Jurnal Evaluasi Pendidikan
Volume 4, No 1, Maret 2016 (80-89)
Online: http://journal.student.uny.ac.id/ojs/index.php/jep

APLIKASI MODEL PENSKORAN *EQUAL WEIGHTING* DAN *DIFFERENTIAL WEIGHTING* UNTUK MENGESTIMASI SKOR KIMIA SISWA

[1)]Rizki Nor Amelia, [2)]Farida Agus Setiawati
[1)]Magister Penelitian dan Evaluasi Pendidikan, Universitas Negeri Yogyakarta
[2)]Jurusan Psikologi, Fakultas Ilmu Pendidikan, Universitas Negeri Yogyakarta
[1)]rizkinoramelia@gmail.com, [2)]faridaagus@yahoo.co.id

Abstrak

Penelitian ini bertujuan untuk: (1) mendeskripsikan karakteristik distribusi skor kimia siswa hasil estimasi dari model penskoran *Equal Weighting* (EW) dan *Differential Weighting* (DW), dan (2) melihat kesesuaian skor kimia yang dihasilkan dari penerapan kedua model penskoran tersebut. Penelitian ini merupakan penelitian deskriptif. Data yang diperoleh berupa pola respon siswa SMA terhadap tes *try out* Ujian Nasional mata pelajaran kimia Tahun Ajaran 2015/2016 di Kota Yogyakarta. Analisis data menggunakan Rasch Model dan korelasi intraklass. Sampel penelitian sejumlah 358 siswa diambil menggunakan teknik *cluster random sampling*. Hasil analisis menunjukkan bahwa: (1) rerata skor DW lebih tinggi daripada rerata skor EW, namun skor yang dihasilkan lebih menyebar dari reratanya, keduanya memiliki harga skew negatif sehingga distribusinya juling ke kiri dan kurtosis yang platikurtik; dan (2) penerapan model penskoran merubah peringkat siswa karena kesesuaian skor antarmodel penskoran masuk dalam kategori *fair agreement.*

Kata kunci: *kimia, model penskoran, rasch model*

THE PPLICATION OF SCORING MODEL USING EQUAL WEIGHTING AND DIFFERENTIAL WEIGHTING TO ESTIMATE STUDENTS CHEMISTRY SCORE

[1)]Rizki Nor Amelia, [2)]Farida Agus Setiawati
[1)]Magister Penelitian dan Evaluasi Pendidikan, Universitas Negeri Yogyakarta
[2)]Jurusan Psikologi, Fakultas Ilmu Pendidikan, Universitas Negeri Yogyakarta
[1)]rizkinoramelia@gmail.com, [2)]faridaagus@yahoo.co.id

Abstract

This study aimed to describe: (1) the characteristics of students' chemistry scores distribution based on the estimation of scoring models using Equal Weighting (EW) and Differential Weighting (DW), and (2) the suitability of students' chemistry scores distribution based on both scoring model. This study was a descriptive research. The data obtained from students respons pattern on try out test of National Examination on chemistry subject in the academic year 2015/2016 in Yogyakarta. The analysis in this study used the Rasch model and Intraclass Correlation. The sample of this research was 358 students taken by cluster random sampling technique. The analysis shows that: (1) the mean score of WD is higher than the mean score of NR, but the WD score is more spread out from their mean. Both scores have negative skew so that their distribution is left-tailed and has platycurtic kurtosis; (4) the application of scoring models causes the changes of students' rank because the consistency of students' scores is in the category of fair agreement.

Keywords*: chemistry, scoring model, rasch model*

Pendahuluan

Terdapat empat kompetensi yang wajib dimiliki oleh guru-guru Indonesia, yaitu kompetensi pedagogik, kompetensi profesional, kompetensi sosial, dan kompetensi kepribadian. (Permendiknas No.16, 2007, p.3). Secara khusus, kompetensi pedagogi guru yang harus dilaksanakan adalah kemampuan menyelenggarakan penilaian proses dan hasil belajar yang terdiri dari: (1) memahami prinsip-prinsip penilaian, evaluasi proses, dan evaluasi hasil belajar sesuai dengan karakteristik mata pelajaran yang diampu; (2) menentukan aspek-aspek penilaian dan hasil belajar yang penting untuk dinilai dan dievaluasi sesuai dengan karakteristik mata pelajaran yang diampu; (3) menentukan prosedur penilaian, evaluasi proses, dan evaluasi hasil belajar; (4) mengembangkan instrumen penilaian, evaluasi proses, dan evaluasi hasil belajar; (5) mengadministrasikan penilaian proses dan hasil belajar secara berkesinambungan dengan menggunakan berbagai instrumen; (6) menganalisis hasil penilaian proses dan hasil belajar untuk berbagai tujuan; serta (7) melakukan evaluasi proses dan hasil belajar (Permendiknas No.16, 2007, p.18).

Menyadari kewajiban tersebut, maka dalam pengukuran pendidikan kimia, para pengembang tes (terutama guru kimia) harus mengukur kemampuan, keberhasilan belajar, sikap, minat atau *latent trait* (ciri terpendam) lainnya yang terdapat pada siswa. Namun karena terpendam, berbagai ciri yang akan diukur tidak dapat dilihat maupun diamati. Oleh karena itu, guru kimia memerlukan cara tidak langsung untuk mengukur berbagai ciri yang terdapat pada siswa. Salah satu upaya yang dapat dilakukan adalah pemberian sejumlah stimulus, baik dalam bentuk tes atau kuesioner (Naga, 1992, p.2), meskipun sulit memperoleh alat ukur yang stabil untuk mengukur karakteristik seseorang (Mehrens & Lehmann, 1973, p.103). Jika stimulus itu tepat mengenai sasaran, maka tanggapan atau respons terhadap stimulus yang kelihatan itu dapat menunjukkan kemampuan, keberhasilan belajar, sikap, minat, atau ciri lainnya yang ingin diukur dari para siswa. Respons yang kelihatan itu selanjutnya dapat ditafsirkan melalui pemberian skor yang memadai.

Informasi yang didapatkan dalam kegiatan pra-survey menunjukkan bahwa hampir sebagian besar guru kimia di kota Yogyakarta dalam melaksanakan tes, tidak terkecuali pada *try out* Ujian Nasional (UN), lebih banyak menggunakan bentuk tes pilihan ganda dengan menggunakan model penskoran *Number Right (NR)* secara klasik. Dari hal tersebut dapat dipahami bahwa guru sama sekali tidak memperhatikan tingkat kesukaran dari masing-masing butir. Akibatnya, siswa yang menjawab sepuluh butir "sukar" dengan benar, akan memiliki skor yang sama dengan siswa yang menjawab sepuluh butir "mudah" dengan benar. Hal ini tentu saja belum mencerminkan keadilan. Berbagai alasan yang dikemukakan antara lain kemudahan dalam mengoreksi jawaban siswa jika dibandingkan dengan model penskoran lain dan tidak ada insentif dalam mengoreksi lembar jawaban siswa. Sehingga bentuk tes pilihan ganda dengan model penskoran *NR* secara klasik menjadi satu-satunya alternatif yang paling ekonomis dan tidak merepotkan dalam pengkoreksian pekerjaan siswa.

Kenyataan tersebut semakin memperkuat isu dibidang pengukuran terutama dalam hal *scoring* yang membutuhkan solusi agar skor yang diperoleh siswa benar-benar merefleksikan kemampuannya. Estimasi kemampuan dapat dilakukan dengan pendekatan klasik (*Classical Test Theory*, CTT) maupun pendekatan modern (*Item Response Theory*, IRT). Dalam praktiknya, CTT lebih umum digunakan karena perhitungannya yang relatif lebih sederhana karena kemampuan siswa diukur dari skor yang merupakan akumulasi jumlah butir yang dijawab benar (*NR scoring* klasik). Skor ini selanjutnya diolah oleh guru menggunakan Penilaian Acuan Patokan (PAN) atau Penilaian Acuan Normatif (PAK). Berdasarkan kelemahan tersebut, kemudian dilakukan perbaikan melalui IRT dengan berbagai variasi parameter logistiknya (PL), salah satunya adalah model 1-PL yang dikembangkan menjadi model Rasch. Pendekatan yang dilakukan melalui model Rasch adalah berbeda. Tujuan utama dari pemodelan Rasch adalah membuat skala pengukuran dengan interval yang sama. Hal ini dikarenakan skor (*raw score*) tidak memiliki sifat keintervalan, sehingga skor tidak dapat digunakan secara langsung untuk memberikan penafsiran kemampuan siswa (Sumintono & Widhiarso, 2015, p.37). Jadi, estimasi skor, baik berdasarkan CTT maupun IRT, dapat dilakukan dengan berbagai model penskoran.

Estimasi skor menurut IRT didasarkan pada peluang menjawab benar pada kemampuan θ terhadap masing-masing item soal. Artinya, skor yang diestimasi menggunakan IRT adalah *true score*-nya (Crocker & Algina, 2008,

p.351). *True score* dapat diperoleh dengan mengaplikasikan model-model penskoran. Model penskoran bagi bentuk pilihan ganda sendiri, menurut para ahli dapat dibagi menjadi beberapa macam (Rudner, 2000, p.3; Crocker & Algina, 2008, pp.400-407; Naga, 2013, p.331; Frary, 1989, pp.81-87; Stanley & Wang, 1970, pp.667-676). Pada penelitian ini, model penskoran yang diaplikasikan adalah model penskoran *Equal Weighting* (EW) dan *Differential Weighting* (DW). Salah satu bentuk model penskoran EW adalah *Number Right* (NR), sedangkan salah satu bentuk model penskoran DW adalah *Weighting by Difficulty* (WD) .

Number right scoring disebut juga *adding raw score* (Rudner, 2000, p.3). Model penskoran ini merupakan model penskoran yang paling sederhana karena hanya menjumlahkan peluang menjawab benar pada setiap butir soal dan menganggap setiap butir soal tersebut memiliki bobot yang sama (Lord, 1980, p.230; Baker, 2001, p.66; Crocker & Algina, 2008, p.403; Naga, 1992, p.494). Menurut Rudner (2000, p.3), model penskoran jumlah benar sesungguhnya termasuk model penskoran yang mengakomodasi pembobotan secara implisit dan menganggap semua butir memiliki bobot yang setara (*equal weighting*). Butir soal yang tidak diisi (*omitted*) oleh siswa diberi skor nol, sedangkan butir soal yang dijawab salah tidak mendapatkan hukuman pengurangan nilai.

Apabila metode pembobotan setara (*equal weighting*) memberikan distribusi bobot yang sama besar pada masing-masing butir, maka untuk pembobotan tidak setara (*differential weighting*) memberikan bobot yang berbeda berdasarkan kriteria tertentu yang digunakan. Kriteria tertentu yang biasa dijadikan bobot tersebut adalah panjang butir, validitas butir, dan tingkat kesukaran butir (Stanley & Wang, 1970, pp.675-676). Senada dengan hal tersebut, Mardapi (2008, p.133) menyatakan bahwa bobot setiap soal ujian yang ada dalam suatu perangkat tes ditentukan dengan mempertimbangkan faktor-faktor yang berkaitan dengan materi dan karakteristik soal itu sendiri, seperti luas lingkup materi yang hendak dibuatkan soalnya, esensialitas dan tingkat kedalaman materi yang ditanyakan, serta tingkat kesukaran soal tersebut. Faktor tingkat kesukaran merupakan salah satu faktor yang penting untuk dipertimbangkan karena konsep tingkat kesukaran pada tes objektif lebih bermakna daripada tingkat kesukaran pada tes esai maupun tes lisan (Mehrens & Lehmann, 1973, p.194).

Pada model penskoran *WD*, pembuat soal memberikan bobot pada masing-masing butir berdasarkan intuisinya terhadap level kesukaran butir yang telah dibuatnya atau seberapa berharganya butir tersebut dibanding butir lain (Stanley & Wang, 1968). Cara pembobotan seperti itu disebut juga "*a priori*" (*subjective weighting)*. Oleh karena itu, penskoran model ini tergolong dalam penskoran dengan pendekatan eksplisit (Rudner, 2000, p.3). Guilford & Digmann (1954, p.409), berpendapat bahwa pembobotan berdasarkan "*a priori*" sangat tergantung pada bias pribadi (subjektivitas) sehingga cenderung membahayakan keandalan dan validitas tes kecuali kriteria untuk pembobotan butir telah ditetapkan secara konsisten dan ketat. Meskipun begitu, menurut Feldt (2004) sebagaimana dikutip Sanghoon Mun (2014, p.1), terdapat dua keunggulan yang ditawarkan oleh model pembobotan DW. Pertama, model DW berasumsi bahwa model EW tidak dapat merefleksikan pentingnya konten uji. Kedua, guru dapat meminimalisir sejumlah siswa yang mendapatkan skor sama sehingga siswa-siswa tersebut dapat dibedakan dengan berdasarkan kapasitasnya.

Dari sekian banyak model penskoran yang dicetuskan oleh para ahli, ternyata penelitian tentang model penskoran yang ada di Indonesia sebagian besar hanya berkutat pada *reward scoring* dan *punishment scoring*, misalnya penelitian Slamet & Maarif (2014, p.54), Kurniawan (2012, p.1), maupun Wijaya (2005, p.ii). Sementara itu, penelitian di luar negeri memang sudah lebih variatif dan aplikatif, namun fokus penelitiannya tidak melihat efek model penskoran yang digunakan terhadap estimasi skor siswa, misalnya penelitian yang dilakukan oleh Lau, et.al. (2011, p.99); Hoe, et.al (2009, p.51-57); maupun Yen, et.al (2010, p.174).

Sebenarnya penelitian tentang efek model penskoran terhadap estimasi skor siswa pernah dilakukan oleh Musmuliadi (2009, p.ii) dan Huda (2015, p.ii). Akan tetapi kedua penelitian tersebut memiliki beberapa keterbatasan, salah satu diantaranya adalah pembobotan yang digunakan masih merupakan *equal weighting*. Oleh karena itu, penelitian ini berusaha mengakomodasi *differential weighting* untuk melihat efeknya terhadap skor kimia yang dihasilkan. Selama ini, di Indonesia khususnya, DW hanya diterapkan bagi bentuk tes non-

objektif (essay). Padahal secara teoritis, pembobotan tipe tersebut dapat diterapkan juga bagi bentuk tes pilihan ganda melalui pembobotan berdasarkan tingkat kesukaran butir (*weighting by difficulty*) yang dilakukan secara *a priori* atau *logical judgement* (Rudner, 2000, p.3; Stanley & Wang, 1968, p.29). Berdasarkan latar belakang yang telah diuraikan tersebut, maka peneliti tertarik untuk menyelidiki skor kimia siswa hasil estimasi dari penerapan model penskoran yang berbeda berdasarkan *IRT*.

Metode Penelitian

Penelitian ini termasuk jenis penelitian deskriptif dengan pendekatan kuantitatif. Penelitian dilaksanakan di SMA Negeri Kota Yogyakarta Tahun Ajaran 2015/2016 pada bulan Februari-April 2016. Populasi dalam penelitian ini sebanyak 2.101 siswa, sedangkan sampel yang digunakan sebanyak 358 siswa. Sampel tersebut diambil menggunakan kriteria Tabel Morgan dengan teknik *cluster random sampling* berdasarkan area atau wilayah.

Data berupa pola respon siswa dalam menjawab instrumen tes *try out* UN Kimia Tahun Ajaran 2015/2016 dalam penelitian ini diambil menggunakan teknik dokumentasi. Data yang diperoleh dianalisis menggunakan Rasch Model dengan bantuan program WIN-STEPS dan Korelasi Intraklas dengan bantuan program SPSS. Hasil analisis data berupa distribusi skor hasil estimasi menggunakan dua model penskoran yang berbeda dipaparkan secara deskriptif melalui nilai *mean*, *standard deviation*, *skewness*, dan *kurtosis*. Adapun persamaan matematis yang digunakan untuk melakukan estimasi skor kimia siswa, yaitu:

*Equal Weighting (Number Right) (*Crocker & Algina, 2008, p.403)

$$\boldsymbol{\xi_a = \Sigma^{(a)} P_g(\theta)} \text{ ..... (1)}$$

dengan:

ξ_a = *number right score* siswa
$\Sigma^{(a)}$ = banyaknya butir yang dijawab oleh siswa a
$P_g(\theta)$ = peluang siswa a dengan kemampuan θ dalam menjawab butir g dengan benar

*Differential Weighting (Weighting by Difficulty) (*Lord, 1980, p.73)

$$y = \sum_{i=1}^{n} w_i P_i(\theta) \text{ ..... (2)}$$

dengan :

y = skor berdasarkan model penskoran pembobotan
w_i = bobot butir ke-i
$P_i(\theta)$ = peluang siswa dengan kemampuan θ dalam menjawab butir *I* dengan benar

Dikarenakan salah satu tujuan penelitian ini adalah menyelidiki skor hasil model penskoran *differential weighting* berdasarkan tingkat kesukaran butir dari *logical judgement* terhadap estimasi skor siswa, maka diperlukan guru-guru yang berkompeten untuk melakukan pembobotan. Ada sembilan guru yang dipilih untuk menetapkan bobot butir. Guru-guru tersebut adalah guru-guru yang berkompeten dalam kepenulisan soal sesuai rekomendasi Dinas Pendidikan Kota Yogyakarta. Adapun besarnya bobot yang diberikan pada masing-masing butir soal tercantum dalam Tabel 1.

Tabel 1. Besarnya Bobot Tiap Tingkat Kesukaran Butir Soal

Tingkat Kesukaran Butir	Bobot
Mudah	1
Sedang	2
Sukar	3

Hasil Penelitian dan Pembahasan

Deskripsi Karakteristik Distribusi Skor Kimia Siswa Hasil Estimasi dari Model Penskoran *Equal Weighting* (EW) dan *Differential Weighting* (DW)

Estimasi skor kimia dalam penelitian ini dilakukan berdasarkan pendekatan modern (IRT) menggunakan Rasch Model. Ada dua model penskoran yang diaplikasikan yaitu model penskoran *Equal Weighting* (EW) dan *Differential Weighting* (DW). Estimasi skor berdasarkan kedua model penskoran tersebut berturut-turut diperoleh dari persamaan (1) dan (2). Tabel 2 merupakan statistik deskriptif skor kimia hasil estimasi menggunakan masing-masing model penskoran dengan bentuk distribusi

skor seperti ditunjukkan oleh Gambar 1 dan Gambar 2.

Sebelum melakukan estimasi skor kimia menggunakan kedua model penskoran, terlebih dahulu dilakukan analisis kecocokan model berupa kecocokan butir (*item fit*) dan kecocokan responden (*person fit*). Kecocokan model pada output program WINSTEPS dapat dilihat dari tingkat kesukaran butir atau *item measure* dan abilitas peserta tes atau *person measure*. Tingkat kesukaran butir maupun abilitas peserta tes dikatakan cocok dengan model jika harga OUTFIT MNSQ berada dalam kisaran 0,5-1,5 (Linacre, 2002, p.878). Berdasarkan kriteria tersebut, maka semua butir dalam instrumen *try out* UN Kimia cocok dengan model Rasch dan 29 responden (peserta tes) tidak cocok (*misfit*) dengan model Rasch karena berada diluar range OUTFIT MNSQ yang telah ditetapkan. Adapun peserta tes yang tidak cocok adalah peserta tes dengan kode 22, 34, 35, 48, 49, 53, 55, 77, 90, 97, 103, 110, 118, 125, 128, 189, 195, 220, 228, 236, 241, 275, 276, 278, 280, 281, 302, 333, dan 337.

Hasil analisis kecocokan *person* menyimpulkan bahwa terdapat 29 *person misfit*, sehingga hanya 329 respon yang dianalisis menggunakan kedua model penskoran. Selain menggunakan kriteria OUTFIT MNSQ, cara mengidentifikasi *person misfit* adalah melalui matriks Guttman atau *scalograms*. Matriks Guttman mampu memberikan informasi yang berharga karena butir soal telah diurutkan dari butir termudah (nomor 18) hingga butir tersukar (nomor 17). Matriks ini juga bisa menunjukkan unidimensionalitas data (Hambleton & Swaminathan, 1985, p.22).

Tabel 2. Statistik Deskriptif Skor

Statistik	EW	DW
N	329	329
Mean	24,99	48,09
Median	24,78	47,54
Mode	31,58	61,68
Std. Deviation	8,28	17,12
Skewness	-0,09	-0,06
Std. Error of Skewness	0,13	0,13
Kurtosis	-1,00	-1,03
Std. Error of Kurtosis	0,27	0,27
Minimum	5,40	8,95
Maximum	39,81	79,50

Tabel 2 menunjukkan bahwa rerata (*mean*) skor EW siswa adalah 24,99 dengan nilai simpangan baku (*standard deviation*) sebesar 8,28 dan rentang skor dari 5,40 sampai 39,81. Sementara itu, rerata (*mean*) skor DW siswa adalah 48,09 dengan nilai simpangan baku (*standard deviation*) sebesar 17,12 dan rentang skor dari 8,95 sampai 79,50. Berdasarkan paparan tersebut, dapat disimpulkan bahwa rerata skor DW lebih tinggi daripada rerata skor EW. Jika dilihat berdasarkan nilai simpangan baku, skor DW relatif lebih menyebar dari reratanya dibandingkan skor EW.

Selanjutnya, untuk melihat bentuk distribusi *skewness* (kepencongan) dan *kurtosis* (kepuncakan), maka digunakan pedoman dari Sulistyo (2010, p.72). Jika rasio *skewness* maupun *kurtosis* berada di antara -2 sampai 2, maka distribusi data adalah normal. Rasio *skewness* dapat dihitung dengan cara membagi nilai *skewness* terhadap standar errornya, demikian pula rasio *kurtosis* juga dihitung dengan cara membagi nilai *kurtosis* terhadap standar *error*-nya. Hasil perhitungan pada Tabel 3 menunjukkan bahwa tidak terpenuhinya kriteria *kurtosis* yang baik, maka dapat disimpulkan bahwa distribusi skor tampak dari peserta tes tidak normal.

Tabel 3. Ringkasan Hasil Perhitungan bagi Rasio Skewness dan Rasio Kurtosis

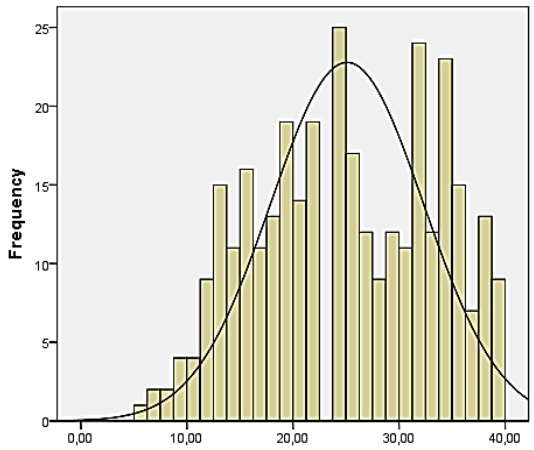
Data	**Rasio Skewness**	**Rasio Kurtosis**
Skor EW	-0,716	-3,753
Skor DW	-0,470	-3,854

Selain menggunakan kriteria rasio *skewness* dan rasio *kurtosis*, bentuk distribusi data juga dapat disimpulkan dari hasil uji normalitas skor menggunakan statistik one-Sample Kolmogorov-Smirnov pada Tabel 4. Hasil uji terhadap kedua data skor hasil model penskoran menunjukkan signifikansi yang lebih kecil dari α (0,003 < 0,05). Selanjutnya, dikarenakan kedua model penskoran menghasilkan nilai rerata yang lebih besar dari median dan harga *skew* yang negatif, maka distribusi skor hasil estimasi menggunakan kedua model penskoran tersebut membentuk kurva yang sedikit juling ke kiri. Kurva yang juling ke kiri menunjukkan bahwa sebagian besar peserta tes mendapatkan skor yang relatif tinggi. Sementara itu, baik skor EW, maupun DW memiliki harga *kurtosis* yang

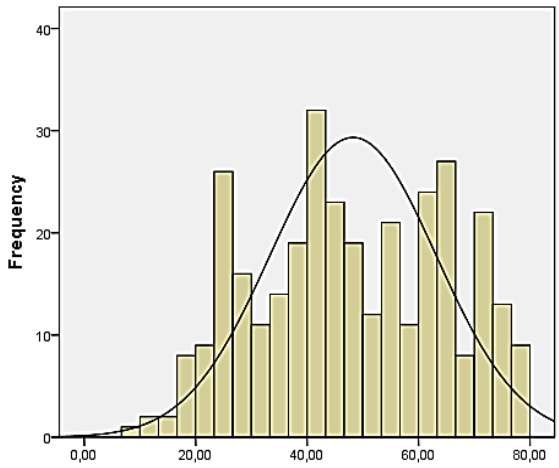
negatif atau platikurtik. Hal ini mengindikasikan bahwa distribusi data pada median dan di sekitar median sama banyaknya (cenderung merata atau datar).

Tabel 4. Uji Normalitas Skor

One-Sample Kolmogorov-Smirnov Test			
		EW	DW
Normal Parameters	Mean	24,99	48,09
	Std. Deviation	8,22	17,12
Kolmogorov-Smirnov Z		1,809	1,802
Asymp. Sig. (2-tailed)		0,003	0,003



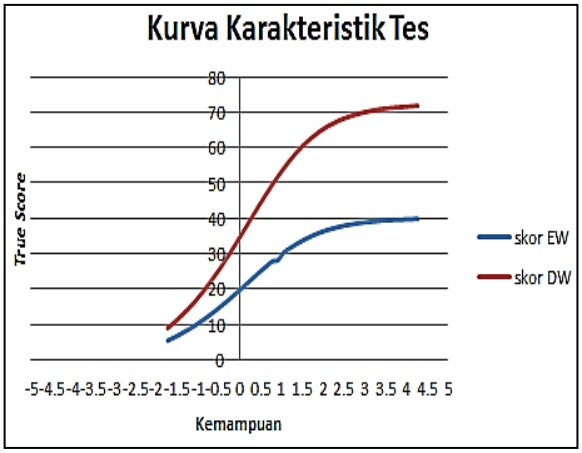
Gambar 1. Distribusi Skor EW



Gambar 2. Distribusi Skor DW

Skor kimia hasil estimasi menggunakan kedua model penskoran menunjukkan bahwa bentuk distribusinya tidak normal atau cenderung juling ke kiri. Bentuk distribusi ini menyiratkan bahwa rerata skor yang dihasilkan lebih tinggi daripada mediannya atau akibat penerapan kedua model penskoran ini menyebabkan skor siswa menjadi lebih tinggi daripada skor tampak yang diestimasi secara klasik. Hal ini terjadi karena kedua model penskoran yang diaplikasikan telah berbasis IRT, sehingga yang diakumulasikan adalah peluang menjawab benarnya atau $P(\theta)$. Dikarenakan perhitungan skor didasarkan pada kemampuan siswa, maka dapat dikatakan bahwa distribusi yang sedikit juling ke kiri ini menunjukkan bahwa sebagian siswa memiliki kemampuan kimia yang tinggi.

Kemampuan (θ) yang dihasilkan oleh IRT diukur pada skala yang kurang lumrah karena dapat bernilai negatif, sehingga mempersulit orang untuk melakukan interpretasi (Hambleton & Swaminathan, 1985, p.56). Agar lebih mudah diinterpretasi, kemampuan tersebut selanjutnya dinyatakan dalam bentuk skor sesungguhnya atau *true score*. Konsep kemampuan maupun *true score* pada dasarnya sama kecuali pada skala pengukurannya. Kemampuan didefinisikan pada interval ($-\infty$ sampai $+\infty$), sedangkan *true score* didefinisikan pada interval (0,n) (Hambleton & Swaminathan, 1985, p.62; Lord, 1980: 46). Berdasarkan persamaan yang dituliskan Crocker & Algina (2008, p.352), skor hasil estimasi, baik menggunakan model penskoran EW maupun DW, dapat disebut dengan skor sesungguhnya (*true score*). Hubungan antara *true score* dengan kemampuan (θ) dinyatakan oleh persamaan (1) dan (2) yang lebih umum disebut Kurva Karakteristik Tes (*Test Characteristic Curve*, *TCC*). Kurva ini menggambarkan hubungan fungsional antara antara *true score* dengan kemampuan (Baker, 2001, p.70) dan merupakan fungsi yang naik secara monoton (Hambleton & Swaminathan, 1985, p.62). Dalam Gambar 3 disajikan *TCC* masing-masing *true score* terhadap θ.



Gambar 3. Kurva Karakteristik Tes Model EW dan DW

Gambar Kurva Karakteristik Tes tersebut sekaligus berfungsi sebagai visualisasi tentang skor DW yang lebih tinggi daripada skor EW. Menurut Baker (2001. pp.71-72), pada beberapa kasus bentuk kurva ini hampir berbentuk garis lurus, namun pada banyak tes umumnya berbentuk tidak linear; pada kasus yang lain fungsi ini naik secara halus, kemudian memiliki sedikit kelandaian sebelum naik lagi sampai membentuk asimtot. Berdasarkan pernyataan tersebut, dapat disimpulkan bahwa bentuk kurva karakteristik tes model DW mengadopsi kasus pertama sedangkan bentuk kurva karakteristik tes model EW mengadopsi kasus kedua. Jadi, secara umum, dapat disimpulkan bahwa semakin tinggi kemampuan (θ) maka akan semakin tinggi pula *true score* siswa pada kedua model penskoran.

Kesesuaian Skor Kimia yang Dihasilkan dari Penerapan Model Penskoran *Equal Weighting* (EW) dan *Differential Weighting* (DW)

Penerapan kedua model penskoran untuk mengestimasi skor kimia siswa berefek pada hasil estimasi skor yang diperoleh. Efek yang dimaksud dalam penelitian ini adalah apakah terdapat perbedaan peringkat siswa berdasarkan skor yang hasilkan. Dengan diketahui adanya efek dari kedua model penskoran terhadap hasil estimasi skor kimia siswa SMA, maka guru dapat mempertimbangkan penggunaan model penskoran pada bentuk tes pilihan ganda, sehingga skor yang dihasilkan benar-benar adil dan mampu merefleksikan kemampuan siswa yang sesungguhnya.

Penerapan korelasi intraklas digunakan sebagai ukuran konsistensi skor yang dihasilkan oleh masing-masing model penskoran (Prihoda, et.al, 2006, p.379). Penggunaan korelasi produk Momen Pearson dinilai kurang tepat karena korelasi tersebut digunakan untuk mengukur derajat hubungan dua variabel (Muller & Buttner, 1994, p.2465; Widhiarso, n.d., p.1), bukan mengukur derajat konsistensi skor sehingga tidak diketahui sejauh mana kekonsistensian skor yang dihasilkan oleh masing-masing model penskoran. Berdasarkan Tabel 5, didapatkan koefisien korelasi intraklas (*Intraclass Correlation Coefficient*, ICC) antarmodel penskoran sebesar 0,377. Angka tesebut menunjukkan bahwa tingkat konsistensi skor antara skor hasil estimasi menggunakan model penskoran EW dan skor hasil estimasi menggunakan model penskoran DW masuk dalam kategori *fair agreement* (Chang, n.d., p.1).

Tabel 5. Koefisien Korelasi Intraklas antarmodel Penskoran

Measure	ICC	F Test with True Value			
		Value	df1	df2	Sig
Single	0,317	8,255	328	328	0,00
Average	0,481	8,255	328	328	0,00

Setidaknya terdapat dua alasan yang dapat menjelaskan mengapa model penskoran yang digunakan tidak memberikan kesesuaian skor secara signifikan. Alasan pertama, butir yang tidak dijawab (*omitted response*) maupun butir yang tidak sempat dijawab (*not reached response*) diberi skor nol. Hal ini sesuai dengan prosedur Lord (1980, pp.226-227) yang menyatakan bahwa "*if number right scored used, then there will be no omitted and not reached responses"*. Pemberian skor nol ini tentu saja sangat merugikan, baik bagi siswa dengan abilitas (θ) tinggi maupun siswa dengan abilitas rendah. Padahal jika siswa menjawab butir dengan salah, tidak ada hukuman berupa pengurangan nilai. Oleh karenanya, strategi paling dominan dilakukan siswa adalah menjawab semua butir soal. Hal inilah yang juga merupakan salah satu kelemahan model penskoran EW maupun DW, dimana kedua model penskoran ini memang tidak sensitif terhadap tebakan yang dilakukan siswa. Berikut adalah beberapa contoh estimasi skor pada siswa dengan kode 237, 268, 334 dan 336.

Tabel 6. Contoh Estimasi Skor NR dan WD

Kode	X	θ	EW	DW
S_268	24	0.52	25,198	48,371
S_237	24	0.52	24,484	46,9313
S_336	18	-0.24	17,1765	31,8279
S_334	18	-0.24	14,7056	26,2057

Tabel 6 memperlihatkan bahwa siswa dengan kode 237 (S_237) memiliki skor tampak (X) yang sama dengan siswa dengan kode 268 (S_268), sedangkan siswa dengan kode 334 (S_334) memiliki skor tampak yang sama dengan siswa dengan kode 336 (S_336). Namun, S_237 dan S_334 memiliki hasil estimasi skor berdasarkan kedua model penskoran yang relatif lebih rendah daripada S_268 dan S_336.

Hal ini disebabkan karena dalam pola respon S_237 terdapat dua butir soal omit yaitu butir nomor 5 dan 17, sedangkan S_334 memiliki 7 butir omit yaitu butir nomor 16, 20, 28, 32, 34, 35, dan 38. Jadi, dapat disimpulkan bahwa keberadaan omit menyebabkan skor EW dan DW siswa tersebut menjadi lebih rendah daripada yang seharusnya, sehingga pada akhirnya mempengaruhi urutan peringkat siswa.

Alasan kedua, digunakannya model Rasch sebagai dasar estimasi kemampuan siswa. Model Rasch melakukan estimasi kemampuan hanya mendasarkan pada banyaknya butir yang dijawab benar oleh siswa. Hal ini dikarenakan kesempatan untuk menyelesaikan satu soal dengan benar hanya bergantung pada rasio kemampuan seseorang dan tingkat kesukaran soal (Sumintono & Widhiarso, 2015, p.44). Dengan demikian, siswa yang memiliki jumlah jawaban benar yang sama (skor tampak sama), maka akan dihasilkan estimasi kemampuan yang sama pula, terlepas dari karakteristik butir tertentu yang dijawab dengan benar (Masters, 1988, p.16; DRC, 2016, p.4). Akibatnya, siswa dengan level abilitas (θ) yang sama, akan memiliki skor NR yang sama pula (Lord, 1980, p.45).

Hasil penelitian Huda (2014, p.ii) menggunakan model penskoran NR 2-PL dan NR 3-PL menunjukkan bahwa skor tampak yang sama dapat memberikan abilitas (θ) yang berbeda, sehingga hasil skor NR juga berbeda. Selain itu, koefisien korelasi intraklas pada model penskoran NR 2-PL dan NR 3-PL sebesar 0,996. Artinya kedua model penskoran menghasilkan skor yang memiliki urutan *ranking* dan menunjukkan kesesuaian skor yang relatif sama. Penggunaan model 3-PL sebagai model yang baik untuk melakukan estimasi kemampuan juga didukung oleh DRC (2016, p.4) yang menyatakan bahwa model 3-PL memberikan hasil yang lebih akurat pada pengukuran kemampuan individu siswa jika dibandingkan dengan model yang lainnya.

Berdasarkan paparan dari alasan kedua tersebut, maka dapat disimpulkan bahwa dalam penelitian ini, penggunaan Rasch Model memberikan data yang kurang sensitif. Namun, dengan bantuan *output scalogram*, tetap dapat diidentifikasi siswa mana yang memiliki kemampuan lebih tinggi meskipun secara numerik abilitasnya sama. Siswa teridentifikasi memiliki kemampuan tinggi manakala siswa tersebut konsisten dalam mengerjakan soal-soal yang sukar (Sumintono & Widhiarso, 2015: 45). Gambar 4 adalah contoh identifikasi abilitas siswa menggunakan *scalogram*.

Gambar 4 berisi pola respon bagi siswa dengan kode 64, 110, 122, 126, 145, 221, dan 252 yang sama-sama memiliki skor tampak sebesar 36. Dengan melihat konsistensi responden untuk menjawab soal yang sukar, maka akan terlihat bahwa siswa dengan kode 64 memiliki kemampuan yang lebih tinggi daripada siswa-siswa di bawahnya. Hal ini dikarenakan siswa kode 64 lebih sukses mengerjakan soal yang tingkat kesukarannya lebih tinggi dibanding siswa-siswa di bawahnya. Sebenarnya, selain mengurutkan dari butir termudah hingga butir tersukar (dari kiri ke kanan), scalogram juga mengurutkan siswa dari abilitas tertinggi hingga abilitas terendah (dari atas ke bawah). Jadi dengan mudah disimpulkan bahwa meskipun memiliki skor tampak yang sama, siswa 64 memiliki abilitas yang lebih tinggi daripada siswa 110, 122, 126, 145, 221, dan 252.

```
GUTTMAN SCALOGRAM OF RESPONSES:
Person |Item
       |1223  2  3231  1 2  13231241323321123311
       |82902178753249604453686931096843051571 27
       |----------------------------------------
    64 +11111111111111111111101111100111111111 0
   110 +11110111111111111110111011101111111111 1
   122 +11111111111110111111111111111111110110 110
   126 +11111111111111111111010111110111111110 1
   145 +11111111111111111111101110111111111011 0
   221 +11111111111111110111111111011101010111 1
   252 +11111111101111111111111101110011111111 1
```

Gambar 4. Skalogram

Simpulan dan Saran

Simpulan

Berdasarkan deskripsi hasil penelitian dan pembahasan yang telah dijelaskan, maka dapat diambil kesimpulan sebagai berikut. Pertama, rerata skor DW lebih tinggi daripada rerata skor EW, namun skor yang dihasilkan lebih menyebar dari reratanya. Berdasarkan harga *skew* dan *kurtosis*nya, dapat disimpulkan bahwa distribusinya tidak normal atau juling ke kiri. Bentuk distribusi ini menyiratkan bahwa rerata skor yang dihasilkan lebih tinggi daripada mediannya atau akibat penerapan kedua model penskoran ini menyebabkan skor siswa menjadi lebih tinggi daripada skor tampak yang diestimasi secara klasik. Baik skor EW, maupun DW memiliki harga *kurtosis* yang negatif atau platikurtik, sehingga distribusi datanya cenderung merata atau datar.

Kedua, penerapan model penskoran yang berbeda menghasilkan skor yang berbeda, sehingga berefek pada peringkat siswa yang berbeda pula. Konsistensi skor antara skor hasil estimasi menggunakan model penskoran EW dan skor hasil estimasi menggunakan model penskoran DW masuk dalam kategori *fair agreement*.

Saran

Berdasarkan simpulan dan keterbatasan penelitian ini, dapat diberikan saran sebagai berikut. Pertama, penelitian yang terkait dengan estimasi skor (dalam hal ini *true score*) masih perlu ditindaklanjuti. Kedua, pada penelitian ini, data set yang digunakan masih relatif terbatas serta mata pelajaran yang dianalisis hanya satu. Terkait dengan hal tesebut, perlu ada penelitian lebih lanjut yang menggunakan data set yang lebih bervariasi agar diperoleh yang lebih komprehensif.. Ketiga, Lembaga Penjaminan Mutu Pendidikan (LPMP) DIY agar lebih aktif dalam mensosialisasikan kepada guru-guru tentang perlunya analisis butir soal ketika mengembangkan butir soal dalam menyusun tes yang akan digunakan untuk menguji kompetensi siswa.

Daftar Pustaka

Baker, F.B. (2001). *The basics of item response theory (2nd Ed*). USA: ERIC Clearinghouse on Assessment and Evaluation.

Chang, A. (n.d). Intraclass correlation coefficient explained. Diambil tanggal 20 April 2016 pada http://www.statstodo.com/ICC_Exp.php.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. New York: Holt, Reinhart, and Winston, Inc.

Data Recognition Corporation (DRC). (2016). Accuracy of test scores: why IRT models matter. DRC CTB.

Frary, R.B. (1989). Partial-credit scoring methods for multiple-choice test. *Applied Measurement in Education*. 2(1). 79-96.

Guilford, J.P. & Dingman, H.F. (1954). A validation study of ratio-judgement methods. *The American Journal of Psychology*. 67 (3). 395-410.

Hambleton, R.K., & Swaminathan, H. (1985). *Items response theory: principles and application.* Boston: Kluwer-Nijjhoff Publish.

Hoe, L.S. et.al. (2009). Improving educational assessment: a computer-adaptive multiple choice assessment using NRET as the scoring method. *US-China Education Review*. 6(5). 51-60.

Huda, N. (2015). *Komparasi model penskoran berdasarkan teori respon butir pada soal ujian nasional mata pelajaran matematika SMA/MA program IPA*. Tesis magister, tidak diterbitkan, Universitas Negeri Yogyakarta, Yogyakarta.

Joko Sulistyo. (2012). *6 hari jago SPSS 17*. Yogyakarta: Cakrawala.

Kurniawan, F. (2012). *Pengaruh penskoran sistem denda (penskoran dengan koreksi atas jawaban tebakan) terhadap motivasi belajar siswa*. Skripsi, tidak diterbitkan, Universitas Negeri Malang, Malang.

Linacre. (2002). What do infit and outfit mean-square and standardized mean?. *Rasch Measurement Transaction*, 16, 878.

Lau, P.N.K., et.al. (2011). Guessing, partial knowledge, and misconceptions in multiple-choice tests. *Educational Technology & Society*. 14(4). 99-110.

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

Menteri Pendidikan dan Kebudayaan. (2007). *Permendiknas No.16, Tahun 2007, tentang Standar Kualifikasi Akademik dan Kompetensi Guru.*

Naga, D.S. (1992). *Teori sekor pada pengukuran pendidikan*. Jakarta: Gunadarma Mehrens, W.A., & Lehmann, J.L. (1973). *Measurement and evaluation in education and psychology*. New York: Holt, Rinehart, and Winston, Inc.

Masters, G.N. (1988). Item discrimination: when more is worse. *Journal of Education Measurement*. 23 (1). 15-29.

Muller, R., & Buttner, P. (1994). Critical disscussion of intraclass correlation coefficients. *Statistics in Medicine*. 13. 2465-2476.

Musmuliadi, M. (2013). Hubungan model penskoran terhadap estimasi skor sesungguhnya berdasarkan teori respon butir. *Jurnal Penelitian dan Evaluasi Pendidikan*. 13 (2). Retrieved from http://journal.uny.ac.id/index.php/jpep/article/view/1412/1199

Naga, D. S. (1992). *Teori sekor pada pengukuran pendidikan*. Jakarta: Gunadarma Mehrens, W.A., & Lehmann, J.L. (1973). *Measurement and evaluation in education and psychology*. New York: Holt, Rinehart, and Winston, Inc.

Prihoda, T.J., A., et.al. (2006). Correction for guessing increase validity in multiple-choice examination in an oral and maxillofacial pathology course. *Journal of Dental Education*. 70. 378-386.

Rudner, L.M. (2000). *Informed test component weighting*. USA: Educational Resources Information Center (ERIC), US Department of Education.

Sanghoon Mun. (2014). *A study on teachers' item weighting and the rasch model: summative test items' difficulty logits calibration using the rasch model.* Disertasi doktor, tidak diterbitkan, University of Bath, Thailand.

Slamet & Maarif, S. (2014). Pengaruh bentuk tes formatif assosiasi pilihan ganda dengan reward dan punishment score pada pembelajaran matematika siswa SMA. *Jurnal Ilmiah Program Studi Matematika STKIP Siliwangi Bandung*. 3 (1). Februari 2014.

Stanley, J.C., & Wang M.D. (1968). *Differential weighting: a survey of methods and empirical studies*. USA: Departmen of Health, Education, & Welfare.

Stanley, J.C., & Wang M.D. (1970). Differential weighting: a review of methods and empirical studies. *Review of Educational Research*. 40 (5). 663-705.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi permodelan rasch pada assessment pendidikan*. Cimahi: Trim Komunikata

Widhiarso, W. (n.d). Berkenalan dengan korelasi intraklas. Diambil tanggal 23 April 2016 dari http://elisa1.ugm.ac.id/files/wahyu_psy/NLSQ3Vmj/Berkenalan%20dengan%20Korelasi%20Intrakelas.pdf

Wijaya, Y.S. (2005). P*erbandingan fungsi informasi butir model logistik dua parameter ditinjau dari model penskoran tes pilihan ganda pada siswa SMA DKI jakarta tahun 2004*. Disertasi doktor, tidak diterbitkan, Universitas Negeri Jakarta, Jakarta.

Yen, Y.C., et.al. (2010). Development and evaluation of a confidence-weighting computerized adaptive testing. *Educational Technology & Society*. 13(3). 163-176.