, mengintegrasikan tiga dimensi karakteristik butir secara simultan dalam satu fungsi yang terdefinisi secara matematis dengan baik (Birnbaum, 1968). Integrasi ini memungkinkan setiap butir soal dideskripsikan secara holistik, bukan sebagai kumpulan indeks yang terpisah.

Parameter *daya pembeda* (a) merupakan dimensi pertama yang menjadi kunci dalam menilai utilitas psikometrik suatu butir. Secara geometris, parameter *a* merepresentasikan

kecuraman *Item Characteristic Curve* (ICC) pada titik infleksinya, yaitu ketika θ = b. Kajian literatur mengkonfirmasi bahwa butir dengan nilai a ≥ 0,65 mampu membedakan peserta berkemampuan tinggi dan rendah secara bermakna, sementara butir dengan a < 0,35 praktis tidak memberikan kontribusi diferensial yang berarti terhadap keputusan pengukuran (Baker, 2001). Temuan konseptual ini memiliki implikasi langsung: penyusunan soal yang menghasilkan butir-butir dengan a rendah secara konsisten mengindikasikan permasalahan mendasar dalam kualitas konten, ambiguitas kunci jawaban, atau kehadiran *construct-irrelevant variance* yang melemahkan validitas butir (Haladyna & Rodriguez, 2013).

Parameter *tingkat kesukaran* (b) dalam IRT 3PL memberikan informasi yang lebih kaya dibandingkan indeks kesukaran konvensional p dalam CTT. Berbeda dari p yang merupakan proporsi sederhana dari jawaban benar dalam satu sampel, parameter b merepresentasikan posisi titik infleksi ICC dalam skala kemampuan θ yang bersifat kontinu dan invarian. Butir dengan b dalam rentang −1,0 hingga 1,0 dipandang berada pada zona kesukaran ideal karena memberikan informasi pengukuran terbanyak pada segmen kemampuan yang paling banyak dihuni peserta dalam distribusi normal standar (de Ayala, 2009). Distribusi nilai b yang terlalu terkonsentrasi pada ujung ekstrem skala menghasilkan perangkat tes yang secara psikometrik tidak efisien: terlalu banyak butir yang terlalu mudah atau terlalu sukar hanya memberikan kontribusi informasi pengukuran yang kecil bagi sebagian besar peserta.

Parameter *pseudo-guessing* (c) merupakan kontribusi paling distinktif model 3PL dibandingkan dengan model IRT yang lebih sederhana. Kajian ini menegaskan pemahaman konseptual kritis bahwa c bukanlah representasi langsung dari probabilitas tebakan acak murni (yang untuk soal lima opsi seharusnya 0,20), melainkan merupakan estimasi dari keseluruhan mekanisme yang memungkinkan peserta berkemampuan sangat rendah menjawab benar termasuk eliminasi parsial distraktor, respons bias, dan tebakan terarah (Lord, 1980). Implikasinya, nilai c yang jauh di atas 1/k bukan sekadar masalah teknis, melainkan merupakan sinyal diagnostik tentang lemahnya kualitas distraktor yang memerlukan revisi substansial.

Tabel 5. Sintesis Temuan Konseptual Kajian terhadap Model IRT 3PL dan Implikasi Praktisnya

Dimensi Kajian	Temuan Konseptual IRT 3PL	Implikasi Praktis	Referensi Utama
Parameter Daya Pembeda (a)	Merepresentasikan slope ICC; butir dengan a ≥ 0,65 berkontribusi signifikan terhadap informasi pengukuran	Butir dengan a rendah perlu direkonstruksi atau diganti; investigasi kualitatif konten wajib dilakukan	Baker (2001); Haladyna & Rodriguez (2013)

Dimensi Kajian	Temuan Konseptual IRT 3PL	Implikasi Praktis	Referensi Utama
Parameter Tingkat Kesukaran (b)	Dinyatakan dalam skala θ; distribusi b yang merata di sekitar nol mengoptimalkan informasi pengukuran	Perakitan tes harus mempertimbangkan kesesuaian distribusi b dengan kemampuan populasi sasaran	de Ayala (2009); Lord (1980)
Parameter Pseudo-Guessing (c)	Nilai c > 0,25 mengindikasikan distraktor tidak berfungsi; bukan sekadar tebakan acak murni	Rekonstruksi distraktor diperlukan; c tinggi menandai kelemahan konstruksi soal pilihan ganda	Lord (1980); Haladyna & Rodriguez (2013)
Fungsi Informasi Butir/Tes (IIF/TIF)	IIF = $D^2a^2[(P(\theta)-c)^2/((1-c)^2P(\theta)Q(\theta))]$; TIF bersifat aditif dan menghasilkan profil presisi diferensial	TIF memungkinkan identifikasi rentang kemampuan yang kurang terlayani; tidak tersedia dalam CTT	Embretson & Reise (2000); Baker & Kim (2004)
Verifikasi Asumsi	Tiga asumsi kritis: unidimensionalitas (CFI ≥ 0,95; RMSEA ≤ 0,08), independensi lokal (Q3 < 0,20), dan fit model	Kegagalan verifikasi asumsi menghasilkan estimasi parameter yang bias dan interpretasi yang menyesatkan	Hambleton et al. (1991); Yen (1984)

Sumber: Disintesis dari Baker (2001), de Ayala (2009), Lord (1980), Hambleton et al. (1991), dan Haladyna & Rodriguez (2013)

Tabel 5 merekam sintesis temuan konseptual kajian ini secara komprehensif. Yang perlu ditekankan adalah bahwa ketiga parameter model 3PL tidak berdiri sendiri sebagai indeks yang independen, melainkan saling berinteraksi dalam membentuk profil psikometrik holistik setiap butir soal. Butir dengan a tinggi tetapi c juga tinggi, misalnya, menghadirkan situasi diagnostik yang kompleks: butir tersebut mampu membedakan peserta berkemampuan sedang hingga tinggi secara efektif, namun distraktornya tidak berfungsi memadai sehingga bahkan peserta berkemampuan sangat rendah pun memiliki peluang tebakan yang substansial. Keputusan revisi atau retensi butir seperti ini memerlukan pertimbangan yang memadukan informasi dari ketiga parameter secara bersamaan.

Perbandingan Kapasitas Diagnostik IRT dan CTT

Komparasi kritis antara kerangka IRT dan CTT mengungkapkan bahwa perbedaan antara keduanya bukan semata-mata terletak pada kompleksitas komputasional, melainkan pada perbedaan epistemologis yang lebih mendasar tentang apa yang diukur dan bagaimana hasil

pengukuran diinterpretasikan. CTT, dengan modelnya X = T + E, mendefinisikan kemampuan peserta hanya dalam konteks tes tertentu yang digunakan, sehingga estimasi kemampuan secara inheren terikat pada karakteristik perangkat tes yang spesifik (Lord & Novick, 1968). IRT, sebaliknya, mendefinisikan kemampuan sebagai konstruk laten θ yang bersifat universal dan dapat diestimasi secara invarian terlepas dari perangkat tes yang digunakan, sepanjang asumsi model terpenuhi (Embretson & Reise, 2000).

Tabel 6. Perbandingan Kapasitas Diagnostik CTT dan IRT Model 3PL dalam Konteks Analisis Butir Soal UTS

Kapasitas Diagnostik	CTT (Konvensional)	IRT Model 3PL
Invariansi parameter butir	Tidak tersedia; terikat pada sampel	Tersedia; invarian secara teoretis lintas kelompok
Profil presisi pengukuran	Satu indeks global (Cronbach Alpha)	TIF diferensial di setiap titik θ
Kuantifikasi efek tebakan	Tidak dimodelkan secara eksplisit	Parameter c memodelkan pseudo-guessing
Deteksi bias butir (DIF)	Tidak tersedia secara langsung	Tersedia melalui analisis DIF berbasis IRT
Dukungan bank soal & CAT	Terbatas; memerlukan ekuating yang kompleks	Langsung; semua butir dalam skala θ yang sama
Kebutuhan sampel minimum	Relatif kecil (30–50 peserta)	200–500 peserta untuk model 3PL
Kompleksitas komputasi	Rendah; dapat dihitung manual	Tinggi; memerlukan perangkat lunak khusus

Sumber: Diadaptasi dari Embretson & Reise (2000), Baker & Kim (2004), dan Hambleton et al. (1991)

Tabel 6 mengkristalisasi perbedaan kapasitas diagnostik antara kedua pendekatan secara sistematis. Tiga perbedaan yang paling signifikan bagi konteks evaluasi UTS di sekolah menengah patut mendapat perhatian khusus. Pertama, invariansi parameter: dalam CTT, indeks daya beda D dan tingkat kesukaran p suatu butir soal dapat berubah secara substansial apabila diestimasi dari kelompok peserta yang berbeda — misalnya, dari kelas yang memiliki kemampuan rata-rata lebih tinggi. Dalam IRT, parameter a, b, dan c secara teoretis tetap konstan lintas kelompok yang berbeda, yang merupakan syarat fundamental untuk perbandingan butir dan pengembangan bank soal yang valid (Baker & Kim, 2004).

Kedua, profil presisi pengukuran: Cronbach Alpha dalam CTT menghasilkan satu angka reliabilitas yang mengasumsikan ketepatan pengukuran yang homogen di seluruh rentang

kemampuan. Asumsi ini secara empiris tidak realistis — setiap tes, karena distribusi nilai b butir-butirnya yang spesifik, cenderung memberikan pengukuran yang lebih presisi pada rentang kemampuan tertentu dan kurang presisi pada rentang lainnya. TIF dalam IRT secara eksplisit menampilkan profil presisi ini, memungkinkan evaluator untuk mengidentifikasi dengan tepat di mana tes mereka efektif dan di mana tidak, serta merencanakan penyempurnaan yang terarah (Hambleton & Swaminathan, 1985).

Ketiga, kemampuan CTT yang terbatas dalam menangani tebakan menjadi isu yang sangat relevan untuk soal pilihan ganda di tingkat UTS. Ketika peserta menebak jawaban, skor observasi mereka menjadi over-estimasi terhadap skor benar, dan CTT tidak memiliki mekanisme untuk mengkoreksi bias ini secara sistematis. Model 3PL, melalui parameter c, secara eksplisit memodelkan dan mengoreksi pengaruh tebakan dalam estimasi kemampuan, menghasilkan estimasi θ yang lebih akurat khususnya untuk peserta berkemampuan rendah yang paling rentan terhadap strategi tebakan (Lord, 1980).

Meskipun demikian, kajian ini tidak memandang IRT sebagai pengganti mutlak CTT dalam setiap konteks. Untuk evaluasi formatif skala kecil dengan jumlah peserta terbatas atau untuk tes-tes pendek yang tidak memerlukan ekuating lintas perangkat, CTT tetap merupakan pendekatan yang pragmatis dan memadai. Keunggulan IRT paling menonjol dan paling dapat dibenarkan dalam konteks yang memerlukan: pengembangan bank soal yang terkalibrasi, pembanding skor lintas perangkat tes, pengukuran presisi dalam rentang kemampuan yang luas, atau identifikasi bias butir secara sistematis.

Verifikasi Asumsi sebagai Prasyarat Implementasi IRT yang Valid

Salah satu temuan konseptual yang paling penting dari kajian ini adalah penekanan bahwa penerapan IRT tanpa verifikasi asumsi yang memadai bukan hanya tidak disarankan, melainkan secara aktif berbahaya dari perspektif pengukuran. Estimasi parameter yang diperoleh dari data yang melanggar asumsi unidimensionalitas atau independensi lokal secara sistematis tidak valid dan dapat menghasilkan keputusan tentang kualitas butir yang sepenuhnya menyesatkan (Hambleton, Swaminathan & Rogers, 1991). Ini membedakan IRT dari CTT dalam hal persyaratan analitik: CTT relatif robust terhadap pelanggaran asumsinya, sementara IRT lebih sensitif terhadap kepatuhan asumsi.

Kajian literatur mengidentifikasi bahwa prosedur verifikasi unidimensionalitas yang paling direkomendasikan saat ini adalah kombinasi dari *Confirmatory Factor Analysis* (CFA) dengan model satu faktor dan analisis komponen utama pada matriks korelasi tetrakori. Kriteria kecocokan CFI $\geq 0{,}95$ dan RMSEA $\leq 0{,}08$ untuk model satu faktor, dikombinasikan dengan rasio

eigenvalue $\lambda_1/\lambda_2 \geq 3{,}0$, memberikan bukti konvergen yang kuat tentang *essential unidimensionality* sebagaimana dikonseptualisasikan oleh Stout (1987). Penting untuk dicatat bahwa kegagalan memenuhi salah satu kriteria ini tidak secara otomatis mendiskualifikasi penerapan IRT, melainkan memerlukan investigasi lebih lanjut tentang sumber dimensionalitas tambahan — apakah berasal dari faktor metode, kelompok konten, atau konstruk yang berbeda secara substantif.

Untuk verifikasi independensi lokal, statistik Q3 yang dikembangkan oleh Yen (1984) merupakan alat diagnostik yang paling operasional dan paling mudah diinterpretasikan. Korelasi residual $Q3 > 0{,}20$ antara dua butir mengindikasikan bahwa keduanya berbagi varians yang tidak dapat dijelaskan oleh θ, yang paling sering disebabkan oleh dependensi konten (jawaban satu butir memberi petunjuk langsung untuk butir lain) atau *testlet dependency* (sekelompok butir merujuk pada stimulus bersama). Implikasi praktis bagi penyusun soal UTS sangat jelas: butir-butir dalam satu paket soal yang merujuk pada satu teks bacaan, tabel, atau diagram yang sama perlu diperlakukan sebagai *testlet* dan dianalisis menggunakan model yang mengakomodasi dependensi tersebut, atau salah satu dari pasangan butir yang berkorelasi residual tinggi perlu dihapus dari perangkat.

Temuan konseptual ini memiliki implikasi langsung bagi pelatihan guru dalam konstruksi soal: pemahaman tentang asumsi IRT seharusnya menginformasikan proses penulisan soal bahkan sebelum analisis statistik dilakukan. Guru yang memahami asumsi independensi lokal, misalnya, akan secara alami menghindari penyusunan rangkaian butir yang jawaban benarnya saling bergantung, bahkan tanpa perlu menunggu hasil analisis statistik formal.

Fungsi Informasi sebagai Keunggulan Unik IRT dalam Manajemen Kualitas Tes

Konsep *Item Information Function* (IIF) dan *Test Information Function* (TIF) merupakan kontribusi paling transformatif IRT terhadap teori dan praktik evaluasi pendidikan. Dalam CTT, tidak ada padanan konseptual untuk fungsi informasi ini: reliabilitas Cronbach Alpha adalah skalar tunggal yang tidak memberikan informasi tentang variasi presisi pengukuran di sepanjang kontinum kemampuan. TIF, yang merupakan penjumlahan aditif dari seluruh IIF butir, memberikan profil presisi diferensial yang memungkinkan evaluator untuk menjawab pertanyaan-pertanyaan diagnostik yang tidak dapat dijawab oleh CTT: *Di titik kemampuan mana tes ini paling akurat? Pada rentang kemampuan mana tes ini memberikan informasi yang tidak memadai? Apabila butir baru ditambahkan, di mana persisnya presisi pengukuran akan meningkat?*

Hubungan matematis $SE(\theta) = 1/\sqrt{I(\theta)}$ mengandung implikasi praktis yang sangat penting: *standard error* pengukuran bukan merupakan konstanta yang berlaku seragam untuk semua peserta seperti yang diasumsikan secara implisit dalam CTT, melainkan merupakan fungsi dari kemampuan θ peserta. Peserta yang kemampuannya berada di wilayah di mana TIF bernilai rendah akan mendapatkan estimasi kemampuan dengan galat yang lebih besar, dan keputusan berbasis skor mereka secara inheren lebih tidak pasti. Transparansi informasi ini memungkinkan pengambil keputusan pendidikan untuk mengkalibrasi kepercayaan mereka terhadap skor secara tepat, bukan mengasumsikan kepastian yang merata di seluruh rentang skor.

Dalam konteks pengembangan bank soal untuk UTS di NTT, TIF memberikan panduan yang sangat berharga untuk proses perakitan tes. Apabila analisis TIF mengungkapkan bahwa tes yang ada memberikan informasi yang tidak memadai ($I(\theta)$ rendah) pada rentang kemampuan rendah misalnya, karena terlalu sedikit butir dengan $b < -0{,}5$ maka solusi yang tepat sasaran adalah menambahkan butir-butir dengan tingkat kesukaran rendah hingga sedang, bukan menambah jumlah butir secara sembarangan. Pendekatan berbasis informasi ini menghasilkan perangkat tes yang lebih efisien secara psikometrik lebih sedikit butir untuk tingkat presisi yang sama.

Kerangka Implementasi IRT dalam Konteks Pendidikan NTT

Pembahasan tentang kerangka implementasi IRT untuk sekolah menengah di NTT menghasilkan sejumlah temuan konseptual yang penting tentang kesenjangan antara kapasitas teoritis pendekatan ini dan realitas struktural di lapangan. Tiga tantangan yang diidentifikasi kapasitas teknis sumber daya manusia, ukuran sampel yang terbatas, dan infrastruktur teknologi merupakan hambatan nyata yang tidak dapat diabaikan dalam perencanaan implementasi, namun ketiganya juga tidak bersifat insurmountable apabila ditangani dengan strategi yang tepat.

Tantangan kapasitas teknis dapat diatasi melalui pendekatan pelatihan bertingkat yang dimulai dari pemahaman konseptual sebelum beralih ke implementasi teknis. Hasil kajian ini menunjukkan bahwa pemahaman konseptual tentang model 3PL termasuk interpretasi parameter, kriteria kualitas, dan logika verifikasi asumsi dapat dipelajari oleh guru dengan latar belakang pendidikan matematika atau sains tanpa harus menguasai statistik psikometrik tingkat lanjut terlebih dahulu. Investasi dalam pemahaman konseptual ini akan mempercepat kurva pembelajaran ketika pelatihan teknis berbasis perangkat lunak dilaksanakan.

Tantangan ukuran sampel merupakan tantangan yang paling fundamental secara statistik. Model 3PL secara konseptual memerlukan sampel minimal 200–500 peserta untuk estimasi parameter yang stabil, sementara sekolah-sekolah di NTT terutama di daerah terpencil sering kali

hanya memiliki 30–80 peserta per angkatan. Kajian literatur menunjukkan dua pendekatan yang dapat ditempuh. Pertama, *pooling* data lintas sekolah dalam satu rayon atau kabupaten, yang memungkinkan akumulasi ukuran sampel yang memadai sambil meningkatkan representativitas kalibrasi. Kedua, penggunaan model IRT yang lebih parsimonious seperti model 1PL (Rasch) yang memerlukan sampel yang lebih kecil (sekitar 100–150 peserta) karena hanya mengestimasi satu parameter per butir (Hambleton et al., 1991). Meskipun model 1PL tidak mengakomodasi variasi daya pembeda dan efek tebakan, ia tetap memberikan keunggulan invariansi parameter yang merupakan keunggulan paling fundamental IRT atas CTT.

Dari perspektif jangka panjang, implementasi IRT di NTT perlu dipandang sebagai *trajectory* bertahap, bukan transformasi instantaneous. Fase pertama dapat difokuskan pada pembangunan kapasitas konseptual dan pengumpulan data kalibrasi melalui kolaborasi lintas sekolah. Fase kedua dapat beralih ke estimasi parameter menggunakan model yang lebih sederhana (1PL atau 2PL) sambil mengembangkan bank soal awal. Fase ketiga, apabila kapasitas dan ukuran sampel kumulatif telah memadai, dapat melibatkan transisi ke model 3PL yang lebih komprehensif. Pendekatan bertahap ini memungkinkan manfaat IRT mulai dinikmati dalam jangka pendek tanpa harus menunggu semua prasyarat ideal terpenuhi terlebih dahulu.

Akhirnya, kajian ini menegaskan bahwa peningkatan kualitas instrumen evaluasi berbasis IRT di NTT bukan sekadar agenda teknis-psikometrik, melainkan merupakan agenda keadilan pendidikan yang lebih luas. Instrumen yang valid, reliabel, dan mampu mendeteksi bias memberikan gambaran yang lebih akurat tentang kemampuan setiap peserta didik terlepas dari faktor-faktor non-kognitif seperti kecemasan tes, familiaritas dengan format tes, atau aksesibilitas bahan ajar. Dalam konteks NTT di mana heterogenitas kondisi belajar antar wilayah sangat substansial, kontribusi IRT terhadap ekuitas pengukuran memiliki nilai strategis yang melampaui pertimbangan psikometrik semata.

4. PENUTUP

Artikel konseptual ini telah menyajikan kajian yang komprehensif tentang penerapan Teori Respons Butir model logistik tiga parameter (IRT 3PL) sebagai kerangka analisis kualitas butir soal Ujian Tengah Semester di jenjang sekolah menengah atas, dengan perhatian khusus pada relevansinya bagi konteks pendidikan di Nusa Tenggara Timur. Secara keseluruhan, kajian ini menegaskan bahwa IRT bukan sekadar alternatif teknis yang lebih canggih terhadap CTT, melainkan merepresentasikan pergeseran paradigma yang fundamental dalam cara kita mengkonseptualisasikan dan mengoperasionalisasikan pengukuran kemampuan dari pendekatan

yang berpusat pada tes menuju pendekatan yang berpusat pada konstruk kemampuan laten yang bersifat invarian.

Model 3PL, dengan ketiga parameternya yang saling melengkapi daya pembeda (a), tingkat kesukaran (b), dan pseudo-guessing (c) memberikan deskripsi psikometrik yang jauh lebih kaya dan diagnostik tentang karakteristik setiap butir soal dibandingkan dengan indeks konvensional dalam CTT. Parameter daya pembeda memungkinkan identifikasi butir-butir yang secara efektif membedakan peserta berkemampuan tinggi dan rendah; parameter tingkat kesukaran memberikan informasi tentang kesesuaian antara tuntutan kognitif butir dan kemampuan populasi sasaran; sementara parameter pseudo-guessing secara eksplisit mengkuantifikasi efektivitas konstruksi distraktor dan mengidentifikasi butir-butir yang rentan terhadap strategi tebakan oportunistik. Bersama-sama, ketiga parameter ini membentuk basis diagnostik yang komprehensif untuk keputusan revisi, retensi, atau penggantian butir soal.

Kajian ini juga menggarisbawahi bahwa implementasi IRT yang efektif memerlukan pemenuhan tiga asumsi dasar unidimensionalitas, independensi lokal, dan kecocokan model yang harus diverifikasi secara empiris sebelum interpretasi parameter dilakukan. Kegagalan dalam memverifikasi asumsi-asumsi ini dapat menghasilkan estimasi parameter yang bias dan menyesatkan. Prosedur verifikasi yang direkomendasikan CFA untuk unidimensionalitas dan statistik Q3 untuk independensi lokal telah tersedia dalam perangkat lunak yang dapat diakses secara bebas, sehingga hambatan teknis untuk menerapkannya relatif dapat diatasi dengan pelatihan yang memadai. Kerangka kerja implementasi enam tahap yang dirumuskan dalam artikel ini diharapkan dapat berfungsi sebagai panduan operasional yang konkret bagi institusi pendidikan yang ingin mengadopsi analisis IRT secara sistemik. Akhirnya, artikel ini menekankan bahwa peningkatan kualitas instrumen evaluasi berbasis IRT di sekolah-sekolah menengah di NTT dan kawasan Indonesia Timur secara lebih luas bukanlah semata-mata agenda teknis-psikometrik, melainkan merupakan agenda keadilan pendidikan. Instrumen yang valid, reliabel, dan bebas bias merupakan komponen kritis dari sistem evaluasi yang adil sistem yang memberikan gambaran akurat tentang kemampuan setiap peserta didik terlepas dari latar belakang sosial, geografis, atau demografisnya. Investasi dalam pengembangan kapasitas analisis IRT di kalangan pendidik merupakan investasi dalam ekuitas pengukuran yang memiliki dampak jangka panjang terhadap kualitas keputusan pedagogis di seluruh jenjang sistem pendidikan.

5. SARAN

Berdasarkan kajian konseptual yang telah dikembangkan, beberapa rekomendasi berikut diajukan kepada pemangku kepentingan yang relevan. Bagi guru dan tim penyusun soal, disarankan untuk mulai membiasakan diri dengan konsep dan terminologi IRT melalui literatur yang tersedia dalam bahasa Indonesia, serta memanfaatkan panduan interpretasi parameter yang disajikan dalam artikel ini sebagai referensi praktis dalam mengevaluasi kualitas butir soal yang disusun. Kesadaran tentang pentingnya variasi tingkat kesukaran dan efektivitas distraktor dalam konstruksi soal pilihan ganda merupakan langkah awal yang tidak memerlukan komputasi IRT formal namun sudah diinformasi oleh prinsip-prinsip IRT.

Bagi kepala sekolah dan manajemen institusi, direkomendasikan untuk memfasilitasi program pelatihan analisis butir soal berbasis IRT bagi tim guru secara berkala, serta mengalokasikan sumber daya minimal akses komputer dengan perangkat lunak R yang terinstal untuk mendukung praktik analisis butir yang lebih sistematis. Pengembangan bank soal tervalidasi yang dapat digunakan bersama dalam satu rayon atau wilayah juga sangat dianjurkan sebagai proyek kolaboratif jangka menengah.

Bagi Dinas Pendidikan Kota Kupang dan Provinsi NTT, kajian ini merekomendasikan perumusan kebijakan yang secara eksplisit mendorong dan memberikan insentif bagi adopsi pendekatan psikometrik berbasis IRT dalam pengembangan instrumen evaluasi di satuan pendidikan menengah. Kolaborasi strategis dengan perguruan tinggi setempat khususnya program studi Pendidikan Matematika, Statistika, dan Psikologi Pendidikan untuk mengembangkan kapasitas lokal dalam analisis IRT merupakan langkah yang paling cost-effective dan berkelanjutan dalam jangka panjang. Terakhir, penelitian empiris lanjutan yang menerapkan kerangka konseptual yang dibangun dalam artikel ini pada data respons aktual siswa di NTT sangat diperlukan untuk memvalidasi relevansi dan adaptabilitas pendekatan ini dalam konteks lokal yang spesifik.

DAFTAR PUSTAKA

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Addison-Wesley.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

Lord, F. M. (1952). *A theory of test scores*. Psychometric Monographs, 7, 1–84.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Parama Publishing.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589–617. https://doi.org/10.1007/BF02294821

Van Der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. Springer.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125–145. https://doi.org/10.1177/014662168400800201