Available at http://ejournal.unp.ac.id/index.php/jipe

JIPE
Jurnal Inovasi Pendidikan Ekonomi
Fakultas Ekonomi, Universitas Negeri Padang

PRINTED ISSN 2302-898X
ONLINE ISSN 2621-5624
Vol. 8 No. 1, April 2018
Page 22 - 31

ESTIMASI PARAMETER TES DENGAN PENSKORAN POLITOMUS MENGGUNAKAN GRADED RESPONSE MODEL PADA SAMPEL KECIL

Friyatmi
Program Studi S3 Universitas Negeri Yogyakarta/Dosen Universitas Negeri Padang
Email: friyatmi.spd2015@student.uny.ac.id

Abstract

Multiple choice tests with dichotomous scores are very often used by educators because it is easy to apply. Unfortunately, this scoring lacks an opportunity for educators to diagnose students' mistakes. This study aimed to analyze the items test using graded responses model (GRM). The data in this study was the response of the economics test that have been tested to 96 first-year students of economic education. The test form was multiple choice item using politomous scoring with ordinal scale 1-2-3-4-5. Test scoring use polytomous scores. Data was analyzed with descriptive quantitative techniques using PARSCALE application. The results showed that there are 50% items test that fit with the GRM model. Furthremore, the application of GRM model in this study seems less suitable because only 20% that have good quality. Big sample size is recommended when using the GRM model to obtain accurate estimation.

Keyword*: polytomous, graded responses model, IRT.*

PENDAHULUAN

Tes merupakan salah satu alat bantu yang sering digunakan pendidik untuk mengukur ketercapaian kompetensi peserta didik. Terdapat beberapa model pengukuran yang biasanya dipakai dalam melakukan analisis butir tes, baik menggunakan teori tes klasik maupun *Item Responses Theory*. Pemilihan model yang tepat akan mengungkap keadaan yang sebenarnya dari data tes sebagai hasil pengukuran. Salah satu bentuk tes yang familiar digunakan oleh pendidik pada level sekolah menengah adalah tes berbentuk pilihan berganda dengan penskoran dikotomi. Bentuk penskoran dikotomi memiliki skor yang ekstrim dimana jawaban yang benar diberi skor 1 dan jawaban yang salah diberi 0 (Bond & Fox, 2007; DeMars, 2010). Kelebihan penskoran dikotomi ini adalah memberi kemudahan bagi pendidik dalam pemeriksaan dan penskoran tes, namun kurang memberi kesempatan bagi pendidik untuk mendiagnosis kesalahan konsep yang dilakukan oleh peserta didik (Isgiyanto, 2011). Kelemahan ini dapat diminimalisir dengan mengembangkan penskoran politomus. Penskoran politomus pada tes berarti bahwa skor jawaban dikembangkan menjadi lebih dari dua kategori, bukan hanya benar atau salah (Demars, 2003). Teknik penskoran ini memungkinkan guru untuk mendiagnosis kesalahan peserta tes sehingga dapat dijadikan rekomendasi untuk perbaikan pembelajaran.

Model analisis IRT yang biasanya digunakan untuk menganalisis butir data berbentuk politomus diantaranya *Nominal Respons Model* (NRM), *Rating Scale Model* (RSM), *Partial Credit Model* (PCM), *Graded Respon Model* (GRM) dan *Generalized Partial Credit Model* (GPCM) (Retnawati, 2014). Setiap model memiliki karakteristik khas sendiri yang membedakan satu dengan yang lainnya. Salah satu model yang dapat memberikan informasi maksimal dari parameter butir adalah model GRM. Model GRM digunakan untuk respon yang bersifat kategorikal dan merupakan pengembangan model IRT 2-PL yang mampu mengungkap dua parameter butir, yaitu tingkat kesukaran dan daya pembeda. Model GRM memungkinkan diterapkan pada tes yang pilihan jawabannya memiliki gradasi. Beberapa penelitian memperlihatkan bahwa penggunaan model GRM pada sampel yang besar dapat memberikan estimasi parameter yang akurat (Demars, 2003; Edelen & Reeve, 2007). Apakah keakuratan juga akan diperoleh jika model GRM digunakan untuk menganalisis *classroom test* dengan sampel yang kecil? Penelitian ini bertujuan untuk menganalisis tes dengan penskoran politomus menggunakan model *Graded Responses* pada sampel kecil. Keefektifan model GRM akan dievaluasi saat menggunakan sampel kecil. Analisis tes bertujuan untuk mengetahui karakteristik/kualitas butir yang baik secara empiris menggunakan *Graded Response Model*.

METODE PENELITIAN

Penelitian ini termasuk penelitian eksploratif yang menganalisis karakteristik butir tes berdasarkan pendekatan IRT menggunakan model GRM. Penelitian ini menggunakan 20 item tes ekonomi berbentuk *multiple choice* yang diujikan kepada 96 mahasiswa tahun pertama Pendidikan Ekonomi FE UNP. Penskoran tes menggunakan teknik penskoran politomus dengan skala ordinal 1-2-3-4-5.

Analisis item menggunakan IRT haruslah memenuhi asumsi yang disyaratkan. Asumsi yang umum digunakan secara luas oleh model-model IRT ialah asumsi *unidimensional*, *local independent* dan *invarian parameter*. Menurut Brennan dan Kolen (2004: 156) asumsi unidimensi artinya setiap butir tes hanya mengukur satu kemampuan. Ini bermakna bahwa probabilitas suatu respon butir adalah sebagai suatu fungsi karakteristik laten tunggal peserta ujian. Tes yang telah diukur diharapkan hanya mengukur satu karakter atau kemampuan saja. Untuk memenuhi asumsi unidimensi faktor yang paling dominan mempengaruhi kinerja tes dibandingkan dengan tujuan disusunnya suatu tes. Apabila faktor dominan yang muncul sudah sesuai dengan tujuan disusunnya suatu tes maka asumsi unidimensi telah terpenuhi. Pengujian asumsi unidimensional dianalisis menggunakan *exploratory factor analysis* dengan bantuan program SPSS.

Lokal independen dipahami sebagai independensi semua peserta ujian dari butir tes di dalam subpopulasi. Independensi lokal dengan demikian dipahami sebagai skor komposit suatu butir yang diberikan oleh subpopulasi peserta ujian yang homogen yang independen. Ini berarti pula bahwa respon terhadap dua butir tes tidak saling berkorelasi di dalam subpopulasi homogen. Menurut Retnawati (2014) asumsi unidimensi dapat dideteksi dengan membuktikan asumsi unidimensional.

Asumsi invariansi parameter dipahami sebagai sebuah fungsi dari karakteristik parameter peserta ujian atau butir tes yang tidak akan berubah di dalam subpopulasi meskipun subpopulasi tersebut berubah. Sebagaimana yang diungkapkan oleh Naga (1992: 173) bahwa fungsi atau karakteristik butir adalah tetap sekalipun kelompok peserta yang menjawab butir yang sama itu berubah-ubah, dan untuk kelompok yang sama, ciri mereka adalah tetap sekalipun butir yang mereka jawab berubah-ubah. Hal ini bermakna bahwa parameter-parameter yang menjadi ciri suatu butir tes tidak bergantung pada distribusi karakteristik peserta ujian dan parameter yang menjadi ciri peserta ujian tidak bergantung pada perangkat butir tes.

Analisis terhadap butir tes menggunakan bantuan program PARSCALE yang memuat dua parameter, yaitu parameter butir dan parameter peserta. Parameter butir menggunakan model GRM yang mengungkap dua pameter butir yaitu tingkat kesukaran (*threshold*) dan daya pembeda (*slope*). Hasil estimasi parameter butir dapat dilihat pada output program PARSCALE fase 2, sedangkan estimasi parameter peserta dapat dilihat pada output Parscale fase 3.

Untuk mengetahui kualitas item dapat digunakan ketentuan dari Hulin, Drasgow dan Parsons (1983: 16-25) dengan kriteria baik, kurang baik, dan tidak baik. Kriteria masing-masing parameter sebagai berikut:

a. Item tergolong baik apabila item cocok dengan model, memiliki indeks kesukaran -2.0 s/d 2.0, indeks daya pembeda butir 0.0 s/d 2.0,.
b. Item tergolong kurang baik apabila item cocok dengan model, memiliki indeks kesukaran <-2.0 atau >2.0 dan indeks daya pembeda butir >2.0.
c. Item tergolong tidak baik, jika tidak cocok dengan model.

HASIL DAN PEMBAHASAN

Langkah awal sebelum dilakukan estimasi parameter tes adalah melakukan pengujian asumsi IRT yang meliputi uji asumsi unidimensi, independensi lokal, dan invariansi parameter. Pengujian unidimensi dilakukan untuk mengetahui apakah tes yang digunakan mengukur satu macam trait. Uji asumsi unidimensi dilakukan melalui analisis faktor menggunakan program SPSS. Salah satu hal yang perlu diperhatikan dalam melakukan analisis faktor adalah terpenuhinya kecukupan sampel. Untuk mengetahui kecukupan sampel dapat dilihat dari nilai Chi-Square pada uji Barlett.

Tabel 1. Hasil uji *Bartlett test of sphericity* dan KMO-MSA

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.811
Bartlett's Test of Sphericity	Approx. Chi-Square	769.896
	Df	190
	Sig	0.000

Sumber : Data diolah 2017

Berdasarkan hasil analisis, dengan nilai Chi-square 769,896 dengan df 190 dan signifikansi <0,01 hasil ini menunjukkan ukuran sampel sebanyak 96 pada analisis ini telah mencukupi. Untuk mendapatkan item-item yang mengukur dimensi yang sama, dilakukan proses ekstraksi sehingga dihasilkan beberapa faktor. Banyak faktor yang terbentuk ditunjukkan oleh komponen yang mempunyai eigenvalue >1 yang terlihat dalam tabel berikut:

Tabel 2. Hasil Ekstraksi Analisis Faktor

Total Variance Explained

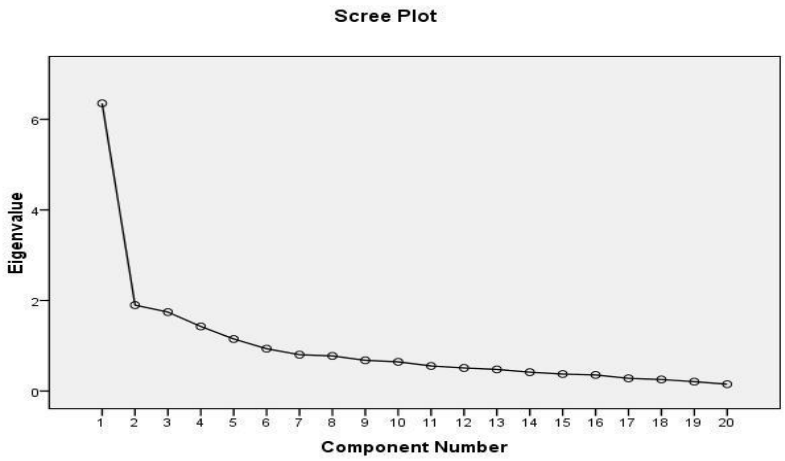
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.353	31.764	31.764	6.353	31.764	31.764	3.416	17.081	17.081
2	1.898	9.491	41.255	1.898	9.491	41.255	2.452	12.261	29.342
3	1.744	8.718	49.973	1.744	8.718	49.973	2.421	12.104	41.446
4	1.427	7.136	57.109	1.427	7.136	57.109	2.390	11.948	53.395
5	1.150	5.749	62.858	1.150	5.749	62.858	1.893	9.463	62.858
6	.935	4.677	67.535						
7	.803	4.016	71.551						
8	.777	3.884	75.435						
9	.679	3.397	78.831						
10	.646	3.232	82.063						
11	.554	2.771	84.834						
12	.511	2.555	87.389						
13	.478	2.390	89.779						
14	.417	2.086	91.865						
15	.375	1.877	93.743						
16	.356	1.778	95.521						
17	.280	1.402	96.923						
18	.256	1.278	98.201						
19	.208	1.040	99.241						
20	.152	.759	100.000						

Extraction Method: Principal Component Analysis.

Sumber : Data diolah 2017

Hasil analisis faktor menunjukkan bahwa terdapat 5 faktor yang nilai eigennya lebih dari 1, sehingga dapat dikatakan bahwa dari 20 item yang dianalisis mengelompok ke dalam 5 faktor. Kelima faktor tersebut menjelaskan sekitar 62,858% dari total varians. Hasil analisis juga menunjukkan bahwa faktor pertama dapat menjelaskan 31,764% dari total varians. Eigenvalue faktor pertama nilainya lebih dari dua kali eigenvalue faktor kedua, sehingga dapat dikatakan bahwa faktor-faktor tersebut telah membentuk faktor yang dominan. Sebagaimana yang dinyatakan oleh Naga (1992: 297) kalau eigenvalue faktor pertama nilai beberapa kali nilai eigenvalue faktor kedua, sedangkan eigenvalue faktor kedua dan seterusnya adalah hampir sama maka dapat dikatakan bahwa syarat unidimensi sudah terpenuhi.

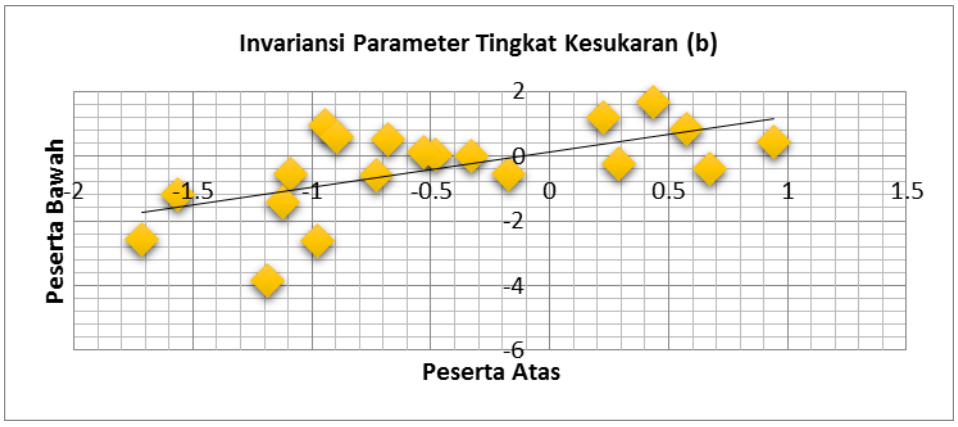
Apabila dilihat dari component matrix maka jumlah item yang berkumpul pada faktor pertama sudah dominan yaitu sebanyak 15 item dari 20 item yang dianalisis (75%). Gambaran yang lebih jelas dari sifat unidimensi perangkat ini dapat dilihat pada scree plot berikut:



Gambar 1. Scree Plot Analisis Faktor

Berdasarkan hasil analisis faktor dan diperjelas dengan scree plot di atas, maka dapat disimpulkan bahwa asumsi unidimensi telah dapat terpenuhi pada perangkat yang dianalisis ini, meskipun dengan standar yang tidak terlalu ketat. Karena pada dasarnya sangat sulit memenuhi syarat unidimensi secara ketat, sebagaimana yang dinyatakan oleh Hambleton & Swaminathan (1985: 17) "pada praktiknya asumsi unidimensi sulit untuk dipenuhi secara ketat karena adanya faktor lain seperti faktor kognitif, personality, faktor administrasi dalam tes, seperti kecemasan, dan motivasi". Terpenuhinya asumsi unidimensi melalui analisis faktor di atas, maka secara tidak langsung juga sudah membuktikan terpenuhinya asumsi independensi lokal (Retnawati, 2014).

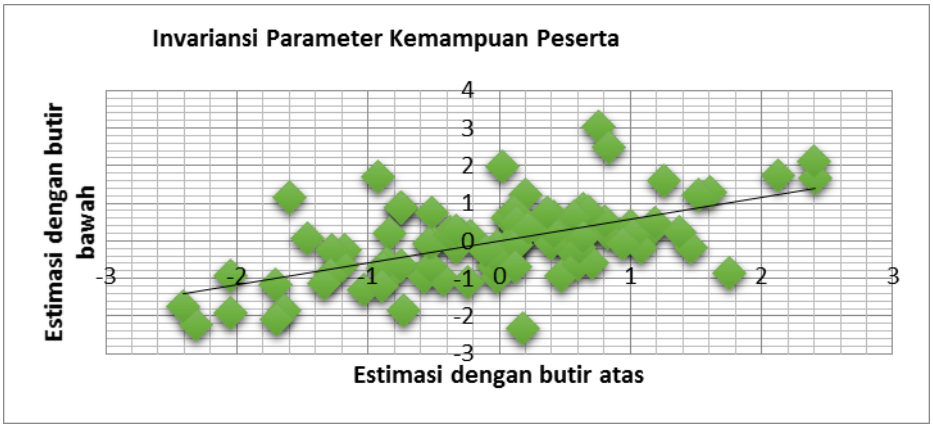
Uji invariansi parameter bertujuan untuk mengetahui apakah karakteristik item tidak berubah meskipun dijawab oleh kelompok siswa yang berbeda, begitu juga sebaliknya. Sehingga terdapat dua uji invariansi parameter yakni invariansi parameter butir dan invariansi parameter kemampuan peserta. Pengujian invariansi ini dapat dapat dilihat melalui diagram pencar untuk tiap parameter butir dan kemampuan peserta tes. Pengujian invariansi parameter butir dilihat dengan mengelompokan peserta tes menjadi 2 yaitu kelompok siswa peserta atas dan bawah. Begitu juga dengan pengujian invariansi parameter kemampuan peserta menggunakan separo butir bagian atas dan separo butir bagian bawah.



Gambar 2. Invariansi Parameter Butir

Estimasi parameter tes dengan penskoran politomus Menggunakan graded response model pada sampel kecil

Berdasarkan Gambar 2 terlihat bahwa masing-masing titik berada relatif dekat dengan garis linear, hal ini menjukkan bahwa tidak terjadi variasi parameter butir pada 2 kelompok peserta tes tersebut. Demikian juga dengan invariansi parameter kemampuan peserta.



Gambar 3. Invariansi Parameter Kemampuan Peserta

Gambar 3 membuktikan bahwa estimasi kemampuan peserta menggunakan belahan separo butir atas dan bawah memperlihatkan kemampuan peserta mengumpul relatif dekat dengan garis yang berarti tidak terjadi variasi parameter kemampuan peserta. Berdasarkan hasil analisis, maka diperoleh hasil bahwa invariansi parameter butir dan parameter kemampuan peserta telah terbukti.

Terpenuhinya pengujian ketiga asumsi tersebut berimplikasi bahwa tes ini layak dianalisis menggunakan pendekatan teori respon butir (IRT). Teori respon butir memuat dua parameter, yaitu parameter butir dan parameter peserta. Parameter ciri peserta θ menyatakan ciri peserta dengan kemampuan θ, sedangkan parameter butir dinyatakan melalui model yang digunakan dalam analisis. Dalam kasus ini, sesuai dengan model data yang berbentuk ordinal, maka model yang digunakan adalah *Graded Response* Model dengan bantuan program PARSCALE. Model ini memberikan estimasi parameter butir dalam bentuk daya pembeda (a_i) dan indeks kesukaran item (b_i). Hasil estimasi parameter menggunakan program PARSCALE dapat dilihat pada tabel berikut.

Tabel 3. Analisis Statistik Estimasi Parameter Butir

PARAMETER	MEAN	STN DEV	N
Slope	0,438	0,122	20
Log (slope)	-0,864	0,293	20
Threshold	2,120	1,620	20
Guessing	0,000	0,000	0

Sumber : Data diolah 2017

Data tabel di atas memperlihatkan bahwa secara keseluruhan perangkat yang dianalisis memiliki rata-rata estimasi parameter daya pembeda sebesar 0,438 dan estimasi tingkat kesukaran 2,120. Secara umum, hasil ini memperlihatkan bahwa estimasi daya pembeda item berada pada kategori yang baik, sedangkan estimasi tingkat kesukaran berada pada kategori yang sukar karena indeksnya >2. Namun, untuk lebih jelasnya berikut dideskripsikan masing-masing estimasi parameter butir untuk setiap item.

Tingkat kesukaran butir merupakan fungsi dari kemampuan seseorang (Mardapi, 1991: 11). Seseorang yang memiliki kemampuan tinggi akan merasa mudah mengerjakan butir soal, sebaliknya mereka yang memiliki kemampuan rendah akan merasa sulit menjawab butir soal. Tingkat kesukaran butir bergerak dari skala $-\infty \leq b \leq \infty$ pada teori respon butir. Tapi pada prakteknya butir yang dinyatakan baik adalah butir yang memiliki tingkat kesukaran (bi) berkisar diantara $-2 \leq b \leq +2$. Butir yang memiliki tingkat kesukaran dekat atau di bawah skala -2 menunjukkan butir soal tersebut termasuk kategori mudah. Sedangkan butir yang memiliki

tingkat kesukaran (b) dekat atau terletak di atas skala +2,00 menunjukkan butir soal tersebut termasuk kategori sukar (Hambleton, Swaminathan,& Rogers, 1991: 13).

Tabel 4. Hasil Analisis Parameter Indeks Kesukaran Item

Tingkat kesukaran	Jumlah Butir
Sukar	10
Baik	10
Mudah	0
	20

Sumber : Data diolah 2017

Berdasarkan hasil analisis, maka terdapat 10 butir soal (50%) yang memiliki karakteristik baik. Sisanya sebanyak 10 butir soal memiliki karakteristik yang kurang baik karena memiliki indeks kesukaran >2, yang termasuk soal yang tergolong sulit.

Parameter indeks daya beda (a_i) adalah kemiringan kurva karakteristik butir di titik bi pada skala kemampuan tertentu. Karena merupakan kemiringan, berarti semakin besar kemiringannya maka semakin besar indeks daya beda butir tersebut. Secara teoritis daya beda butir terletak pada skala $-\infty \leq a \leq \infty$. Namun dalam prakteknya nilai ai terletak antara 0 sampai 2 (Hambleton, Swaminathan & Rogers, 1991: 15). Berdasarkan hasil analisis, maka semua butir soal memiliki daya pembeda yang baik karena semua item memiliki indeks daya pembeda antara 0 – 2 sebagaimana terlihat pada Tabel 5.

Tabel 5. Hasil Analisis Indeks Daya Beda Item

Daya Pembeda	Jumlah Butir
Baik	20
Kurang Baik	0
Jumlah	20

Sumber : Data diolah 2017

Prosedur selanjutnya dalam analisis tes adalah melakukan pengujian kecocokan model. Item yang cocok dengan model adalah item dengan nilai probabilitas chi-square yang signifikan, yaitu item yang memiliki probabilitas chi-square ≥0.05. Berdasarkan nilai probabilitas chi-square maka dapat disimpulkan jumlah item yang cocok pada masing-masing model. Berdasarkan hasil analisis, hanya terdapat 10 butir yang fit dengan model GRM, sebagaimana tersaji pada tabel berikut.

Tabel 6. Hasil Analisis Kecocokan Model

Kecocokan Model	Jumlah Butir
Cocok	10
Tidak cocok	10

Sumber : Data diolah 2017

Berdasarkan analisis fit model dan estimasi parameter di atas, maka dapat ditentukan berapa banyak item yang baik yang memenuhi kriteria IRT dengan menggunakan Graded Response Model. Untuk mengetahui kualitas item dapat digunakan ketentuan dari Hulin, Drasgow dan Parsons (1983: 16-25) dengan kriteria baik, kurang baik, dan tidak baik. Kriteria masing-masing parameter adalah item tergolong baik apabila item cocok dengan model, memiliki indeks kesukaran -2.0 s/d 2.0, indeks daya pembeda butir 0.0 s/d 2.0. Item tergolong kurang baik apabila item cocok dengan model, memiliki indeks kesukaran <-2.0 atau >2.0 dan indeks daya

pembeda butir >2.0. Item tergolong tidak baik, jika tidak cocok dengan model. Hasil analisis kualitas butir soal sesuai dengan kriteria tersebut disajikan pada tabel berikut.

Tabel 7. Kesimpulan Hasil Analisis Estimasi Parameter Butir

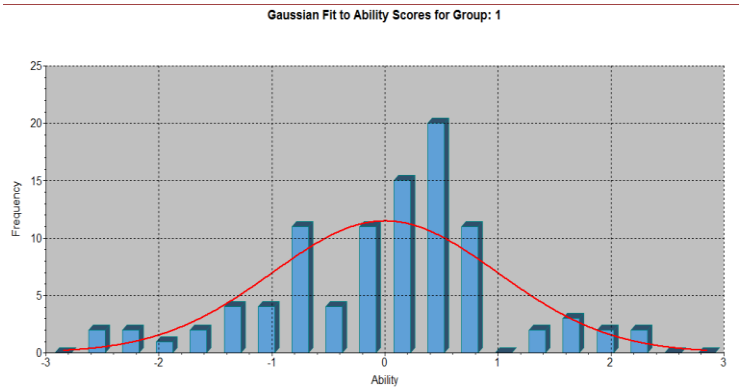
No. Item	Tingkat sukaran	Daya Pembeda	Model Fit	Simpulan
1	0.649	0.613	Cocok	Baik
2	2.372	0.464	Tidak Cocok	Tidak Baik
3	2.046	0.468	Cocok	Kurang Baik
4	4.94	0.261	Tidak Cocok	Tidak Baik
5	-0.135	0.456	Tidak Cocok	Tidak Baik
6	6.326	0.732	Tidak Cocok	Tidak Baik
7	0.517	0.398	Tidak Cocok	Tidak Baik
8	3.411	0.392	Tidak Cocok	Tidak Baik
9	0.052	0.478	Tidak Cocok	Tidak Baik
10	3.504	0.305	Cocok	Kurang Baik
11	2.857	0.393	Cocok	Kurang Baik
12	1.591	0.559	Tidak Cocok	Tidak Baik
13	2.224	0.500	Cocok	Kurang Baik
14	0.648	0.479	Cocok	Baik
15	1.49	0.422	Tidak Cocok	Tidak Baik
16	1.968	0.537	Cocok	Baik
17	1.948	0.455	Cocok	Baik
18	2.745	0.219	Cocok	Kurang Baik
19	2.653	0.298	Cocok	Kurang Baik
20	0.598	0.335	Tidak Cocok	Tidak Baik

Sumber : Data diolah 2017

Berdasarkan hasil analisis, dari 20 butir soal yang dianalisis maka hanya terdapat 4 item (20%) yang memiliki kualitas item yang baik. 10 item (50%) masuk kategori tidak baik karena memang tidak fit dianalisis menggunakan Graded Response Model. Sedangkan 6 item sisanya (30%) masuk kategori kurang baik karena fit dengan model GRM namun memiliki indeks tingkat kesukaran >2. Item yang tergolong baik kualitasnya adalah item nomor 1, 14, 16, dan 17.

Sedikitnya item yang berkualitas baik dapat disebabkan oleh banyaknya butir yang tidak fit dengan model GRM. Banyaknya item yang tidak fit memperlihatkan bahwa model GRM kurang cocok digunakan pada sampel kecil. Linacre (1994) menyatakan bahwa jumlah sampel kecil dari 100 lebih stabil dalam mengestimasi parameter untuk model Rasch. Model Rasch untuk skor politomus merujuk pada model PCM. Perbedaan model PCM dengan GRM terletak pada jumlah parameter butir yang dapat diestimasi. Model GRM dapat memberi informasi parameter tingkat kesukaran dan daya pembeda item, sedangkan model PCM hanya mampu mengestimasi parameter tingkat kesulitan. Implikasi dari hal ini berarti bahwa semakin banyak parameter yang ingin diestimasi maka akan dibutuhkan jumlah sampel yang semakin besar agar mampu memberikan hasil yang lebih stabil dan akurat (Edelen & Reeve, 2007). Meskipun tidak ada takaran pasti berapa jumlah sampel yang tepat untuk model GRM, Tsutakawa dan Johnson (1990) menyarankan sampel sebanyak 500 untuk analisis yang lebih kompleks agar diperoleh estimasi parameter yang akurat. Hal ini membuktikan bahwa model GRM kurang tepat digunakan untuk analisis *classroom test*.

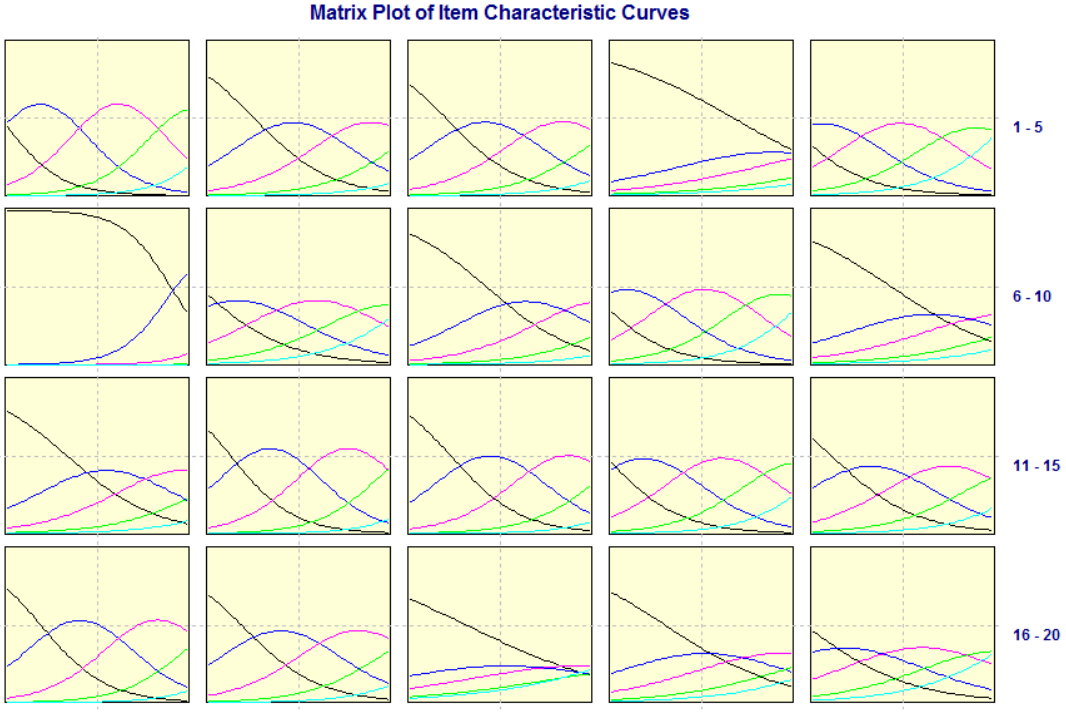
Disamping menghasilkan parameter butir, parameter kemampuan peserta (θ) dapat dilihat pada output Parscale fase 3. Berdasarkan output 3 program Parscale diperoleh informasi bahwa rerata kemampuan siswa adalah sebesar 0,59913 dengan standar deviasi 1,37045, sebagaimana yang terlihat dalam Gambar 4.



Gambar 4. Kurva Estimasi Kemampuan Peserta

Berdasarkan grafik tersebut terlihat bahwa kecendrungan proporsi siswa yang memiliki kemampuan rendah lebih besar dibandingkan yang memiliki kemampuan tinggi. Berdasarkan data tersebut dapat disimpulkan bahwa wajar saja cenderung tidak ada soal yang mudah bagi siswa karena kemampuan mereka sendiri relatif tidak tinggi.

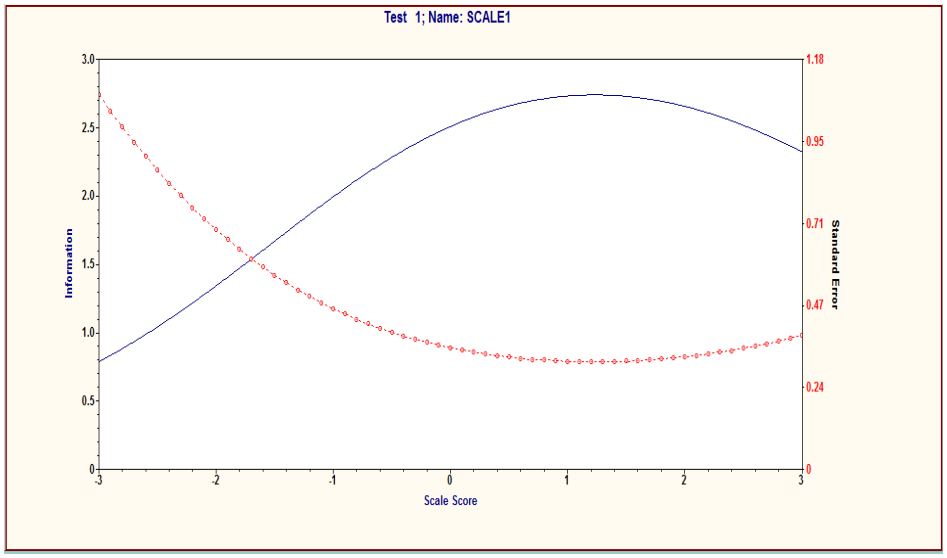
Hubungan performansi subyek pada suatu butir dan perangkat laten yang mendasarinya dapat dilihat dari kurva karakteristik butir atau *item characteristic curve* (ICC). Menurut Mardapi (1991: 6) kurva karaktersitik butir merupakan fungsi matematika yang menyatakan hubungan antara peluang sukses menjawab suatu butir dengan kemampuan yang diukur oleh butir. Hubungan ini merupakan fungsi regresi nonlinier skor butir terhadap trait atau kemampuan yang diukur oleh tes. Bentuk kurva ini pada prinsipnya merupakan kurva ogive, yaitu kurva frekuensi kumulatif. Gabungan kurva karakteristik butir seluruh item dapat dilihat pada gambar berikut.



Gambar 5. Kurva Karakteristik Butir

Gambar 5 di atas menggambarkan karakteristik 20 butir soal yang dianalisis dengan model GRM. Ciri kurva ICC yang bagus untuk data politomus seperti ICC butir soal nomor 1, 14, 16, dan 17 karena memiliki parameter butir yang baik dan cocok dengan model sehingga dalam kurva tersebut terlihat dengan jelas batas setiap kategorikal yang digunakan. Sementara ICC soal no.4 dan 6 merupakan contoh yang ICC yang tidak baik karena bentuk kurva tidak jelas yang disebabakan juga oleh tidak cocok dengan model dan parameter butir juga kurang baik.

Secara keseluruhan keandalan dan keakuratan suatu pengukuran tes dapat dilihat berdasarkan nilai informasi. Besarnya nilai fungsi informasi dari suatu tes secara kasat mata dapat kita lihat pada Gambar 6. Garis melengkung ke atas pada kurva menggambarkan informasi tes sedangkan garis putus-putus menunjukkan kesalahan pengukuran. Berdasarkan Gambar 6 maka dapat diprediksi instrumen ini cocok untuk mengukur kemampuan peserta antara -1,7 s/d >3, sedangkan nilai fungsi informasi tes maksimal tercapai pada kemampuan siswa sebesar 1,3.



Gambar 6. Kurva Informasi Tes

Secara garis besar dapat disimpulkan dari kurva informasi tes ini bahwa instrumen ini lebih cocok diujikan pada peserta tes dengan kemampuan yang sedang hingga yang berkemampuan tinggi. Apabila diuji pada peserta dengan ability di bawah -1,7 maka kesalahan pengukuran lebih tinggi daripada fungsi informasi yang bisa diperoleh dari tes tersebut.

SIMPULAN

Berdasarkan hasil analisis data maka disimpulkan bahwa meskipun pengujian asumsi IRT terpenuhi untuk perangkat tes ini, namun penggunaan Graded Response Model pada analisis perangkat tes ini kelihatannya kurang cocok karena hanya sebesar 50% butir soal yang fit dengan model GRM. Hasil analisis estimasi paramater butir menggunakan GRM mengungkapkan bahwa hanya terdapat 4 butir soal (20%) yang memiliki kualitas yang baik, yaitu butir soal yang cocok dengan model, memiliki daya pembeda dan tingkat kesulitan yang baik. Sebanyak 6 item soal (30%) memiliki karakteristik kurang baik, karena fit dengan model GRM namun memiliki tingkat kesukaran yang tidak baik. Sisanya 10 soal berkualitas tidak baik karena tidak fit dengan model GRM. Hasil analisis estimasi parameter kemampuan peserta memperlihatkan bahwa pada umumnya peserta tes memiliki rata-rata ability sebesar 0,59913 yang dapat diklasifikasikan pada abiliti yang sedang. Berdasarkan hasil penelitian maka penggunaan sampel kecil pada model GRM memperlihatkan hasil yang kurang efektif sehingga kurang cocok digunakan untuk menganalisis *classroom test.* Model GRM lebih disarankan untuk diujikan pada tes dengan sampel yang besar agar diperoleh estimasi parameter yang lebih akurat dan stabil.

DAFTAR PUSTAKA

Bond, T. G., & Fox, C. M. (2007). *Applying the rasch model: fundamental measurement in the human sciences.* (2nd Ed.). Mahwah: Lawrence Erlbaum Associates, Publishers.

Brennan, R.L & Kolen. (2004). Test Equating, Scaling, and Linking. New York: Springer.

DeMars, C.E. (2010). *Item response theory*. New York: Oxford University Press, Inc

DeMars, C.E. (2003). Sample Size and the Recovery of Nominal Response Model Item Parameters. *Applied Psychological Measurement*, Vol. 27 No. 4

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*(1), 5.

Hambleton, R.K & Swaminathan, H. & Rogers. (1991). *Fundamental of Item Response Theory*. Newbury Park: Sage Publication Inc.

Hambleton, R.K & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer-Nijhoff Publishing.

Hullin C.L et al. (1983*). Item response theory application to psychological measurement*. Homewood: Dow Jones-Irwin.

Isgiyanto, Awal. (2011). Diagnosis Kesalahan Siswa Berbasis Penskoran Politomus Model Partial Credit Pada Matematika. *Jurnal Penelitian dan Evaluasi Pendidikan* Tahun 15, Nomor 2

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7(4), 328.

Mardapi, Djemari. (1991). *Konsep Dasar Teori Respon Butir*. Perkembangan dalam bidang Pengukuran Pendidikan. Yogyakarta. *Jurnal Cakrawala* Pendidikan Nomor 3 th X.

Naga, D.S. (1992). *Pengantar Teori Sekor Pada Pengukuran Pendidikan*. Jakarta: Penerbit Gunadharma.

Retnawati, Heri. (2014). *Teori Respons Butir dan Penerapannya. Untuk Peneliti, Praktisi Pengukuran dan Pengujian, Mahasiswa Pascasarjana*. Yogyakarta: Parama Publishing

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika*, 55, 371–390.

.