JKPI: Jurnal Kajian Pendidikan IPA
Program Studi Pendidikan IPA, Universitas Garut

p-ISSN 2798-5636
e-ISSN 2798-7043
Vol. 6 No. 1 Tahun 2026

Rasch Model Dalam Evaluasi Instrumen *Extraneous Cognitive Load* Pada Media Pembelajaran Berbantuan *Artificial Intelligence*

Iffa Ichwani Putri [a, 1*], Nurkhairo Hidayati [a, 2], Ummi Kalsum [a, 3], Suryanti [a, 4], Nurul Fauziah [a, 5], Sepita Ferazona [a, 6], Andri Hendrizal [b, 7], Wiwit Yuli Lestari [c, 8]

[a] Pendidikan Biologi, FKIP, Universitas Islam Riau, Pekanbaru, Riau, 28284
[b] Manajemen Sumber Daya Perairan, FPK, Universitas Riau, 28293
[c] Pendidikan Ilmu Pengetahuan Alam, Universitas Garut, 44151
[1] iffa.ichwani@edu.uir.ac.id *; [2] khairobio@edu.uir.ac.id; [3] ummi.kalsum@edu.uir.ac.id; [4] yantibio@uir.ac.id; [5] fauziahnurul@edu.uir.ac.id; [6] sepitabio@edu.uir.ac.id; [7] andri.h@lecturer.unri.ac.id; [8] wiwit@uniga.ac.id
*korespondensi penulis

ARTICLE HISTORY
Received: 30 November 2025 Revised: 31 December 2025 Accepted: 17 January 2026

ABSTRAK

Pengukuran *extraneous cognitive load* (ECL) penting untuk mengevaluasi kualitas desain pembelajaran digital, namun instrumen *self-report* perlu diuji kualitas psikometriknya. Penelitian ini bertujuan mengevaluasi instrumen ECL mahasiswa dengan menggunakan model Rasch. Data dikumpulkan melalui survei pada 38 mahasiswa menggunakan kuesioner ECL berisi 17 butir dengan skala Likert 1–5, kemudian dianalisis dengan *Rasch Rating Scale Model* untuk menilai reliabilitas–separasi, *fit item* dan *person*, serta targeting person–item *melalui Wright map*, Hasil menunjukkan pemetaan *person–item* pada skala logit dapat dilakukan, namun *targeting* kurang optimal karena rerata person jauh di bawah rerata item, mengindikasikan dominasi ECL rendah dan keterbatasan cakupan item pada rentang rendah. Diagnostik kategori memperlihatkan distribusi respons sangat timpang (kategori 4–5 hampir tidak digunakan), dan kurva probabilitas kategori yang tidak membentuk puncak jelas pada kategori atas, sehingga fungsi skala 5 kategori belum efektif. Disimpulkan bahwa instrumen dapat digunakan sebagai pengukuran awal ECL, namun perlu perbaikan pada beberapa item dan penambahan item untuk meningkatkan presisi serta interpretabilitas.

Kata kunci : AI, Biologi, ECL, Media Pembelajaran, Teknologi.

ABSTRACT

Rasch Model in Extraneous Cognitive Load Instrument Evaluation on Artificial Intelligence-Assisted Learning Media. Extraneous cognitive load (ECL) measurements are important to evaluate the quality of digital learning design, but self-report instruments need to be tested for psychometric quality. This study aims to evaluate students' ECL instruments using the Rasch model. Data was collected through a survey of 38 students using an ECL questionnaire containing 17 items with a Likert scale of 1–5, then analyzed with the Rasch Rating Scale Model to assess reliability-separation, fit items and persons, as well as targeting of persons through the Wright map, The results showed that mapping of person-items on a logit scale could be done, but the targeting was less optimal because the average person was far below the average of items. indicates low ECL dominance and limited item coverage at low range. Category diagnostics show a very uneven distribution of responses (categories 4–5 are hardly used), and the probability curve of categories that do not form a clear peak in the upper categories, so the 5 category scale function is not yet effective. It was concluded that the instrument could be used as an initial measurement of ECL, but it needed to be improved on some items and the addition of items to improve precision and interpretability.

Keywords: AI, Biology, ECL, Learning Media, Technology.

Pendahuluan

Pembelajaran di pendidikan tinggi semakin terintegrasi dengan lingkungan digital atau media pembelajaran digital. Diantaranya adalah *Learning Management System* (LMS), video

pembelajaran, modul interaktif, hingga media berbasis presentasi. Transformasi ini membuka peluang peningkatan kualitas belajar, namun juga menghadirkan risiko meningkatnya beban kognitif yang tidak relevan akibat desain penyajian informasi yang kurang efisien (Mayer & Moreno, 2003).

Berdasarkan kerangka *Cognitive Load Theory* (CLT), proses belajar sangat dipengaruhi oleh keterbatasan kapasitas memori kerja. Ketika tuntutan pemrosesan informasi melampaui kapasitas memori kerja, pembelajaran menjadi tidak efektif karena sumber daya kognitif tersita untuk aktivitas yang tidak berkontribusi pada pemahaman (Sweller, 1988; Sweller et al., 2019). Sehingga evaluasi beban kognitif menjadi penting sebagai dasar perbaikan desain pembelajaran, terutama pada konteks multimedia dan e-learning yang memadukan banyak representasi (teks, gambar, audio, navigasi antarmuka).

CLT membedakan beban kognitif ke dalam beberapa komponen. Secara umum, beban kognitif berkaitan dengan kompleksitas materi (*Intrinsic Cognitive Load/ICL*) dan kualitas desain instruksional (*Extraneous Cognitive Load/ECL*), sementara konstruksi skema pengetahuan sering dibahas sebagai *Germane Cognitive Load* dalam literatur CLT klasik (Sweller, 2011, 2018, 2020). Di antara komponen tersebut, ECL memiliki relevansi praktis yang kuat karena ECL bersumber dari karakteristik presentasi dan aktivitas belajar yang tidak secara langsung mendukung tujuan belajar, misalnya informasi redundan, tata letak yang memicu split-attention, atau kebutuhan berpindah-pindah sumber secara tidak perlu.

ECL merupakan indikator sensitif terhadap kualitas desain instruksional pembelajaran. Penelitian pendidikan membutuhkan instrumen yang mampu mengukur ECL secara akurat dan stabil. Dalam perspektif pembelajaran multimedia, strategi untuk menurunkan ECL antara lain meliputi pengurangan informasi yang tidak relevan, integrasi sumber (menghindari split-attention), serta desain yang meminimalkan beban pemrosesan non-esensial sehingga kapasitas memori kerja dapat diarahkan pada pemahaman (Mayer & Moreno, 2003). Pengukuran beban kognitif secara luas mengandalkan self-report karena mudah diterapkan pada konteks kelas, efisien untuk sampel besar, dan dapat dikaitkan dengan desain pembelajaran tertentu (Ayres, 2017; Korbach et al., 2018; Leeuwen et al., 2015; Mavilidi et al., 2021; Sweller et al., 2019).

Model Rasch merupakan pendekatan pengukuran instrumen yang relevan untuk evaluasi instrumen ECL pada media pembelajaran. Model Rasch memposisikan responden dan butir instrumen pada skala yang sama (logit) melalui pemodelan probabilistik, sehingga memungkinkan pemeriksaan yang lebih mendalam. Analisis dengan model ini menelaah item bekerja konsisten dengan konstruk pengukuran, responden menunjukkan pola jawaban yang menyimpang (*person misfit*), dan menilai instrumen menarget sebaran responden secara memadai (Putri et al., 2024; Sumintono, 2018; Zhang et al., 2023). Keunggulan penting lainnya adalah Rasch menyediakan dasar untuk mengubah skor ordinal (*Likert*) menjadi ukuran interval, yang lebih tepat untuk analisis komparatif dan pemodelan lanjutan jika asumsi terpenuhi (Sumintono, 2016, 2017, 2018).

Instrumen ECL umumnya menggunakan kategori berurutan, evaluasi Rasch yang relevan adalah model polytomous dan pendekatan *rating scale analysis*. Dalam kerangka ini, kualitas kategori skala dinilai melalui beberapa indikator: kategori harus digunakan cukup sering, rata-rata ukuran kategori meningkat seiring naiknya kategori, fit kategori tidak ekstrem, dan ambang kategori (*thresholds*) berurutan secara logis (Putri et al., 2024; Sumintono, 2018;

Zhang et al., 2023). Efektivitas kategori bergantung pada keteraturan *threshold* dan penggunaan kategori yang memadai; jika *threshold* tidak berurutan atau kategori sangat jarang digunakan, maka kategori tersebut tidak memberikan informasi pengukuran yang berarti dan interpretasi respon menjadi lemah. Kondisi ini berdampak langsung pada kualitas estimasi *person*, ketelitian pengukuran, serta kemampuan instrumen membedakan tingkat ECL antar responden.

Selain fungsi kategori, Rasch juga menekankan pentingnya targeting instrumen. Melalui *person–item* map (*wright map*), peneliti dapat menilai apakah sebaran kesulitan item menutupi rentang trait responden (Boone et al., 2014). Pada ECL, pengukuran ideal adalah instrumen yang memiliki butir yang merepresentasikan ECL rendah hingga tinggi. Tanpa targeting yang memadai, instrumen cenderung tidak sensitif pada rentang tertentu, misalnya ketika sebagian besar mahasiswa berada pada ECL rendah tetapi item-item yang tersedia lebih banyak mengukur ECL sedang–tinggi, akibatnya pengukuran pada rentang rendah menjadi kurang presisi dan muncul gejala *floor effect*.

Berdasarkan urgensi substantif (ECL sebagai indikator kualitas desain instruksional) dan kebutuhan metodologis (jaminan fungsi kategori, fit, dan targeting), artikel ini bertujuan mengevaluasi kualitas psikometrik instrumen pengukuran *Extraneous Cognitive Load* (ECL) pada media pembelajaran berbantuan AI mahasiswa menggunakan model Rasch. Secara khusus, evaluasi difokuskan pada: kesesuaian respons person dan item terhadap model (*person-fit* dan *item-fit*), kesesuaian targeting person–item melalui Wright map, dan diagnostik kategori skala rating serta kurva probabilitas kategori. Dengan pendekatan ini, hasil penelitian diharapkan memberikan dasar yang lebih kuat untuk menyimpulkan kelayakan instrumen ECL pada media pembelajaran berbantuan AI, agar instrumen menjadi lebih valid untuk evaluasi pembelajaran biologi dengan memanfaatkan media pembelajaran-AI di pendidikan tinggi.

Metode

Penelitian ini menggunakan desain kuantitatif–survei dengan tujuan utama evaluasi psikometrik instrumen (validasi empiris) menggunakan model Rasch. Model Rasch dipilih karena menyediakan kerangka fundamental measurement yang memodelkan hubungan probabilistik antara respons responden dan kesulitan item pada satu skala logit, sehingga memungkinkan evaluasi ketepatan *item/person* (fit), reliabilitas–separasi, targeting (kesesuaian sebaran item–responden), serta fungsi kategori skala rating (Sumintono, 2018).

Partisipan penelitian berjumlah 38 mahasiswa, yang seluruhnya mengisi instrumen ECL. Unit analisis adalah respons individu terhadap butir (*person-by-item*). Jumlah butir yang dianalisis adalah 17 item, sehingga total observasi respons berjumlah 646 (38×17). Instrumen yang dievaluasi adalah kuesioner *Extraneous Cognitive Load* (ECL) yang terdiri dari 17 butir dengan skala respons Likert 5 kategori (1–5). Skor yang lebih tinggi merepresentasikan kecenderungan ECL yang lebih tinggi (misalnya semakin banyak gangguan desain/format instruksi yang tidak relevan dengan tujuan belajar). Pernyataan pada instrumen yang digunakan untuk pengukuran ECL pada media pembelajaran-AI dapat dilihat pada Tabel 1. Seluruh data respons dikompilasi dalam satu set data untuk memudahkan proses pengolahan dan analisis selanjutnya.

Tabel 1. Uraian Pernyataan Pengukuran ECL pada Media Pembelajaran-AI

No	Pernyataan
1.	Multimedia AI yang digunakan membantu dalam memahami materi perkuliahan.
2.	Penggunaan media pembelajaran AI membantu dalam pemahaman konten materi yang rumit.
3.	Media yang digunakan dalam perkuliahan memudahkan memahami istilah baru dalam materi perkuliahan.
4.	Latihan soal yang diberikan membantu memahami materi dengan dengan baik.
5.	Strategi penyampaian materi dengan multimedia AI meningkatkan keinginan mempelajari materi perkuliahan.
6.	Video animasi pada multimedia AI membantu dalam pemahaman mekanisme yang terjadi dalam konsep materi perkuliahan.
7.	Kegiatan diskusi membantu memahami konsep materi perkuliahan.
8.	Instruksi atau arahan pada perkuliahan dapat diikuti dengan mudah dalam kegiatan perkuliahan.
9.	Latihan yang diberikan memotivasi untuk memahami berbagai konsep dalam materi perkuliahan.
10.	Tahapan kegiatan dapat memudahkan pemahaman sesuai dengan waktu yang telah ditentukan.
11.	Aktivitas belajar membantu mengembangkan pemahaman terkait konsep materi perkuliahan.
12.	Media yang digunakan dalam perkuliahan membantu dalam memahami konsep yang sulit.
13.	Diskusi membantu dengan lebih baik untuk memahami materi yang sulit.
14.	Jumlah materi yang diberikan sesuai dengan waktu pada setiap pertemuan.
15.	Media interaktif yang digunakan sangat membantu dalam penjelasan materi perkuliahan.
16.	Aplikasi media menarik dan memotivasi untuk mengikuti perkuliahan dengan antusias.
17.	Media yang digunakan dalam perkuliahan genetika membantu dalam memahami konsep materi yang detail.

Kuesioner diberikan kepada mahasiswa setelah mereka mengikuti aktivitas pembelajaran pada konteks perkuliahan yang relevan dengan penggunaan media AI. Seluruh respons dicatat dalam format matriks *person–item* dan diekspor untuk analisis Rasch. Data diperiksa untuk memastikan tidak terjadi kesalahan pengodean kategori dan konsistensi arah skoring.

Analisis Rasch dilakukan menggunakan *software Winsteps*. Hasil yang dievaluasi meliputi:

1. Statistik ringkasan *person* dan *item*
   Ringkasan mencakup *mean* dan sebaran *measure* (logit), *standard error*, serta indeks *separation* dan *reliability* pada level *person* dan *item*, sebagaimana standar pelaporan di analisis Rasch.
2. Kesesuaian item dan person terhadap model (fit)
   Kesesuaian dievaluasi menggunakan Infit MNSQ dan Outfit MNSQ beserta ZSTD untuk mengidentifikasi pola respons yang terlalu acak (*underfit*) atau terlalu terprediksi (*overfit*). Interpretasi dan diagnosis misfit mengikuti panduan Winsteps dan literatur Rasch terapan, termasuk kehati-hatian pada interpretasi ZSTD yang sensitif terhadap ukuran sampel/observasi. Dalam studi ini, nilai MNSQ dipakai sebagai indikator utama karena bersifat lebih stabil secara praktis untuk keputusan revisi instrumen dibanding ZSTD pada ukuran sampel tertentu (Winsteps misfit diagnosis).

3. Evaluasi targeting melalui *Wright map* (*Person–Item Map*)
   Peta person–item digunakan untuk menilai apakah tingkat kesulitan item menutupi rentang trait responden (targeting), serta untuk mengidentifikasi gap rentang logit yang minim item.
4. Scalogram/Guttman scalogram
   Scalogram digunakan sebagai diagnostik tambahan untuk mengidentifikasi pola respons yang tidak konsisten (*aberrant response pattern*) pada level person, yang biasanya terkait dengan misfit person.

Hasil dan Pembahasan

Pembelajaran berbantuan media AI dirancang untuk menyajikan materi pada perkuliahan dengan cara yang lebih interaktif, visual, dan terstruktur. Mahasiswa berinteraksi dengan berbagai fitur seperti multimedia, video animasi, latihan soal, diskusi, serta instruksi yang jelas. Setiap komponen ini berfungsi untuk mengurangi beban kognitif yang tidak relevan (ECL), sehingga mahasiswa dapat fokus pada pemahaman konsep inti.

Melalui pengalaman belajar, mahasiswa diminta menilai sejauh mana media AI membantu mahasiswa memahami materi, mengurangi kesulitan, dan meningkatkan motivasi. Respon kuesioner yang terdiri dari 17 pernyataan mencerminkan persepsi mahasiswa terhadap efektivitas media AI pada pembelajaran. Respon mahasiswa pada kuesioner akan menunjukkan apakah media AI benar-benar berfungsi sebagai desain instruksional yang berkualitas dengan beban kognitif yang lebih efisien

Tabel 2. Ringkasan *Person*/Responden

Statistik	Total Score	Count	Measure (logit)	Model S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
Mean	30.7	17.0	-2.93	0.58	—	—	—	—
SEM	0.9	0.0	0.29	0.04	—	—	—	—
P.SD	5.8	0.0	1.75	0.23	—	—	—	—
S.SD	5.8	0.0	1.77	0.23	—	—	—	—
Max	45.0	17.0	0.43	1.84	—	—	—	—
Min	17.0	17.0	-8.44	0.34	—	—	—	—

Komponen	RMSE	True SD	Separation	Person Reliability
Real	0.69	1.61	2.34	0.85
Model	0.63	1.63	2.61	0.87

Indikator	Nilai	Catatan
Person raw score-to-measure correlation	0.98	"approximate due to missing data"
Cronbach Alpha (KR-20) – Person raw score "test" reliability	0.89	SEM = 1.92
Standardized (50 item) reliability	0.95	Berdasarkan standarisasi panjang tes 50 item

Model Rasch digunakan untuk mengubah skor mentah ordinal menjadi ukuran interval dalam satuan logit, sehingga kemampuan responden dan kesulitan butir dapat ditempatkan pada skala yang sama dan dianalisis secara lebih tepat dibanding pendekatan klasik. Analisis berdasarkan Tabel Ringkasan mencakup mean dan sebaran *measure* (logit), *standard error*, serta indeks *separation* dan *reliability* pada level person dan item, sebagaimana standar pelaporan di analisis Rasch (Tabel 2).

Tabel 2 menggambarkan distribusi kemampuan responden, besarnya kesalahan pengukuran, serta daya pembeda instrumen terhadap variasi kemampuan tersebut. Rata-rata skor total responden adalah 30,7 dari 17 butir, dengan rata-rata ukuran kemampuan sebesar −2,93 logit dan simpangan baku sekitar 1,75 logit. Nilai rata-rata yang berada jauh di bawah titik nol logit (rata-rata kesulitan butir) menunjukkan bahwa secara keseluruhan kemampuan responden berada di bawah tingkat kesulitan instrumen, sehingga tes cenderung relatif sulit bagi sampel ini.

Rentang ukuran kemampuan cukup lebar, dari −8,44 hingga 4,13 logit, yang mengindikasikan heterogenitas kemampuan responden dan memberikan dasar bahwa instrumen berpotensi membedakan beberapa tingkat kemampuan yang berbeda. Jika dilihat dari sisi ketepatan pengukuran, nilai *standard error* rata-rata sekitar 0,58 logit dan standar error rata-rata person (REAL RMSE) 0,69 logit menunjukkan bahwa estimasi kemampuan individu dilakukan dengan tingkat ketelitian yang memadai; semakin kecil nilai kesalahan pengukuran, semakin presisi estimasi kemampuan pada skala logit.

Nilai separation person sebesar 2,34 (berbasis Real RMSE) dan 2,61 (berbasis Model RMSE) menunjukkan bahwa instrumen mampu mengelompokkan responden menjadi kurang lebih 3–4 strata kemampuan yang berbeda, dihitung dengan rumus strata = (4×separation + 1)/3. Menurut pedoman Rasch, separation di atas 2 mengindikasikan bahwa alat ukur cukup sensitif untuk membedakan kelompok responden dengan kemampuan tinggi dan rendah.

Indeks reliabilitas person yang diperoleh sebesar 0,85 (Real) dan 0,87 (Model) berada di atas batas minimal 0,80 yang umumnya digunakan sebagai kriteria reliabilitas yang baik pada analisis Rasch. Hal ini berarti konsistensi jawaban responden pada instrumen ini tinggi dan perbedaan kemampuan yang terukur lebih banyak disebabkan oleh perbedaan sejati antar responden dibandingkan oleh kesalahan pengukuran.

Koefisien Cronbach Alpha (KR-20) sebesar 0,89 dengan SEM 1,92, serta reliabilitas terstandar (50 butir) sebesar 0,95. Nilai alpha di atas 0,70 secara umum diinterpretasikan sebagai menunjukkan konsistensi internal yang baik, dan nilai mendekati 0,90 menggambarkan reliabilitas yang sangat tinggi. Konsistensi antara reliabilitas Rasch dan Cronbach Alpha menguatkan temuan bahwa instrumen memiliki stabilitas pengukuran yang baik dan layak digunakan untuk tujuan penelitian maupun evaluasi.

Berdasarkan keseluruhan indikator pada tabel, dapat diketahui bahwa instrumen yang dianalisis dengan model Rasch memiliki kualitas psikometrik yang baik dari sisi reliabilitas dan kemampuan membedakan beberapa tingkatan kemampuan responden.

Beberapa studi pengembangan instrumen menggunakan Rasch menempatkan reliabilitas sekitar 0,80–0,90 sebagai indikator bahwa instrumen cukup stabil untuk keperluan penelitian dan evaluasi program, termasuk pada skala sikap, kecemasan, maupun literasi sains. Tingginya korelasi antara skor mentah dan ukuran Rasch (0,98) menunjukkan bahwa skoring Likert yang digunakan cukup linear terhadap trait ECL yang diukur, sesuai dengan sifat skor total sebagai *sufficient statistic* dalam model Rasch (Yasin et al., 2018).

Di sisi lain, koefisien Cronbach Alpha (KR-20) sebesar 0,89 menguatkan bukti bahwa instrumen ECL memiliki konsistensi internal yang tinggi. Aturan praktis banyak teks pengukuran menyebutkan bahwa alpha ≥ 0,80 umumnya dipandang menunjukkan reliabilitas yang baik, sementara nilai mendekati 0,90 mengindikasikan homogenitas butir yang kuat

terhadap konstruk yang sama. Studi penggunaan model Rasch di berbagai konteks pendidikan menunjukkan pola serupa, yaitu bahwa skala dengan Cronbach Alpha > 0,80 dan person reliability > 0,80 pada umumnya dipandang cukup andal untuk mengukur konstruk laten seperti kecemasan statistik, kemampuan pemecahan masalah, atau literasi (Mokhsein & Akhmad, 2019).

Kesesuaian antara reliabilitas Rasch dan *Cronbach Alpha* pada instrumen ini menunjukkan bahwa interaksi mahasiswa–butir ECL berlangsung secara konsisten, sehingga skor yang dihasilkan dapat dipercaya sebagai refleksi stabil dari tingkat ECL yang dialami. Dari perspektif penelitian cognitive load, hasil ini menempatkan instrumen ECL yang dianalisis sejajar dengan skala-skala beban kognitif lain yang telah divalidasi secara luas.

Secara keseluruhan, analisis Rasch atas instrumen ECL mahasiswa menunjukkan bahwa skala memiliki reliabilitas tinggi, pemisahan responden yang memadai, dan struktur kemampuan yang cukup menyebar untuk membedakan beberapa level ECL. Hal ini memberikan dasar psikometrik yang kuat bagi penggunaan instrumen tersebut dalam riset desain pembelajaran maupun evaluasi intervensi pengurangan ECL, misalnya melalui perbaikan desain materi, penyederhanaan tampilan, atau pengurangan redundansi informasi—sebagaimana direkomendasikan dalam literatur *Cognitive Load Theory*.

Kesesuaian item dan person terhadap model (fit) dievaluasi menggunakan Infit MNSQ dan Outfit MNSQ beserta ZSTD untuk mengidentifikasi pola respons yang terlalu acak (underfit) atau terlalu terprediksi (overfit) (Tabel 3).

Tabel 3. Kesesuaian item dan person terhadap model (fit)

Entry (Item)	Total Score	Total Count	JMLE Measure	Model S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	PTMEA Corr.	PTMEA Exp.	Exact Match OBS%	Exact Match EXP%
10	61	38	1.04	0.37	1.45	1.73	1.56	1.62	0.36	0.56	59.5	74.6
2	62	38	0.90	0.37	0.99	0.05	0.93	-0.11	0.65	0.57	73.0	74.8
8	62	38	0.90	0.37	1.34	1.34	1.40	1.24	0.49	0.57	67.6	74.8
6	66	38	0.35	0.37	1.01	0.14	1.01	0.15	0.71	0.59	75.7	74.7
13	66	38	0.35	0.37	0.75	-0.93	0.68	-1.05	0.56	0.59	75.7	74.7
11	68	38	0.08	0.37	0.91	-0.24	1.28	0.92	0.38	0.59	67.6	74.6
14	68	38	0.08	0.37	0.72	-1.01	0.69	-1.03	0.60	0.59	78.4	74.6
15	68	38	0.08	0.37	0.89	-0.32	0.85	-0.40	0.60	0.59	73.0	74.6
3	69	38	-0.06	0.37	0.83	-0.54	0.82	-0.51	0.67	0.60	81.1	74.5
4	69	38	-0.06	0.37	0.91	-0.21	0.88	-0.28	0.65	0.60	75.7	74.5
16	70	38	-0.19	0.36	1.00	0.11	0.95	-0.04	0.62	0.60	83.8	74.3
17	70	38	-0.19	0.36	0.70	-1.03	0.58	-1.48	0.63	0.60	83.8	74.3
1	71	38	-0.32	0.36	0.63	-1.33	0.55	-1.62	0.70	0.60	86.5	74.1
7	72	38	-0.45	0.36	1.41	1.30	1.31	1.02	0.65	0.60	64.9	74.0
5	73	38	-0.57	0.35	1.79	2.21	2.10	2.79	0.50	0.60	78.4	74.0
9	74	38	-0.69	0.35	0.83	-0.47	0.86	-0.39	0.71	0.60	73.0	73.8
12	79	38	-1.26	0.32	1.05	0.25	0.93	-0.15	0.59	0.61	70.3	70.4
MEAN	**68.7**	**38.0**	**0.00**	**0.36**	**1.01**	**0.06**	**1.02**	**0.04**	—	—	**74.6**	**74.2**
P.SD	**4.5**	**0.0**	**0.58**	**0.01**	**0.30**	**1.00**	**0.39**	**1.12**	—	—	**7.0**	**1.0**

Pada kolom MEASURE, tingkat kesulitan butir berada pada rentang sekitar +1,04 logit (butir 10) sampai −1,26 logit (butir 12), dengan rata-rata 0,00 logit dan simpangan baku 0,58 logit. Distribusi ini menunjukkan bahwa butir-butir ECL tersebar cukup seimbang di sekitar pusat skala, sehingga secara teoritis mampu menangkap variasi ECL dari kondisi relatif tinggi (butir dengan logit positif) sampai relatif rendah (butir dengan logit negatif).

Butir dengan measure positif (misalnya butir 10, 2, 8, 6) dapat diinterpretasikan sebagai pernyataan yang lebih sulit disetujui dan cenderung merepresentasikan situasi ECL yang berat.

Sebaliknya, butir dengan measure negatif (misalnya butir 1, 5, 9, 12) relatif lebih mudah disetujui, menggambarkan situasi ECL yang lebih rendah. Dalam kerangka *Cognitive Load Theory*, variasi kesulitan butir semacam ini penting agar instrumen mampu membedakan pengalaman ECL mahasiswa pada berbagai tingkat kompleksitas desain pembelajaran.

Kolom INFIT dan OUTFIT mean-square (MNSQ) dan ZSTD memberikan informasi kesesuaian butir dengan model Rasch. Secara rata-rata, nilai INFIT MNSQ = 1,01 dan OUTFIT MNSQ = 1,02 dengan simpangan baku masing-masing 0,30 dan 0,39, mendekati nilai ekspektasi 1,0 yang menandakan bahwa pola respons untuk sebagian besar butir tidak menyimpang secara sistematis dari prediksi model dan dalam kategori fit.

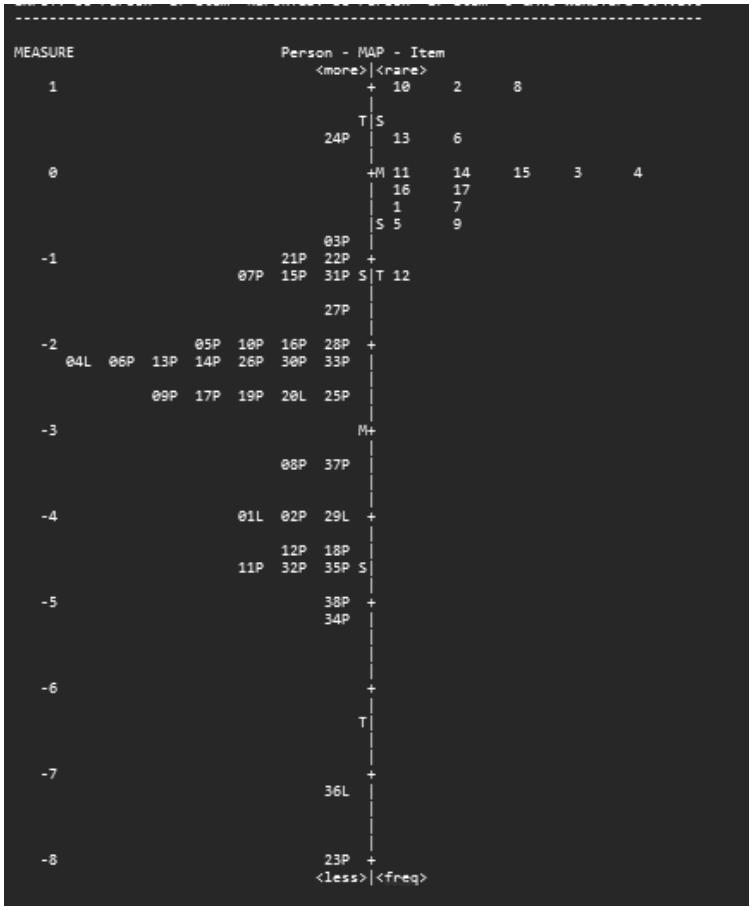
Beberapa butir seperti 1 dan 17 menunjukkan INFIT MNSQ di bawah 0,70 dengan OUTFIT MNSQ sekitar 0,55–0,58 disertai ZSTD negatif. Butir-butir ini dikategorikan sebagai overfit. Menurut Bond & Fox, butir overfit tidak merusak skala, namun cenderung kurang informatif karena tidak menambahkan variasi baru di luar pola umum yang sudah ditangkap butir lain. Dalam konteks skala ECL, butir-butir tersebut mungkin berisi pernyataan yang sangat langsung atau umum sehingga hampir semua responden dengan kecenderungan ECL tertentu merespons dengan pola yang sama.

Korelasi point–measure (PTMEASURE-AL CORR) berkisar antara 0,36 sampai 0,71 dengan rata-rata sekitar 0,59. Semua nilai korelasi positif dan berada di atas 0,30, yang lazim dijadikan batas minimal bahwa suatu butir berkontribusi secara konsisten terhadap konstruk yang sama dengan keseluruhan skala. Butir 10 memiliki korelasi terendah (0,36), konsisten dengan indikasi misfit pada MNSQ, sehingga patut menjadi kandidat utama untuk ditinjau ulang redaksinya. Sebaliknya, butir 1 dan 9 menunjukkan korelasi tinggi (0,70–0,71), menandakan daya diskriminasi yang kuat terhadap perbedaan tingkat ECL antar mahasiswa.

Kolom EXACT MATCH menunjukkan persentase kecocokan persis antara respons yang diamati dan yang diprediksi model. Rata-rata kecocokan persis 74,6% sangat dekat dengan nilai ekspektasi 74,2%, yang berarti secara global model Rasch merepresentasikan pola respons responden dengan baik. Akan tetapi, variasi antarbutir cukup nyata: butir 10 dan 7 memiliki kecocokan observasi (59,5% dan 64,9%) yang jauh di bawah ekspektasi (74%), selaras dengan indikasi misfit pada MNSQ; sebaliknya, butir 1 mencapai kecocokan 86,5% yang lebih tinggi dari ekspektasi 74,1% dan kembali konsisten dengan pola overfit. Secara substantif, misfit pada butir dapat merefleksikan beberapa kemungkinan: redaksi yang ambigu atau terlalu panjang, konten butir yang memuat lebih dari satu aspek sekaligus (misalnya sekaligus menyentuh aspek navigasi dan tampilan visual), atau konteks pengalaman belajar mahasiswa yang sangat heterogen sehingga respons terhadap butir tersebut menjadi tidak stabil (Rahayah et al., 2010; Sumintono, 2018).

Berdasarkan data dapat diketahui bahwa instrumen ECL ini secara umum telah memenuhi kriteria psikometrik Rasch. Distribusi kesulitan cukup seimbang, korelasi point–measure seluruh butir positif dan memadai, serta rata-rata indeks fit dekat dengan nilai ekspektasi. Instrumen ini karenanya layak digunakan untuk mengukur extraneous cognitive load mahasiswa dalam konteks perkuliahan digital atau berbasis multimedia. Namun, butir 10 dan 5 ditinjau ulang melalui revisi redaksi atau uji coba lanjutan, agar struktur skala menjadi lebih fokus pada satu dimensi ECL yang seragam.

Peta Person–Item (Wright map) menempatkan responden (kiri) dan butir (kanan) pada skala logit yang sama, sehingga dapat dievaluasi targeting (korespondensi antara sebaran trait responden dan kesulitan butir), hierarki butir, serta area skala yang miskin informasi (Gambar 1).



Gambar 1. Peta Person–Item (Wright map)

Pada *map* (Gambar 1) tampak rerata person (M) berada jauh di bawah rerata item (item mean secara konvensi dipusatkan di sekitar 0 logit). Ini konsisten dengan ringkasan sebelumnya (mean person ≈ −2,93 logit) dan menunjukkan bahwa, untuk sampel ini, butir-butir ECL relatif lebih sulit disetujui dibanding level trait yang dimiliki mayoritas responden. Secara substantif, ini dapat dibaca sebagai indikasi bahwa mahasiswa cenderung melaporkan ECL yang rendah, atau bahwa konten butir lebih banyak merepresentasikan ECL tingkat sedang–tinggi sehingga kurang menarget pengalaman ECL yang lebih rendah.

Dari sisi hierarki item instrumen, bagian atas map menunjukkan butir 10, 2, dan 8 berada pada logit tertinggi (paling sulit), sehingga butir-butir ini merepresentasikan indikator ECL yang hanya cenderung disetujui oleh responden dengan ECL tinggi. Sebaliknya, butir 12 tampak sebagai butir paling mudah (logit terendah), menjadi jangkar pada sisi ECL rendah. Wright map membantu menguji koherensi teoretik: bila butir yang secara konseptual menggambarkan ECL paling berat memang berada di puncak skala, maka struktur kontinum item mendukung validitas konstruk.

Namun, *map* juga memperlihatkan kesenjangan cakupan (*coverage gaps*) pada rentang logit rendah: banyak responden berkelompok sekitar −2 hingga −4 logit, sementara sebaran

item relatif padat di sekitar 0 hingga −1 logit dan tampak minim butir pada rentang lebih rendah. Konsekuensinya, instrumen berpotensi kurang informatif untuk membedakan mahasiswa dengan ECL rendah–sangat rendah, karena perbedaan antar responden pada area itu tidak diikat oleh butir yang sesuai tingkatannya.

Urutan responden dan item dapat terlihat pada skala Guttman/scalogram of responses (Gambar 2). *Observed* data matrix yang sudah diurutkan: responden tersusun dari yang paling tinggi (baris atas) ke paling rendah (baris bawah), sedangkan item tersusun dari paling mudah diendorse (kolom kiri) ke paling sulit (kolom kanan). Tujuannya untuk melihat sejauh mana respons mendekati pola Guttman (semakin tinggi trait, semakin konsisten memberikan kategori lebih tinggi pada item yang lebih mudah, dan menurun saat item makin sulit).

```
GUTTMAN SCALOGRAM OF RESPONSES:
Person |Item
       |1    11  111 1  1
       |29571673414563280
       |-----------------
    24 +23243333223232323  24P
     3 +32232222322222232  03P
    21 +22332123222222331  21P
    22 +33222322222222222  22P
     7 +23222222322332211  07P
    15 +22223223222222222  15P
    31 +32222322222222222  31P
    27 +33232222221222221  27P
     5 +32222222222222212  05P
    10 +22222222222222222  10P
    16 +22222222222222222  16P
    28 +22222222222222222  28P
     4 +22222222222222212  04L
     6 +23312222222231211  06P
    13 +32222123222132121  13P
    14 +22222222222222122  14P
    26 +32222222212222221  26P
    30 +22222221222222222  30P
    33 +22532222112212112  33P
     9 +22222222222222211  09P
    17 +22112222222222222  17P
    19 +22222222221221222  19P
    20 +22222222122322121  20L
    25 +22222222222212122  25P
     8 +23222321122112111  08P
    37 +22212221222221112  37P
     1 +22212212212111122  01L
     2 +12222222211121211  02P
    29 +21223112211212211  29L
    12 +21231231112111111  12P
    18 +22111212221211121  18P
    11 +32111111121112122  11P
    32 +21221111212112112  32P
    35 +21111111122212212  35P
    38 +11112121121212121  38P
    34 +11111122122111112  34P
    36 +11111111121111111  36L
    23 +11111111111111111  23P
       |-----------------
       |1    11  111 1  1
       |29571673414563280
```

Gambar 2. Scalogram

Pola umum dari gambar 2 menjelaskan bahwa responden paling atas (24P) menunjukkan banyak respons pada kategori 3–4 (bahkan ada 5 pada beberapa responden), sedangkan responden paling bawah (23P) menunjukkan deret 1 hampir di semua item. Ini konsisten dengan kontinum trait yang diukur: semakin ke bawah, kecenderungan mengendorse pernyataan ECL (kategori tinggi) semakin rendah. Namun, yang juga mencolok adalah

dominasi respons kategori 2 pada banyak responden menengah (222222). Secara psikometrik, dominasi kategori tengah seperti ini sering menjadi gejala targeting item yang kurang pas (item relatif terlalu sulit sehingga banyak responden jatuh pada kategori rendah–tengah), dan dapat mengurangi informasi pengukuran pada rentang trait rendah.

Dari perspektif fungsi kategori skala rating, scalogram menunjukkan bahwa kategori ekstrem (khususnya 5) tampaknya jarang digunakan (muncul pada sedikit responden), sedangkan kategori 2 sangat dominan. Dalam model rating scale/polytomous Rasch, kategori yang jarang terpakai atau tidak naik secara teratur dapat mengindikasikan masalah pada struktur kategori (misalnya label kategori tidak dipahami konsisten, atau ambang/threshold tidak terurut) (Yasin et al., 2018).

Nilai utama scalogram adalah menyediakan bukti visual tentang inversi atau Guttman reversals/errors yaitu respons yang muncul tidak sesuai terhadap urutan kesulitan item. Pada baris 33P memperlihatkan pola seperti 255, 21112, yakni kemunculan kategori sangat tinggi (5) pada beberapa posisi, tetapi kemudian jatuh ke kategori rendah dan bercampur tidak stabil. Pola semacam ini adalah ciri klasik aberrant response behavior (misalnya careless responding, misunderstanding pada sebagian butir, atau interaksi person–item yang tidak termodelkan), yang pada output Rasch biasanya juga muncul sebagai person underfit (INFIT/OUTFIT MNSQ tinggi).

Secara keseluruhan, scalogram menunjukkan bahwa terdapat struktur kontinum yang secara umum konsisten (indikasi unidimensionalitas operasional melalui kecenderungan *stochastic Guttman order*), tetapi terdapat keterbatasan targeting pada rentang trait rendah dan sejumlah kecil pola respons menyimpang (*misfit person*) yang perlu ditangani/diargumentasikan. Dengan mengintegrasikan scalogram bersama statistik fit dan *Wright map*, instrumen ECL sudah berfungsi baik secara menyeluruh.

Implikasi utama dari temuan Rasch ini adalah bahwa skor ECL yang dihasilkan instrumen sudah dapat diperlakukan sebagai ukuran pada satu kontinum (logit) sehingga perbandingan antar responden/kelompok secara umum bermakna. Dalam praktik pengukuran, kondisi ini berarti instrumen mampu membedakan mahasiswa dengan ECL relatif tinggi dibanding membedakan mahasiswa pada level rendah–menengah. Prinsip penguatan instrumen melalui analisis Rasch untuk meningkatkan ketepatan konstruksi dan kualitas butir sejalan dengan praktik pengembangan instrumen berbasis Rasch yang menekankan pemetaan butir–responden dan kontrol mutu butir/kategori sebelum instrumen dipakai untuk keputusan yang lebih konsekuensial.

Pada level implikasi pembelajaran digital, hasil ini mendukung CLT. Instrumen dapat menjadi dasar audit desain untuk menurunkan beban non-esensial (*extraneous*) yang berasal dari cara penyajian, navigasi, dan format tugas—bukan dari kompleksitas materi itu sendiri. Ketika ECL terukur tinggi, intervensi desain (mereduksi elemen distraktor, memperjelas alur tugas, mengurangi pencarian informasi yang tidak perlu) menjadi prioritas karena CLT menekankan bahwa ECL terutama dipicu oleh desain instruksional dan upaya reduksinya paling berdampak ketika tuntutan kognitif materi (*intrinsic*) relatif tinggi.

Temuan tentang fungsi kategori skala rating mengimplikasikan bahwa interpretasi skor mentah (Likert) belum sepenuhnya stabil sebagai indikator tingkat ECL. Karena itu, rekonstruksi kategori (menggabungkan kategori yang tidak terpakai/ambigu, memperjelas

label respons, atau menata ulang jumlah kategori) perlu dilakukan dan diverifikasi dengan kriteria fungsi kategori (frekuensi memadai, average measure meningkat, fit kategori memadai, dan *Andrich thresholds* berurutan). Setelah kategori stabil, instrumen akan lebih akurat sebagai dasar keputusan perbaikan desain pembelajaran digital dan evaluasi upaya menurunkan beban non-esensial sesuai CLT.

Simpulan

Berdasarkan analisis Rasch, instrumen telah menunjukkan struktur pengukuran yang secara umum dapat dipetakan pada kontinum logit, namun masih memiliki keterbatasan penting pada aspek targeting dan fungsi kategori skala rating. Secara keseluruhan, instrumen ECL ini dapat digunakan sebagai alat ukur awal untuk mengevaluasi beban kognitif ekstraneous mahasiswa, tetapi peningkatan kualitas pengukuran masih diperlukan. Selanjutnya disarankan untuk menambah atau merevisi beberapa butir yang lebih mudah untuk menutup rentang ECL rendah–sedang sehingga targeting membaik, dan melakukan rekonstruksi skala kategori, fit kategori, dan presisi pengukuran. Penguatan ini penting agar instrumen lebih sensitif sebagai dasar evaluasi desain pembelajaran digital dan upaya menurunkan beban non-esensial sesuai prinsip CLT.

Ucapan terima kasih

Penulis mengucapkan terima kasih kepada dukungan program studi, fakultas dan lembaga Universitas Islam Riau.

Referensi

Ayres, P. (2017). Subjective measures of cognitive load: What can they reliably measure? In *Cognitive Load Measurement and Application: A Theoretical Framework for Meaningful Research and Practice*. https://doi.org/10.4324/9781315296258

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Item Measures. In *Rasch Analysis in the Human Sciences*. https://doi.org/10.1007/978-94-007-6857-4_5

Korbach, A., Brünken, R., & Park, B. (2018). Differentiating Different Types of Cognitive Load: a Comparison of Different Measures. *Educational Psychology Review*, *30*(2), 503–529. https://doi.org/10.1007/s10648-017-9404-8

Leeuwen, A. Van, Janssen, J., Erkens, G., & ... (2015). Teacher regulation of cognitive activities during student collaboration: Effects of learning analytics. ... *& Education*. https://www.sciencedirect.com/science/article/pii/S0360131515300439

Mavilidi, M. F., Ouwehand, K., Schmidt, M., Pesce, C., Tomporowski, P. D., Okely, A., & Paas, F. (2021). Embodiment as a pedagogical tool to enhance learning. In The Body, Embodiment, and Education: An Interdisciplinary Approach (pp. 183-203). Taylor & Francis Ltd (UK). https://doi.org/10.4324/9781003142010-10

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*(1). https://doi.org/10.1207/S15326985EP3801_6

Mokshein, S. E., Ishak, H., & Ahmad, H. (2019). The Use Of Rasch Measurement Model In English Testing. *Jurnal Cakrawala Pendidikan*, *38*(1), 16–32.

https://doi.org/10.21831/cp.v38i1.22750

Putri, I. I., Rahmat, A., & Riza, L. S. (2024). *Assessing Undergraduate Cognitive System Thinking Instruments in Genetics Lectures : A Rasch Model Analysis*. *16*, 3924–3937. https://doi.org/10.35445/alishlah.v16i3.5621

Rahayah, S., Omar, B., & Sharif, S. (2010). *Validity and reliability multiple intelligent item using rasch measurement model*. *9*, 729–733. https://doi.org/10.1016/j.sbspro.2010.12.225

Sumintono, B. (2016). Aplikasi Pemodelan Rasch pada asesmen pendidikan: Implementasi penilaian formatif (assessment for learning). *Makalah Dipresentasikan Dalam Kuliah Umum Pada Jurusan Statistika, Institut Teknologi Sepuluh November, Surabaya, 17 Maret 2016.*, *March*.

Sumintono, B. (2017). *Rasch model measurement as tools in assessment for learning*. eprints.um.edu.my.

Sumintono, B. (2018). *Rasch Model Measurements as Tools in Assessment for Learning*. https://doi.org/10.2991/icei-17.2018.11

Sweller, J. (1988). *Cognitive Load During Problem Solving : Effects on Learning*. *285*, 257–285. https://doi.org/10.1207/s15516709cog1202

Sweller, J. (2011). *Cognitive Load Theory and E-Learning*. https://doi.org/10.1007/978-3-642-21869-9_3

Sweller, J. (2018). Measuring cognitive load. *Perspectives on Medical Education*, *7*(1). https://doi.org/10.1007/s40037-017-0395-4

Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development*. https://doi.org/10.1007/s11423-019-09701-3

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive Architecture and Instructional Design: 20 Years Later. *Educational Psychology Review*, *31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Yasin, S., Yunus, M. F. M., & Ismail, I. (2018). The use of rasch measurement model for the validity and reliability. *Journal of Counseling and Educational Technology, 1(2)*, 22-27. https://doi.org/10.32698/0111

Zhang, L., Liu, X., & Feng, H. (2023). Development and validation of an instrument for assessing scientific literacy from junior to senior high school. *Disciplinary and Interdisciplinary Science Education Research*, *5*(1). https://doi.org/10.1186/s43031-023-00093-2