Sultra Educational Journal (Seduj)
Website: http://jurnal-unsultra.ac.id/index.php/seduj
Email: seduj.journal@gmail.com

Peningkatan Presisi Penskoran Hasil Belajar: Analisis Fungsi Informasi Tes (TIF) dan Standard Error Menggunakan Software Xcalibre untuk Model 3-Parameter Logistik (3PL)

INFO PENULIS

Muhammad Tahir
STKIP YPUP Makassar
tahirjakarta@gmail.com

Mulyati
STKIP YPUP Makassar
mulyatiypup@gmail.com

INFO ARTIKEL

ISSN: 2807-7474
Vol. 5, No. 3, Desember 2025
http://jurnal-unsultra.ac.id/index.php/seduj

Saran Penulisan Referensi:

Tahir, M., & Mulyati. (2025). Peningkatan Presisi Penskoran Hasil Belajar: Analisis Fungsi Informasi Tes (TIF) dan Standard Error Menggunakan Software Xcalibre untuk Model 3-Parameter Logistik (3PL). *Sultra Educational Journal, 5* (3), 753-761.

Abstrak

Penelitian ini bertujuan untuk mengevaluasi presisi pengukuran instrumen hasil belajar dengan memanfaatkan keunggulan metrik Item Response Theory (IRT). Berbeda dengan Classical Test Theory (CTT) yang mengasumsikan kesalahan pengukuran konstan, penelitian ini menganalisis variabilitas Fungsi Informasi Tes (TIF) dan Standard Error of Measurement (SE(θ)) di sepanjang kontinum kemampuan. Metode penelitian yang digunakan adalah kuantitatif deskriptif-evaluatif dengan melibatkan 620 responden yang mengerjakan 45 butir soal pilihan ganda dikotomus. Kalibrasi dilakukan menggunakan Model 3-Parameter Logistik (3PL) melalui Software Xcalibre pada lingkungan R untuk mengestimasi parameter daya beda (a), tingkat kesulitan (b), dan tebakan semu (c). Hasil analisis menunjukkan bahwa model 3PL memiliki kesesuaian yang sangat baik dengan data ($RMSEA$ = 0,051; $SRMR$ = 0,042). Temuan utama mengindikasikan bahwa TIF mencapai nilai maksimum sebesar 25,80 pada tingkat kemampuan θ = +0,25, yang berkorespondensi dengan nilai SE(θ) minimum sebesar 0,197. Sebaliknya, pada rentang kemampuan ekstrem (θ < -2.0 and θ > +2.5), nilai SE(θ) meningkat tajam hingga mencapai 0,392, yang menunjukkan penurunan presisi pengukuran secara signifikan. Simpulan dari penelitian ini menegaskan bahwa presisi penskoran bersifat heteroskedastik atau bergantung pada tingkat kemampuan individu. Oleh karena itu, institusi pendidikan direkomendasikan untuk beralih dari pelaporan skor tunggal ke pelaporan skor yang disertai dengan Interval Kepercayaan 95% guna meminimalisir bias interpretasi. Selain itu, pengembangan tes di masa depan perlu menambah butir soal dengan tingkat kesulitan ekstrem untuk meratakan distribusi informasi tes.

Kata Kunci: Item Response Theory (IRT), Model 3-Parameter Logistik (3PL), Software Xcalibre, Fungsi Informasi Tes (TIF), Standard Error of Measurement.

Abstract

This study aims to evaluate the measurement precision of learning outcome instruments by leveraging the advantages of Item Response Theory (IRT) metrics. Unlike Classical Test Theory (CTT), which assumes a constant measurement error, this study analyzes the variability of the Test Information Function (TIF) and the Standard Error of Measurement (SE(θ)) across the ability continuum. The research method employed is descriptive-evaluative quantitative, involving 620 respondents who completed 45 dichotomous multiple-choice items. Calibration was performed using the 3-Parameter Logistic (3PL) Model via Xcalibre Software in the R environment to estimate discrimination (a), difficulty (b), and pseudo-guessing (c) parameters. The analysis results demonstrate that the 3PL model has an excellent fit with the data (RMSEA = 0.051; SRMR = 0.042). The primary findings indicate that the TIF reaches a maximum value of 25.80 at an ability level of θ = +0.25, corresponding to a minimum SE(θ) value of 0.197. Conversely, at extreme ability ranges (θ < -2.0 and θ > +2.5), the SE(θ) value increases sharply to 0.392, indicating a significant decrease in measurement precision. The study concludes that scoring precision is heteroscedastic, meaning it depends on the individual's ability level. Therefore, educational institutions are recommended to shift from reporting single scores to reporting scores accompanied by a 95% Confidence Interval to minimize interpretation bias. Furthermore, future test development should incorporate items with extreme difficulty levels to achieve a more uniform test information distribution.

Keywords:
Item Response Theory (IRT), 3-Parameter Logistic (3PL) Model, Xcalibre Software, Test Information Function (TIF), Standard Error of Measurement.

A. Pendahuluan

Pentingnya presisi dalam pengukuran hasil belajar, pengukuran hasil belajar merupakan komponen fundamental dalam sistem pendidikan dan asesmen. Keputusan krusial mengenai kelulusan, penempatan siswa, dan efektivitas kurikulum sangat bergantung pada skor tes yang dihasilkan. Oleh karena itu, tuntutan terhadap kualitas skor, khususnya aspek presisi atau keandalan (*reliability*), menjadi sangat tinggi.

Secara tradisional, kualitas skor hasil belajar dianalisis menggunakan Classical Test Theory (CTT). Dalam CTT, presisi tes diukur dengan koefisien tunggal, seperti ∝ Cronbach, yang dianggap berlaku secara seragam untuk semua peserta (Lord, 1980). Walaupun sederhana dan luas digunakan, asumsi CTT mengenai error pengukuran standar (Standard Error of Measurement - SEM) yang konstan di seluruh rentang kemampuan telah lama dikritik karena tidak realistis (Hambleton & Swaminathan, 1985). Kenyataannya, tes sering kali dirancang untuk mengukur kemampuan rata-rata dengan lebih baik daripada kemampuan yang ekstrem (sangat rendah atau sangat tinggi). Konsekuensinya, skor yang diberikan kepada peserta berprestasi tinggi atau rendah mungkin memiliki tingkat kesalahan yang jauh lebih besar daripada skor peserta rata-rata, namun hal ini tidak tercermin dalam laporan reliabilitas CTT.

Pergeseran paradigma ke Item Response Theory (IRT). Keterbatasan CTT dalam menangani presisi pengukuran yang bervariasi telah mendorong pergeseran paradigma menuju Item Response Theory (IRT). IRT menawarkan model matematis yang lebih canggih untuk menjelaskan hubungan antara tingkat kemampuan laten peserta θ dan probabilitas respons benar terhadap suatu item (Embretson & Reise, 2000). Keunggulan utama IRT terletak pada sifatnya yang item-invariant (parameter item tidak bergantung pada sampel) dan sample-invariant (estimasi kemampuan peserta tidak bergantung pada item mana yang mereka jawab).

Satu diantara berbagai model IRT, Model 3-Parameter Logistik (3PL) dipilih ketika asumsi tebakan (guessing) dianggap signifikan dalam format tes pilihan ganda. Model 3PL mengestimasi tiga parameter untuk setiap item: daya beda (*a*), tingkat kesulitan (*b*), dan kemungkinan tebakan (*c*). Penggunaan model 3PL menjadi penting dalam asesmen hasil belajar di tingkat yang lebih tinggi, di mana peserta mungkin memiliki insentif untuk menebak jawaban yang tidak mereka ketahui (Birnbaum, 1968; Lord, 1980).

Presisi IRT: Fungsi Informasi Tes (TIF) dan Standard Error (SE). Metrik presisi IRT jauh lebih informatif daripada α Cronbach. Dalam IRT, presisi diukur oleh Fungsi Informasi Tes (TIF).

1. Fungsi Informasi Item (IIF): Setiap item dalam tes berkontribusi pada presisi pengukuran. IIF menunjukkan seberapa efektif suatu item dalam membedakan antar peserta pada tingkat

kemampuan θ tertentu. Item dengan daya beda (a) tinggi dan tingkat kesulitan (b) yang dekat dengan θ peserta akan memberikan informasi paling banyak.

2. Fungsi Informasi Tes (TIF): TIF adalah penjumlahan dari IIF semua item dan menunjukkan kontribusi informasi pengukuran dari seluruh tes pada setiap tingkat kemampuan θ (Hambleton & Swaminathan, 1985).

Hubungan TIF dengan kesalahan pengukuran diwujudkan dalam Standard Error of Measurement SE(θ) yang spesifik untuk setiap tingkat kemampuan θ. TIF dan $SE(\theta)$ berhubungan secara invers:

$$SE(\theta) = \frac{1}{\sqrt{TIF(\theta)}}$$

Ketika nilai TIF tinggi, maka SE(θ) akan rendah, menunjukkan presisi yang tinggi dan kesalahan pengukuran yang kecil pada tingkat kemampuan tersebut. Sebaliknya, TIF yang rendah menunjukkan bahwa skor peserta di rentang θ tersebut harus diinterpretasikan dengan hati-hati karena memiliki kesalahan pengukuran yang besar.

Analisis TIF dan $SE(\theta)$ memberikan pemahaman yang mendalam tentang: (1) di mana letak kekuatan pengukuran tes (tingkat θ dengan TIF maksimum), dan (2) batasan dan kelemahan tes (tingkat θ dengan $SE(\theta)$ maksimum). Informasi ini memungkinkan pengembang tes untuk:

1. Mengarahkan perbaikan item agar berfokus pada rentang kemampuan yang memiliki presisi rendah.
2. Melaporkan skor peserta dengan interval kepercayaan yang akurat.

Peran Software Xcalibre dibutuhkan untuk implementasi IRT, khususnya Model 3PL, Software Xcalibre menjadi alat yang ideal untuk penelitian akademik dan institusi pendidikan (Chalmers, 2012). Fleksibilitas Software Xcalibre mendukung berbagai model IRT, termasuk 3PL, model respon bertingkat (Graded Response Model), hingga model multidimensi (MIRT). Fungsi Analisis Komprehensif: Paket ini menyediakan fungsi bawaan untuk menghitung dan memplot TIF, $SE(\theta)$, Kurva Karakteristik Item (ICC), dan melakukan uji fit model yang canggih. Penggunaan Software Xcalibre dalam penelitian ini memastikan bahwa analisis TIF dan $SE(\theta)$ dilakukan dengan metode estimasi yang mutakhir dan transparan (misalnya, menggunakan algoritma EM atau MHRM), menghasilkan estimasi parameter yang lebih andal dibandingkan dengan software yang kurang fleksibel.

Permasalahan Penelitian. Berdasarkan latar belakang di atas, penelitian ini diarahkan untuk menjawab pertanyaan-pertanyaan utama:

1. Seberapa stabil dan fit Model 3PL IRT dalam memodelkan data respons tes hasil belajar?
2. Pada tingkat kemampuan θ manakah tes hasil belajar mencapai puncak presisi pengukuran (TIF maksimum)?
3. Bagaimana variasi Standard Error of Measurement $SE(\theta)$ di sepanjang kontinum kemampuan θ, dan apa implikasinya terhadap pemberian dan pelaporan skor yang kredibel?

Signifikansi Penelitian. Hasil penelitian ini memberikan kontribusi signifikan, terutama dalam ranah praktis:

1. Peningkatan Akuntabilitas Skor: Dengan mengidentifikasi $SE(\theta)$ yang spesifik, lembaga pendidikan dapat melaporkan skor peserta dengan interval kepercayaan yang lebih akurat, meningkatkan akuntabilitas dan validitas interpretasi skor (Thissen & Wainer, 2001).
2. Panduan Pengembangan Tes: Analisis TIF secara eksplisit memberikan panduan bagi pengembang tes untuk merevisi item atau menambahkan item baru yang secara strategis dapat meningkatkan presisi pengukuran di rentang kemampuan yang dianggap penting.

Secara keseluruhan, penelitian ini tidak hanya mengaplikasikan teknik psikometri tingkat lanjut, tetapi juga menggunakan keunggulan Software Xcalibre untuk menghasilkan data presisi yang komprehensif, yang merupakan langkah esensial menuju penskoran hasil belajar yang lebih kredibel dan informatif.

B. Metodologi

Desain Penelitian

Penelitian ini mengadopsi pendekatan kuantitatif non-eksperimental dengan fokus pada psikometri dan evaluasi instrumen. Desain ini bertujuan utama untuk mengkalibrasi item tes hasil belajar dan menganalisis properti pengukuran yang dihasilkan oleh Model 3-Parameter Logistik (3PL) Item Response Theory (IRT), khususnya dalam konteks Fungsi Informasi Tes (TIF) dan Standard Error of Measurement ($\text{SE}(\theta)$). Penelitian ini bersifat deskriptif

evaluatif karena mendeskripsikan dan mengevaluasi kualitas dan presisi skor berdasarkan model statistik canggih, bukan menguji hubungan kausal antarvariabel (Embretson & Reise, 2000).

Populasi dan Sampel

Populasi penelitian, populasi target penelitian ini adalah seluruh peserta didik tingkat SMA/MA. Siswa Kelas XII pada mata pelajaran Matematika pada Tahun Akademik, 2024/2025. Populasi ini dipilih karena merupakan subjek utama yang hasil belajarnya dievaluasi oleh instrumen yang sedang dikaji.

Prosedur dan ukuran sampel, pengambilan sampel dilakukan menggunakan teknik Convenience Sampling. Penentuan Ukuran Sampel dalam IRT:Salah satu pertimbangan krusial dalam penelitian IRT adalah ukuran sampel. Model 3PL, yang melibatkan estimasi tiga parameter per item (daya beda *a*, kesulitan *b*, dan tebakan *c*), dikenal membutuhkan ukuran sampel yang lebih besar dibandingkan Model Rasch (1PL) atau 2PL untuk memastikan stabilitas dan akurasi estimasi parameter (Lord, 1980).Menurut studi simulasi yang dilakukan oleh Hambleton dan Swaminathan (1985), untuk Model 3PL dengan panjang tes sekitar 40 item, ukuran sampel minimal yang direkomendasikan adalah 500 peserta untuk mendapatkan estimasi parameter yang relatif stabil dan bias yang rendah. Fan (1998) lebih lanjut menyarankan bahwa sampel di atas 500 akan secara signifikan mengurangi Standard Error of Parameter Estimates $SE(a, b, c)$. Oleh karena itu, penelitian ini menetapkan target jumlah responden sebanyak 620 peserta. Jumlah ini diharapkan memadai untuk memenuhi asumsi asimtotik yang diperlukan oleh prosedur estimasi Model 3PL yang digunakan dalam paket mirt (Chalmers, 2012), sehingga menjamin reliabilitas temuan TIF dan $SE(\theta)$.

Instrumen pengukuran,

Instrumen yang digunakan adalah Tes Hasil Belajar (THB) mata pelajaran matematika yang dirancang untuk mengukur capaian pembelajaran yang spesifik.

- Format Tes: Tes ini terdiri dari 45 item berbentuk pilihan ganda dengan lima opsi jawaban (A, B, C, D, E).
- Penskoran Dikotomus: Respons peserta diberi skor secara dikotomus: 1 untuk jawaban benar dan 0 untuk jawaban salah. Model 3PL dirancang khusus untuk data dikotomus seperti ini, di mana peluang tebakan (*c*) dapat dimodelkan (Lord, 1980).
- Validitas Konten: Sebelum analisis psikometri, instrumen telah melalui proses validasi konten yang ketat, melibatkan tinjauan ahli (expert review) oleh tiga ahli kurikulum dan dua guru pengampu mata pelajaran untuk memastikan kesesuaian item dengan kurikulum dan tujuan pembelajaran.

Prosedur Pengumpulan Data

Peserta tes diberikan *Informed Consent* untuk memastikan partisipasi mereka bersifat sukarela dan anonimitas data dijamin. Administrasi tes, tes dilaksanakan secara online melalui platform LMS di bawah pengawasan ketat. Waktu pengerjaan ditetapkan selama 90 menit. Pengumpulan data respons, data dikumpulkan dan direkapitulasi dalam format matriks respons peserta × item. Matriks ini hanya berisi angka 1 dan 0, yang merupakan format input standar untuk kalibrasi IRT.

Teknik Analisis Data IRT Menggunakan Software Xcalibre

Seluruh analisis statistik dan psikometri dilakukan menggunakan Software Xcalibre.

- Uji asumsi Model IRT, sebelum kalibrasi utama, dua asumsi mendasar IRT diuji, yaitu: (1) Asumsi Unidimensionalitas: Asumsi ini menyatakan bahwa kinerja peserta pada tes hanya dapat dijelaskan oleh satu konstruk laten (kemampuan, θ). Uji dilakukan dengan Analisis Faktor Eksploratori (EFA) atau Analisis Faktor Konfirmatori (CFA). Indeks seperti Root Mean Square Error of Approximation (RMSEA) dan Standardized Root Mean Square Residual (SRMR) akan dievaluasi. Di samping itu, Analisis Paralel (Parallel Analysis) juga digunakan untuk menentukan jumlah faktor yang paling tepat (Hayton, Allen, & Scarpello, 2004). (2) Asumsi Independensi Lokal: Asumsi ini menyatakan bahwa respons terhadap item bersifat independen statistik ketika kemampuan laten (θ) telah dikontrol. Pelanggaran dapat diidentifikasi melalui analisis residual atau uji Q3 Yen, yang menunjukkan korelasi sisa antara item yang tidak dijelaskan oleh θ.
- Kalibrasi Model 3-Parameter Logistik (3PL). Data respons dianalisis menggunakan Software Xcalibre dengan spesifikasi Model 3PL.

Persamaan 3PL:

$$P(\mathrm{X_{ij}} = 1|\theta_j) = \mathrm{c_i} + (1 - \mathrm{c_i})\frac{1}{1 + \mathrm{e}^{-\mathrm{a_i}(\theta_j - \mathrm{b_i})}}$$

di mana:

- $P(X_{ij} = 1|\theta_j)$: Probabilitas peserta *j* menjawab benar item *i*, diberikan kemampuan θ_j.
- a_i: Parameter Daya Beda (Discrimination) item *i*. Menunjukkan seberapa baik item membedakan peserta dengan kemampuan tinggi dan rendah.
- b_i: Parameter Tingkat Kesulitan (Difficulty/Location) item *i*. Menunjukkan tingkat kemampuan θ di mana probabilitas jawaban benar adalah $(1 + c_i)/2$.
- c_i: Parameter Tebakan (Guessing/Pseudo-Chance Level) item *i*. Probabilitas peserta dengan kemampuan sangat rendah menjawab benar.
- θ_j: Kemampuan Laten (Latent Trait/Ability) peserta *j*.

Metode Estimasi: Estimasi parameter dilakukan menggunakan metode yang disiapkan pada Software Xcalibre, yaitu Maximum Likelihood Estimation (MLE).

- Evaluasi Item dan *Model Fit*: (1) Uji Goodness-of-Fit Model: Kesesuaian model secara keseluruhan (global fit) diuji menggunakan statistik M_2 (Modifikasi dari M_2 yang dikembangkan oleh Maydeu-Olivares) atau χ^2 yang disediakan oleh Xcalibre. Indeks fit model seperti RMSEA dan SRMR dievaluasi untuk menentukan kecocokan model yang memadai (Hu & Bentler, 1999). (2) Item Fit: Uji kecocokan item (local item fit) dilakukan untuk mengidentifikasi item yang tidak sesuai dengan model 3PL. Uji χ^2 S-X2 (Orlando & Thissen, 2003) dan analisis residual grafis digunakan. Item yang menunjukkan misfit yang signifikan akan dicatat untuk rekomendasi revisi.
- Analisis Kunci: Fungsi Informasi Tes (TIF) dan Standard Error (SE(θ)). Setelah model dikalibrasi dan dianggap fit secara memadai, analisis TIF dan SE yang menjadi fokus utama penelitian dapat dilakukan: (1) Fungsi Informasi Item (IIF): Fungsi iteminfo() di Xcalibre digunakan untuk mendapatkan kontribusi informasi spesifik setiap item di seluruh rentang θ. (2) Fungsi Informasi Tes (TIF): Fungsi testinfo() digunakan untuk menghitung TIF, yang merupakan penjumlahan IIF di seluruh item. Plot TIF akan menjadi output visual kunci untuk mengidentifikasi tingkat kemampuan di mana tes memberikan presisi tertinggi. (3) Standard Error of Measurement (SE(θ)): Nilai SE(θ) dihitung sebagai akar kebalikan dari TIF pada setiap titik θ. Plot SE(θ) akan memvisualisasikan bagaimana error pengukuran bervariasi sepanjang kontinum kemampuan.

  Penentuan Presisi: Tingkat presisi diinterpretasikan berdasarkan titik di mana TIF mencapai puncaknya. Secara umum, presisi dianggap memadai ketika $\mathrm{SE}(\theta) < 0.32$ (setara dengan reliabilitas 0.90) pada rentang kemampuan yang ditargetkan (Lord, 1980).
- Pemberian dan Pelaporan Skor: (1) Estimasi Kemampuan (θ): Skor kemampuan peserta diestimasi menggunakan Metode EAP (Expected A Posteriori) yang merupakan default yang andal untuk IRT. (2) Pelaporan Presisi: Untuk pelaporan, SE(θ) akan digunakan untuk menghitung Interval Kepercayaan (CI) 95% untuk skor kemampuan setiap peserta:

$$\theta_{\mathrm{CI}} = \theta_{\mathrm{EAP}} \pm 1.96 \times \mathrm{SE}(\theta)$$

  Perhitungan ini memberikan representasi yang lebih jujur tentang ketidakpastian pengukuran dibandingkan dengan hanya melaporkan skor mentah atau skor θ tunggal.

C. Hasil dan Pembahasan

Statistik Deskriptif dan Uji Asumsi Awal

Statistik Deskriptif Hasil Tes. Data respons dari 620 peserta terhadap 45 item. Tes Hasil Belajar dianalisis. Skor mentah peserta dari 12 hingga 43, dengan skor rata-rata (Mean) : 28.5 dan Standard Deviation (SD) : 6.2.

Tabel 1. Statistik Deskriptif Skor Mentah Tes Hasil Belajar

Statistik	Nilai	Interpretasi
Jumlah Responden (N)	620	Memadai untuk Kalibrasi 3PL
Panjang Tes (Item)	45	

Skor Minimum	12	
Skor Maksimum	43	
Mean	28.5	
Standard Deviation (SD)	6.2	
Skewness	0.35	Sedikit cenderung positif (ke kiri)
Kurtosis	-0.15	Distribusi mendekati normal
Reliabilitas CTT (α Cronbach)	0.84	Presisi CTT yang memadai

Nilai reliabilitas CTT (α Cronbach = 0.84) menunjukkan bahwa tes secara keseluruhan memiliki konsistensi internal yang baik menurut standar tradisional (Nunnally & Bernstein, 1994). Namun, nilai ini tidak memberikan informasi tentang variasi error pengukuran pada tingkat kemampuan yang berbeda, yang merupakan fokus utama penelitian ini.

Uji Asumsi Unidimensionalitas. Asumsi kunci dari IRT unidimensional (seperti Model 3PL) adalah bahwa tes harus mengukur satu konstruk dominan. Uji unidimensionalitas dilakukan menggunakan Analisis Faktor Eksploratori (EFA) pada matriks korelasi item respons dikotomus dengan Principal Axis Factoring.

Analisis menunjukkan bahwa Faktor 1 (Kemampuan Laten) menjelaskan Persentase Varians: 38.2% dari total varians, sementara Faktor 2 hanya menjelaskan Persentase Varians: 4.8% (lihat Tabel 2).

Tabel 2. Hasil Analisis Faktor Eksploratori (EFA)

Faktor	*Eigenvalue*	Varians Dijelaskan (%)
1	17.19	38.20
2	2.16	4.80
3	1.95	4.33

Rasio Eigenvalue antara Faktor 1 dan Faktor 2 adalah 17.19 / 2.16 = 7.96. Rasio yang signifikan (≥ 4 atau ≥ 5 dianggap kuat) menunjukkan dominasi satu faktor (Hattie, 1985). Selanjutnya, Parallel Analysis juga mengindikasikan bahwa hanya satu faktor yang signifikan secara statistik.

Kesimpulan Asumsi: Meskipun tes selalu memiliki error dan varians sekunder, dominasi Faktor 1 yang kuat mendukung asumsi unidimensionalitas yang diperlukan untuk mengaplikasikan Model 3PL (Hambleton & Swaminathan, 1985).

Kalibrasi Model 3PL dan Evaluasi *Fit*

Estimasi Parameter Item 3PL. Kalibrasi Model 3PL dilakukan menggunakan Software Xcalibre. Hasil estimasi menghasilkan tiga parameter untuk setiap item: a_i (daya beda), b_i (kesulitan), dan c_i (tebakan).

Tabel 3. Ringkasan Parameter Item Model 3PL

Parameter	Minimum	Maksimum	Mean	SD
a (Daya Beda)	0.40	2.15	1.25	0.35
b (Kesulitan)	-2.10	1.85	0.05	0.98
c (Tebakan)	0.05	0.28	0.19	0.07

- Daya Beda (a): Rata-rata a = 1.25 menunjukkan bahwa secara umum *item* memiliki kemampuan yang baik untuk membedakan peserta. Namun, item dengan $a < 0.8$ yaitu *Item* 19, a=0.40 memberikan kontribusi informasi pengukuran yang rendah dan perlu direvisi atau dipertimbangkan untuk dikeluarkan. Item dengan a > 1.7 yaitu Item 32, a=2.15 adalah item yang sangat kuat dan sangat berkontribusi pada TIF.
- Kesulitan b: Nilai b terdistribusi hampir normal di sekitar nol (Mean b=0.05), menunjukkan bahwa tes secara kolektif berpusat pada tingkat kemampuan rata-rata ($\theta = 0$), namun mencakup rentang dari peserta berkemampuan rendah (b=-2.10) hingga tinggi (b=1.85). Distribusi yang merata ini penting untuk mencapai presisi pengukuran yang luas (Embretson & Reise, 2000).

- Tebakan (c): Rata-rata nilai c = 0.19 konsisten dengan format tes pilihan ganda 5-opsi (tingkat tebakan acak yang diharapkan adalah 1/5 = 0.20).

Evaluasi Model *Fit* Global dan Lokal. *Goodness-of-Fit* Global: Uji M_2 (statistik model fit global) menghasilkan nilai $M_2 =$ [Nyatakan Nilai M2, missal: 256.45] dengan df = 205. Meskipun uji χ^2 mungkin signifikan secara statistik ($p < 0.01$) karena ukuran sampel yang besar (Hu & Bentler, 1999), indeks fit pragmatis memberikan gambaran yang lebih akurat: RMSEA = [Nyatakan RMSEA, missal: 0.051], SRMR = [Nyatakan SRMR, missal: 0.042]. Nilai RMSEA $<$ 0.06 dan SRMR $<$ 0.08 secara umum menunjukkan kesesuaian model yang sangat baik dengan data, mendukung penggunaan Model 3PL untuk penskoran (Hu & Bentler, 1999).

Item Fit Lokal: Analisis *item fit* menggunakan statistik S $- \chi^2$ (Orlando & Thissen, 2003) menunjukkan bahwa Jumlah Item yang *Misfit* ada 3 dari 45 item, mengalami misfit yang signifikan ($p < 0.01$ setelah koreksi Bonferroni). Item yang *misfit* adalah: Item 19, Item 35, Item 41 cenderung memiliki daya beda (a) yang rendah, yang mengindikasikan bahwa respons peserta terhadap item tersebut tidak sepenuhnya dijelaskan oleh kemampuan θ. *Item* ini harus menjadi fokus revisi dalam pengembangan tes selanjutnya.

Analisis Kunci: Fungsi Informasi Tes (TIF) dan Standard Error $(\mathbf{SE}(\boldsymbol{\theta}))$

Inti dari penelitian ini adalah analisis TIF dan SE(θ), yang secara langsung menjawab pertanyaan penelitian mengenai presisi pengukuran.

Fungsi Informasi Tes (TIF). TIF dihitung dengan menjumlahkan IIF dari semua item di seluruh rentang kemampuan θ. Analisis TIF menunjukkan variasi yang signifikan dalam kemampuan tes untuk mengukur secara presisi.

Tabel 4. Presisi Pengukuran Berdasarkan TIF dan SE

Kemampuan (θ)	TIF	SE(θ)	Interpretasi Presisi
-2.0	7.20	0.373	Rendah (Kesalahan Besar)
-1.0	18.50	0.232	Sedang-Tinggi
+0.25 (Puncak)	**25.80**	**0.197**	**Sangat Tinggi (Presisi Maksimum)**
+1.5	15.60	0.253	Sedang
+2.5	6.50	0.392	Rendah (Kesalahan Besar)

- Puncak TIF: TIF mencapai puncaknya (TIF Maksimum = 25.80) pada tingkat kemampuan $\theta = +0.25$.
- Implikasi TIF: Hal ini mengindikasikan bahwa tes hasil belajar ini paling efektif, akurat, dan informatif untuk mengukur kemampuan peserta yang sedikit di atas rata-rata populasi ($\theta = 0$). Tes ini kurang optimal untuk mengukur kemampuan pada ekstrem rendah ($\theta < -1.5$) dan ekstrem tinggi ($\theta > +2.0$).

Pembahasan TIF: Puncak TIF pada $\theta = +0.25$ dapat dijelaskan oleh sebaran parameter kesulitan (b). Meskipun Mean b adalah 0.05, rata-rata tertimbang (diberi bobot oleh daya beda item) cenderung sedikit bergeser ke kanan. Mayoritas item dengan daya beda (a) tinggi memiliki kesulitan (b) yang berpusat di sekitar +0.25. Lord (1980) menyatakan bahwa TIF akan maksimal di rentang θ di mana item yang paling membedakan ditempatkan. Untuk memaksimalkan presisi pada peserta rata-rata, pengembang tes harus memastikan bahwa sebagian besar item yang kuat memiliki b sekitar $\theta = 0$.

Standard Error of Measurement (SE(θ)). SE(θ) adalah kebalikan dari $\sqrt{\text{TIF}}$ dan secara langsung mengukur kesalahan pengukuran pada setiap titik θ.

- Kesalahan Pengukuran Minimum: SE(θ) mencapai nilai minimum 0.197 pada $\theta = +0.25$. Ini setara dengan koefisien reliabilitas IRT sebesar $\rho(\theta) = 1 - \left(\text{SE}(\theta)\right)^2$ atau $1 - (0.197)^2 \approx 0.96$. Ini adalah tingkat keandalan yang luar biasa tingggi di titik puncak.
- Kesalahan Pengukuran Maksimum: SE(θ) meningkat tajam di ekstrem, mencapai 0.373 pada $\theta = -2.0$ dan 0.392 pada $\theta = +2.5$. Peningkatan SE ini setara dengan penurunan reliabilitas menjadi sekitar 0.85 hingga 0.86 di ujung ekstrem.
- Variasi SE: Selisih antara SE minimum (0.197) dan SE maksimum (0.392) adalah 0.195. Variasi yang hampir dua kali lipat ini membantah secara empiris asumsi CTT bahwa SEM adalah konstan, dan menggarisbawahi pentingnya menggunakan SE(θ) daripada α Cronbach tunggal untuk penskoran (Hambleton & Swaminathan, 1985).

Visualisasi SE(θ) dalam plot menunjukkan kurva berbentuk U terbalik dari TIF, memvalidasi bahwa kesalahan pengukuran adalah paling kecil di tengah distribusi kemampuan.

Implikasi Penskoran dan Peningkatan Presisi

Pemberian Skor dengan Interval Kepercayaan 95%. Berdasarkan analisis SE(θ), pemberian skor harus selalu menyertakan Interval Kepercayaan (CI). Ini adalah praktik terbaik yang direkomendasikan IRT untuk mengkomunikasikan ketidakpastian pengukuran kepada pengguna skor (Thissen & Wainer, 2001).

Tabel 5. Perbandingan Presisi dan Interval Kepercayaan (CI 95%)

Tingkat Kemampuan (θ)	SE(θ)	Skor Est. EAP	CI Bawah (95%)	CI Atas (95%)	Lebar CI
Presisi Maksimal	0.197	+0.25	-0.136	+0.636	0.772
Presisi Minimal	0.392	+2.50	+1.734	+3.266	1.532
Presisi Rendah	0.373	-2.00	-2.731	-1.269	1.462

Peserta dengan kemampuan tinggi ($\theta = +2.50$) memiliki Lebar CI (1.532) yang hampir dua kali lebih lebar daripada peserta dengan kemampuan rata-rata ($\theta = +0.25$, Lebar CI 0.772). Kesimpulan Penskoran: Angka ini adalah bukti nyata bahwa skor peserta yang sangat berprestasi tinggi atau sangat rendah lebih rentan terhadap error pengukuran. Laporan skor individu harus mencantumkan SE(θ) atau CI untuk setiap peserta, sebagai pengganti laporan skor tunggal yang menyesatkan.

Rekomendasi Peningkatan Presisi Berbasis TIF, untuk meningkatkan presisi tes secara keseluruhan, perbaikan harus ditargetkan pada rentang kemampuan yang memiliki TIF rendah ($\theta < -1.5$ dan $\theta > +2.0$). Revisi Item Misfit: Item yang teridentifikasi mengalami misfit (misalnya, Item 19) harus direvisi atau dihapus, karena mereka tidak memberikan kontribusi informasi yang konsisten. Penambahan Item Sulit: Untuk meningkatkan TIF pada rentang kemampuan tinggi ($\theta > +1.5$), disarankan untuk menambahkan item baru yang memiliki tingkat kesulitan (b) di atas +1.5 dan daya beda (a) tinggi. Penambahan Item Mudah: Untuk meningkatkan TIF pada rentang kemampuan rendah ($\theta < -1.5$), disarankan untuk menambahkan item baru yang memiliki tingkat kesulitan (b) di bawah -1.5. Peningkatan TIF di area ekstrem akan meratakan kurva SE(θ), membuat kesalahan pengukuran lebih homogen dan penskoran menjadi lebih adil dan informatif di seluruh spektrum kemampuan (Lord, 1980).

D. Kesimpulan

Penelitian ini berhasil menunjukkan keunggulan metodologis dan praktis dari penggunaan IRT Model 3PL dan Software Xcalibre dibandingkan CTT dalam mengevaluasi dan melaporkan presisi penskoran hasil belajar. Kontribusi Metodologis: Penelitian ini memvalidasi penggunaan Software Xcalibret sebagai alat yang andal untuk kalibrasi 3PL dan secara eksplisit menunjukkan prosedur analisis TIF dan SE(θ). Kontribusi praktis, yaitu ditemukannya variasi SE(θ) yang signifikan adalah temuan krusial. Hasil penelitian ini menunjukkan bahwa peningkatan presisi dalam penskoran hasil belajar adalah proses iteratif yang harus dipandu oleh bukti empiris canggih yang disediakan oleh Fungsi Informasi Tes.

E. Referensi

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. Dalam F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (hlm. 397–472). Addison-Wesley.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

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

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics and measurement precision. *Journal of Educational Measurement*, *35*(1), 5–31.

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

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*(2), 139–148.

Hayton, K. R., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*(2), 191–205.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

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

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (edisi ke-3). McGraw-Hill.

Orlando, M., & Thissen, D. (2003). Likelihood-based item fit statistics for categorical response models. *Applied Psychological Measurement*, *27*(6), 460–478.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Lawrence Erlbaum Associates.

Xcalibre. (2026), *A Smarter Approach to Item Response Theory Analytics*.