JagoMIPA: Jurnal Pendidikan Matematika dan IPA
p-ISSN: 2797-6475, e-ISSN: 2797-6467
Volume 4, nomor 4, 2024, hal. 718-733
Doi: https://doi.org/10.53299/jagomipa.v4i4.927

Analisis Kualitas Butir Soal Ujian Tengah Semester Biologi Umum Menggunakan Model Rasch

Ernie Novriyanti*, Riyan Arthur
Universitas Negeri Jakarata, Jakarta, Indonesia

*Coresponding Author: novriyantiernie@gmail.com
Dikirim: 26-11-2024; Direvisi: 28-11-2024; Diterima: 01-12-2024

Abstrak: Penelitian ini bertujuan mengevaluasi kualitas butir soal Ujian Tengah Semester (UTS) Biologi Umum menggunakan analisis Model Rasch. Instrumen terdiri dari 75 soal pilihan ganda yang mencakup berbagai materi Biologi Umum, dianalisis berdasarkan item fit, tingkat kesulitan, dan fungsi informasi tes. Hasil menunjukkan bahwa reliabilitas item sangat tinggi (0,97), menandakan konsistensi dalam mengukur kemampuan mahasiswa. Sebagian besar item memiliki nilai Infit dan Outfit dalam rentang toleransi (0,5–1,5), mendukung validitas model Rasch, meskipun beberapa item menunjukkan misfit, seperti Item 13 dan 54. Analisis Wright Map mengindikasikan cakupan instrumen yang baik pada kemampuan menengah, namun terdapat celah pada tingkat kemampuan ekstrem. Distribusi tingkat kesulitan item menunjukkan variasi yang cukup, tetapi beberapa materi, seperti Makhluk hidup dan metode ilmiah, didominasi item sangat mudah, yang kurang optimal untuk pengukuran kemampuan mahasiswa. Temuan ini menggarisbawahi pentingnya distribusi tingkat kesulitan yang seimbang dan revisi pada item misfit untuk meningkatkan kualitas instrumen. Rekomendasi utama adalah penambahan item untuk rentang kemampuan ekstrem dan penyesuaian distribusi tingkat kesulitan demi pengukuran yang lebih komprehensif dan akurat.

Kata Kunci: Analisis Rasch; Biologi Umum; reliabilitas; tingkat kesulitan; evaluasi pendidikan

Abstract: This study aims to evaluate the quality of midterm exam items in General Biology using the Rasch Model analysis. The instrument consists of 75 multiple-choice questions covering various General Biology topics, analyzed for item fit, difficulty levels, and test information function. Results show a high item reliability (0.97), indicating consistent measurement of student ability. Most items have Infit and Outfit values within the acceptable range (0.5–1.5), supporting the validity of the Rasch model, though some items, such as Items 13 and 54, exhibit misfit. The Wright Map reveals good instrument coverage at intermediate ability levels, though gaps exist at extreme ability levels. Item difficulty distribution shows sufficient variation, but some topics, such as Living Beings and Scientific Methods, are dominated by very easy items, which are suboptimal for assessing student capabilities. These findings highlight the importance of balanced item difficulty distribution and revision of misfit items to improve instrument quality. Key recommendations include adding items for extreme ability ranges and adjusting item difficulty distribution for more comprehensive and accurate measurement.

Keywords: Rasch analysis; General Biology; reliability; difficulty level; educational assessment

PENDAHULUAN

Evaluasi kualitas butir soal ujian merupakan elemen penting dalam sistem penilaian pendidikan, khususnya dalam konteks perguruan tinggi. Soal ujian berfungsi untuk mengukur sejauh mana kompetensi yang diharapkan telah dicapai oleh

mahasiswa (Sainuddin, 2018). Jika soal yang digunakan tidak memenuhi kriteria kualitas yang memadai, hasil penilaian dapat menjadi bias dan tidak mencerminkan kemampuan mahasiswa yang sebenarnya (Bond & Fox, 2015). Oleh karena itu, pengembangan dan evaluasi soal ujian perlu dilakukan secara sistematis untuk memastikan instrumen tersebut valid, andal, dan sesuai dengan tujuan pembelajaran.

Dalam konteks Ujian Tengah Semester (UTS) Biologi Umum, analisis kualitas soal menjadi penting untuk mengidentifikasi kekuatan dan kelemahan instrumen evaluasi yang digunakan. Pendekatan analisis yang sering digunakan adalah Teori Tes Klasik (*Classical Test Theory*/CTT) dan Teori Respons Butir (*Item Response Theory*/IRT) (Naga, 1992). CTT, meskipun sederhana dan mudah diterapkan, memiliki beberapa keterbatasan, seperti ketergantungan pada karakteristik populasi sampel tertentu dan kurangnya kemampuan untuk memberikan informasi rinci tentang parameter item (Crocker & Algina, 1986). Sebaliknya, Rasch Model dalam IRT mampu mengatasi keterbatasan tersebut dengan memberikan estimasi parameter yang bebas dari karakteristik sampel, sehingga lebih akurat dalam mengevaluasi kualitas butir soal (Linacre, 2024).

Analisis menggunakan Rasch Model memberikan keunggulan, termasuk kemampuan untuk mengevaluasi item fit dan person fit serta memetakan distribusi kemampuan peserta terhadap tingkat kesulitan item menggunakan Wright Map (Hambleton & Swaminathan, 2013; Linacre, 2010). Selain itu, pendekatan ini memungkinkan evaluasi asumsi unidimensionalitas dan independensi lokal yang esensial untuk validitas instrumen (Sumintono & Widhiarso, 2014). Dalam penelitian ini, analisis kualitas soal UTS Biologi Umum dilakukan untuk mengidentifikasi item yang tidak sesuai dengan model Rasch, serta memberikan rekomendasi untuk meningkatkan efektivitas soal dalam mengukur pencapaian mahasiswa.

KAJIAN TEORI

Teori Respons Butir dan Rasch Model

Teori Respons Butir (IRT) adalah pendekatan yang berbasis probabilitas untuk mengevaluasi kualitas butir soal (Fox, 2020). Salah satu model dalam IRT yang paling sering digunakan adalah Rasch Model. Model ini dirancang untuk memprediksi probabilitas seorang peserta menjawab benar pada suatu item, berdasarkan tingkat kemampuan peserta (*person ability*) dan tingkat kesulitan item (*item difficulty*) (Bond & Fox, 2015). Berbeda dengan CTT, Rasch Model memiliki keunggulan dalam memberikan parameter yang bebas dari distribusi sampel, sehingga hasil analisisnya lebih generalisabel (Linacre, 2024).

Unidimensionalitas dan Local Independence

Dua asumsi utama dalam Rasch Model adalah unidimensionalitas dan independensi lokal (*local independence*) (Andrich et al., 2019). Unidimensionalitas mengacu pada fakta bahwa semua item dalam tes harus mengukur satu konstruk psikologis yang sama, sementara independensi lokal berarti jawaban pada satu item tidak memengaruhi jawaban pada item lainnya (Aryadoust, 2017). Pelanggaran terhadap salah satu asumsi ini dapat menyebabkan bias dalam interpretasi hasil tes.

Kriteria Evaluasi Kualitas Item

1. Item Fit: Statistik *infit mean square* (MNSQ) dan *outfit MNSQ* digunakan untuk mengevaluasi sejauh mana respons pada item sesuai dengan prediksi model Rasch.

Rentang nilai ideal untuk kedua statistik ini adalah 0,5 hingga 1,5 (Linacre, 2024). Item dengan nilai *fit* di luar rentang ini dianggap tidak sesuai dan memerlukan revisi atau penghapusan.

2. Point-Measure Correlation (PTMEA): Korelasi antara skor item dan kemampuan peserta seharusnya bernilai positif. Item dengan nilai PTMEA negatif menunjukkan bahwa item tersebut tidak mendukung pengukuran konstruk yang dimaksud (Sumintono & Widhiarso, 2014).
3. Reliabilitas dan Separation Index: Reliabilitas butir dan responden dievaluasi menggunakan statistik reliabilitas dan *separation index*. Reliabilitas di atas 0,8 dianggap baik, sedangkan nilai *separation index* ≥ 2 menunjukkan kemampuan instrumen dalam memisahkan tingkat kemampuan peserta (Elvira et al., 2023; Zhou et al., 2017).

Wright Map

Wright Map adalah alat visualisasi dalam Rasch Model yang memetakan distribusi kemampuan peserta terhadap tingkat kesulitan item. Peta ini membantu dalam mengidentifikasi apakah soal memiliki tingkat kesulitan yang sesuai dengan kemampuan peserta. Item yang terlalu mudah atau terlalu sulit cenderung tidak efektif dalam membedakan tingkat kemampuan peserta (Bond & Fox, 2015).

Meskipun Rasch Model memberikan banyak keunggulan, tantangan dalam penerapannya meliputi kebutuhan akan sampel yang cukup besar dan kompleksitas perhitungan statistik. Namun, perangkat lunak seperti Winsteps telah mempermudah implementasi model ini dalam berbagai konteks pendidikan, termasuk analisis kualitas soal ujian di perguruan tinggi (Aryadoust, 2017).

METODE PENELITIAN

Desain Penelitian

Penelitian ini menggunakan desain penelitian deskriptif kuantitatif untuk mengevaluasi kualitas butir soal Ujian Tengah Semester (UTS) Biologi Umum berdasarkan Rasch Model. Desain ini bertujuan untuk memberikan gambaran sistematis tentang validitas, reliabilitas, serta distribusi tingkat kesulitan dan kemampuan mahasiswa dalam menjawab soal-soal UTS.

Subjek Penelitian

Subjek penelitian ini adalah 75 butir soal pilihan ganda UTS Biologi Umum yang diberikan kepada mahasiswa semester I di Universitas Negeri Padang. Soal tersebut mencakup lima pilihan jawaban (satu kunci jawaban benar dan empat pengecoh) yang dirancang untuk mengukur pemahaman mahasiswa terhadap berbagai konsep dalam Biologi Umum, seperti metabolisme, struktur sel, biodiversitas, dan metode ilmiah.

Populasi penelitian adalah 135 mahasiswa yang mengikuti UTS. Data dari hasil ujian ini digunakan untuk analisis Rasch. Respons mahasiswa dikodekan sebagai 1 untuk jawaban benar dan 0 untuk jawaban salah, sesuai dengan analisis respons dikotomi.

Instrumen Penelitian

Instrumen yang digunakan dalam penelitian ini adalah soal-soal UTS Biologi Umum dalam format pilihan ganda yang dikemabangkan berdasarakn materi tyang

telah disajikan ecara lengkap pada tabel 1. Setiap soal terdiri dari satu kunci jawaban yang benar dan empat pengecoh. Data yang dianalisis mencakup proporsi mahasiswa yang menjawab setiap opsi dalam soal, termasuk kunci jawaban dan pengecoh. Data ini digunakan untuk menghitung indeks kesukaran, validitas, dan reliabilitas instrumen soal yang ada.

Tabel 1. Materi dan Sub Materi UAS Biologi Umum

No	Materi	Sub Materi	Nomor Soal	Σ
1	Struktur dan organisasi tubuh hewan	Echinodermata	1	1
		Arthropoda	2,3	2
		Moluska	4,5	2
		Annelida	6	1
		Nemathelminthes	7	1
		Platyhelminthes	8	1
		Coelenterata	9,10	2
		Porifera	11,12	2
		Ciri-ciri Hewan	13	1
2	Struktur dan organisasi tubuh tumbuhan	Ciri-ciri, contoh,dan manfaat tumbuhan lumut (bryophyta).	14,15,16,17	4
		Ciri-ciri, contoh, dan manfaat tumbuhan biji (spermatophyta).	18,19,20,21, 22,23	6
		Tumbuhan Biji (mono/dikotil)	24,25,26,27, 28,29	6
3	Biodiversitas	Keanekaragam ekosistem	30,,31,32,33 ,34	5
		Endemk: anoa, babirusa, dan maleo, cendrawasih, Komodo, bangkai, rafflesia(parasite)	35	1
		Urutan hirarki	36	1
		Aturan tata nama	37	1
		Perubahan Iklim, Penggunaan Lahan, Deforestasi, urbanisasi, dan konversi lahan untuk pertanian Eksploitasi Berlebihan polusi	38,39	2
4	Metabolisme	Ciri-ciri Makhluk Hidup	40,41,42,43, 44,45	6
		Mekanisme fotosintesis	46,47,48,49	4
		Mekanisme respirasi seluler	50,51,52	3
5	Sel sebagai dasar kehidupan	Sel sebagai dasar kehidupan	53	1
		Metode Pengamatan sel.	54	1
		Sifat fisik dan sifat kimia sel	55	1
		Struktur dan fungsi sel	56,57,58,59, 60,61	6
		transport zat pada sel	62,63,64	3
6	Makhluk hidup dan metode ilmiah	Kedudukan biologi dalam sains	65	1
		Objek biologi	66,67	2
		Cabang-cabang biologi	68,69,70	3
		Langkah metode ilmiah	71,72,73	3
		Teori asal-usul kehidupan (abio, bio dan modern)	74,75	2
Total				75

Teknik Analisis Data

Data yang telah dikumpulkan dianalisis menggunakan statistik deskriptif. Hasil analisis disajikan dalam bentuk tabel dan grafik yang mencakup distribusi item fit, reliabilitas, Wright Map, dan klasifikasi tingkat kesulitan item. Interpretasi hasil dilakukan dengan membandingkan nilai statistik terhadap kriteria yang telah

ditentukan, seperti rentang MNSQ, PTMEA, dan asumsi unidimensionalitas. Data juga digunakan untuk memberikan rekomendasi perbaikan pada butir soal yang tidak memenuhi kriteria kualitas berdasarkan Rasch Model.

Analisis data dalam penelitian ini menggunakan pendekatan Rasch Model, yang dilakukan melalui perangkat lunak Winsteps. Proses ini diawali dengan pengujian unidimensionalitas untuk memastikan bahwa seluruh butir soal mengukur satu konstruk utama, yaitu pemahaman konseptual mahasiswa dalam Biologi Umum. Sesuai dengan teori Rasch, *raw variance explained by measures* harus melebihi 20% untuk menunjukkan bahwa sebagian besar varians dalam data berasal dari konstruk yang diukur (Linacre, 2024). Selain itu, *unexplained variance in 1st contrast* diharapkan berada di bawah 15% untuk kategori baik atau di bawah 5% untuk kategori sangat baik (Bond & Fox, 2015). Analisis ini penting untuk memastikan validitas internal instrumen dan untuk mendeteksi potensi pelanggaran asumsi unidimensionalitas.

Langkah berikutnya adalah analisis *item fit* dan *person fit*, yang bertujuan untuk mengevaluasi sejauh mana pola respons peserta sesuai dengan prediksi model Rasch. Statistik *infit mean square (MNSQ)* dan *outfit MNSQ* digunakan dengan rentang ideal antara 0,5 hingga 1,5, sebagaimana direkomendasikan oleh Linacre (2024). Item dengan nilai *fit* di luar rentang ini dianggap tidak sesuai, sehingga perlu direvisi atau dihapus. Selain itu, point-*measure correlation* (PTMEA) dianalisis untuk memastikan bahwa setiap item memiliki korelasi positif dengan kemampuan peserta, yang menunjukkan bahwa item tersebut mendukung pengukuran konstruk. Hasil ini dilengkapi dengan pemetaan menggunakan Wright Map, yang memberikan visualisasi distribusi kemampuan mahasiswa terhadap tingkat kesulitan item. Data yang diperoleh dari analisis ini tidak hanya membantu dalam mengevaluasi validitas dan reliabilitas instrumen, tetapi juga memberikan rekomendasi yang terfokus untuk peningkatan kualitas soal berdasarkan prinsip-prinsip psikometri (Sumintono & Widhiarso, 2014).

HASIL DAN PEMBAHASAN

Analisis prasyarat

Sebelum melakukan analisis IRT dengan Winstep ada beberapa langkah penting yang harus dilakukan untuk memastikan validitas dan keandalan hasil. Pertama sintak yang digunakan harus sesuai dengan model. Berikut sintaks untuk analysis IRT untuk data ujian mata kuliah biologi umum dengan respons berupa data dikotomi dan menggunakan analisis rasch dengan winstep.

Berdasarkan hasil analisis pada Gambar 1 menggunakan perangkat lunak Winsteps, uji asumsi unidimensionalitas menunjukkan bahwa instrumen yang digunakan memenuhi kriteria untuk analisis Rasch. Nilai *raw variance explained by measures* pada kolom empirical sebesar 27.1%, melebihi ambang batas minimum 20% sebagaimana disarankan oleh Linacre (2011) untuk data dengan respons dikotomi. Nilai ini menunjukkan bahwa mayoritas varians dalam data berasal dari konstruk utama yang diukur oleh instrumen, yang memberikan bukti kuat bahwa instrumen ini unidimensional. Selain itu, nilai *unexplained variance in the 1st contrast* adalah 2.0%, yang berada dalam kategori sangat baik (<5%) menurut (Holster & Lake, 2016) Dimensi residual lainnya, yaitu dari *2nd contrast* hingga *5th contrast*, masing-masing memiliki nilai varians residual dalam rentang 2.4%-3.0%, yang juga berada dalam kategori sangat baik berdasarkan kriteria Linacre (2011). Dengan demikian, instrumen

ini terbukti unidimensional dan valid dalam mengukur satu konstruk utama, yaitu kemampuan mahasiswa pada mata kuliah Biologi Umum.

```
TABLE 23.0 Rasch.xlsx                          ZOU306WS.TXT  Nov 25 16:16 2024
INPUT: 135 PERSON  75 ITEM  REPORTED: 135 PERSON  75 ITEM  2 CATS  WINSTEPS 3.73
--------------------------------------------------------------------------------

     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                                -- Empirical --    Modeled
Total raw variance in observations     =        102.9 100.0%         100.0%
  Raw variance explained by measures   =         27.9  27.1%          26.7%
    Raw variance explained by persons  =          3.5   3.4%           3.4%
    Raw Variance explained by items    =         24.4  23.7%          23.3%
  Raw unexplained variance (total)     =         75.0  72.9% 100.0%   73.3%
    Unexplned variance in 1st contrast =          3.1   3.0%   4.1%
    Unexplned variance in 2nd contrast =          2.8   2.7%   3.7%
    Unexplned variance in 3rd contrast =          2.6   2.6%   3.5%
    Unexplned variance in 4th contrast =          2.5   2.5%   3.4%
    Unexplned variance in 5th contrast =          2.4   2.4%   3.2%
```

Gambar 1. Analisis Varians Residu Terstandarisasi Memvalidasi Unidimensionalitas Instrumen

Selain uji unidimensionalitas, pengujian independensi lokal juga dilakukan untuk memastikan bahwa respons terhadap satu item tidak memengaruhi respons terhadap item lainnya, kecuali melalui konstruk yang sama. Dari hasil analisis korelasi residual pada Tabel 23.99, ditemukan pasangan item dengan korelasi tertinggi, yaitu item 13 dan item 24 dengan nilai korelasi residual sebesar 0.38, yang melampaui ambang batas 0.30 sebagaimana direkomendasikan oleh Aryadoust et al. (2020). Korelasi tinggi juga ditemukan pada pasangan item lain, seperti item 35 dan item 54 dengan nilai korelasi residual 0.35. Korelasi residual yang tinggi ini dapat mengindikasikan adanya potensi pelanggaran independensi lokal. Namun, setelah dilakukan analisis lebih lanjut terhadap isi item, diketahui bahwa item-item tersebut berasal dari materi yang berbeda, sehingga keterkaitan tersebut tidak dianggap signifikan secara substantif.

```
TABLE 23.99 Rasch.xlsx                         ZOU306WS.TXT  Nov 25 16:16 2024
INPUT: 135 PERSON  75 ITEM  REPORTED: 135 PERSON  75 ITEM  2 CATS  WINSTEPS 3.73
--------------------------------------------------------------------------------

     LARGEST STANDARDIZED RESIDUAL CORRELATIONS
     USED TO IDENTIFY DEPENDENT ITEM
-------------------------------------
|CORREL-| ENTRY       | ENTRY       |
|  ATION|NUMBER ITEM  |NUMBER ITEM  |
|-------+-------------+-------------|
|   .38 |    13 It_13 |    24 It_24 |
|   .35 |    35 It_35 |    54 It_54 |
|   .30 |    10 It_10 |    31 It_31 |
|   .29 |    13 It_13 |    55 It_55 |
|   .27 |    14 It_14 |    58 It_58 |
|   .27 |    18 It_18 |    28 It_28 |
|-------+-------------+-------------|
|  -.30 |     4 It_04 |    46 It_46 |
|  -.29 |     1 It_01 |    36 It_36 |
|  -.28 |    43 It_43 |    44 It_44 |
|  -.27 |     7 It_07 |    37 It_37 |
-------------------------------------
```

Gambar 2. Korelasi Residu Terstandarisasi untuk Mengidentifikasi Ketergantungan Antar-Item

Menurut Retnawati (2014), apabila asumsi unidimensionalitas terpenuhi, maka asumsi independensi lokal cenderung terpenuhi pula. Dengan demikian, meskipun terdapat beberapa pasangan item dengan korelasi residual tinggi, secara keseluruhan tidak ditemukan pelanggaran independensi lokal yang signifikan. Hal ini memperkuat validitas data dan mendukung penerapan model Rasch pada instrumen ini.

Berdasarkan analisis korelasi residual yang dilakukan dan disajikan pada Gambar 2, ditemukan bahwa angka 0.38 menunjukkan korelasi residual antara pasangan item 13 dan item 24 dengan arah positif. Sementara itu, pasangan item 4 dan item 46 memiliki korelasi negatif. Menurut kriteria local independence yang dikemukakan oleh Aryadoust et al. (2020), local independence dianggap terlanggar apabila korelasi residual antar pasangan item memiliki nilai positif dan lebih besar dari 0.30, dalam hal ini beberapa item yang perlu di hilangkan yakni pasangan item 13 dan 24, 35 dan 54 karena nilainya memiliki nilai korelasi residu yang sangat besar. Namun menurut pendapat lain dari Retnawati (2014) jika asumsi unidimensi terpenuhi, maka independesni lokal juga terpenuhi. Dari kasus ini penulis merasa dalam hal keterkaitan antar kedua pasangan item ini tidak ada karena berasal dari materi yang berbeda. Oleh karena itu, dapat disimpulkan bahwa asumsi local independence terpenuhi. Analisis ini mengindikasikan bahwa respon terhadap satu item tidak tergantung pada respon terhadap item lainnya.

Rasch Analisis

Hasil analisis rasch menunjukkan hasil seperti pada Tabel 2.

Tabel 2. Item Statistics

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
13	5	135	2.72	.46	1.03	.2	1.52	1.1	-.05	.10	96.3	96.3	It_13
54	6	135	2.52	.42	1.05	.3	1.64	1.4	-.13	.11	95.6	95.5	It_54
6	7	135	2.36	.39	1.01	.1	1.33	.9	.04	.11	94.8	94.8	It_06
25	7	135	2.36	.39	1.03	.2	1.31	.9	-.02	.11	94.8	94.8	It_25
44	7	135	2.36	.39	1.00	.1	1.11	.4	.08	.11	94.8	94.8	It_44
10	9	135	2.09	.35	1.02	.2	1.31	.9	.01	.13	93.3	93.3	It_10
74	9	134	2.08	.35	1.09	.4	1.85	2.1	-.25	.13	93.3	93.3	It_74
31	10	135	1.97	.33	1.05	.3	1.28	.9	-.03	.13	92.6	92.6	It_31
45	10	135	1.97	.33	1.03	.2	1.09	.4	.06	.13	92.6	92.6	It_45
55	12	135	1.77	.31	.94	-.2	.92	-.2	.27	.14	91.1	91.1	It_55
24	16	135	1.44	.27	1.04	.3	1.24	1.0	.01	.16	88.1	88.1	It_24
35	16	135	1.44	.27	1.05	.3	1.21	.9	.01	.16	88.1	88.1	It_35
38	20	135	1.18	.25	1.04	.3	1.08	.5	.09	.18	85.2	85.2	It_38
58	21	135	1.12	.24	1.04	.3	1.08	.5	.10	.18	84.4	84.4	It_58
28	23	135	1.00	.23	1.07	.5	1.16	.9	.03	.19	83.0	82.9	It_28
27	24	135	.95	.23	.95	-.3	.92	-.4	.29	.19	82.2	82.2	It_27
65	24	134	.94	.23	.99	.0	.96	-.2	.21	.19	82.1	82.1	It_65
52	25	135	.90	.23	1.00	.0	.98	-.1	.20	.19	81.5	81.5	It_52
53	28	135	.75	.22	.92	-.6	.87	-.9	.37	.20	79.3	79.2	It_53
57	28	134	.74	.22	1.19	1.5	1.40	2.5	-.23	.20	79.1	79.1	It_57
1	30	135	.66	.21	1.11	1.0	1.17	1.2	-.02	.21	77.8	77.8	It_01
14	30	135	.66	.21	.97	-.3	.89	-.8	.30	.21	77.8	77.8	It_14
47	31	135	.62	.21	.95	-.4	.96	-.2	.29	.21	77.0	77.0	It_47
63	31	135	.62	.21	.92	-.7	.88	-.9	.37	.21	77.0	77.0	It_63
49	32	135	.57	.21	.97	-.2	.90	-.8	.30	.21	76.3	76.3	It_49
33	33	135	.53	.21	.98	-.1	.94	-.4	.26	.21	74.8	75.5	It_33
8	34	135	.49	.20	.96	-.3	.91	-.7	.30	.22	75.6	74.8	It_08
9	37	135	.37	.20	1.01	.2	.98	-.1	.21	.22	71.9	72.6	It_09
23	38	135	.33	.20	1.08	1.0	1.17	1.6	.03	.22	72.6	71.9	It_23
30	39	135	.29	.19	1.02	.3	1.01	.1	.19	.22	69.6	71.3	It_30
12	41	135	.22	.19	.97	-.4	.98	-.2	.28	.23	71.9	70.0	It_12
66	41	135	.22	.19	.99	-.1	.99	-.1	.24	.23	67.4	70.0	It_66
68	41	135	.22	.19	.95	-.6	.93	-.7	.33	.23	71.9	70.0	It_68
62	42	135	.18	.19	1.12	1.6	1.14	1.5	-.01	.23	65.2	69.3	It_62
5	44	135	.11	.19	1.04	.6	1.05	.6	.15	.23	64.4	68.1	It_05
7	45	135	.07	.19	.98	-.3	1.01	.2	.26	.23	69.6	67.5	It_07
22	45	135	.07	.19	1.02	.3	1.01	.1	.21	.23	63.7	67.5	It_22
60	46	135	.04	.19	1.06	.9	1.06	.8	.12	.23	64.4	66.9	It_60

11	47	135	.00	.19	1.01	.1	1.03	.4	.21	.23	66.7	66.4	It_11
39	49	135	-.07	.18	.96	-.6	.95	-.6	.31	.24	66.7	65.3	It_39
26	50	135	-.10	.18	.86	-2.4	.84	-2.4	.51	.24	71.9	64.8	It_26
42	50	135	-.10	.18	1.02	.4	1.03	.5	.19	.24	60.0	64.8	It_42
16	52	135	-.17	.18	1.03	.5	1.02	.3	.19	.24	58.5	64.0	It_16
51	56	135	-.30	.18	1.06	1.2	1.06	1.1	.13	.24	56.3	62.5	It_51
50	57	135	-.33	.18	1.02	.5	1.02	.4	.20	.24	59.3	62.1	It_50
18	59	135	-.39	.18	1.03	.6	1.03	.6	.19	.24	57.8	61.5	It_18
20	60	135	-.43	.18	.92	-1.9	.91	-1.8	.40	.24	68.1	61.2	It_20
21	60	135	-.43	.18	.96	-.8	.96	-.7	.31	.24	65.2	61.2	It_21
37	61	135	-.46	.18	.98	-.4	.98	-.4	.28	.24	63.0	61.0	It_37
71	63	135	-.52	.18	1.08	1.9	1.09	1.9	.09	.24	55.6	60.5	It_71
3	64	135	-.55	.18	.90	-2.4	.89	-2.5	.44	.24	66.7	60.3	It_03
19	65	135	-.58	.18	1.00	-.1	.99	-.2	.25	.24	60.0	60.2	It_19
61	66	135	-.62	.18	1.03	.8	1.04	.9	.18	.24	55.6	60.1	It_61
40	72	135	-.80	.18	.98	-.4	.97	-.5	.28	.24	63.7	60.3	It_40
41	72	135	-.80	.18	1.01	.3	1.00	.1	.22	.24	56.3	60.3	It_41
36	73	135	-.84	.18	.97	-.8	.96	-.8	.31	.24	65.2	60.5	It_36
67	74	135	-.87	.18	.92	-1.9	.90	-2.0	.40	.24	65.9	60.6	It_67
46	77	135	-.96	.18	.93	-1.6	.93	-1.4	.37	.24	70.4	61.3	It_46
15	78	135	-1.00	.18	.93	-1.5	.92	-1.4	.38	.24	63.7	61.6	It_15
43	78	135	-1.00	.18	.98	-.4	.99	-.2	.27	.24	66.7	61.6	It_43
70	81	135	-1.09	.18	.99	-.1	.98	-.2	.25	.24	60.7	62.6	It_70
59	84	135	-1.19	.18	.98	-.4	.97	-.4	.28	.23	60.7	63.9	It_59
34	87	135	-1.29	.18	1.02	.3	1.02	.3	.20	.23	66.7	65.5	It_34
73	87	135	-1.29	.18	.99	-.2	1.00	.1	.25	.23	68.1	65.5	It_73
29	90	135	-1.40	.19	.94	-.9	.90	-1.2	.36	.23	67.4	67.2	It_29
32	91	135	-1.43	.19	.92	-1.1	.89	-1.3	.39	.23	70.4	67.9	It_32
17	100	135	-1.77	.20	1.01	.1	1.00	.1	.20	.21	72.6	74.1	It_17
69	101	135	-1.81	.20	.92	-.8	.87	-1.1	.38	.21	76.3	74.8	It_69
48	102	135	-1.85	.20	.98	-.2	.93	-.5	.27	.21	75.6	75.5	It_48
2	105	135	-1.98	.21	.97	-.2	.96	-.2	.25	.20	77.8	77.8	It_02
72	108	135	-2.12	.22	.97	-.2	.95	-.3	.26	.19	80.0	80.0	It_72
56	111	135	-2.27	.23	.98	-.1	.91	-.5	.26	.18	82.2	82.2	It_56
4	114	135	-2.44	.24	1.03	.3	1.11	.6	.09	.18	84.4	84.4	It_04
64	118	135	-2.69	.26	1.04	.3	1.06	.4	.08	.16	87.4	87.4	It_64
75	122	135	-3.00	.29	1.00	.1	1.02	.2	.13	.14	90.4	90.4	It_75
MEAN	51.6	135.0	.00	.22	1.00	-.1	1.04	.0			74.2	74.1	
S.D.	31.4	.2	1.31	.06	.06	.8	.17	1.0			11.5	11.4	

Item Fit

Item fit adalah elemen penting dalam analisis Rasch untuk mengevaluasi seberapa baik setiap item dalam instrumen pengukuran sesuai dengan prediksi model Rasch. Dua indikator utama yang digunakan dalam analisis ini adalah Infit Mean Square (MNSQ) dan Outfit Mean Square (MNSQ) sebagaimana disajikan pada kolom 5 dan 8 pada tabel 2 , dengan nilai ideal berada dalam rentang 0.5 hingga 1.5 (Linacre, 2011). Selain itu, nilai ZSTD juga digunakan sebagai pendukung dalam mengevaluasi kesesuaian, dengan nilai ideal berada dalam rentang -1.96 hingga 1.96. Nilai di luar rentang ini mengindikasikan kemungkinan pola respons yang tidak sesuai dengan model. Namun, terdapat beberapa item yang menunjukkan Misfit, di antaranya adalah Item 3, Item 13, Item 26, Item 54, Item 57, Item 67, dan Item 74. Contohnya adalah Item 13 dari materi Struktur dan organisasi tubuh hewan, yang memiliki nilai Outfit MNSQ sebesar 1.52 dengan ZSTD sebesar 1.1, menunjukkan adanya pola respons yang tidak konsisten dengan prediksi model. Selain itu, Item 54 dari materi Sel sebagai dasar kehidupan juga menunjukkan nilai Outfit MNSQ sebesar 1.64 dan ZSTD sebesar 1.4, mengindikasikan potensi masalah dalam desain item atau respons peserta. Misfit ini mengindikasikan bahwa item-item tersebut mungkin mengukur sesuatu di luar konstruk utama atau memiliki tingkat kesulitan yang tidak sesuai dengan kemampuan mayoritas peserta.

Item measure

Dari segi tingkat kesulitan item (*Item Measure*), instrumen ini mencakup spektrum kesulitan yang cukup luas, mulai dari item yang sangat mudah hingga sangat

sulit. Berdasarkan hasil analisis yang disajikan pada tabel 2 kolom4, Item seperti Item 4 dari materi Struktur dan organisasi tubuh hewan (β = -2.44) dan Item 29 dari materi Struktur dan organisasi tubuh tumbuhan (β = -1.40) tergolong sangat mudah, memungkinkan hampir semua peserta menjawabnya dengan benar. Sebaliknya, Item 6 dan Item 13 dari materi Struktur dan organisasi tubuh hewan, dengan masing-masing nilai β sebesar 2.36 dan 2.72, termasuk dalam kategori sangat sulit dan hanya dapat dijawab oleh peserta dengan kemampuan tinggi. Distribusi tingkat kesulitan ini menunjukkan bahwa instrumen memiliki variasi yang cukup baik untuk mengukur kemampuan peserta dengan berbagai tingkatan, meskipun beberapa item dengan tingkat kesulitan ekstrem, seperti Item 6 dan Item 13, memerlukan revisi untuk meningkatkan keseimbangan instrumen.

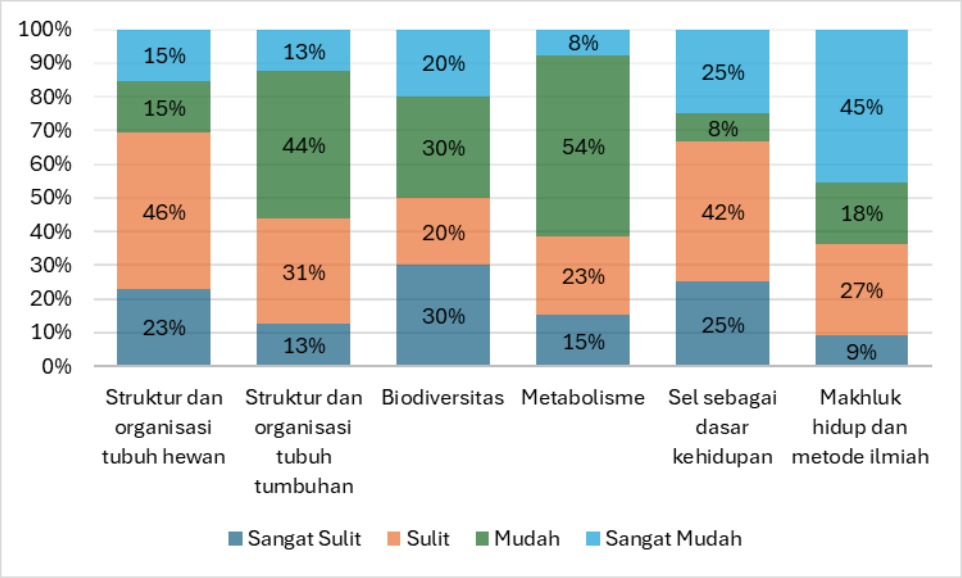
Tabel 3. Rekapitulasi dan persentase indeks kesulitan berdasarkan materi

Difficulty (β)	Keseluruhan	Struktur dan organisasi tubuh hewan	Struktur dan organisasi tubuh tumbuhan	Biodiversitas	Metabolisme	Sel sebagai dasar kehidupan	Makhluk hidup dan metode ilmiah
Sangat Sulit	14	23%	13%	30%	15%	25%	9%
Sulit	24	46%	31%	20%	23%	42%	27%
Mudah	22	15%	44%	30%	54%	8%	18%
Sangat Mudah	15	15%	13%	20%	8%	25%	45%

Tabel 3 menunjukkan distribusi tingkat kesulitan item berdasarkan kategori kesulitan dari masing-masing materi yang diuji. Secara keseluruhan, item dalam kategori sulit memiliki persentase terbesar, yaitu 46%, diikuti oleh kategori mudah (15%), sangat sulit (14%), dan sangat mudah (15%). Pada materi struktur dan organisasi tubuh tumbuhan, proporsi terbesar berada pada kategori mudah, yaitu 44%, yang menunjukkan bahwa mayoritas peserta dapat menjawab item dari materi ini dengan baik. Sebaliknya, pada materi sel sebagai dasar kehidupan, 42% item berada pada kategori sulit, tetapi juga ada 25% pada kategori sangat mudah, yang menunjukkan distribusi tingkat kesulitan yang kurang merata. Materi makhluk hidup dan metode ilmiah cenderung lebih didominasi oleh kategori sangat mudah dengan persentase tertinggi sebesar 45%, mengindikasikan item yang terlalu mudah bagi peserta. Keseluruhan hasil ini menggambarkan bahwa instrumen memiliki distribusi tingkat kesulitan yang cukup beragam, tetapi beberapa materi memerlukan revisi untuk menciptakan keseimbangan yang lebih baik.

Gambar 3 memperlihatkan distribusi tingkat kesulitan item dalam bentuk persentase untuk masing-masing materi. Struktur dan organisasi tubuh tumbuhan memiliki distribusi yang merata, dengan fokus tertinggi pada kategori mudah sebesar 44%. Di sisi lain, materi biodiversitas menunjukkan dua puncak distribusi, yaitu pada kategori mudah (30%) dan sangat sulit (30%), yang menandakan bahwa materi ini dapat mengukur peserta di berbagai tingkatan kemampuan. Untuk materi makhluk hidup dan metode ilmiah, dominasi kategori sangat mudah (45%) menunjukkan bahwa banyak item dalam materi ini terlalu sederhana bagi peserta. Sebaliknya, metabolisme memiliki fokus yang lebih besar pada kategori mudah (54%), mencerminkan pengukuran yang lebih efektif pada tingkat kemampuan menengah. Secara

keseluruhan, variasi distribusi ini memberikan gambaran tentang efektivitas instrumen dalam mengukur kemampuan peserta, tetapi juga menunjukkan perlunya penyesuaian pada beberapa materi agar distribusi tingkat kesulitan lebih seimbang.



Gambar 3. Sebaran kesulitan dari masing-masing materi

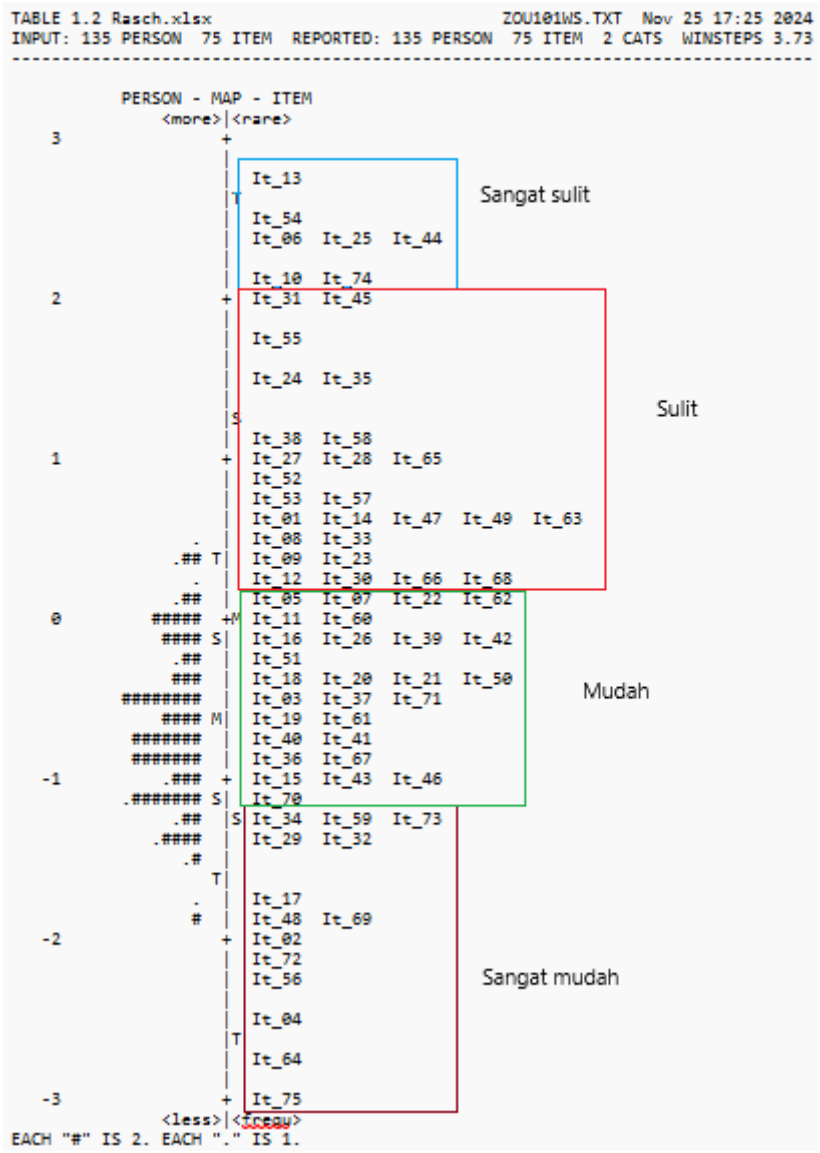
Wright Map

Wright Map merupakan konsep yang penting dalam analisis Item Response Theory (IRT), yang pertama kali digunakan secara signifikan oleh Wilson dan Draney pada tahun 2002. Wright Map menyajikan perbandingan visual langsung antara distribusi kemampuan responden dan distribusi tingkat kesulitan (difficulty level) dari item soal dalam satu grafik atau garis yang sama.

Gambar Wright Map ini, terlihat bahwa item tersebar di empat kategori kesulitan: sangat sulit, sulit, mudah, dan sangat mudah. Item dengan tingkat kesulitan sangat sulit, seperti Item 13, Item 54, dan Item 44, hanya dapat dijawab oleh peserta dengan kemampuan sangat tinggi. Sebaliknya, item sangat mudah, seperti Item 17, Item 64, dan Item 75, dapat dijawab oleh hampir semua peserta, termasuk yang memiliki kemampuan rendah. Distribusi ini menunjukkan bahwa instrumen mencakup spektrum tingkat kesulitan yang cukup luas, yang penting untuk mengukur peserta dengan berbagai tingkat kemampuan.

Berdasarkan Gambar 4 diperoleh informasi bahwa sebagian besar peserta berada di sekitar nilai logit 0, yang menunjukkan bahwa mayoritas peserta memiliki kemampuan menengah. Di sisi lain, item dengan tingkat kesulitan berada dalam rentang logit -1 hingga +1 mendominasi kategori mudah dan sulit, seperti Item 20 dan Item 33. Hal ini menunjukkan bahwa sebagian besar item berada pada tingkat kesulitan yang sesuai dengan kemampuan mayoritas peserta, yang mendukung keakuratan pengukuran.

Namun, terdapat beberapa celah pada Wright Map. Misalnya, terdapat sedikit item pada rentang logit di bawah -2 atau di atas +2. Hal ini mengindikasikan bahwa instrumen kurang mampu mengukur peserta dengan kemampuan ekstrem, baik yang sangat rendah maupun sangat tinggi. Untuk meningkatkan akurasi pengukuran, direkomendasikan untuk menambahkan item pada kedua ujung rentang kesulitan ini.



Gambar 4. Wrigh Map

Reliabilitas Item

Berdasarkan Tabel 4 ringkasan hasil analisis untuk 75 item yang diukur, instrumen ini menunjukkan reliabilitas item yang sangat tinggi. Nilai reliabilitas item adalah 0.97, yang berada di atas ambang batas ideal (≥0.80), menunjukkan bahwa item-item dalam instrumen ini sangat konsisten dalam mengukur konstruk yang diinginkan. Nilai reliabilitas ini dihitung menggunakan formula yang mempertimbangkan rasio antara variabilitas item yang sebenarnya (true variance) dengan total variabilitas item, termasuk error measurement. Dengan reliabilitas setinggi ini, dapat disimpulkan bahwa instrumen memiliki kemampuan yang sangat baik untuk membedakan tingkat kesulitan antar item.

Selain itu, nilai *item separation* sebesar 5.45 mengindikasikan bahwa item dalam instrumen ini memiliki penyebaran tingkat kesulitan yang cukup luas dan mampu membedakan berbagai level kemampuan mahasiswa. Nilai ini jauh di atas batas minimum 2.0 yang disarankan untuk instrumen yang dapat diandalkan. Penyebaran tingkat kesulitan item ini juga diperkuat oleh nilai standar deviasi (SD) dari tingkat

kesulitan item sebesar 1.31, yang menunjukkan bahwa variasi tingkat kesulitan item cukup tinggi, memberikan cakupan yang baik untuk mengukur kemampuan peserta dari berbagai tingkatan.

Tabel 4. Summary of 75 measured item

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	51.6	135.0	.00	.22	1.00	-.1	1.04	.0
S.D.	31.4	.2	1.31	.06	.06	.8	.17	1.0
MAX.	122.0	135.0	2.72	.46	1.19	1.9	1.85	2.5
MIN.	5.0	134.0	-3.00	.18	.86	-2.4	.84	-2.5

REAL RMSE	.24	TRUE SD	1.29	SEPARATION	5.45	ITEM	RELIABILITY	.97
MODEL RMSE	.23	TRUE SD	1.29	SEPARATION	5.52	ITEM	RELIABILITY	.97
S.E. OF ITEM MEAN = .15								

UMEAN=.0000 USCALE=1.0000
ITEM RAW SCORE-TO-MEASURE CORRELATION = -.98
10122 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 10486.15 with 9913 d.f. p=.0000
Global Root-Mean-Square Residual (excluding extreme scores): .4148
Capped Binomial Deviance = .2250 for 10122.0 dichotomous observations

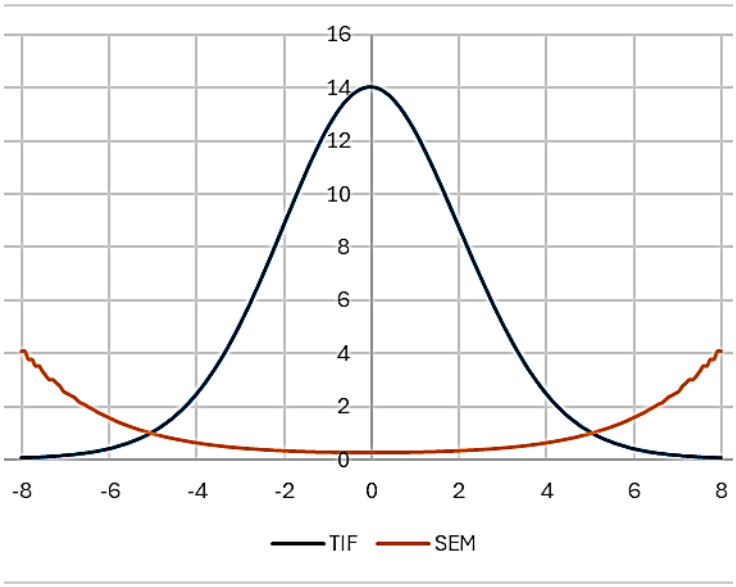
Secara umum, nilai rata-rata Infit MNSQ adalah 1.00 dan Outfit MNSQ adalah 1.04, dengan standar deviasi masing-masing 0.06 dan 0.17. Nilai-nilai ini menunjukkan bahwa mayoritas item dalam instrumen memiliki pola respons yang sesuai dengan prediksi model Rasch. Nilai maksimum Infit ZSTD sebesar 1.9 dan Outfit ZSTD sebesar 2.5 menunjukkan adanya beberapa item yang sedikit menyimpang dari model, tetapi secara keseluruhan, item-item ini tetap berada dalam batas toleransi yang dapat diterima.

Hasil lain yang mendukung reliabilitas instrumen adalah nilai *Root Mean Square Error of Measurement (RMSE)* sebesar 0.24 untuk real data dan 0.23 untuk model data, yang menunjukkan bahwa error measurement relatif kecil dibandingkan dengan variabilitas item yang sebenarnya. Hal ini memperkuat validitas hasil analisis, karena error yang rendah memastikan bahwa hasil pengukuran sebagian besar mencerminkan kemampuan sebenarnya.

Fungsi Informasi Tes

Berdasarkan hasil analisis dan gambar 5, perpotongan TIF dan SEM terjadi pada rentang kemampuan sekitar -5.04 hingga 5.04 logit. Pada rentang ini, TIF mencapai puncaknya, menunjukkan bahwa instrumen ini memberikan informasi yang paling akurat tentang kemampuan individu dengan tingkat tersebut. Nilai SEM yang rendah pada titik ini mengindikasikan bahwa kesalahan pengukuran juga minimal, yang berarti hasil tes memberikan estimasi kemampuan individu yang sangat akurat. Dalam praktiknya, hal ini menunjukkan bahwa instrumen ini sangat baik dalam mengukur kemampuan peserta dengan tingkat kemampuan menengah hingga ekstrem.

Pada puncak TIF, yaitu sekitar logit 0, tes memberikan informasi maksimal (TIF $\approx$ 14), menunjukkan bahwa instrumen ini paling akurat dalam mengukur kemampuan peserta pada tingkat kemampuan menengah. Sebaliknya, nilai SEM berada pada titik terendah di area yang sama, menunjukkan bahwa kesalahan pengukuran sangat kecil, sehingga hasil tes pada rentang ini memiliki tingkat keandalan yang tinggi. Hal ini sesuai dengan prinsip bahwa semakin tinggi informasi yang diberikan oleh tes, semakin rendah tingkat kesalahan pengukurannya.



Gambar 5. Fungis Informasi tes dan standar kesalahan pengukuran

Pembahasan

Hasil penelitian ini menunjukkan bahwa Analisis unidimensionalitas menunjukkan bahwa instrumen memenuhi asumsi dasar model Rasch, dengan raw variance explained by measures sebesar 27.1%. Hal ini sejalan dengan kriteria yang diajukan oleh Holster dan Lake (2016), yang menyatakan bahwa nilai di atas 20% cukup untuk memastikan bahwa instrumen mengukur satu konstruk utama. Nilai unexplained variance dalam 1st contrast sebesar 3.1% berada dalam kategori sangat baik, menandakan bahwa data memiliki struktur unidimensional yang kuat (Bond & Fox, 2015).

Selain itu, instrumen yang dianalisis memiliki reliabilitas yang tinggi (0.97), sejalan dengan temuan Sumintono dan Widhiarso (2014), yang melaporkan bahwa instrumen berbasis Rasch cenderung memberikan hasil yang konsisten dalam memisahkan peserta berdasarkan kemampuan mereka. Nilai reliabilitas ini mengindikasikan bahwa instrumen dapat diandalkan untuk mengukur kemampuan mahasiswa dalam memahami materi Biologi Umum secara konsisten. Selain itu, temuan ini diperkuat oleh Hambleton et al. (2013), yang menekankan bahwa reliabilitas tinggi memungkinkan pengukuran yang akurat pada berbagai tingkat kemampuan.

Sebagian besar item menunjukkan nilai Infit dan Outfit MNSQ dalam rentang toleransi model Rasch (0.5 hingga 1.5), menandakan kesesuaian yang baik antara pola respons aktual dengan model yang diharapkan. Hal ini mendukung hasil Aryadoust et al. (2017), yang menyoroti bahwa distribusi tingkat kesulitan yang merata mampu mengukur kemampuan peserta secara efektif. Namun, beberapa item seperti Item 13 dan Item 54 menunjukkan nilai misfit, mengindikasikan perlunya revisi pada konstruk atau redaksi butir soal untuk memastikan kesesuaian yang lebih baik dengan model Rasch.

Distribusi tingkat kesulitan item menunjukkan bahwa instrumen mencakup rentang kemampuan dari sangat mudah hingga sangat sulit. Materi seperti Struktur dan organisasi tubuh tumbuhan memiliki distribusi yang seimbang, mencerminkan instrumen yang baik untuk mengukur peserta dengan berbagai tingkat kemampuan.

Sebaliknya, materi seperti Makhluk hidup dan metode ilmiah didominasi oleh item sangat mudah (45%), yang kurang optimal untuk mengukur mahasiswa dengan kemampuan menengah hingga tinggi. Temuan ini sesuai dengan rekomendasi Hambleton et al. (2013), yang menyarankan bahwa instrumen harus memiliki distribusi tingkat kesulitan yang mencerminkan populasi peserta.

Dalam hal proporsi materi, hasil menunjukkan bahwa materi seperti Metabolisme memiliki distribusi tingkat kesulitan yang ideal dengan mayoritas item berada dalam kategori mudah (54%). Namun, materi seperti Sel sebagai dasar kehidupan memiliki proporsi item sulit yang dominan (42%) dengan sedikit item dalam kategori sedang, mengindikasikan kebutuhan revisi untuk menciptakan keseimbangan yang lebih baik. Penelitian Retnawati (2014) mendukung temuan ini, yang menyatakan bahwa proporsi tingkat kesulitan yang tidak seimbang dapat memengaruhi validitas isi instrumen.

Analisis Wright Map menunjukkan bahwa instrumen ini paling efektif untuk mengukur kemampuan mahasiswa pada tingkat menengah, tetapi terdapat celah pada ujung spektrum kemampuan. Hal ini konsisten dengan rekomendasi Linacre (2024), yang menyarankan penambahan item pada rentang kemampuan ekstrem untuk meningkatkan cakupan pengukuran. Penambahan item ini diperlukan agar instrumen dapat memberikan informasi yang lebih akurat di seluruh tingkat kemampuan (HOANG, 2021; Mardapi, 2018).

Dibandingkan dengan penelitian sebelumnya, hasil ini menunjukkan konsistensi dalam efektivitas model Rasch untuk mengevaluasi kualitas butir soal. Penelitian Sumintono dan Widhiarso (2014) serta Aryadoust et al. (2017) menegaskan bahwa distribusi tingkat kesulitan yang baik dan reliabilitas tinggi adalah kunci dalam menghasilkan instrumen yang valid dan andal (Setiawan et al., 2019). Namun, temuan ini juga menyoroti pentingnya revisi item misfit dan redistribusi tingkat kesulitan untuk meningkatkan kualitas pengukuran secara keseluruhan.

Secara keseluruhan, penelitian ini menekankan pentingnya distribusi tingkat kesulitan yang seimbang dan cakupan materi yang merata dalam menciptakan instrumen yang efektif untuk evaluasi pembelajaran. Temuan ini memberikan kontribusi signifikan untuk mendukung pengembangan instrumen yang valid, andal, dan relevan dengan kebutuhan pendidikan Biologi Umum saat ini.

KESIMPULAN

Penelitian ini berhasil mengevaluasi kualitas butir soal Ujian Tengah Semester Biologi Umum menggunakan analisis Rasch, yang memberikan wawasan mendalam tentang reliabilitas, tingkat kesulitan, dan kesesuaian item dengan model. Dengan nilai reliabilitas item yang tinggi (0.97), instrumen ini terbukti konsisten dalam mengukur kemampuan mahasiswa, sesuai dengan prinsip pengukuran yang andal. Sebagian besar item menunjukkan nilai Infit dan Outfit dalam batas toleransi, menunjukkan kesesuaian yang baik dengan model Rasch, meskipun beberapa item misfit memerlukan revisi. Distribusi tingkat kesulitan item menunjukkan cakupan yang cukup baik, tetapi terdapat kebutuhan untuk menyeimbangkan proporsi tingkat kesulitan pada beberapa materi, seperti Makhluk hidup dan metode ilmiah, yang didominasi oleh item sangat mudah. Analisis unidimensionalitas menunjukkan bahwa instrumen ini valid dalam mengukur satu konstruk utama, dengan dimensi tunggal yang kuat. Selain itu, Wright Map mengungkapkan efektivitas instrumen dalam

mengukur kemampuan mahasiswa pada tingkat menengah, meskipun diperlukan penambahan item untuk kemampuan ekstrem guna meningkatkan cakupan pengukuran.

Temuan ini memiliki implikasi penting bagi pengembangan instrumen evaluasi pendidikan, terutama dalam mata kuliah Biologi Umum. Instrumen yang lebih seimbang dalam proporsi tingkat kesulitan dan cakupan materi tidak hanya meningkatkan validitas dan reliabilitas pengukuran tetapi juga mendukung pembelajaran yang lebih bermakna. Oleh karena itu, rekomendasi utama dari penelitian ini adalah melakukan revisi pada item misfit untuk memastikan kesesuaian yang lebih baik dengan model Rasch, serta menambahkan item pada rentang kemampuan ekstrem untuk meningkatkan cakupan pengukuran. Penelitian lanjutan juga disarankan untuk menguji efektivitas instrumen yang telah direvisi dalam konteks populasi yang lebih luas atau pada berbagai jenjang pendidikan. Dengan demikian, instrumen ini dapat menjadi alat yang lebih komprehensif dan akurat dalam mengevaluasi kemampuan mahasiswa, sekaligus mendukung pengambilan keputusan yang berbasis data dalam pendidikan tinggi.

DAFTAR PUSTAKA

Andrich, D., Marais, I., Andrich, D., & Marais, I. (2019). Violations of the assumption of Independence I—multidimensionality and response dependence. *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences*, 173–185.

Aryadoust, V. (2017). Understanding the Role of Likeability in the Peer Assessments of University Students' Oral Presentation Skills: A Latent Variable Approach. *Language Assessment Quarterly*, *14*(4), 398–419. https://doi.org/10.1080/15434303.2017.1393820

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3th ed.). Psychology Press.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Elvira, M., Retnawati, H., Rohaeti, E., & Sainuddin, S. (2023). Measurement of Students' Chemistry Practicum Skills Using Many Facets Rash Model. *European Journal of Educational Research*, *12*(3).

Fox, J.-P. (2020). Special issue on item response theory in medical studies. In *Statistical methods in medical research* (Vol. 29, Issue 4, pp. 959–961). SAGE Publications Sage UK: London, England.

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.

HOANG, N. T. (2021). An Instrument For Assessing Local Enterprise's Internal Capacity In Linkage With Foreign Direct Investment. *Journal of Contemporary Issues in Business and Government*, *27*(4), 50–60. https://doi.org/10.47750/cibg.2021.27.04.008

Holster, T. A., & Lake, J. (2016). Guessing and the Rasch model. *Language Assessment Quarterly*, *13*(2), 124–141.

Linacre, J. M. (2010). Predicting responses from rasch measures. *Journal of Applied Measurement*, *11*(1), 1–10.

Linacre, J. M. (2024). *A User's Guide to WINSTEPS MINISTEP Rasch-Model Computer Programs*. winsteps.com.

Mardapi, D. (2018). *Teknik penyusunan Instrumen tes dan nontes*. Parama Publisihing.

Naga, D. S. (1992). *Pengantar teori sekor pada pengukuran pendidikan*. Gunadarma.

Retnawati, H. (2014). *Teori Respons Butir dan Penerapannya: untuk Peneliti, Praktisi Pengukuran, dan Mahasiswa.* (Vol. 15). Parama Publishing.

Sainuddin, S. (2018). Analisis Karakteristik Butir Tes Matematika Berdasarkan Teori Modern (Teori Respon Butir). *Jurnal Penelitian Matematika Dan Pendidikan Matematika*, *1*(1), 1–12.

Setiawan, A., Mardapi, D., Supriyoko, & Andrian, D. (2019). The development of instrument for assessing students' affective domain using self- and peer-assessment models. *International Journal of Instruction*, *12*(3), 425–438. https://doi.org/10.29333/iji.2019.12326a

Sumintono, Bambang&Widhiarso, W. (2014). *Aplikasi Model Rasch untuk Penelitian Ilmu-Ilmu Sosial* (Edisi Nove). Trim Komunikata Publishing House.

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi Model Rasch : untuk Penelitian Ilmu-Ilmu Sosial (Revisi) [Application of the Rasch Model: For Social Sciences Research (Revised Edition)]*. Trim Komunikata Publishing House.

Zhou, L., Almutairi, A. R., Alsaid, N. S., Warholak, T. L., & Cooley, J. (2017). Establishing the validity and reliability evidence of preceptor assessment of student tool. *American Journal of Pharmaceutical Education*, *81*(8), 10–20. https://doi.org/10.5688/ajpe5908