Jurnal Evaluasi Pendidikan
Volume 3, No 1, Maret 2015 (56-66)
Online: http://journal.student.uny.ac.id/ojs/index.php/jep

KOMPARASI MODEL PENSKORAN BERDASARKAN TEORI RESPONS BUTIR PADA SOAL UJIAN NASIONAL MATA PELAJARAN MATEMATIKA

Nuril Huda, Djemari Mardapi
Prodi PEP PPs UNY, Universitas Negeri Yogyakarta
nurilhuda894@yahoo.co.id, djemarimardapi@gmail.com

Abstrak

Penelitian ini bertujuan mendeskripsikan: 1) karakteristik perangkat tes Ujian Nasional (UN) paket E03 mata pelajaran matematika SMA/MA program IPA tahun pelajaran 2013/2014, 2) distribusi skor dari empat model penskoran, 3) hubungan kesesuaian skor dari empat model penskoran, dan 4) komparasi skor dari empat model penskoran. Data penelitian ini berupa respons j peserta didik SMA/MA program IPA yang menempuh tes UN paket E03 mata pelajaran matematika tahun pelajaran 2013/2014 dipropinsi Daerah Istimewa Yogyakarta (DIY). Analisis data dilakukan berdasarkan teori respons butir dengan menggunakan program Bilog-MG versi 3.0. Hasil penelitian ini adalah 1) perangkat tes UN paket E03 memiliki daya beda, tingkat kesukaran, dan *pseudo-guessing* yang baik, dengan kesalahan baku pengukuran sebesar 0,22 logit, 2) distribusi skor keempat model penskoran memiliki *skewness* dan *kurtosis* yang berbeda, 3) hubungan kesesuaian skor antarmodel penskoran yang paling tinggi terjadi pada model penskoran NW_2PL dan NW_3PL, 4) hasil uji perbandingan menunjukkan bahwa ada perbedaan signifikan (p value < 0,05) rerata skor dari keempat model penskoran an.
***Kata kunci**: karakteristik butir soal, teori respon butir, model penskoran*

A COMPARISON OF SCORING MODELS BASED ON ITEM RESPONSE THEORY IN THE MATHEMATICS NATIONAL EXAMINATION

Nuril Huda, Djemari Mardapi
Prodi PEP PPs UNY, Universitas Negeri Yogyakarta
nurilhuda894@yahoo.co.id, djemarimardapi@gmail.com

Abstract

This study aims to describe: 1) the characteristics of the mathematics national examination of senior high school natural science in the academic year of 2013/2014, 2) the scores distribution of four scoring models, 3) the score suitability relationship of four scoring models, and 4) a comparison of score based on four scoring models. The data this study was computerized answer sheets in Daerah Istimewa Yogyakarta. The data analysis was based on item response theory with Bilog-MG version 3.0 program. The results of this study are 1) the items are categorized in the good item characteristics based on item difficulty, item discrimination, and pseudo-guessing, with standard error measurement in the interval 0.22 logit, 2) the score distribution based on NW_2PL, NW_3PL, W_2PL and W_3PL model have different skewness and kurtosis, 3) the score suitability relationship of scoring models NW_2PL and NW_3PL, 4) the comparison mean scoring models have p-value < 0.05 (significant), which the mean scores are different.
***Keywords**: item characteristics, item response theory, scoring models*

Pendahuluan

Pendidikan merupakan suatu proses, paling tidak di dalamnya harus terdapat tiga unsur pokok yang saling berkaitan, yaitu tujuan belajar, pengalaman belajar, dan prosedur evaluasi. Prosedur evaluasi akan menetapkan langkah-langkah sistematis untuk mengetahui ketercapaian tujuan belajar dan langkah awal dari prosedur evaluasi dengan mengadakan pengukuran. Menurut Mardapi (2012, p.1), pengukuran merupakan kegiatan melakukan kuantifikasi gejala atau objek. Objek ini bisa berupa motivasi, hasil belajar, atau hasil yang semuanya dinyatakan dalam bentuk angka. Dalam kaitannya dengan hasil belajar, pengukuran merupakan proses pemberian angka yang diharapkan dapat menunjukan kemampuan peserta didik mengenai suatu mata pelajaran. Dalam pengukuran diperlukan alat ukur yang mampu memberikan informasi tentang posisi seseorang dalam atribut yang diukur.

Alat ukur yang digunakan dalam pendidikan terdiri atas tes dan nontes. Alat ukur berupa tes umumnya memberikan informasi tentang karakteristik kognitif dari peserta tes. Alat ukur non-tes dapat berupa angket, lembar observasi, dan pedoman wawancara yang umumnya memberikan informasi tentang karakteristik afektif atau psikomotorik dari peserta tes. Alat ukur tes tersebut harus memenuhi prasyarat sebagai alat ukur baik dan diharapkan dapat memberikan gambaran atau informasi yang akurat, serta dapat dipercaya.

Menurut Mardapi (2012, p.109), tes diklasifikasikan menjadi dua, yaitu tes objektif dan tes non-objektif. Objektif yang maksud adalah sistem penskorannya, siapa saja yang memeriksa lembar jawaban tes menghasilkan skor yang sama. Tes yang non-objektif dengan sistem penskorannya dipengaruhi oleh pemberi skor. Dengan kata lain bahwa tes yang objektif adalah tes yang sistem penskorannya objektif, sedangkan tes yang non-objektif sistem penskorannya dipengaruhi subjektivitas pemberi skor. Bentuk tes objektif yang sering digunakan adalah bentuk pilihan ganda, benar salah, menjodohkan, dan uraian objektif, sedangkan bentuk non objektif adalah uraian objektif dan uraian non-objektif.

Keuntungan yang diperolah dengan penggunaan bentuk tes pilihan ganda diantaranya adalah materi yang diujikan dapat mencakup ruang lingkup yang luas, dapat mengukur kemampuan yang bermacam-macam dari yang paling sederhana sampai yang paling kompleks, dan sistem penskoran yang lebih cepat dan mudah.

Salah satu kelemahan tes pilihan ganda adalah memberikan peluang lebih besar bagi peserta tes untuk melakukan kecurangan (*cheating*) dalam tes (Satoridona, Van der Linden, & Meijer, 2006, p.412). Misalnya, seorang peserta tes dapat dengan sekilas melihat dan menyalin jawaban peserta lain atau peserta tes saling bekerja sama dengan kode-kode tertentu. Perilaku curang ini bisa menghasilkan skor yang tidak valid dan tidak reliabel pada suatu tes pilihan ganda.

Kelemahan lain dari tes pilihan ganda adalah tes ini sangat peka terhadap kecurangan dan tidak sensitif terhadap perbedaan tingkat pengetahuan peserta tes (Simon, Budescu, & Nevo, 1997, p.65). Ketika menghadapi suatu pertanyaan, seorang peserta kemungkinan akan memiliki satu dari dua kondisi yaitu (1) peserta tes memiliki pengetahuan atas pertanyaan tersebut dan yakin terhadap jawabannya secara sempurna, (2) peserta tes hanya mengetahui sebagaian dari jawaban dan tidak yakin terhadap jawaban dari pertanyaan itu. Jika kondisi kedua terjadi pada peserta tes, maka kemungkinan adalah peserta akan melakukan kecurangan dengan kerjasama antarpeserta tes. Kecurangan tersebut dapat dilakukan dengan mencari lengahan-lengahan pengawas tes. Peserta akan leluasa dalam melakukan kecurangan pada item soal yang tidak bisa dijawab jika tidak ada hukuman bagi jawaban yang salah seperti yang digunakan pada perangkat tes Ujian Nasional (UN) di tingkat SMA/MA pada mata pelajaran matematika program IPA, sehingga skor yang diperoleh tidak menggambarkan kemampuan peserta tes sesungguhnya.

Estimasi skor kemampuan dapat dilakukan melalui pendekatan teori tes klasik (*Classical Test Theory*, selanjutnya ditulis CTT) maupun teori respons butir(*Item Response Theory*, selanjutnya ditulis IRT). Pendekatan teori tes klasik lebih umum digunakan dalam praktik karena kesederhanaan dalam perhitungan. Berdasarkan pendekatan ini kemampuan seorang peserta tes terhadap materi yang diukur oleh suatu tes pilihan ganda diestimasi dengan menggunakan jumlah butir yang dijawab benar dan sering dinyatakan sebagai skor tampak (X). Estimasi kemampuan berdasarkan pendekatan ini tidak memperhatikan pola respons peserta tes yang menjawab. Hasil estimasi kurang sensitif terhadap karakteristik butir. Karakteristik

butir seperti tingkat kesulitan butir, daya pembeda butir, dan efek tebakan tidak dipertimbangkan dalam mengestimasi kemampuan siswa. Selain itu, bobot tiap butir soal dianggap sama. Menurut Garcí–Pérez & Frary (1989, p.403), skor jumlah benar pada tespilihan ganda tidak beralasan diklaim sebagai estimasi kemampuan yang dimiliki oleh siswa. Metode ini hanya menunjukkan informasi tentang ranking siswa.

Berbeda dengan teori tes klasik, estimasi kemampuan peserta pada IRT ditentukan berdasarkan pola responsnya, tidak ditentukan berdasarkan jumlah butir yang dijawab dengan benar. Jadi, pola respons siswa yang bervariasi menunjukkan variasi kemampuan peserta tes (Mardapi, 1999, pp.9-10). Estimasi kemampuan peserta berdasarkan IRT sering dinyatakan sebagai θ. Skor kemampuan atau θ yang dihasilkan oleh IRT diukur pada skala yang kurang lazim karena skor tersebut dapat bernilai negatif sehingga mempersulit sebagian orang untuk menginterpretasi.

Model-model penskoran tes pilihan ganda pada CTT adalah model penskoran jumlah benar, model penskoran koreksi *guessing*, dan model penskoran pembobotan (Rofieq, 2010, pp.3-5). Pemberian bobot setiap butir soal suatu perangkat tes ditentukan dengan mempertimbangkan faktor-faktor yang berkaitan dengan materi dan karakteristik butir soal itu sendiri (Mardapi, 2008, p.133). Biasanya didasarkan pada ruang lingkup materi yang hendak dibuatkan soalnya, esensialitas dan tingkat kedalaman materi yang ditanyakan, dan tingkat kesukaran soal tersebut.

Beberapa ahli telah mengembangkan model penskoran, Lord (1980, p.45) menjelaskan tentang model penskoran jumlah benar. Estimasi skor berdasarkan model ini diperoleh dengan menjumlahkan peluang menjawab benar pada setiap butir soal dan menganggap setiap butir soal memiliki bobot yang sama. Crocker & Algina (2008, p.400) menjelaskan model penskoran *Correction for Guessing*, model penskoran ini mengoreksi jumlah skor benar yang disebabkan oleh unsur tebakan dengan memberikan hukuman pengurangan skor pada soal yang dijawab salah karena jawaban yang salah dianggap sebagai jawaban hasil tebakan. Selanjutnya Lord (1980, p.74) menjelaskan tentang model penskoran pembobotan optimal, pada model ini masing-masing butir soal diberi bobot optimum yang berbeda sesuai dengan karakteristik butir soal dan berdasarkan model IRT yang digunakan.

Menurut Musmuliadi (2009, pp.246-267), rerata skor sesungguhnya yang paling tinggi diperoleh pada model penskoran jumlah benar, sedangkan rerata paling kecil terjadi pada model penskoran *Correction for Guessing*. Berdasarkan hal ini model penskoran yang digunakan ada dua yaitu model penskoran jumlah benar dan model penskoran pembobotan. Dari dua model penskoran diperinci menjadi empat model penskoran tergantung pada parameter yang digunakan. Penelitian ini menggunakan empat model penskoran berdasarkan IRT yaitu, model penskoran jumlah benar (*number of right score*) berdasarkan model IRT 2PL, model penskoran jumlah benar (*number of right score*) berdasarkan model IRT 3PL, model penskoran pembobotan optimal (*optimal weighting*) berdasarkan model IRT 2PL, dan model penskoran pembobotan optimal (*optimal weighting*) berdasarkan model IRT 3PL.

Sebelum tahap penskoran hasil tes berdasarkan IRT perlu diketahui karakteristik butir soal dan karakteristik peserta tes. IRT membangun suatu model yang menghubungkan karakteristik peserta tes dengan karakteristik item soal yang berlaku bagi semua kelompok item dan semua kelompok peserta tes tanpa ada saling ketergantungan satu sama lain.

Masalah utama yang akan diungkap dalam penelitian ini adalah bagaimanakah karakteristik perangkat tes ujian nasional mata pelajaran matematika SMA/MA program IPA tahun pelajaran 2013/2014 berdasarkan teori respons butir dan bagaimanakah komparasi skor keempat model penskoran.

Sejalan dengan rumusan masalah yang akan diselesaikan, maka tujuan penelitian ini adalah: (1) mendeskripsikan karakteristik perangkat tes UN mata pelajaran matematika SMA/MA program IPA tahun pelajaran 2013/2014 berdasarkan teori respons butir; (2) mendeskripsikan distribusi skor ke empat model penskoran; (3) mendeskripsikan hubungan kesesuaian skor dari ke empat model penskoran; (4) mendeskripsikan komparasi skor dari ke empat model penskoran.

Secara umum manfaat penelitian ini adalah (1) memberikan karakteristik perangkat tes UN mata pelajaran matematika SMA/MA program IPA berdasarkan teori respons butir; (2) memberikan penerapan IRT dalam model penskoran pada tes pilihan ganda; (3) memberikan gambaran tentang ragam model pen-

skoran untuk menentukan estimasi skor pada tes pilihan ganda; (4) memberikan penerapan model penskoran berdasarkan kemampuan yang dimiliki oleh masing-masing peserta tes (azas keadilan).

Metode Penelitian

Penelitian ini merupakan penelitian deskriptif eksploratif dengan pendekatan kuantitatif, yaitu dengan menganalisis lembar jawaban siswa kelas XII SMA/MA program IPA yang telah mengikuti ujian nasional mata pelajaran matematika tahun ajaran 2013/2014 di Propinsi Daerah Istimewa Yogyakarta (DIY). Penelitian ini mendeskripsikan empat model penskoran unurk mengestimasi skor berdasarkan teori respons butir.

Penelitian in dilakukan di Daerah Istimewa Yogyakarta (DIY) yang dilakukan pada bulan Januari sampai Maret 2015. Data set dalam penelitian ini adalah lembar jawaban peserta didik tes ujian nasional mata pelajaran matematika tingkat SMA/MA program IPA tahun ajaran 2013/2014 di Propinsi DIY. Berdasarkan informasi yang diperolah dari Dinas Pendidikan Pemuda dan Olahraga DIY siswa yang mengikuti UN adalah 9883 siswa tingkat SMA/MA yang tersebar dalam 69 sekolah negeri dan 27 sekolah swasta pada 4 kabupaten dan 1 kota madya.

Pemilihan kode E03 ini didasarkan pada alasan bahwa jumlah data set yang diperoleh lebih banyak yaitu 589 respons daripada paket E01 sebanyak 525 respons, E02 sebanyak 533 respons, dan E16 sebanyak 513 respons. Data set tersebut diperoleh dari Badan Penelitian dan Pengembangan Pusat Penilaian Pendidikan (Puspendik Jakarta).

Pengumpulan data dilakukan dengan teknik dokumentasi dengan cara mengutip respons peserta didik dari lembar jawaban peserta didik kelas XII SMA/MA program IPA yang telah mengikuti ujian nasional mata pelajaran matematika tahun ajaran 2013/2014 di propinsi DIY. Adapun pemindain lembar jawab komputer tersebut dilakukan oleh tim Universitas Negeri Yogyakarta yang kemudian di kirim ke Badan Penelitian dan Pengembangan Pusat Penilaian Pendidikan (Puspendik Jakarta).

Teknik analisis data yang dilakukan dalam penelitian ini adalah analisis karakteristik butir soal berdasarkan pendekatan IRT menggunakan program komputer *Bilog-MG* versi 3.0 sebagai berikut: (1) tingkat kesukaran soal dapat dilihat pada kolom *Threshold*, (2) daya beda soal dapat dilihat pada kolom *Slope*, (3) efektivitas distraktor dapat dilihat pada kolom *Asymptote*. Kriteria yang digunakan untuk melihat butir yang baik dalam teori respons butir diginakan pendapat yang dikemukan oleh Hadi (2011, p.3) yaitu sebagai berikut.

Tabel 1. Kriteria Butir Soal yang Baik Berdasarkan Teori Respons Butir

Parameter	Nilai	Keterangan
Daya beda (a)	0,4 s/d 2	Baik
Tingkat kesukaran (b)	-2 s/d 2	Baik
Psedo guessing (c)	0 s/d $\frac{1}{k}$ (k=jumlah alternatif jawaban)	Baik
Uji cocok model	> 0,05	Fit model

Hasil estimasi parameter butir dan parameter kemampuan digunakan untuk menghitung fungsi informasi dan estimasi skor berdasarkan masing-masing model penskoran. Model penskoran yang digunakan adalah pertama model penskoran jumlah benar (*number of right score*) berdasarkan model IRT 2 PL (NW_2PL), kedua model penskoran jumlah benar (*number of right score*) berdasarkan model IRT 3 PL (NW_3PL). Secara matematis dinyatakan sebagai berikut (Lord, 1980, p. 230)

$$T_{NC} = \sum_{i=1}^{n} P_i(\theta)$$

Keterangan rumus :

T_{NC} : skor jumlah benar dari seseorang dengan kemampuan θ

P_i : Probabilitas/kemungkinan peserta tes yang memiliki kemampuan/ability (θ) menjawab butir ke-i

n : Banyaknya butir ke-i

Ketiga penskoran pembobotan optimum (*optimal weighting*). Pembobotan terhadap respons peserta didik dalam penelitian ini dilakukan secara implisit dengan besar bobot bervariasi berdasarkan model IRT yang digunakan (Rudner, 2001, p.3). Masing-masing butir soal diberi bobot optimum yang berbeda sesuai dengan karakteristik butir soal dan berdasarkan model IRT 2PL (W_2PL), keempat

model penskoran pembobotan optimum (*optimal weighting*) berdasarkan model IRT 3PL (W_3PL). Secara umum skor komposit terbobot secara matematis dirumuskan sebagai berikut (Lord, 1980, p.73):

$$T_{OW} = \sum_{i=1}^{n} w_i \; P_i(\theta)$$

Keterangan rumus:

T_{OW} : estimasi skor berdasarkan pembobotan optimum dengan kemampuan θ

P_i : Probabilitas/kemungkinan peserta tes yang memiliki kemampuan/ability (θ) menjawab butir ke-i

w_i : nilai bobot butir ke-i

Lord (Rudner, 2001, p.17) menyatakan bahwa berdasarkan pendekatan teori respons butir, pembobotan implisit bervariasi berdasarkan model IRT yang digunakan. Menurut Hamblenton & Swaminthan (1985, p.104) model 3PL bobot setiap soal sebagai berikut:

$$w_i = \frac{Da_i}{1 + c_i \; e^{-D(\theta - b_i)}}$$

Dari model pembobotan di atas, apabila model 1PL (dengan $a_i = 1$ dan $c_i = 0$) maka pembobotanya menjadi D, dan model 2PL (dengan $c_i = 0$) maka pembobotanya menjadi Da_i. Berdasarkan pertanyaan di atas bahwa pembobotan dengan menggunakan model 1PL dan 2PL tidak tergantung pada kemampuan, sedangkan pembobotan dengan menggunakan model 3PL tergantung pada kemampuan setiap peserta ujian.

Analisis dilanjutkan dengan mendeskripsikan distribusi skor dari masing-masing model penskoran, menentukan hubungan kesesuaian skor dengan korelasi intraklas antara model penskoran, serta membandingkan hasil skor dari keempat model penskoran. Hubungan kesesuaian skor dihasilkan dengan analisis korelasi intraklas. Perbandingan skor yang dihasilkan oleh masing-masing model penskoran dilakukan dengan analisis varinas (ANOVA) dengan pengukuran berulang (*repeated measure*).

Hasil Penelitian dan Pembahasan

Analisis empiris karakteristik tes dan butir soal dilakukan dengan menggunakan program *Bilog-MG* versi 3.0. Analisis tersebut meliputi tingkat kesukaran, daya beda, efektivitas pengecoh (*psedo-guessing*), fungsi informasi, dan kesalahan baku pengukuran. Sebelum analisis dilakukan harus membuktikan asumsi-asumsi teori respons butir.

Asumsi-Asumsi Teori Respons Butir (IRT)

Asumsi Unidimensi

Asumsi unidimensi dilakukan untuk mengetahui apakah perangkat tes yang digunakan mengukur satu macam ciri (*trait*) atau kemampuan. Prasyarat asumsi unidimensi dapat ditunjukkan dengan menggunakan analisis faktor menggunakan program SPSS 16. Sebelum melakukan analisis faktor dilakukan pengujian kelayakan analisis dengan menggunakan uji Kaiser-Meyer Olkin – Measure of Sampling Adequacy (KMO-MSA) dan uji Bartles's pada tiap tes.

Menurut Hair et al (1998, p.88-89), syarat analisis faktor adalah Kaiser-Meyer Olkin (KMO)-MSA > 0,5 dan signifikan uji Barlett's kurang dari 0,05. Uji KMO-MSA untuk melihat kecukupan sampel, sedang uji Barlett's untuk homogenitas data yang digunakan. Hasil analisis empiris uji sumsi unidimensi dicantumkan pada Tabel 2 dan Tabel 3.

Tabel 2. Uji KMO dan Bartlett

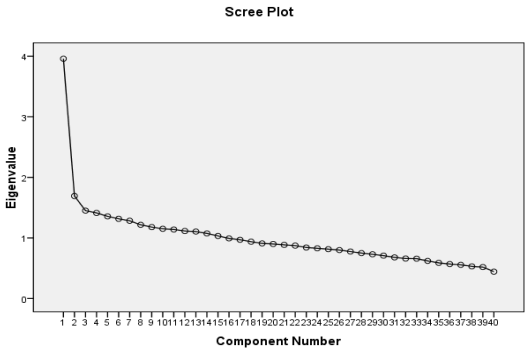
Kesesuaian Ukuran Sampel Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0,739
Bartlett's Test of Sphericity	Approx. Chi-Square	2,012E3
	df	780
	Sig.	0,000

Berdasarkan hasil analisis empiris pada Tabel 2 didapatkan nilai KMO-MSA sebesar 0,739 atau mendekati 1 dan nilai signifikan uji Bartlett sebesar 0,000 < 0,05. Hal ini berarti ada 15 faktor yang diukur pada tes ujian nasional mata pelajaran matematika SMA/MA program IPA tahun pelajaran 2013/2014. Hasil tersebut menunjukkan bahwa sampel 589 respons yang digunakan telah memenuhi jumlah kecukpan sampel dan memiliki data yang homogen. Hal ini menunjukkan bahwa tes memenuhi prasyarat analisis faktor. Untuk mendapatkan item-item yang mengukur dimensi yang sama, dilakukan proses ekstraksi sehingga dihasilkan beberapa faktor. Tiap faktor yang terbentuk

mempunyai nilai eigen, dan faktor yang memiliki nilai eigen di atas 1 dipertahakan.

Tabel 3. *Eigen valu*e dan Komponen Varian 15 komponen

Component	Total	% of Variance	Cumulative %
1	3,960	9,900	9,900
2	1,693	4,231	14,131
3	1,450	3,625	17,757
4	1,412	3,531	21,287
5	1,356	3,389	24,676
6	1,316	3,289	27,965
7	1,283	3,208	31,174
8	1,216	3,040	34,214
9	1,178	2,946	37,160
10	1,151	2,878	40,037
11	1,139	2,847	42,884
12	1,114	2,784	45,668
13	1,104	2,761	48,429
14	1,074	2,684	51,114
15	1,032	2,581	53,695



Gambar 1. *Scree Plot Eigenvalue*

Hasil analisis pada Tabel 3 menunjukkan ada 15 *eigen value* yang lebih dari 1. Secara lebih jelas dapat diperhatikan gambar 1 di mana didalamnya terdapat plot nomor komponen hasil ekstraksi dan nilai eigen. Kelima belas faktor tersebut dapat menjelaskan sekitar 53,695% dari total varians.

Secree plot yang membuktikan bahwa tes bersifat unidimensi dapat dilihat pada Gambar 1. Jika diamati terlihat adanya penurunan drastis antara faktor ke-1 dan faktor ke-2. Nilai eigen mulai landai pada faktor ke-2 atau menurun tajam dari faktor pertama ke faktor ke-2. Sebagaimana dijelaskan di atas, faktor pertama terhitung 9,90%.

Asumsi Independensi lokal

Asumsi independensi lokal berarti respons peserta ujian terhadap sebuah item butir dan item butir yang lain bersifat independen setelah *latent trait* dikontol (Hambleton, 1989, p.150). *Latent traits* yang dimaksud adalah kemampuan matematika. Dominansi satu faktor yang ada berdasarkan analisis faktor telah mengarahkan pada terpenuhinya bukti bahwa data yang dimiliki bersifat unidimensional, hanya ada satu faktor yang mempengaruhi respons peserta tes. Berdasarkan pernyataan tersebut, dapat disebutkan juga bahwa karena data yang dimiliki bersifat unidimensional, maka respons yang diberikan para peserta tes bersifat independen, kondisional terhadap kemampuan mereka masing-masing. Jika kemampuan para perserta tes sudah diketahui, maka perilaku respons terhadap satu item butir tidak berpengaruh terhadap perilaku respons terhadap item yang lain. Jadi dapat simpulkan bahwa asumsi independensi lokal otomatis terpenuhi jika asumsi unidimensi terpenuhi.

Uji Kecoccokan Model

Uji kecocokan model untuk setiap butir soal menggunakan α= 5 % yang merupakan nilai *default* dari program *Bilog-MG* versi 3.0 dengan derajat bebas (*degree of freedom*, df) yang sudah ditetapkan oleh program tersebut sesuai dengan banyaknya kategoti interval kemampuan hasil estimasi θ. Hasil Uji kecocokan model menggunakan uji statistik adalah 12 (30,00%) butir soal fit model IRT 1PL, 35 (87,50%) butir soal fit model IRT 2PL, dan 36 (90,00%) butir fit model IRT 3PL.

Karakteristik Perangkat Tes

Hasil analisis estimasi parameter butir perangkat tes UN mata pelajaran matenatika SMA/MA program IPA provinsi DIY menggunakan model IRT 3PL dapat dilihat pada o*utput* program *Bilog-MG* versi 3.0 phase 2. Ringkasan hasil analisis diperoleh bahwa nilai parameter daya beda (a) ada 38 butir soal (95%) termasuk kriteria baik yaitu interval 0,40 sampai 2,00, 1 butir soal (2,5%) termasuk kriteria tidak baik yaitu kurang dari 0,40, dan 1 butir soal (2,5%) termasuk tidak baik yaitu lebih dari 2. Tingkat kesukaran (b) ada 34 butir soal (85%) termasuk kriteria baik yaitu interval –2 logit sampai +2 logit, 1 butir soal (2,5%) termasuk kriteria mudah karena kurang dari –2 logit, 5 butir soal (12,5%) termasuk kriteria

sukar karena lebih dari +2 logit. p*seudo-guessing* (c) ada 25 butir soal (62,5%) termasuk kriteria baik yaitu interval 0 sampai 0,20 dan 15 butir soal (37,5%) termasuk kriteria tidak baik karena lebih dari 0,20.

Berdasarkan karakteristik butir soal yang memiliki daya beda, tingkat kesukaran, pengecoh, dan fit model maka terdapat 18 (45%) butir soal yang baik. Nilai fungsi informasi dari model IRT 3PL dengan kemampuan -3 logit sampai dengan +3 logit adalah 21,51 dan nilai ini tercapai θ sebesar 0,55 logit serta kesalahan baku pengukuran adalah 0,22 logit.

Deskripsi Estimasi Skor dari Model Penskoran

Estimasi skor dalam penelitian ini dilakukan kemampuan dalam penelitian ini dilakukan berdasarkan IRT model 2PL dan 3PL. Ada 4 model penskoran yang akan digunakan untuk mengestimasi skor yaitu model penskoran jumlah benar berdasarkan model IRT 2PL (NW_2PL) dan model IRT 3PL (NW_3PL) dan model penskoran pembobotan optimal berdasarkan model IRT 2PL (W_2PL) dan model IRT 3PL (W_3PL).

Tabel 4. Statistik Deskriptif Skor

Statistik	Model penskoran			
	Non Weight		Weight	
	2PL	3PL	2PL	3PL
Rerata	13,334	13,243	19,011	23,129
Simpangan Baku	8,558	8,738	12,182	19,739
Skewness	0,498	0,597	0,341	0,708
Kurtosis	–0,906	–0,790	– 1,138	–0,715
Minimum	0,700	1,074	0,536	0,124
Maksimum	33,721	35,588	43,924	72,790

Mencermati hasil pada Tabel 4 tampak bahwa rerata skor NW_2PL hampir sama dengan rerata skor NW _3PL dan rerata skor W_3PL lebih tinggi daripada rerata skor W_2PL. Jika dilihat berdasarkan nilai simpangan baku, penyebaran skor berdasarkan NW_2PL dan NW_3PL relatif sama, sedangkan penyebaran skor berdasarkan W_2PL dan W_3PL lebih menyebar dari reratanya.

Distribusi skor berdasarkan keempat model penskoran menunjukan nilai skewness positif, artinya distribusi skor juling ke kanan yang menunjukan bahwa sebagian besar peserta memperoleh skor yang rendah. Nilai *kurtosis* negatif, artinya distribusi skor cenderung datar atau menyebar. Pada Tabel 4 dapat diperhatikan model penskoran yang paling menceng ke kanan adalah W_3PL.dan model penskoran yang paling datar adalah W_2PL. Selain itu, jika diperhatikan dengan seksama tampak skor NW_2PL dan NW_3PL menunjukkan nilai *skewnees* dan *kurtosis* relatif sama dan skor NW_3PL dan W_3PL menunjukkan nilai *skewnees* dan *kurtosis* relatif sama. Hal ini menunjukkan bahwa distribusi skor yang dihasilkan oleh kedua model penskoran tersebut tidak memiliki perbedaan yang signifikan.

Hubungan Kesesuaian Skor dengan Model Penskoran

Konsistensi atau kesesuaian skor antar-masing-masing model penskoran ditunjukkan oleh besarnya koefisien korelasi antarmodel penskoran. Penerapan korelasi intraklas digunakan sebagai ukuran kesesuaian skor yang dihasilkan oleh masing-masing model penskoran (Prihoda, Pinckard, Mc Mahan, et al, 2006, p.379). Koefisien korelasi intraklas (*intraclass*) yang menyatakan kesesuian skor (*agreement*) bahwa secara umum koefisien korelasi intraklas antarmodel penskoran menunjukkan nilai yang cukup tinggi karena lebih dari 0,62 (Thonkine, 2005, p.53). Hal ini berarti bahwa tingkat kesesuian antarmodel penskoran cukup tinggi. Koefisien korelasi intraklas paling tinggi pada NW_2PL dan NW_3PL sebesar 0,996, sehingga model penskoran NW_2PL dan NW_3PL menghasilkan skor yang memiliki urutan rangking yang sama dan menunjukan kesesuaian skor yang relatif sama.

Komparasi Model Penskoran

Dalam menentukan model penskoran manakah yang lebih baik digunakan analisis varians (ANOVA) pada pengukuran berulang dengan model-model penskoran sebagai perlakuan. Seperti disebutkan pada metode penelitian penerapan ANOVA dengan pengukuran berulang dengan prasyarat asumsi *Sphericity*. Jika asumsi ini terpenuhi maka kesamaan varian selisih skor antarperlakuan.

Tabel 5 merupakan uji *Mauchy* untuk menguji asumsi S*phericity* dengan tingkat signifikansi = 0,05. Uji ini menguji hipotesis bahwa variansi selisih skor antarperlakuan sama (Field, 2000, p.470). Berdasarkan hasil pada Tabel 4 diperoleh bahwa *p-value* sebesar 0,000 < 0,05 sehingga dapat disimpulkan bahwa ada perbedaan yang signifikan variansi selisih skor

antarperlakuan. Artinya, asumsi *Sphericity* telah dilanggar. Pelanggaran terhadap asumsi ini menyebabkan ada korelasi antara skor yang dihasilan dari keempat model penskoran, sehingga menghasilkan rasio F tidak valid. Jika hasil dari uji *Sphericity* tidak terpenuhi, maka kita harus melihat *Green-house-Geisser, Huynh-Feldt* atau *Lower-Bound*. SPPS menghasilkan tiga koreksi berdasarkan estimasi *Sphericity* yang diberikan oleh *Greenhose-Geisser, Huynh -Feldt,* dan *Lower bound* (Field, 2000, p.474).

Tabel 6 uji effek dalam perlakukan atau *Tests of Within-Subjects Effects* pada kolom model dan sub baris *Greenhouse-Geisser* hasilnya adalah F = 472,475 dan *p-value* sebesar 0,000 < 0,05) yang artinya terdapat interaksi antara skor dengan model penskoran sehingga asumsi *Sphericity* belum terpenuhi.

Hal ini mengindikasikan bahwa estimasi skor memiliki perbedaan yang signifikan antarmodel penskoran. Selajutnya, akan diuji model-model penskoran manakah yang berbeda dengan menggunakan analisis univarians. Berdasarkan hasil sebelumnya diketahui asumsi *Sphericity* tidak terpenuhi sehingga uji perbandingan ganda dilakukan dengan metode *Bonferroni* karena metode ini paling tahan terhadap pelangaran asumsi *Sphericity* (Field, 2000, p.374). Hasil uji *post hoc* ditunjukkan pada Tabel 7.

Hasil uji perbandingan ganda metode *bonferroni* pada Tabel 7 diperoleh bahwa perbandingan rerata antara model penskoran yang signifikan terjadi pada keempat model penskoran karena memiliki p-value paling besar 0,018 < 0,05. Hal ini dapat dilihat pada Tabel 3 menunjukkan bahwa keempat skor memiliki rerata yang berbeda yaitu NW_2PL (13,334), NW_3PL (13,243), W_2PL (19,011), dan W_3PL (23,129).

Tabel 5. Uji Sphericity Mauchly

Mauchly's Test of Sphericity[b]

Measure:MEASURE_1							
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	Df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Model Penskoran	0,001	4162,099	5	0,000	0,351	0,351	0,333

Tabel 6. Uji effek dalam perlakukan

Tests of Within-Subjects Effects

Measure:MEASURE_1						
Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Model Penskoran	Sphericity Assumed	40648,922	3	13549,641	472,475	0,000
	Greenhouse-Geisser	40648,922	1,054	38560,911	472,475	0,000
	Huynh-Feldt	40648,922	1,054	38550,607	472,475	0,000
	Lower-bound	40648,922	1,000	40648,922	472,475	0,000
Error	Sphericity Assumed	50588,018	1764	28,678		
	Greenhouse-Geisser	50588,018	619,839	81,615		
	Huynh-Feldt	50588,018	620,005	81,593		
	Lower-bound	50588,018	588,000	86,034		

Tabel 7. Uji Post Hoc

Pairwise Comparisons

Measure:MEASURE_1						
(I) Model	(J) Model	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1	2	0,094*	0,031	0,018	0,011	0,177
	3	-5,674*	0,155	0,000	-6,084	-5,264
	4	-9,792*	0,468	0,000	-11,030	-8,555
2	1	-0,094*	0,031	0,018	-0,177	-0,011
	3	-5,768*	0,154	0,000	-6,174	-5,361
	4	-9,886*	0,456	0,000	-11,093	-8,679
3	1	5,674*	0,155	0,000	5,264	6,084
	2	5,768*	0,154	0,000	5,361	6,174
	4	-4,118*	0,331	0,000	-4,994	-3,243
4	1	9,792*	0,468	0,000	8,555	11,030
	2	9,886*	0,456	0,000	8,679	11,093
	3	4,118*	0,331	0,000	3,243	4,994

Hasil uji perbandingan ganda metode *bonferroni* pada Tabel 7 diperoleh bahwa perbandingan rerata antara model penskoran yang signifikan terjadi pada keempat model penskoran karena memiliki p-value paling besar 0,018 < 0,05. Hal ini dapat dilihat pada Tabel 3 menunjukkan bahwa keempat skor memiliki rerata yang berbeda yaitu NW_2PL (13,334), NW_3PL (13,243), W_2PL (19,011), dan W_3PL (23,129).

Pembahasan

Hasil analisis perangkat tes UN mata pelajaran matematika SMA program IPA tahun pelajaran 2013/2014 di provinsi DIY berdasarkan teori respons butir dengan kriteria taraf signifikan 5% dari 40 butir soal terdapat 36 butir soal (90%) yang cocok dengan model IRT 3 PL (36 butir soal) atau sebanyak 4 butir soal yang tidak cocok dengan model IRT 3PL. Butir soal yang tidak memenuhi kriteria tersebut adalah butir soal nomor 04, 08, 13, dan 29.

Menurut Meijer (1996) menjelaskan ada 8 penyebab munculnya respons janggal atau respons jawaban tidak fit model yaitu:
1) mencontek, soal ujian sulit namun peserta tes mampu menjawab banyak item dengan benar, 2) beruntung dalam menebak, peserta tidak terduga mampu memberika respons yang benar pada item sulit, 3) peserta tes bingung atau cemas, 4) lamban, respons peserta tes tidak pernah sampai selesai mengerjakan seluruh item, 5) bahasa, menjelaskan minimnya kemampuan peserta tes dalam memahami item atau instruksi, 6) konstans responders, peserta tes merespon tanpa memikirkan dengan kontan item, 7) over kreatif, peserta tes menafsirkan item dalam cara yang tak lazim atau kreatif dan 8) peserta tes yang kurang teliti dalam memberikan respons pada bagian lembar jawaban. Jika 1 dari 8 penyebab model tidak fit dilakukan oleh peserta tes maka diperoleh item yang tidak fit model. semisal peserta tes menyontek atau beruntung dalam menebak secara buta atau sebagaian.

Beberapa butir soal yang memiliki bobot tinggi, besarnya bobot tiap soal dipengaruhi oleh parameter butir yang digunakan. Butir soal yang bobot tinggi yaitu nomor butir 4, 6, 13, 11, 23, 24 dan 28. Berdasarkan persamaan regresi antara parameter butir dengan pembobotan diperoleh bahwa parameter daya beda memiliki kontribusi paling besar. Berikut ini persamaan regresi.

$\hat{Y}_i = -0{,}142 + 1{,}809\, a_i - 0{,}0021\, b_i$ ……
$\hat{Y}_i = -0{,}117 + 1{,}751\, a_i - 0{,}1006\, b_i + 0{,}294\, c_i$

Keterangan:
$\hat{Y}_i$: besarnya pembobotan item ke-i
a_i: nilai daya beda item ke-i
b_i: nilai tingkat kesukaran item ke-i
c_i : nilai *guessing* item ke-i

Pada phase 2 hasil analisis model IRT 3PL dapat dikemukakan bahwa 40 butir soal yang dianalisis terdapat 18 butir soal yang baik (45%) dengan tingkat kesukaran butir soal terletak pada interval -2 sampai dengan +2, daya beda butir soal tertelak pada interval 0 sampai 2, dan *pseudo-guessing* butir soal teretak pada interval 0 sampai 0,20.

Secara keseluruhan butir soal UN matematika tahun pelajaran 2013/2014 menunjukkan bahwa ada 36 (90%) butir soal yang fit model dengan model IRT 3PL. Harga fungsi informasinya sebesar 21,51 pada θ sebesar 0,55 logit dan kesalahan baku pengukuran sebesar 0,22 logit. Bila dilihat harga kesalahan baku pengukuran dan tingkat kemampuan peserta ujian θ dengan taraf kepercayaan 90%, maka internal (θ – (1,650)*(0,22)) logit ≤ θ ≤ (θ + (1,650)*(0,22)) logit, lebar intervalnya 0,363 + 0,363 = 0,726 satuan. Hal ini berarti estimasi terhadap kemampuan peserta yang sesungguhnya berfluktuasi 0,363 angka disekitar angka kemampuan peserta. Jika dilihat dari skor kemampuan yang terletak dari -3 logit sampai +3 logit dan lebar intervalnya hanya 0,363 satuan maka kesalahan baku pengukuran termasuk kecil.

Hasil analisis menunjukkan bahwa skor yang diperoleh dari masing-masing model penskoran memiliki distribusi yang juling ke kanan sehingga sebagian besar siswa memiliki skor rendah, mungkin paket E03 banyak dikerjakan oleh peserta didik dengan kemampuan sedang sampai bawah.

Estimasi skor yang diperoleh melalui keempat model penskoran menunjukkan bahwa rerata skor NW_2PL hampir sama NW_3PL dan rerata skor NW_2PL lebih rendah daripada skor W_2PL dan W_3PL. Jika dilihat berdasarkan nilai simpangan baku, penyebaran skor berdasarkan model penskoran NW_2PL dan NW_3PL relatif sama. Sedangkan skor berdasarkan model penskoran W_3PL lebih menyebar dari reratanya. Kemungkinan hal ini dipengaruhi oleh 1) paket E03 banyak kerjakan oleh peserta tes di kabupaten Sleman, Bantul, Kulon Progo, dan Gunung Kidul, 2) karena terdapat 10% butir soal yang mengacu *higher order thinking* (HOT) tidak semua peserta tes mampu mengerjakannya.

Koefisien korelasi intraklas, kesesuaian skor dari masing-masing model penskoran pada subjek yang sama menunjukkan tingkat kesesuaian skor yang relatif tinggi yaitu lebih dari 0,62. Hal ini disebabkan oleh asumsi awal dalam model penskoran menggunkan pendekatan teori respons butir dan model penskoran yang memiliki koefisien korelasi intraklass paling tinggi adalah model penskoran NW_2PL dengan NW_3PL sehingga hasil skor yang memiliki urutan rangking dan menunjukkan skor yang relatif sama.

Berdasarkan hasil analisis variansi pengukuran berulang diperoleh informasi bahwa varians selisih skor yang dihasilkan oleh empat model penskoran tersebut berbeda secara signifikan. Kemudian dilanjutkan dengan uji perbandingan ganda metode *bonferroni*. Pada Tabel 19 diperoleh bahwa perbandingan rerata antara model penskoran yang signifikan terjadi pada keempat model penskoran karena memiliki p-value paling besar 0,018 < 0,05. Hal ini berarti bahwa keempat skor memiliki rerata yang berbeda yaitu NW_2PL (13,334), NW_3PL (13,243), W_2PL (19,011), dan W_3PL (23,129). Apabila menggunakan nilai signifikan 0,01 maka diperoleh bahwa model penskoran NW_2PL dengan NW_3PL menunjukkan perbandingan rerata yang tidak signifikan karena memiliki p-value sebesar 0,018 > 0,01. Hal ini juga bisa dilihat pada Tabel 3 menunjukkan bahwa kedua rerata skor relatif sama yaitu rerata model penskoran NW_2PL sebesar 13,334 dan rerata model penskoran NW_3PL sebesar 13,243.

Simpulan dan Saran

Simpulan

Berdasarkan analisis hasil penelitian dan pembahasan, maka simpulan yang dapat dibuat sebagai berikut. Karakteristik perangkat tes ujian nasioanl paket E03 mata pelajaran matematika SMA/MA program IPA tahun pelajaran terdiri 40 butir soal memiliki daya beda, tingkat kesukaran, dan *pseudo guessing* yang baik, serta kesalahan baku pengukuran tergolong kecil yaitu sebesar 0,22 logit . Jumlah butir soal yang baik dan fit model ada 18 (45%) butir soal. Distribusi skor berdasarkan model penskoran NW_2PL, NW_3PL, W_2PL dan W_3PL memiliki *skewness* dan *kurtosis* yang berbeda. Hubungan kesesuaian skor antarmodel pernskoran ditunjukan dengan koefisien korelasi intraklas yang paling tinggi pada model penskoran WN_2PL dengan NW_3PL. Hasil uji perbandingan ganda dengan metode *bonferroni* menunjukkan bahwa ada perbedaan rerata skor yang signifikan antarmodel penskoran.

Saran

Berdasarkan hasil penelitian di atas, maka peneliti menyarankan: hendaknya tim penyusun melakukan analisis karaketristik butir dan nilai informasi sesuai dengan karakteristik peserta tes, sehingga tes yang dihasilakan dapat mengukur kemampuan peserta tes sesungguhnya. Lembaga Penjaminan Mutu Pendidikan DIY hendaknya melakukan sosialiasi kepada guru di daerah setempat dan pihak lain yang berkepentingan sebagai alternatif model penskoran untuk perangkat tes pilihan ganda adalah modle penskoran jumlah benar berdasarkan teori respons butir.

Daftar Pustaka

Field, A. (2000). *Discovering statistics using SPSS for windows*. Advanced techniques for the beginner. London: Sage Publication.

Garcí–Pérez, M. A. & Frary, R. B. (1989). Psychometric Properties of Finite-state Scores Versus Number-correct and Formula Scores: A Simulation Study. *Apllied Psychological Measurement*, 13, 403-417.

Hair, J.F., Anderson, R.E., Tatham, R.L., et al. (1998). *Multivariate data analysis* (5thEd). New Jersey: Prentice-Hall, Inc.

Hambleton, R.K. (1989). *Principles and selected applications of item response theory*. New York: Amrican Council on Education and Mcmillan Publising Company

Lord, F. M. (1980). *Application of item response theory to practice testing problem*. New Jersey : Lawrence Elbaum Associates.

Mardapi, D. (1999). *Estimasi kesalahan pengukuran dalam bidang pendidikan dan implikasinyapada ujian nasional*. Pidato pengukuhan guru besar, disampaikan pada rapat senat terbuka IKIP Yogyakarta.

Mardapi, D. (2008). *Teknik penyusunan instrumen tes dan nontes*.Yogyakarta : Mitra Cendikian Press.

Mardapi, D. (2012). *Pengukuran penilaian & evaluasi pendidikan*.Yogyakarta : Nuha Litera.

Meijer, R.R. (1996). Person-Fit Research: An Introduction. *Applied Measurement in Education, 9*(1), 3-8.

Musmuliadi. (2009). Hubungan model penskoran terhadap estimasi skor sesungguhnya berdasarkan teori respons butir. *Jurnal Penelitian dan Evaluasi Pendidikan. 13*, 246-267.

Prihoda, T.J., Pinckand, R.N., Mc Mahan, A., at al. (2006). Correction for guessing increases validity in multiple-choice examination in an oral and maxillofacial pathology cource. *Journal of Dental Education*, 70, 378-386.

Rofieq, Ainur. 2010. Asesmen pembelajaran di sekolah dasar. Diambil tanggal 01 Agustus 2015. http://pjjpgsd.dikti.go.id/file.php/1/repository/dikti/Mata%20Kuliah%20Awal/Assesment%20Pembelajaran/BAC/assessmen_pembelajaran_6.pdf)

Satoridona, L. S., van der Linden, W. J.,& Meijer, R. R.(2006). Detecting answer copying using the Kappa statistic. *Applied Psychologinal Measurement*, 30, 412-431.

Simon, A. B., Budescu, D. V.,& Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement*, 21, 65-88.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education*, 7th Ed. Upper Saddle River: Pearson/Merrill Prentice Hall.