JURNAL STATISTIKA DAN SAINS DATA
Volume 3 Edisi 1, Oktober, 2025, 37-54
https://journal.student.uny.ac.id/index.php/jssd
https://dx.doi.org/10.21831/jssd.v3i1.22498

Analisis soal tes asesmen madrasah Bahasa Indonesia kelas VI menggunakan teori respon butir

Mazidah Azzahra*, Program Studi Statistika, Universitas Negeri Yogyakarta
Heri Retnawati, Program Studi Statistika, Universitas Negeri Yogyakarta
*mazidahazzahra.2020@student.uny.ac.id

Abstrak. Penelitian ini bertujuan untuk mengetahui model yang paling cocok dan mengetahui karakteristik butir-butir soal dalam teori respon butir (IRT) dengan perangkat tes kemampuan siswa pada Asesmen Madrasah (AM) Bahasa Indonesia Kelas VI. Data yang digunakan adalah data dikotomi berupa hasil jawaban AM Bahasa Indonesia Kelas VI yang ada di beberapa madrasah di Kabupaten Wonosobo sebanyak 35 butir soal dengan total 355 siswa. Dilakukan estimasi parameter butir dengan memilih model terbaik di antara model Rasch, 1PL, 2PL, 3PL, dan 4PL. Hasil penelitian menunjukkan bahwa: (1) Model 2PL merupakan model terbaik pada penelitian ini karena memiliki jumlah soal cocok paling banyak yaitu 33 soal, (2) Uji asumsi pada penelitian ini terpenuhi, (3) Hasil perhitungan karakteristik daya beda (*a*) menunjukkan terdapat 1 butir soal yang kurang baik, sedangkan 34 butir soal lainnya sudah baik. Hasil perhitungan karakteristik tingkat kesulitan butir (*b*) menunjukkan terdapat 3 butir soal yang kurang baik, sedangkan 32 butir soal lainnya sudah baik. Hasil pengujian berdasarkan model 2PL IRT menunjukkan terdapat 4 butir soal kurang baik dan 31 butir soal sudah baik.

Kata kunci: butir soal, perangkat tes, asesmen madrasah, bahasa Indonesia, teori respon butir.

Abstract. *This study aims to determine the most suitable model and determine the characteristics of the items in item response theory (IRT) with the student ability test device on the Grade VI Indonesian Language Madrasah Assessment (MA). The data used is dichotomous data in the form of answers to the MA Indonesian Language Grade VI in several madrassas in Wonosobo Regency as many as 35 items with a total of 355 students. Item parameter estimation was carried out by selecting the best model among Rasch, 1PL, 2PL, 3PL, and 4PL models. The results showed that: (1) The 2PL model is the best model because it has the most number of suitable questions, namely 33 questions, (2) The assumption test in this study is fulfilled, (3) The results of the calculation of the characteristics of differentiation (a) show that there is 1 item that is not good, while 34 items are good. The results of the calculation of item difficulty characteristics (b) show that there are 3 items that are not good, while 32 items are good. The test results based on the 2PL IRT model show that there are 4 items that are not good and 31 items are good.*

Keywords: item, test device, madrasah assessment, Indonesian language, item response theory.

PENDAHULUAN

Berdasarkan Undang-Undang No. 20 Tahun 2003 tentang Sistem Pendidikan Nasional dijelaskan bahwa pendidikan merupakan suatu usaha sadar dan terencana untuk mewujudkan suasana belajar dan proses pembelajaran agar peserta didik secara aktif dapat mengembangkan potensi dirinya untuk memiliki kekuatan spiritual, keagamaan, pengendalian diri, kepribadian, akhlak mulia, serta keterampilan yang diperlukan dirinya, masyarakat, bangsa, dan negara. Pendidikan di tingkat madrasah memegang peranan penting dalam pembentukan karakter dan kecerdasan siswa sejak usia dini. Bahasa Indonesia, sebagai mata pelajaran utama, berfungsi tidak hanya untuk meningkatkan kemampuan komunikasi siswa, tetapi juga untuk mengembangkan keterampilan berpikir kritis dan analitis. Oleh karena itu, evaluasi yang efektif terhadap pencapaian kompetensi siswa dalam mata pelajaran ini menjadi krusial.

Selama ini, penilaian hasil belajar di madrasah sering kali dilakukan dengan menggunakan metode tradisional, yang berfokus pada aspek kuantitatif seperti skor dan rata-rata nilai. Namun, metode ini memiliki keterbatasan dalam mengukur kualitas dan efektivitas soal asesmen secara mendalam. Banyak masalah yang muncul, seperti ketidakmampuan soal dalam membedakan tingkat kemampuan siswa, adanya soal yang terlalu mudah atau terlalu sulit, serta kurangnya validitas dan reliabilitas tes. Ada enam hal yang perlu dipertimbangkan dalam perencanaan tes yaitu: pengambilan sampel dan pemilihan butir soal, tipe tes yang digunakan, aspek yang akan diuji, format butir soal, jumlah butir soal, dan distribusi tingkat kesukaran butir soal (Zainul dan Nasoetion, 1997).

Menurut Hambelton dan Van Der Linden (1982), analisis pada perangkat tes dapat terbagi dalam dua teori, yakni teori tes klasik dan teori respon butir. Teori respon butir merupakan teori analisis butir soal yang menjadi perbaikan dari kelemahan yang terdapat dalam teori klasik (Syafii dkk., 2021). Dalam konteks ini, Teori Respon Butir (*Item Response Theory*, IRT) muncul sebagai solusi alternatif yang lebih komprehensif. IRT menawarkan pendekatan yang dapat mengevaluasi kualitas soal secara lebih terperinci dengan mengukur parameter-parameter seperti tingkat kesulitan, daya beda, dan tebakan acak. Metode ini memungkinkan untuk mendapatkan gambaran yang lebih akurat mengenai bagaimana setiap item soal berfungsi dan bagaimana kemampuan siswa dapat diukur secara lebih objektif.

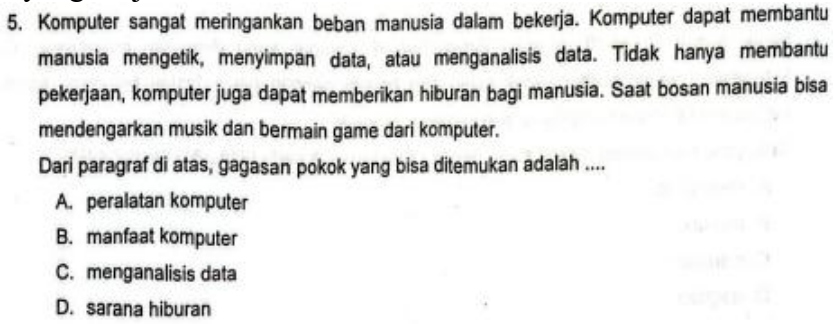
Pada penelitian ini, dilakukan perbandingan kecocokan model analisis respon butir data dikotomi hasil Asesmen Madrasah Bahasa Indonesia Kelas VI dengan model Rasch, 1PL, 2PL, 3PL, dan 4PL. Asesmen Madrasah adalah asesmen sumatif yang diselenggarakan untuk peserta didik kelas akhir jenjang madrasah. Istilah asesmen sendiri muncul belakangan, sebelumnya seringkali para ahli dalam buku-buku evaluasi pembelajaran menggunakan istilah tes bahasa (Susilo dkk, 2021). Hasil perhitungan masing-masing model pada penelitian ini akan dibandingkan berdasarkan jumlah butir cocok yang paling banyak. Model terpilih dalam penelitian ini adalah model 2PL. Hal ini sesuai dengan penelitian yang dilakukan oleh Setiawati dkk (2022) tentang analisis parameter tes Penilaian Akhir Semester Fisika kelas X dengan teori respon butir. Setelah didapatkan model terbaik, dilakukan uji asumsi teori respon butir meliputi uji asumsi unidimensi, invariansi parameter, dan independensi lokal.

Berdasarkan uraian di atas, penelitian ini bertujuan untuk menganalisis soal tes asesmen Bahasa Indonesia kelas VI di madrasah dengan menggunakan IRT. Dengan mengidentifikasi kelemahan dan kekuatan dalam soal asesmen yang ada, penelitian ini diharapkan dapat memberikan rekomendasi yang konstruktif untuk perbaikan dalam penyusunan soal tes di masa depan. Pendekatan ini juga diharapkan dapat meningkatkan efektivitas evaluasi dan memberikan kontribusi positif terhadap peningkatan kualitas pendidikan di madrasah.

METODE

Deskripsi Data

Penelitian ini menggunakan pendekatan kuantitatif. Populasi dalam penelitian ini adalah seluruh lembar jawab peserta Asesmen Madrasah Bahasa Indonesia Kelas VI di Kabupaten Wonosobo tahun ajaran 2023/2024. Studi kasus pada penelitian ini adalah 355 respon jawaban peserta tes terhadap 35 butir soal pilihan ganda Asesmen Madrasah Bahasa Indonesia Kelas VI yang ada di tujuh madrasah yaitu MI An Nur, MI Bendosari, MI Gunungtawang, MI Guppi, MI Klesman, MI Ropoh, dan MI Tumenggungan. Gambar 1 menunjukkan contoh soal tentang pemahaman bacaan yang diujikan dalam Asesmen Madrasah.



5. Komputer sangat meringankan beban manusia dalam bekerja. Komputer dapat membantu manusia mengetik, menyimpan data, atau menganalisis data. Tidak hanya membantu pekerjaan, komputer juga dapat memberikan hiburan bagi manusia. Saat bosan manusia bisa mendengarkan musik dan bermain game dari komputer.

Dari paragraf di atas, gagasan pokok yang bisa ditemukan adalah ....

A. peralatan komputer
B. manfaat komputer
C. menganalisis data
D. sarana hiburan

(Sumber Gambar: Soal Asesmen Madrasah Bahasa Indonesia Kelas VI Kabupaten Wonosobo Tahun Ajaran 2023/2024)

Gambar 1. Contoh Soal Pemahaman Bacaan

Teknik Analisis Data

Analisis data pada penelitian ini menggunakan bantuan program RStudio. R adalah perangkat lunak (*software*) yang berupa bahasa pemrograman yang baik untuk digunakan untuk melakukan komputasi statistika dan pembuatan grafis (R Core Team, 2022). Penelitian ini menggunakan beberapa *package* dari RStudio, seperti *CTT* (Willse, 2018), *mirt* (Chalmers, 2012), *lavaan* (Rosseel, 2012), FactoMineR (Le dkk., 2008), *factoextra* (Kassambra dan Mundt, 2020), dan *ltm* (Rizopoulos, 2006).

Analisis data dilakukan dengan tahapan sebagai berikut: (1) menyiapkan data berupa lembar jawab peserta Asesmen Madrasah Bahasa Indonesia Kelas VI, (2) menguji kecocokan model, yaitu model Rasch, IPL, 2PL, 3PL, dan 4PL, (3) membandingkan hasil kecocokan model dari kelima pemodelan berdasarkan jumlah butir yang cocok paling banyak pada masing-masing model sehingga diperoleh model terbaik, (4) melakukan uji asumsi teori respon butir meliputi uji asumsi unidimensi, invariansi parameter, dan independensi lokal dengan pengambilan keputusan berdasarkan plot pencar dari masing-masing uji asumsi, (5) melakukan analisis karakteristik butir soal berdasarkan model terbaik, dan (6) menghitung nilai fungsi informasi dan SEM pada hasil analisis berdasarkan model terbaik.

Model Rasch adalah suatu model yang memungkinkan untuk menyusun item berdasarkan tingkat kesulitan dan kemampuan peserta (Hambleton, Swaminathan, dan Rogers, 1991).

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

dengan:

i : $1,2,3,\ldots,m$
j : $1,2,3,\ldots,n$
m : Banyaknya butir tes
n : Banyaknya peserta tes
$P(X_{ij} = 1)$: Probabilitas jawaban peserta tes ke-j menjawab benar pada butir ke-i

b_i : Parameter kesulitan butir ke-i
e : Konstanta dengan nilai berkisar 2,718
θ_j : Parameter kemampuan peserta ke-j

Model 1PL menawarkan kerangka yang sederhana untuk analisis data, memungkinkan peneliti untuk memfokuskan perhatian pada kesulitan item tanpa mempertimbangkan diskriminasi (Bond dan Fox, 2015).

$$P(X_{ij} = 1|\theta_j) = \frac{e^{(Da\{\theta_j - b_i\})}}{1+e^{a\{\theta_j - b_i\}}} \quad (2)$$

dengan:
i : $1,2,3,\ldots,m$
j : $1,2,3,\ldots,n$
m : Banyaknya butir tes
n : Banyaknya peserta tes
$P(X_{ij} = 1)$: Probabilitas jawaban peserta tes ke-j menjawab benar pada butir ke-i
b_i : Parameter kesulitan butir ke-i
e : Konstanta dengan nilai berkisar 2,718
D : Konstanta dengan nilai berkisar 1,7
θ_j : Parameter kemampuan peserta ke-j
a : Parameter daya beda butir untuk semua butir (*discrimination*)

Model 2PL mempertimbangkan dua parameter penting, yaitu kemampuan peserta dan diskriminasi item, yang memberikan pemahaman lebih mendalam tentang respons peserta. (Baker dan Kim, 2004).

$$P(X_{ij} = 1|\theta_j) = \frac{e^{(a_i\{\theta_j - b_i\})}}{1+e^{a_i\{\theta_j - b_i\}}} \quad (3)$$

dengan:
i : $1,2,3,\ldots,m$
j : $1,2,3,\ldots,n$
m : Banyaknya butir tes
n : Banyaknya peserta tes
$P(X_{ij} = 1)$: Probabilitas jawaban peserta tes ke-j menjawab benar pada butir ke-i
b_i : Parameter kesulitan butir ke-i
e : Konstanta dengan nilai berkisar 2,718
θ_j : Parameter kemampuan peserta ke-j
a_i : Parameter daya beda butir ke-i

Model 3PL menambahkan parameter ketiga, yaitu tingkat kemungkinan respon acak (c), yang menggambarkan probabilitas dasar bahwa individu dengan kemampuan rendah akan memberikan jawaban yang benar tanpa memperhatikan kemampuan item (Baker dan Kim, 2004).

$$P\left(X_{ij} = 1\middle|\theta_j\right) = c_i + (1 - c_i)\frac{e^{(a_i\{\theta_j - b_i\})}}{1+e^{a_i\{\theta_j - b_i\}}} \quad (4)$$

dengan:
i : $1,2,3,\ldots,m$
j : $1,2,3,\ldots,n$
m : Banyaknya butir tes

n : Banyaknya peserta tes
$P(X_{ij} = 1)$: Probabilitas jawaban peserta tes ke-j menjawab benar pada butir ke-i
b_i : Parameter kesulitan butir ke-i
e : Konstanta dengan nilai berkisar 2,718
θ_j : Parameter kemampuan peserta ke-j
a_i : Parameter daya beda butir ke-i
c_i : Parameter tebakan semu (*pseudo guessing*) butir ke-i

Model 4PL mengintegrasikan empat parameter—kesulitan (b), diskriminasi (a), respon acak (c), dan parameter asimptotik (d)—yang memungkinkan analisis item yang lebih kompleks, terutama untuk data yang mencakup variasi lebih besar dalam respon (Baker dan Kim, 2004).

$$P\left(X_{ij} = 1 \middle| \theta_j\right) = g_i + (u_i - g_i)\frac{e^{(a_i\{\theta_j - b_i\})}}{1+e^{a_i\{\theta_j - b_i\}}} \quad (5)$$

dengan:
i : $1,2,3,\ldots,m$
j : $1,2,3,\ldots,n$
m : Banyaknya butir tes
n : Banyaknya peserta tes
$P(X_{ij} = 1)$: Probabilitas jawaban peserta tes ke-j menjawab benar pada butir ke-i
b_i : Parameter kesulitan butir ke-i
e : Konstanta dengan nilai berkisar 2,718
θ_j : Parameter kemampuan peserta ke-j
a_i : Parameter daya beda butir ke-i
u_i : Batas atas atau nilai maksimum peluang siswa menjawab benar sepanjang skala θ
g_i : Parameter tebakan semu (*pseudo guessing*) butir ke-i

Asumsi unidimensi dapat ditunjukkan dari plot nilai eigen yang menunjukkan satu komponen dominan (Hambleton dkk., 1991). Jika unidimensionalitas terpenuhi maka analisis faktor dapat digunakan untuk memastikan bahwa semua item berkumpul dalam satu faktor tunggal. Jika tidak, hasil pengukuran dapat memberikan informasi yang menyesatkan.Untuk menguji asumsi ini, analisis faktor eksploratori atau konfirmatori sering digunakan. Hasil yang diharapkan adalah satu faktor dominan yang menjelaskan sebagian besar varians item. Memastikan unidimensionalitas sangat penting untuk validitas tes. Jika terdapat lebih dari satu dimensi, interpretasi dari skor total mungkin tidak mencerminkan kemampuan yang ingin diukur.

Asumsi invariansi parameter berarti karakteristik butir soal tidak tergantung pada distribusi parameter kemampuan peserta tes dan parameter yang menjadi ciri peserta tes tidak bergantung dari ciri butir soal (Retnawati, 2014). Dengan kata lain, item diharapkan memberikan informasi yang konsisten, tidak terpengaruh oleh faktor-faktor lain. Jika invariansi parameter terpenuhi, maka model pengukuran dapat digunakan secara adil untuk berbagai kelompok. Jika tidak, interpretasi skor dapat bias dan perbandingan antar kelompok menjadi tidak valid. Pengujian asumsi ini dapat dilakukan dengan menggunakan diagram pencar atau scree plot untuk parameter butir dan kemampuan. Asumsi invariansi parameter terpenuhi apabila titik-titik berpencar mengikuti garis lurus. Asumsi ini penting untuk memastikan bahwa pengukuran berlaku secara universal. Pelanggaran terhadap invariansi parameter dapat menyebabkan kesimpulan yang salah dalam analisis perbandingan antar kelompok.

Asumsi independensi lokal memiliki arti bahwa respon terhadap item tidak bergantung pada respon dari item lainnya (De Ayala, 2008). Artinya, item-item dianggap independen satu sama lain dalam hal pengukuran trait yang sama. Jika asumsi ini dilanggar, hubungan antar item dapat menyebabkan informasi tambahan yang tidak terukur sehingga mengganggu analisis dan interpretasi hasil. Uji ini dapat dilakukan dengan analisis residual. Jika residual menunjukkan pola atau hubungan maka ada kemungkinan bahwa asumsi independensi lokal telah dilanggar. Memastikan independensi lokal penting agar hasil pengukuran dapat diinterpretasikan secara akurat. Jika item-item tidak independent maka hasil dapat terdistorsi dan ini dapat memengaruhi penilaian kemampuan individu.

Tingkat kesukaran butir soal adalah proporsi peserta tes menjawab benar setiap butir soal (Zainul dan Nasoetion, 1997). Biasanya diwakili oleh parameter b, tingkat kesukaran menunjukkan kemampuan di mana probabilitas seorang individu untuk menjawab item dengan benar adalah 50%. Item dengan nilai b yang tinggi dianggap lebih sulit, sementara item dengan nilai b yang rendah dianggap lebih mudah. Pemahaman tentang tingkat kesukaran penting dalam merancang instrumen pengukuran yang efektif. Selain itu, juga berperan dalam tes adaptif, di mana item disesuaikan dengan kemampuan peserta, memastikan bahwa tes tetap menantang, tetapi tidak terlalu sulit.

Daya beda butir soal adalah indeks yang menunjukkan tingkat kemampuan suatu butir soal membedakan kelompok yang memiliki prestasi tinggi dengan kelompok yang memiliki prestasi rendah (Zainul dan Nasoetion, 1997). Dikenal sebagai parameter diskriminasi, biasanya dilambangkan dengan a. Item dengan nilai a yang tinggi berarti efektif dalam melakukan diskriminasi. Sebaliknya, item dengan nilai a yang rendah berarti kurang efektif dalam melakukan diskriminasi. Memahami daya beda sangat penting dalam desain tes karena membantu peneliti memilih item yang tidak hanya relevan, tetapi juga mampu memberikan informasi yang akurat tentang kemampuan individu.

Tebakan semu adalah probabilitas bahwa individu dengan kemampuan sangat rendah dapat menjawab item dengan benar, yang sering terjadi dalam tes dengan pilihan ganda (Embretson dan Reise, 2000). Parameter ini biasanya dilambangkan dengan c. Kehadiran parameter ini penting untuk mengatasi situasi di mana respon tidak sepenuhnya mencerminkan kemampuan, seperti dalam tes pilihan ganda di mana peserta dapat memilih jawaban secara acak. Memahami dan mengestimasi tebakan semu memungkinkan peneliti untuk menghasilkan estimasi yang lebih akurat mengenai kemampuan individu dan meningkatkan validitas keseluruhan instrumen pengukuran.

Fungsi informasi item membantu dalam mengidentifikasi item yang memiliki kemampuan tinggi dalam memberikan informasi, serta item yang mungkin perlu direvisi atau dihapus (DeMars, 2010). Semakin tinggi fungsi informasi suatu item, semakin baik item tersebut dalam mengidentifikasi kemampuan responden. Fungsi informasi biasanya dihitung berdasarkan parameter item, seperti kesulitan dan diskriminasi serta dapat diilustrasikan dalam bentuk kurva. Persamaan fungsi informasi pada butir soal, yaitu:

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \tag{6}$$

dengan:

i : $1, 2, 3, \dots, n$
$I_i(\theta)$: fungsi informasi butir ke- i
$P_i(\theta)$: peluang peserta dengan kemampuan θ menjawab benar pada butir ke- i
$P_i'(\theta)$: turunan fungsi $P_i(\theta)$ terhadap θ
$Q_i(\theta)$: peluang peserta dengan kemampuan θ menjawab benar pada butir ke- i

Menurut Hancock dan Mueller (2013), SEM menyediakan kerangka yang fleksibel untuk menganalisis data pengukuran dan menguji hubungan antara kemampuan latent dan variabel observasi dalam konteks IRT. SEM mengintegrasikan analisis faktor dan analisis regresi sehingga dapat digunakan untuk mengkonfirmasi struktur pengukuran serta menguji model teoritis yang melibatkan beberapa variabel. Persamaan SEM adalah sebagai berikut:

$$SEM(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} \tag{7}$$

dengan:
$SEM(\hat{\theta})$: nilai estimasi SEM
$I(\theta)$: nilai fungsi informasi

HASIL DAN PEMBAHASAN

Hasil

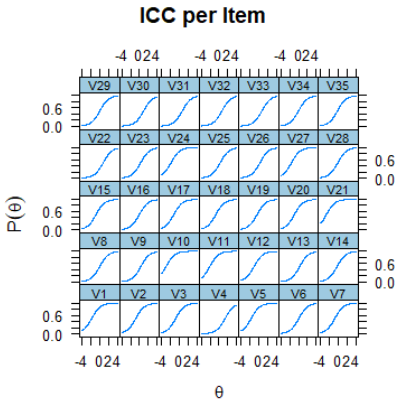
Uji kecocokan model dengan model Rasch, IPL, 2PL, 3PL, dan 4PL digunakan untuk mengestimasi hasil pengukuran kemampuan tes Asesmen Madrasah. Selain itu, data juga akan diestimasi menggunakan grafik ICC (*Item Characteristic Curve*). Semua data dalam penelitian ini layak digunakan karena tidak ada butir soal dengan nilai biserial negatif.

Tabel 1. Hasil Uji Kecocokan Model Rasch

	Butir	Chi-squre	p-value	Keputusan
1	b1	19,796	0,285	Cocok
2	b2	37,956	0,004	Tidak Cocok
3	b3	15,730	0,611	Cocok
4	b4	15,791	0,607	Cocok
5	b5	25,327	0,088	Cocok
6	b6	26,386	0,120	Cocok
7	b7	27,347	0,073	Cocok
8	b8	24,624	0,173	Cocok
9	b9	16,896	0,597	Cocok
10	b10	9,050	0,617	Cocok
11	b11	26,991	0,042	Tidak Cocok
12	b12	19,502	0,362	Cocok
13	b13	26,945	0,106	Cocok
14	b14	15,130	0,714	Cocok
15	b15	24,722	0,133	Cocok
16	b16	17,388	0,564	Cocok
17	b17	23,440	0,174	Cocok
18	b18	36,907	0,005	Tidak Cocok
19	b19	25,536	0,144	Cocok
20	b20	13,940	0,733	Cocok
21	b21	22,925	0,062	Cocok
22	b22	35,142	0,009	Tidak Cocok
23	b23	44,427	0,001	Tidak Cocok
24	b24	34,485	0,007	Tidak Cocok
25	b25	48,611	0,000	Tidak Cocok
26	b26	29,602	0,057	Cocok
27	b27	13,880	0,676	Cocok
28	b28	9,902	0,955	Cocok
29	b29	24,496	0,178	Cocok
30	b30	15,005	0,662	Cocok
31	b31	40,295	0,002	Tidak Cocok
32	b32	17,853	0,465	Cocok
33	b33	37,291	0,005	Tidak Cocok
34	b34	20,046	0,392	Cocok
35	b35	13,492	0,812	Cocok

Model Rasch mengestimasi parameter tingkat kesulitan (*b*) dan kemampuan peserta (Hambleton, Swaminathan, dan Rogers, 1991). Nilai diskriminan atau daya beda (*a*) pada

model Rasch adalah 1. Tabel 1 menunjukkan bahwa terdapat 26 butir soal yang cocok dan 9 butir soal tidak cocok menggunakan model Rasch. Butir soal dikatakan cocok apabila p-value > 0,05, sedangkan butir soal dikatakan tidak cocok apabila p-value < 0,05. Butir soal yang tidak cocok yaitu b2, b11, b18, b22, b23, b24, b25, b31, dan b33.



Gambar 2. Plot ICC Model Rasch Setiap Butir Soal

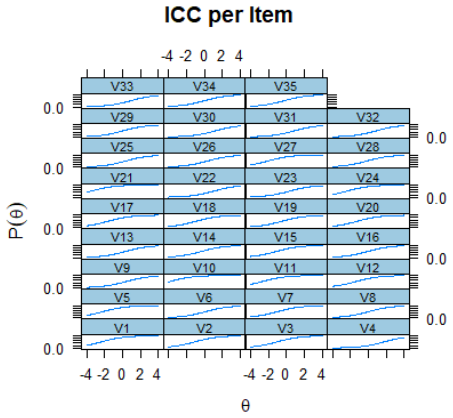
Pada Gambar 2, sumbu *x* mewakili tingkat kemampuan siswa (θ) dan sumbu *y* mewakili probabilitas peserta tes menjawab benar pada butir tertentu. Secara umum, kurva ICC berbentuk S yang menunjukkan bahwa pada tingkat kemampuan rendah, probabilitas siswa menjawab benar sangat rendah lalu meningkat dan mendekati nilai 1 pada kemampuan tinggi. Pada Grafik ICC model Rasch terlihat mayoritas pola yang sama, tetapi terdapat beberapa pola yang tidak sama, namun grafik ICC tersebut masih dapat diterima.

Tabel 2. Hasil Uji Kecocokan Model 1PL

	Butir	Chi-squre	p-value	Keputusan
1	b1	19,797	0,285	Cocok
2	b2	37,956	0,004	Tidak Cocok
3	b3	15,730	0,611	Cocok
4	b4	15,791	0,607	Cocok
5	b5	25,328	0,088	Cocok
6	b6	26,387	0,120	Cocok
7	b7	27,346	0,073	Cocok
8	b8	24,625	0,173	Cocok
9	b9	16,897	0,597	Cocok
10	b10	9,052	0,617	Cocok
11	b11	26,993	0,042	Tidak Cocok
12	b12	19,502	0,362	Cocok
13	b13	26,945	0,106	Cocok
14	b14	15,130	0,714	Cocok
15	b15	24,722	0,133	Cocok
16	b16	17,388	0,564	Cocok
17	b17	23,440	0,174	Cocok
18	b18	36,908	0,005	Tidak Cocok
19	b19	25,536	0,144	Cocok
20	b20	13,940	0,733	Cocok
21	b21	22,927	0,061	Cocok
22	b22	35,143	0,009	Tidak Cocok
23	b23	44,428	0,001	Tidak Cocok
24	b24	34,488	0,007	Tidak Cocok
25	b25	48,612	0,000	Tidak Cocok
26	b26	29,603	0,057	Cocok
27	b27	13,880	0,676	Cocok
28	b28	9,902	0,955	Cocok
29	b29	24,497	0,178	Cocok
30	b30	15,005	0,662	Cocok
31	b31	40,297	0,002	Tidak Cocok
32	b32	17,854	0,465	Cocok

33	b33	37,292	0,005	Tidak Cocok
34	b34	20,046	0,392	Cocok
35	b35	13,492	0,812	Cocok

Model IPL mengestimasi parameter tingkat kesulitan (*b*), tanpa mempertimbangkan diskriminasi (Bond dan Fox, 2015). Tabel 2 menunjukkan bahwa terdapat 26 butir soal yang cocok dan 9 butir soal tidak cocok menggunakan model 1PL. Butir soal dikatakan cocok apabila p-value > 0,05, sedangkan butir soal dikatakan tidak cocok apabila p-value < 0,05. Butir soal yang tidak cocok yaitu b2, b11, b18, b22, b23, b24, b25, b31, dan b33.



Gambar 3. Plot ICC Model 1PL Setiap Butir Soal

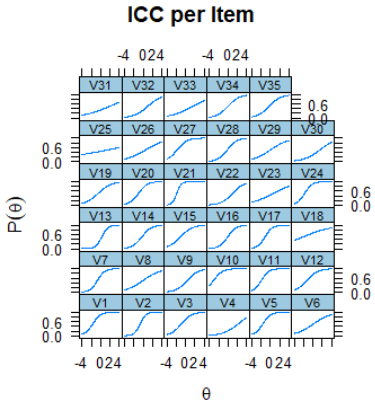
Gambar 3 menunjukkan bahwa grafik pada setiap butir soal ada yang sudah mengikuti ogif normal, tetapi masih sangat terlihat beberapa butir yang tidak mengikuti ogif normal. Hal tersebut berarti belum dapat disimpulkan bahwa model 1PL adalah model terbaik dan harus dilakukan perhitungan dengan model lain terlebih dahulu.

Tabel 3. Hasil Uji Kecocokan Model 2PL

	Butir	Chi-squre	p-value	Keputusan
1	b1	18,807	0,279	Cocok
2	b2	24,597	0,077	Cocok
3	b3	15,428	0,632	Cocok
4	b4	12,200	0,877	Cocok
5	b5	22,311	0,133	Cocok
6	b6	17,830	0,534	Cocok
7	b7	22,019	0,184	Cocok
8	b8	20,455	0,430	Cocok
9	b9	16,455	0,627	Cocok
10	b10	7,322	0,695	Cocok
11	b11	22,741	0,090	Cocok
12	b12	23,388	0,176	Cocok
13	b13	13,675	0,550	Cocok
14	b14	13,336	0,771	Cocok
15	b15	24,190	0,149	Cocok
16	b16	14,182	0,654	Cocok
17	b17	22,583	0,125	Cocok
18	b18	29,666	0,075	Cocok
19	b19	26,220	0,159	Cocok
20	b20	15,150	0,652	Cocok
21	b21	7,340	0,693	Cocok
22	b22	34,014	0,013	Tidak Cocok
23	b23	31,598	0,064	Cocok
24	b24	35,506	0,002	Tidak Cocok
25	b25	24,090	0,343	Cocok
26	b26	23,014	0,288	Cocok
27	b27	14,319	0,644	Cocok
28	b28	11,248	0,915	Cocok

29	b29	20,292	0,440	Cocok
30	b30	12,986	0,839	Cocok
31	b31	21,685	0,418	Cocok
32	b32	17,113	0,646	Cocok
33	b33	27,937	0,142	Cocok
34	b34	18,997	0,457	Cocok
35	b35	13,690	0,801	Cocok

Model 2PL mengestimasi tingkat kesulitan (*b*) dan daya beda (*a*) (Baker dan Kim, 2004). Tabel 3 menunjukkan bahwa terdapat 33 butir soal yang cocok dan 2 butir soal tidak cocok menggunakan model 2PL. Butir soal dikatakan cocok apabila p-value > 0,05, sedangkan butir soal dikatakan tidak cocok apabila p-value < 0,05. Butir soal yang tidak cocok yaitu b22 dan b24.



Gambar 4. Plot ICC Model 2PL Setiap Butir Soal

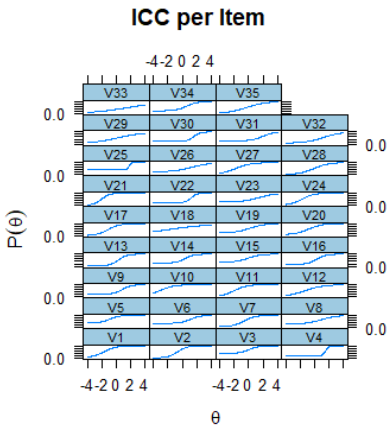
Gambar 4 menunjukkan bahwa grafik pada setiap butir soal ada yang sudah mengikuti ogif normal, tetapi masih sangat terlihat beberapa butir yang berupa garis datar yang berarti memiliki kemungkinan butir tersebut kurang sensitif terhadap tingkat kemampuan peserta tes.

Tabel 4. Hasil Uji Kecocokan Model 3PL

	Butir	Chi-squre	p-value	Keputusan
1	b1	17,410	0,295	Cocok
2	b2	23,753	0,049	Tidak Cocok
3	b3	15,881	0,532	Cocok
4	b4	7,892	0,980	Cocok
5	b5	21,876	0,111	Cocok
6	b6	15,631	0,682	Cocok
7	b7	21,436	0,162	Cocok
8	b8	19,853	0,403	Cocok
9	b9	14,590	0,690	Cocok
10	b10	4,735	0,857	Cocok
11	b11	22,262	0,051	Cocok
12	b12	18,088	0,383	Cocok
13	b13	14,992	0,452	Cocok
14	b14	13,625	0,627	Cocok
15	b15	23,321	0,139	Cocok
16	b16	13,845	0,678	Cocok
17	b17	21,929	0,080	Cocok
18	b18	30,155	0,050	Tidak Cocok
19	b19	26,382	0,120	Cocok
20	b20	13,290	0,651	Cocok
21	b21	6,978	0,639	Cocok
22	b22	28,332	0,057	Cocok
23	b23	32,333	0,040	Tidak Cocok
24	b24	38,119	0,000	Tidak Cocok
25	b25	22,823	0,298	Cocok
26	b26	22,246	0,272	Cocok

27	b27	16,674	0,407	Cocok
28	b28	9,593	0,944	Cocok
29	b29	17,697	0,543	Cocok
30	b30	11,169	0,918	Cocok
31	b31	20,270	0,441	Cocok
32	b32	17,210	0,576	Cocok
33	b33	27,144	0,131	Cocok
34	b34	17,295	0,503	Cocok
35	b35	13,388	0,768	Cocok

Model 3PL pada umumnya mengestimasi daya beda (*a*), tingkat kesulitan (*b*), dan tebakan semu (*pseudo guessing*). Tabel 4 menunjukkan bahwa terdapat 31 butir soal yang cocok dan 4 butir soal tidak cocok menggunakan model 3PL. Butir soal dikatakan cocok apabila p-value > 0,05, sedangkan butir soal dikatakan tidak cocok apabila p-value < 0,05. Butir soal yang tidak cocok yaitu b2, b18, b23, dan b24.



Gambar 5. Plot ICC Model 3PL Setiap Butir Soal

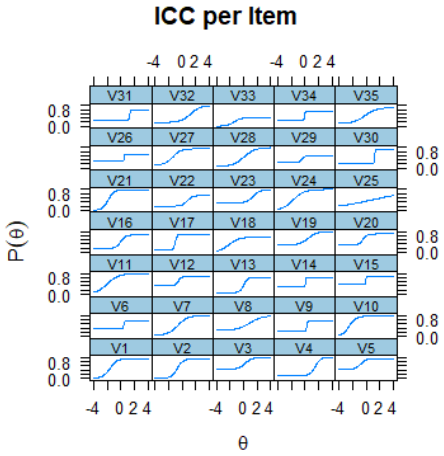
Gambar 5 menunjukkan bahwa grafik pada setiap butir soal ada yang sudah mengikuti ogif normal, tetapi masih sangat terlihat beberapa butir yang terlihat curam yang menunjukkan kemungkinan bahwa butir soal tersebut hanya bisa mendeteksi kemampuan peserta tes pada rentang tertentu saja.

Tabel 5. Hasil Uji Kecocokan Model 4PL

	Butir	Chi-squre	p-value	Keputusan
1	b1	16,363	0,358	Cocok
2	b2	24,967	0,035	Tidak Cocok
3	b3	16,201	0,439	Cocok
4	b4	7,009	0,983	Cocok
5	b5	24,311	0,060	Cocok
6	b6	13,447	0,764	Cocok
7	b7	21,844	0,112	Cocok
8	b8	20,842	0,287	Cocok
9	b9	12,221	0,787	Cocok
10	b10	5,544	0,785	Cocok
11	b11	22,305	0,051	Cocok
12	b12	14,855	0,606	Cocok
13	b13	18,895	0,219	Cocok
14	b14	15,259	0,644	Cocok
15	b15	19,696	0,290	Cocok
16	b16	15,308	0,573	Cocok
17	b17	17,863	0,270	Cocok
18	b18	29,817	0,039	Tidak Cocok
19	b19	27,287	0,054	Cocok
20	b20	14,420	0,567	Cocok
21	b21	5,832	0,666	Cocok
22	b22	30,986	0,029	Tidak Cocok

23	b23	32,424	0,020	Tidak Cocok
24	b24	35,833	0,001	Tidak Cocok
25	b25	24,270	0,186	Cocok
26	b26	22,648	0,205	Cocok
27	b27	14,094	0,592	Cocok
28	b28	12,501	0,709	Cocok
29	b29	18,443	0,427	Cocok
30	b30	12,196	0,788	Cocok
31	b31	24,638	0,135	Cocok
32	b32	17,224	0,439	Cocok
33	b33	26,519	0,088	Cocok
34	b34	18,959	0,394	Cocok
35	b35	16,363	0,684	Cocok

Model 4PL mengestimasi daya beda (*a*), tingkat kesulitan (*b*), tebakan semu (*pseudo guessing*), dan *carelessness*. Model 4PL merupakan pengembangan model 3PL. Tabel 5 menunjukkan bahwa terdapat 30 butir soal yang cocok dan 5 butir soal tidak cocok menggunakan model 4PL. Butir soal dikatakan cocok apabila p-value > 0,05, sedangkan butir soal dikatakan tidak cocok apabila p-value < 0,05. Butir soal yang tidak cocok yaitu b2, b18, b22, b23, dan b24.



Gambar 6. Plot ICC Model 4PL Setiap Butir Soal

Gambar 6 menunjukkan bahwa grafik pada setiap butir soal ada yang sudah mengikuti ogif normal, tetapi masih sangat terlihat beberapa butir yang terlihat curam yang menunjukkan kemungkinan bahwa butir soal tersebut hanya bisa mendeteksi kemampuan peserta tes pada rentang tertentu saja.

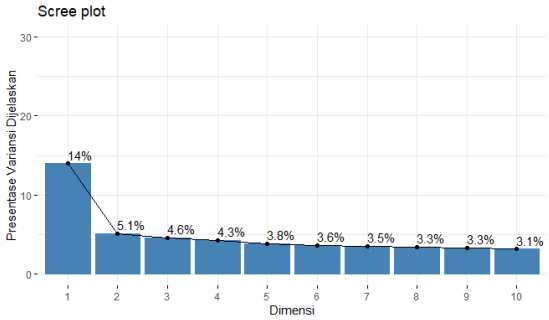
Hasil uji kecocokan model berdasarkan teori respon butir pada penelitian ini menunjukkan hasil yang berbeda antara model Rasch, IPL, 2PL, 3PL, dan 4PL. Model dengan jumlah butir cocok paling banyak dianggap sebagai model terbaik untuk penelitian ini berdasarkan teori respon butir.

Tabel 6. Perbandingan Hasil Uji Kecocokan Model

	Model Rasch	Model 1PL	Model 2PL	Model 3PL	Model 4PL
Cocok	26	26	33	31	30
Tidak Cocok	9	9	2	4	5

Berdasarkan tabel 6, model terbaik menurut teori respon butir yaitu model 2PL dengan jumlah butir soal cocok paling banyak di antara model yang lainnya. Pada model 2PL, terdapat

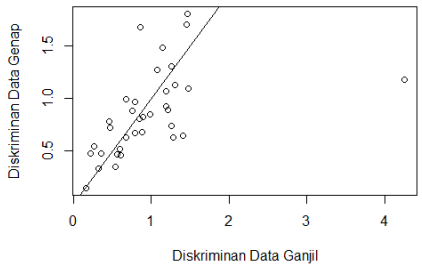
33 butir soal cocok dan 2 butir soal tidak cocok dengan model. Jadi, dapat disimpulkan bahwa model terbaik adalah model 2PL.



Gambar 7. Scree Plot Data

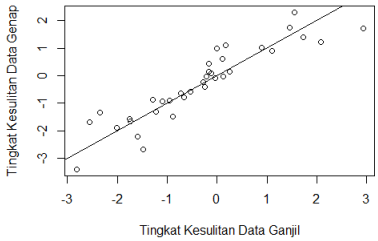
Asumsi unidimensi dapat ditunjukkan dari plot nilai eigen yang menunjukkan satu komponen dominan (Hambleton dkk., 1991). Gambar 7 menunjukkan adanya dominasi satu data dengan data lainnya serta terdapat satu titik elbow di mana terdapat penurunan yang semakin sedikit pada data tersebut. Hal ini menunjukkan bahwa asumsi unidimensi terpenuhi.

Asumsi invariansi parameter berarti karakteristik butir soal tidak tergantung pada distribusi parameter kemampuan peserta tes dan parameter yang menjadi ciri peserta tes tidak bergantung dari ciri butir soal (Retnawati, 2014).



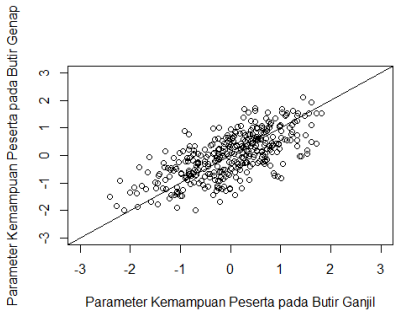
Gambar 8. Invariansi Parameter Butir Daya Beda

Gambar 8 menunjukkan adanya titik-titik yang menyebar mengikuti garis lurus. Hal tersebut menunjukkan bahwa asumsi invariansi parameter butir daya beda terpenuhi.



Gambar 9. Plot Uji Asumsi Invariansi Parameter Tingkat Kesulitan

Gambar 9 menunjukkan adanya titik-titik yang menyebar mengikuti garis lurus. Hal tersebut menunjukkan bahwa asumsi invariansi parameter butir tingkat kesulitan terpenuhi.



Gambar 10. Plot Uji Asumsi Invariansi Parameter Kemampuan

Gambar 10 menunjukkan adanya titik-titik yang menyebar mengikuti garis lurus. Hal tersebut menunjukkan bahwa asumsi invariansi parameter butir kemampuan terpenuhi.

Tabel 7. Nilai Korelasi Butir Asesmen Madrasah Bahasa Indonesia

	b1	b2	b3	b4	b5	b6
b1	1	0,051	-0,092	-0,002	0,047	-0,078
b2	0,051	1	-0,101	0,004	-0,061	-0,143
b3	-0,092	-0,101	1	0,058	-0,069	0,059
b4	-0,002	0,004	0,058	1	0,074	0,080
b5	0,047	-0,061	-0,069	0,074	1	0,027
b6	-0,078	-0,143	0,059	0,080	0,027	1

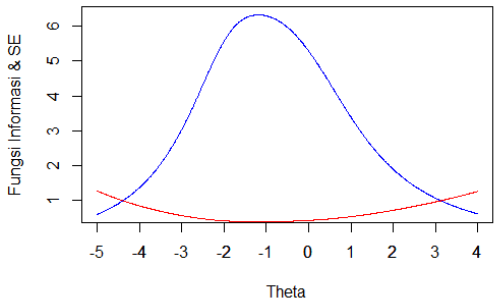
Asumsi independensi lokal memiliki arti bahwa respon terhadap item tidak bergantung pada respon dari item lainnya (De Ayala, 2008). Berdasarkan hasil korelasi pada tabel 7, nilai korelasi pada penelitian ini kurang dari 0,5 sehingga dapat dikatakan bahwa asumsi independensi lokal terpenuhi. Asumsi independensi lokal dapat dikatakan terpenuhi apabila asumsi unidimensi sudah terpenuhi. Jadi, pada penelitian ini, asumsi independensi lokal terpenuhi.

Tabel 8. Karakteristik Butir Berdasarkan Model Terbaik

	a	*b*	Keterangan
b1	1,1915006	-1,73257564	Baik
b2	1,6322495	-0,67710939	Baik
b3	0,9243674	-0,92389761	Baik
b4	0,6197325	2,15197810	Kurang baik
b5	1,1711517	-1,65039225	Baik
b6	0,5423043	0,12012309	Baik
b7	1,2731355	-0,73485683	Baik
b8	0,5415109	-0,03541242	Baik
b9	0,8493798	0,35855467	Baik
b10	1,0277557	-3,10923051	Kurang baik
b11	1,2292192	-1,93924742	Baik
b12	0,8531293	-1,30173119	Baik
b13	1,5506932	0,20596331	Baik
b14	0,9831186	-0,30278581	Baik
b15	0,7994703	-1,08475713	Baik
b16	1,1163946	-0,06305001	Baik
b17	1,2393067	-1,16841209	Baik
b18	0,3232252	-2,09131542	Kurang baik
b19	0,7143415	-0,56237506	Baik
b20	1,0470417	-1,01237072	Baik
b21	2,0258924	-1,89064683	Kurang baik
b22	0,8150895	0,97394561	Baik
b23	0,4092793	0,61747403	Baik
b24	1,2336564	-1,73516800	Baik
b25	0,1514974	1,61594151	Baik
b26	0,4588103	0,52947920	Baik

b27	0,9333543	-1,79931934	Baik
b28	0,9798406	-0,13655845	Baik
b29	0,5234659	0,05534354	Baik
b30	0,5887536	1,53587966	Baik
b31	0,3327606	1,87983717	Baik
b32	0,6664220	0,98279683	Baik
b33	0,4153948	1,58809504	Baik
b34	0,8206879	-0,01189712	Baik
b35	0,7804235	-0,25821355	Baik

Parameter daya beda (a) dikatakan baik apabila berada dalam interval $0 < a < 2$, sedangkan parameter kesukaran butir (b) dikatakan baik apabila nilainya berada dalam interval $-2 < b < 2$ (Hambleton dan Swaminathan, 1985). Jika nilai b_i mendekati 2, tingkat kesulitan soal semakin tinggi. Apabila nilai b_i mendekati -2, tingkat kesulitan soal semakin rendah. Berdasarkan tabel 8, terdapat 4 butir soal yang kurang baik, sedangkan 31 butir soal lainnya sudah baik. Butir soal yang kurang baik yaitu b4, b10, b18, dan b21.



Gambar 11. Plot Fungsi Informasi dan SEM

Fungsi informasi adalah suatu fungsi yang memberikan estimasi mengenai kemampuan responden dari model pada teori respon butir (Mulvia dkk, 2021). *Standard Error of Measurement* (SEM) digunakan untuk mengukur nilai kesalahan dalam suatu pengukuran. Gambar 11 menunjukkan bahwa nilai fungsi informasi maksimum dari penilaian Asesmen Madrasah Bahasa Indonesia dengan 35 butir soal adalah 6,2 pada kemampuan siswa (θ) sekitar -1,1 dan SEM sebesar 0,4. Nilai fungsi informasi akan lebih tinggi dari SEM ketika θ berada di antara -4,4 dan 3,1. Hal ini berarti perangkat tes kemampuan siswa yang digunakan dalam penelitian ini dapat memberikan informasi mengenai kemampuan peserta tes pada kategori sangat rendah hingga tinggi.

Pembahasan

Penelitian ini bertujuan untuk mengestimasi karakteristik butir soal pada perangkat kemampuan siswa pada Asesmen Madrasah menggunakan teori respon butir. Model yang digunakan adalah model *Rasch*, 1PL, 2PL, 3PL, dan 4PL. Dari semua model tersebut, dicari model terbaik dengan melihat jumlah butir yang cocok. Pada penelitian ini, model terpilih adalah model 2PL karena memiliki jumlah soal cocok paling banyak yaitu 33 butir soal. Model lainnya memiliki jumlah butir cocok yang lebih sedikit dibandingkan dengan model 2PL.

Hasil ini sejalan dengan penelitian Setiawati dkk (2022) tentang analisis parameter tes Penilaian Akhir Semester Fisika kelas X dengan teori respon butir dengan hasil bahwa model 2PL merupakan model terbaik karena memiliki jumlah butir cocok terbanyak dan nilai fungsi informasi tertinggi. Namun, hasil penelitian ini tidak sejalan dengan penelitian Astuti (2019) tentang analisis soal tes penilaian akhir semester fisika kelas X menggunakan teori respon butir dengan hasil bahwa model 1PL sebagai model yang fit karena memiliki jumlah butir cocok terbanyak.

Hasil model terbaik dalam teori respon butir dapat bervariasi di setiap penelitian karena beberapa faktor. Pertama, karakteristik sampel yang digunakan dapat memengaruhi bagaimana responden menjawab item-item dalam tes. Kedua, jenis dan jumlah item yang dianalisis juga berperan penting; item yang lebih relevan atau tepat dapat meningkatkan akurasi model. Selain itu, metode estimasi parameter yang dipilih juga dapat menghasilkan hasil yang berbeda. Ketiga, konteks dan tujuan penelitian juga dapat memengaruhi pemilihan model yang dianggap terbaik. Terakhir, variabel eksternal seperti waktu dan lingkungan sosial yang berubah juga dapat memengaruhi hasil.

Setelah semua pengujian asumsi terpenuhi, dilanjutkan pengujian untuk melihat bagaimana karakteristik perangkat tes kemampuan siswa berdasarkan model 2PL teori respon butir. Pada pengujian ini didapati 4 butir soal yang kurang baik dan 31 butir soal yang dapat dikatakan baik. Adanya butir-butir soal yang kurang baik menunjukkan bahwa analisis mengenai karakteristik butir soal pada Asesmen Madrasah penting dilakukan.

Hasil perhitungan daya beda pada penelitian ini menunjukkan terdapat satu butir soal yang kurang baik karena nilai daya beda tidak berada dalam interval $0 < a < 2$. Butir soal tersebut yaitu butir soal nomor b21. Pada butir soal tersebut, nilai $a > 2$ sehingga tingkat daya bedanya masih belum baik. Selain itu, berdasarkan tingkat kesulitannya, penelitian ini menunjukkan terdapat tiga butir soal yang kurang baik karena nilai tingkat kesulitan tidak berada dalam interval $-2 < b < 2$. Butir soal tersebut meliputi butir soal nomor b4, b10, dan b18. Pada butir soal nomor b10 dan b18 terlihat nilai tingkat kesulitan yang cukup rendah dan menunjukkan bahwa soal pada butir-butir tersebut terlalu mudah, sedangkan pada soal nomor b4 terdapat nilai tingkat kesulitan yang cukup tinggi yang menunjukkan soal terlalu sulit. Hal ini sejalan dengan dengan penelitian Setiawati dkk (2022) di mana secara keseluruhan, soal penilaian akhir semester fisika kelas X memiliki nilai parameter a (daya pembeda) yang dapat dikategorikan baik karena berada pada rentang 0 sampai dengan 2 dan nilai parameter b (tingkat kesulitan) yang dapat dikategorikan baik karena berada pada rentang -2 sampai dengan 2.

Nilai fungsi informasi pada penelitian ini adalah 6,2 pada kemampuan siswa (θ) sekitar -1,1 dan SEM sebesar 0,4. Nilai fungsi informasi akan lebih tinggi dari SEM ketika θ berada di antara -4,4 dan 3,1. Hal ini berarti perangkat tes kemampuan siswa yang digunakan dalam penelitian ini dapat memberikan informasi mengenai kemampuan peserta tes pada kategori sangat rendah hingga tinggi. Hal ini sejalan dengan penelitian Wardhani (2024) dimana nilai fungsi informasi tertinggi sebesar 18,51 pada kemampuan siswa (θ) sebesar 1,07 dan SEM sebesar 1. Selain itu, nilai fungsi informasi akan lebih tinggi dari SEM ketika θ berkisar -2,8 hingga 1,8. Hal ini berarti hasil ujian tulis dapat memberikan informasi mengenai kemampuan peserta tes pada kategori sangat rendah hingga tinggi.

SIMPULAN

Berdasarkan hasil dari penelitian ini, diperoleh kesimpulan sebagai berikut: (1) model 2PL merupakan model terbaik pada penelitian ini karena memiliki jumlah soal cocok paling banyak yaitu 33 soal, (2) uji asumsi pada penelitian ini meliputi asumsi unidimensi, invariansi parameter, dan independensi lokal. Semua asumsi terpenuhi sehingga analisis dapat dilakukan, (3) hasil perhitungan karakteristik daya beda (a) menunjukkan terdapat 1 butir soal yang kurang baik, sedangkan 34 butir soal lainnya sudah baik. Hasil perhitungan karakteristik tingkat kesulitan butir (b) menunjukkan terdapat 3 butir soal yang kurang baik, sedangkan 32 butir soal lainnya sudah baik. Hasil pengujian berdasarkan model 2PL teori respon butir menunjukkan bahwa terdapat sebanyak 4 butir soal kurang baik dan 31 butir soal sudah baik. Perhitungan ini diambil dari irisan butir soal yang sudah baik pada dua karakteristik yang diestimasi oleh model 2PL teori respon butir.

DAFTAR PUSTAKA

Astuti, H. L. (2019). Analisis soal tes penilaian akhir semester fisika kelas x menggunakan teori respon butir. (Skripsi Sarjana, Universitas Negeri Yogyakarta). https://eprints.uny.ac.id.

Baker, F. B., & Kim, S.-H. (Eds.). (2004). *Item Response Theory: Parameter Estimation Techniques, Second Edition* (2nd ed.). CRC Press. https://doi.org/10.1201/9781482276725.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge/Taylor & Francis Group.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, 48(6), 1-29. https://doi.org/10.18637/jss.v048.i06.

De Ayala, R. J. (2008). *The theory and practice of item response theory.* The Guilford Press.

DeMars, C. (2010). *Item Response Theory: Understanding Statistic Measurement.* New York: Oxford University Press, Inc.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* London: Lawrence Erlbaum Assciates Publishers.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory.* Boston, MA: Kluwer.

Hambleton, R. K., & van der Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378.

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

Hancock, G. R., & Mueller, R. O. (Eds.). (2013). *Structural equation modeling: A second course* (2nd ed.). IAP Information Age Publishing.

Kassambra, A., Mundt, F. (2020). *factoextra: Extract and visualize the results of multivariate data analyses*. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra.

Le, S., Josse, J., Husson, F. (2008). FactoMineR: An r package for multivariate analysis. *Journal of Statistical Software*, 25(1), 1-18. https://doi.org/10.18637/jss.v025.i01.

Mulvia, R., Ramalis, T.R., & Efendi, R., (2021). Mendeteksi keajegan butir tes dengan fungsi informasi. *Jurnal Pendidikan Indonesia*, 2(1), 72-84. https://doi.org/10.59141/japendi.v2i01.66.

R Core Team. 2022. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Republik Indonesia. (2003). *Undang-Undang RI Nomor 20, Tahun 2003*, tentang Sistem Pendidikan Nasional.

Retnawati, H. (2014). *Teori respons butir dan penerapannya: untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Nuha Medika.

Rizopoulos, D. (2006). Itm: An r package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1-25. URL http://www.jstatsoft.org/v17/i05/.

Rosseel, Y. (2012). *lavaan: An r package for structural equation modeling*. *Journal of Statistical Software*, 48(2), 1-36. https://doi.org/10.18637/jss.v048.i02.

Setiawati, M. N., Purwanto, & Ramalis, T., R. (2022). Analisis parameter tes penilaian akhir semester fisika kelas x dengan teori respon butir. *Wahana Pendidikan Fisika*, 7(1), 1-10. https://doi.org/10.17509/wapfi.v7i1.43089.

Susilo, dkk. (2021). *Teori asesmen dalam pembelajaran bahasa*. Surabaya: Global Aksara Pres.

Syafii, A., dkk. (2021). Analysis of items with item response theory (IRT) approach on final assessment for Al-Quran hadith subjects. *Jurnal Pendidikan Agama Islam*, 18(1), 167-194. https://doi.org/10.14421/jpai.2021.181-09.

Wardhani, N. S. (2024). *Perbandingan kecocokan model analisis butir data dikotomi dengan teori respon butir*. [Skripsi, tidak diterbitkan]. Universitas Negeri Yogyakarta.

Willse, J. T. (2018). *CTT: Classical test theory functions*. R package version 2.3.3. https://CRAN.R-project.org/package=CTT.

Zainul, A., & Nasoetion. (1997). *Penilaian Hasil Belajar*. Pusat Antar Universitas, Direktorat Jenderal Pendidikan Tinggi: Departemen Pendidikan dan Kebudayaan.