MODEL PENSKORAN *PARTIAL CREDIT* PADA BUTIR MULTIPLE TRUE-FALSE BIDANG FISIKA

Wasis
Jurusan Fisika FMIPA UNESA
Gedung D1 Kampus Ketintang Jl Ketintang, 60231
wasisfaa@yahoo.com

Abstrak

Tujuan penelitian ini menghasilkan model penskoran politomus untuk respons butir *multiple true-false*, sehingga dapat mengestimasi secara lebih akurat kemampuan di bidang fisika. Pengembangan penskoran menggunakan *Four-D model* dan diuji akurasinya melalui penelitian empiris dan simulasi. Penelitian empiris menggunakan 15 butir *multiple true-false* yang diambil dari soal UMPTN tahun 1996-2006 dan dikenakan pada 410 mahasiswa baru FMIPA Universitas Negeri Surabaya angkatan tahun 2007. Respons peserta tes diskor dengan tiga model *partial credit* (PCM I; II; dan III) dan secara dikotomus. Hasil penskoran dianalisis dengan program Quest untuk mendapatkan estimasi tingkat kesukaran butir (δ) dan estimasi kemampuan peserta (θ) untuk menentukan nilai fungsi informasi tes dan kesalahan baku estimasi. Penelitian simulasi menggunakan data bangkitan berdasarkan parameter empiris (δ dan θ) memakai program statistik SAS dan akurasi estimasinya dianalisis dengan metode *root mean squared error* (RMSE). Hasil penelitian ini menunjukkan: (i) Penskoran PCM dengan pembobotan mampu mengestimasi kemampuan lebih akurat dibandingkan tanpa pembobotan maupun secara dikotomus; (ii) Semakin banyak jumlah kategori dalam penskoran *partial credit,* semakin akurat.

Kata kunci: *model penskoran partial credit, butir multiple true-false*

THE PARTIAL CREDIT SCORING MODEL FOR THE MULTIPLE TRUE-FALSE BUTIRS IN PHYSICS

Wasis
Surabaya State University
Gedung D1 Kampus Ketintang Jl Ketintang, 60231
wasisfaa@yahoo.com

Abstract

This study is an attempt to overcome the weaknesses. This study aims to produce a polytomous scoring model for responses to multiple true-false butirs in order to get a more accurate estimation of abilities in physics. It adopts the Four-D model and its accuracy is assessed through empirical and simulation studies. The empirical study employed 15 multiple true-false butirs taken from the New Students Entrance Test of State University the year of 1996–2006. It administered to 410 new students enrolled in 2007 of Faculty of Mathematics and Science of Surabaya State University. The testees' responses were scored using the partial credit model (PCM) I; II; and III and also dichotomously scored. The results of the four scoring models were analyzed using the Quest program to obtain the estimation of the butir difficulty level (δ) and that of the testees' abilities (θ). The generating of the simulation data used the SAS statistical program and the estimation accuracy was analyzed by using the root mean squared error (RMSE) method. The results of the study show the following: (i) The scoring with the partial credit model with weighting is capable of estimating abilities more accurate than without weighting and dichotomous scoring; (ii) The more the number of the categories in the partial credit scoring is, the more accurate the result of the ability estimation.

Keywords: *partial credit model scoring, multiple true-false butir*

Pendahuluan

Metode *testing* dengan butir berbentuk pilihan ganda masih dominan dipergunakan dalam berbagai keperluan pengujian, utamanya pengujian dalam skala besar dan hasilnya ingin segera diketahui, misalnya ujian seleksi, ujian sekolah, dan ujian nasional (Dittendik, 2003; Oosterhof, 2003; Rodriguez, 2005). Hal di atas dikarenakan butir bentuk pilihan ganda memiliki beberapa kelebihan, antara lain cakupannya luas, waktu pelaksanaan relatif singkat, dapat diolah dengan cepat, mudah dalam penskoran, dan memiliki objektivitas tinggi.

Namun, untuk mengukur kemampuan pemecahan masalah, seperti kemampuan di bidang fisika, yang umumnya memiliki sejumlah tahapan penyelesaian, butir pilihan ganda menghasilkan respons yang kurang akurat. Pilihan ganda hanya menangkap respons jawaban akhir, sementara tahapan-tahapan menemukan jawaban akhir tersebut tidak terekam secara lengkap. Karena itu diperlukan format butir yang tidak hanya berpeluang digunakan dalam pengujian skala besar, tetapi juga mampu merekam tahap-tahap pemecahan masalah secara rinci. Untuk keperluan di atas, butir *multiple true-false* memenuhi syarat sebagai format butir alternatif.

Di Indonesia, butir *multiple true-false* dikenal dengan nama butir "asosiasi pilihan ganda", dan telah lama digunakan dalam Ujian Masuk Perguruan Tinggi Negeri (UMPTN) atau Seleksi Penerimaan Mahasiswa Baru (SPMB). Namun amat disayangkan, butir yang telah merekam respons peserta tes relatif lebih lengkap tersebut, di lapangan tetap diskor secara dikotomus, yaitu mendapat skor 1 bila benar secara total dan mendapat skor 0 bila salah, seberapapun tingkat kesalahannya. Dengan model penskoran dikotomus tersebut, butir *multiple true-false* atau asosiasi pilihan ganda belum menghasilkan estimasi kemampuan yang maksimal. Kumaidi (1987, 1988) melakukan *classical analysis* terhadap skor dikotomus butir asosiasi pilihan ganda. Hasilnya menunjukkan bahwa penskoran dikotomus pada respons butir asosiasi pilihan ganda tidak mampu membedakan peserta yang berkemampuan tinggi dan rendah.

Di bawah ini disajikan salah satu contoh butir asosiasi pilihan ganda bidang fisika pada UMPTN Tahun 1996, beserta aturan pemberian responsnya.

Rumusan butir:
Seberkas sinar datang dari suatu medium menuju udara. Jika sudut datang lebih besar dari 60^o sinar akan terpantul sempurna. Pernyataan di bawah ini yang benar:
(1). Indeks bias medium $2/\sqrt{3}$
(2). Indeks bias medium lebih besar daripada indeks bias udara
(3). Sudut kritis = 60^o
(4). Sudut kritis tidak bergantung pada indeks bias medium

Aturan pemberian respons:
Pilihlah: A jika hanya (1), (2), dan (3) benar
B jika hanya (1) dan (3) benar
C jika hanya (2) dan (4) benar
D jika hanya (4) yang benar
E jika semuanya benar

Aturan pemberian respons di atas bila dicermati memiliki dua kelemahan mendasar. Kelemahan pertama: *option* (3) tidak berfungsi. Karena, bila *option* (1) salah dan *option* (2) benar, jawabannya pasti C; bila *option* (1) benar dan *option* (2) salah, jawabannya pasti B; bila *option* (1) dan (2) salah, jawabannya pasti D. Bila *option* (1) dan (2) benar, baru menganalisis *option* (4); bila *option* (4) salah, jawabannya A; bila *option* (4) benar, jawabannya E. Jadi, untuk menemukan jawaban yang benar, *option* (3) tidak perlu dianalisis sama sekali. Hal yang demikian tentu tidak dibenarkan dalam kegiatan pengukuran dan pengembangan butir.

Kelemahan kedua: penskoran dikotomus belum memberikan penghargaan yang adil kepada setiap respons peserta. Misalnya, peserta I menjawab A (berarti meyakini *option* (1), (2), dan (3) benar); peserta II menjawab B (meyakini *option* (1) dan (3) benar); peserta III menjawab C (meyakini *option* (2) dan (4) benar); peserta IV menjawab D (meyakini hanya

option (4) yang benar; dan peserta V menjawab E (meyakini semua *option* benar). Berdasar analisis konten, *option* (1), (2), dan (3) benar, sedangkan *option* (4) salah. Sesuai aturan pemberian respons di atas, peserta I diberi skor 1 sedangkan empat peserta yang lain sama-sama mendapat skor 0. Selanjutnya timbul permasalahan II, III, IV, dan V menunjukkan respons berbeda, tetapi diperlakukan sama. Bahkan, peserta IV tidak memilih satu pun *option* yang benar, justru memilih *option* yang salah, tetapi diberi skor sama dengan peserta II, III, dan V yang mampu memilih beberapa *option* yang benar. Permasalahan berikutnya, secara substantif *option* (1), (2), (3), dan (4) menuntut kemampuan penyelesaian yang tidak sama, karena bobot substansinya berbeda, namun dalam analisis respons bobot tersebut tidak diperhitungkan. Bahkan, bila terdapat *option* dengan konten salah, yang memang didesain untuk tidak dipilih, ketika ada peserta yang tetap memilih *option* tersebut tidak menerima konsekuensi apapun, misalnya sanksi *(penalty)* berupa pengurangan skor. Beberapa kelemahan dan permasalahan tersebut memicu dikembangkannya model penskoran yang baru untuk respons butir pilihan.

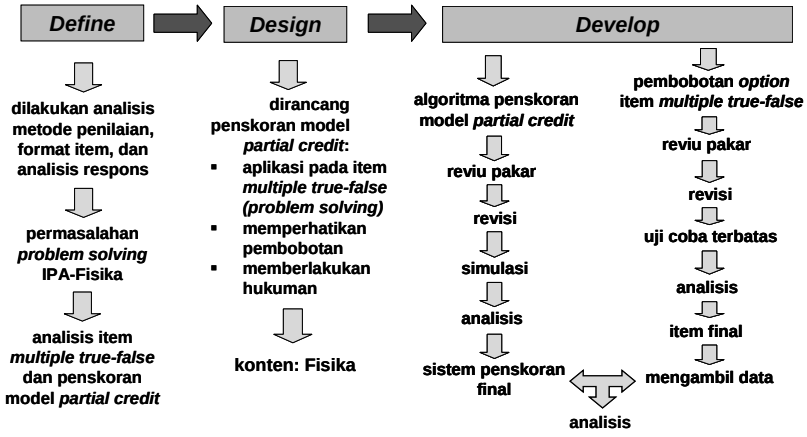
Baker dkk. (2000) dan Tognolini & Davidson (2003) melaporkan bahwa analisis respons secara politomus dapat meningkatkan akurasi pengukuran, karena itu kecenderungan pengembangan penskoran sebaiknya diarahkan pada sistem penskoran politomus, dengan menggunakan banyak kategori. Di antara sejumlah model penskoran politomus, model penskoran *partial credit* memiliki karakteristik penskoran yang sesuai dengan permasalahan bidang fisika. *Threshold* kategori yang lebih tinggi dalam penskoran *partial credit* tidak selalu lebih besar dari *threshold* kategori sebelumnya. Demikian pula dalam permasalahan fisika, tingkat kesukaran untuk mencapai kategori yang lebih tinggi tidak selalu lebih besar dibandingkan tingkat kesukaran untuk mencapai kategori sebelumnya. Sering terjadi, tahapan awal dalam menyelesaikan permasalahan fisika sangat rumit sehingga memerlukan banyak pengetahuan dan keterampilan, tetapi bila sudah sampai di tahapan tersebut untuk mencapai tahapan berikutnya justru lebih sederhana.

Karena itulah model penskoran *partial credit* sesuai untuk digunakan dalam pengukuran kemampuan bidang fisika. Namun perlu diteliti,

bagaimanakah model pengategorian *partial credit* yang dapat diterapkan pada butir *multiple true-false*, khususnya untuk bidang fisika, sehingga menghasilkan taksiran kemampuan (*ability estimate*) yang akurat.

Metode Penelitian

Penelitian ini merupakan penelitian pengembangan dengan hasil berupa model penskoran *partial credit* untuk butir *multiple true-false* pada bidang fisika. Pengembangan mengikuti prosedur yang diajukan Thiagarajan, Semmel & Semmel (1974) yang dikenal dengan *Four-D model*, meliputi *define, design, develop,* dan *desseminate*. Tetapi, penelitian ini belum sampai pada kegiatan diseminasi secara meluas, sehingga hanya meliputi tiga tahapan, yaitu *define* (pendefinisian)*, design* (perancangan)*,* dan *develop* (pengembangan). Langkah-langkah pengembangan secara umum ditunjukkan diagram di bawah ini.



Gambar 1. Tahapan Pengembangan Model Penskoran

Untuk memperoleh hasil pengembangan yang mantap, penelitian ini dibagi menjadi dua tahapan, yaitu tahap penelitian empiris dan tahap penelitian simulasi. Penelitian empiris dilakukan dengan cara memberikan

instrumen tes kepada mahasiswa baru tahun 2007 Fakultas MIPA Universitas Negeri Surabaya, pada tanggal 4-7 September 2007. Sedangkan penelitian simulasi menggunakan program SAS berdasarkan parameter empiris maupun parameter komputasi.

Respons peserta tes hasil penelitian empiris diskor secara dikotomus dan politomus model *partial credit* dengan memperhatikan pembobotan tiap *option* (PCM I), tanpa memperhatikan pembobotan (PCM II), berdasarkan tingkat kesukaran tiap *option* (PCM III).

Hasil penskoran kemudian dianalisis dengan program Quest (Adam & Khoo, 1996) untuk memperoleh estimasi tingkat kesukaran butir (b) dan estimasi kemampuan peserta (θ). Hasil estimasi kesukaran butir (b) dan kemampuan peserta (θ) digunakan untuk menentukan fungsi informasi butir politomus dan dikotomus dengan persamaan (secara berturutan):

$$I_i(\theta) = \sum_{xi=0}^{m_i} \frac{D^2 e^{D(\theta - b_{xi})}}{\left[1 + e^{D(\theta - b_{xi})}\right]^3} \quad \text{dan} \quad I_i(\theta) = \frac{D^2 e^{D(\theta - b)}}{\left[1 + e^{D(\theta - b)}\right]^2}$$

Berdasar fungsi informasi butir di atas, dbutirukan fungsi informasi tes $I(\theta)$ dengan cara menjumlahkannya untuk seluruh butir, dan akhirnya diperoleh kesalahan baku estimasi kemampuan menggunakan persamaan: $\text{SE}(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$. Semakin tinggi fungsi informasi dan semakin kecil kesalahan baku estimasi berarti semakin akurat model penskoran tersebut dalam mengestimasi kemampuan peserta.

Respons hasil penelitian simulasi juga dianalisis menggunakan program Quest untuk memperoleh estimasi kemampuan ($\hat{\theta}_j$). Estimasi kemampuan tersebut kemudian dibandingkan dengan kemampuan sejati (θ_j) yang dibangkitkan secara komputasi menggunakan metode akar dari rerata kesalahan kuadrat (*root mean squared error*) disingkat RMSE. Semakin kecil nilai RMSE berarti estimasi kemampuan yang dilakukan semakin akurat.

Hasil Penelitian dan Pembahasan

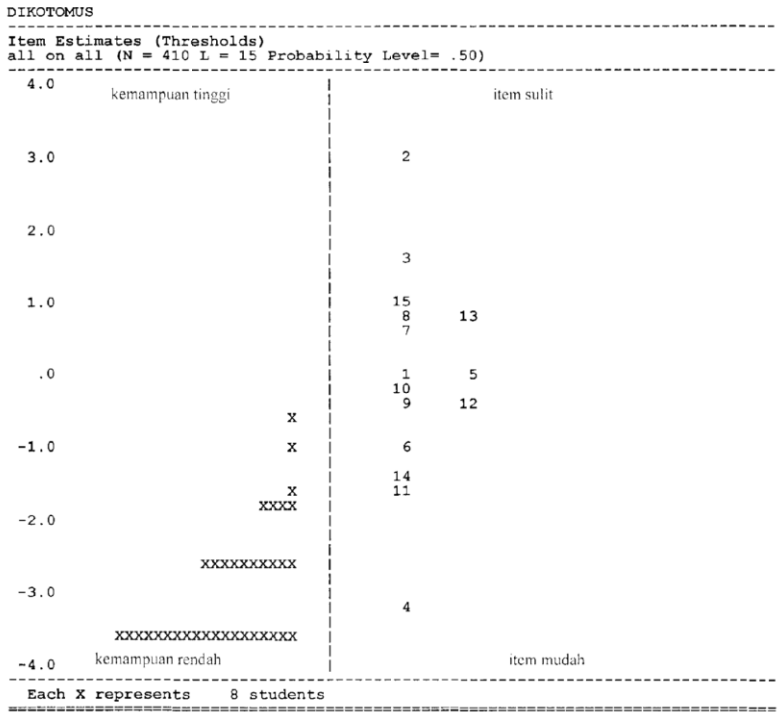
Sebelum dilakukan analisis estimasi kemampuan responden dan estimasi tingkat kesukaran butir, terlebih dahulu dilakukan analisis fit butir menggunakan parameter *infit* dan *outfit* untuk *mean square* (kuadrat rata-rata) dan *t*. Rangkuman nilai *infit mean square*, *outfit mean square*, *infit t*, dan *outfit t* hasil Quest untuk penskoran dikotomus, PCM I, PCM II, dan PCM III dirangkum dalam tabel di bawah ini.

Tabel 1. Parameter Fit Butir

Sistem Penskoran	Infit *Mean square*	Outfit *Mean square*	Infit *t*	Outfit *t*
Dikotomus	0,99	0,87	0,19	-0,12
PCM I	1,00	0,99	-0,01	-0,13
PCM II	1,00	0,98	-0,01	-0,14
PCM III	1,00	0,97	0,02	0,12

Data atau respons dikatakan fit dengan model Rasch, bila harga *infit* dan *outfit mean square*-nya mendekati 1 dan harga *infit* dan *outfit t*-nya mendekati 0 (Adams & Khoo, 1996: 30). Tabel 1 menunjukkan bahwa butir yang digunakan dalam penelitian ini fit dengan model Rasch, baik ketika responsnya diskor secara dikotomus maupun politomus model *partial credit*, tetapi tingkat kecocokan (fit) penskoran dikotomus paling rendah dibandingkan model *partial credit*.

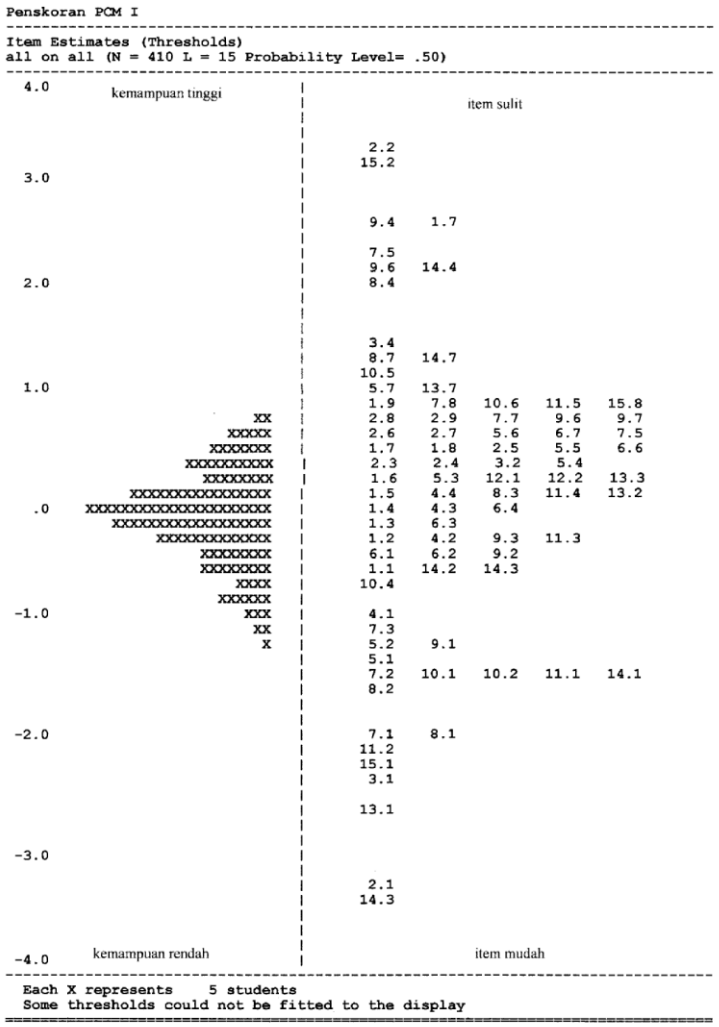
Estimasi tingkat kesukaran butir dan kemampuan peserta hasil analisis Quest ditunjukkan pada Gambar 2 (dikotomus), Gambar 3 (PCM I), Gambar 4 (PCM II) dan Gambar 5 (PCM III).



Gambar 2. Peta Tingkat Kesukaran Butir (Sebelah Kanan) dan Kemampuan Peserta (Sebelah Kiri) Berdasarkan Sistem Penskoran Dikotomus

Gambar 2 menunjukkan bahwa butir tes yang digunakan dalam penelitian ini memiliki estimasi tingkat kesukaran berdistribusi simetris normal, merentang dari tingkat kesukaran rendah hingga tingkat kesukaran tinggi dan dominan di tengah-tengah, dengan rata-rata 0,00 dan simpangan baku 1,42. Butir nomor 2 merupakan butir yang paling sukar, sedangkan butir nomor 4 merupakan butir yang paling mudah. Distribusi kemampuan peserta memiliki rata-rata -2,76, simpangan baku 0,81, dan median -2,43, sehingga memiliki koefisien kemiringan -1,23. Hal ini menunjukkan bahwa soal sangat sulit bagi peserta tes. Skor peserta hanya berada pada rentangan

1 (setara dengan skala -3,41 logits) hingga 8 (0,18 logits), dari skor maksimum 15.

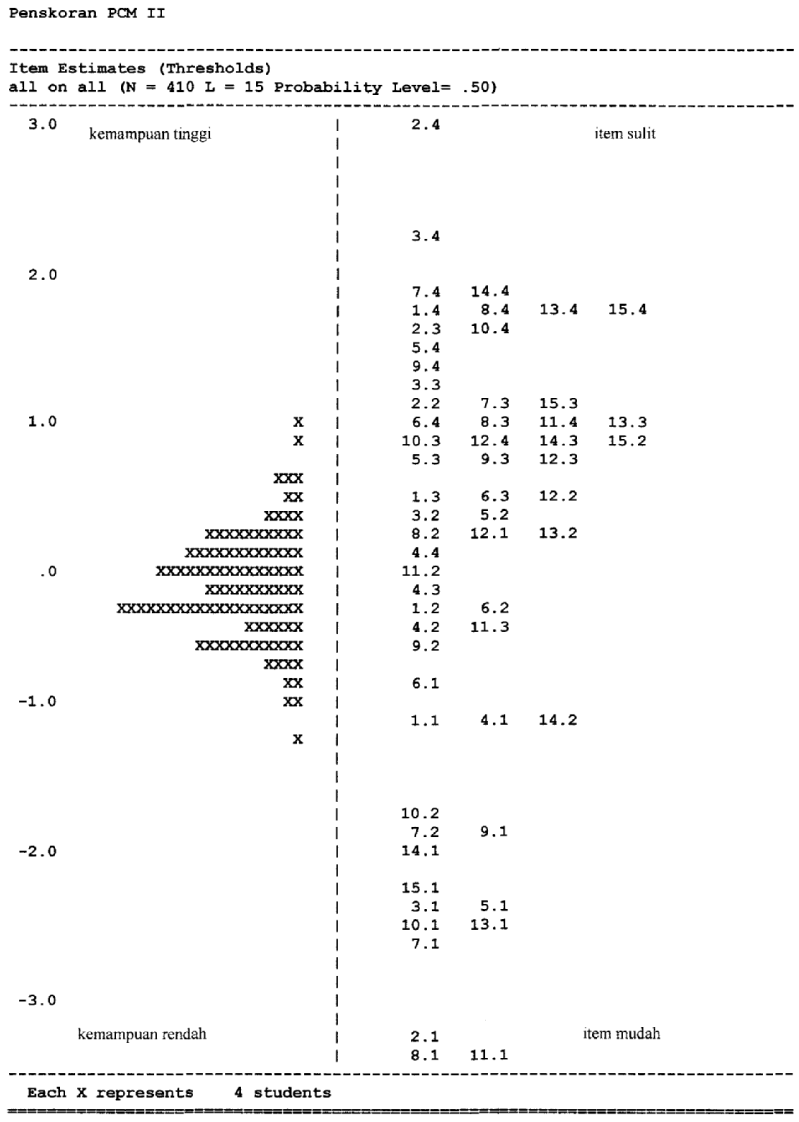


Gambar 3. Peta Tingkat Kesukaran Butir (Sebelah Kanan) dan Kemampuan Peserta (Sebelah Kiri) Berdasarkan Penskoran PCM I
Angka 14.3 menyatakan *threshold* 3 pada butir nomor 14

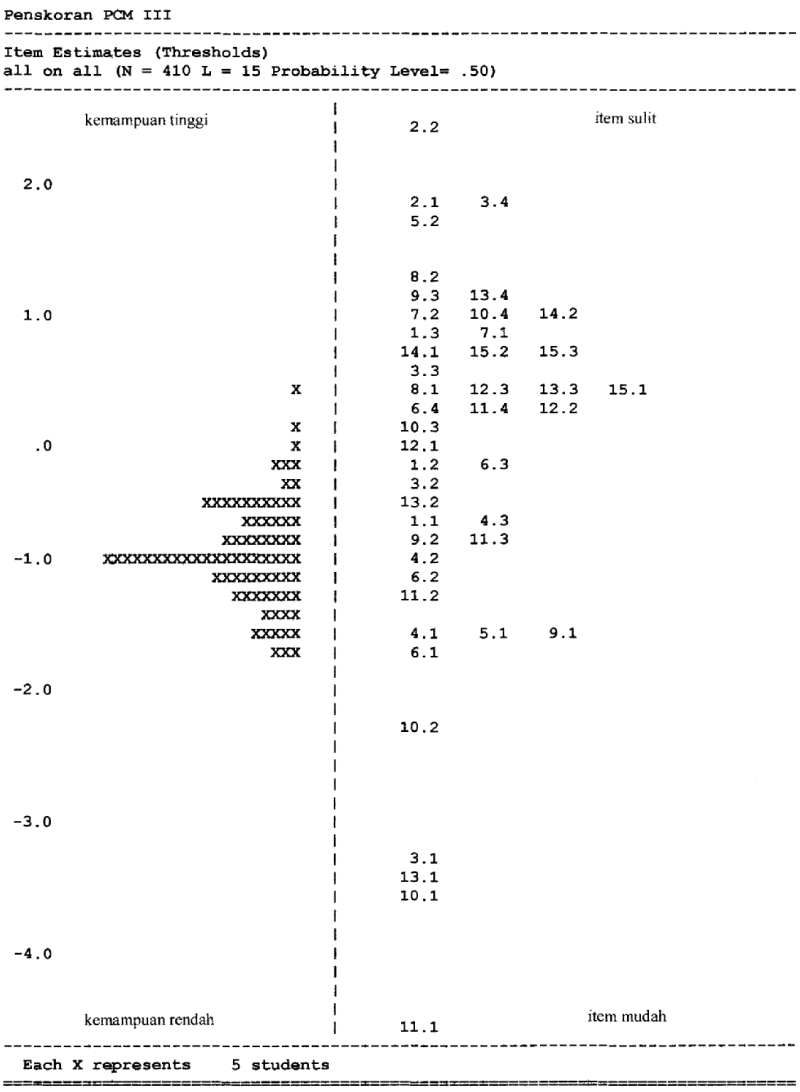
Gambar 3 menunjukkan, untuk mencapai kategori 3 dari kategori 2 pada butir nomor 14 (ditulis 14.3) merupakan *threshold* paling rendah. Sedangkan, untuk mencapai kategori 2 setelah mencapai kategori 1 pada butir nomor 2 (ditulis 2.2) merupakan *threshold* paling tinggi. Gambar 9 juga menunjukkan *threshold* dari seluruh butir merentang dari mudah ke sulit. Rata-ratanya 0,00 dengan simpangan baku 0,64. Sebaran estimasi kemampuan peserta tes hasil penskoran PCM I relatif lebih baik dibandingkan hasil penskoran dikotomus, karena rata-ratanya mendekati 0 dan sebarannya merentang dari nilai negatif ke positif. Peserta mencapai skor dalam rentangan 19 (-0,86 logits) hingga 96 (1,54 logits), dari skor tertinggi 102. Rata-rata estimasi kemampuan peserta adalah -0,08.

Penskoran PCM II menghasilkan jumlah kategori penskoran yang sama untuk semua butir, yaitu 5, meliputi kategori 0, 1, 2, 3, dan 4. Karena itu, semua butir memiliki 4 *threshold* dengan sebaran sebagaimana terlihat pada Gambar 4. Rata-ratanya 0,10 dengan simpangan baku 0,48. Sebaran estimasi kemampuan peserta tes memiliki rata-rata -0,14 dengan simpangan baku 0,38. Peserta mencapai skor dalam rentangan 14 (-1,22 logits) hingga 47 (1,43 logits).

Gambar 5 menunjukkan estimasi *threshold* menyebar dari rendah ke tinggi. Untuk mencapai kategori 2 dari kategori 1 pada butir nomor 2 merupakan *threshold* paling tinggi, sedangkan mencapai kategori 1 pada butir nomor 11 merupakan *threshold* paling rendah. Rata-rata *threshold* pada penskoran PCM III adalah -0,01 dengan simpangan baku 0,97. Sebaran estimasi kemampuan peserta tes memiliki rata-rata -0,88 dengan simpangan baku 0,45. Peserta mencapai skor dalam rentangan 4 (-2,90 logits) hingga 31 (0,69 logits).



Gambar 4. Peta Tingkat Kesukaran Butir (Sebelah Kanan) dan Kemampuan Peserta (Sebelah Kiri) Berdasarkan Sistem Penskoran PCM II



Gambar 5. Peta Tingkat Kesukaran Butir (Sebelah Kanan) dan Kemampuan Peserta (Sebelah Kiri) Berdasarkan Penskoran PCM III

Korelasi keempat estimasi kemampuan hasil Quest berdasarkan penskoran dikotomus, PCM I, PCM II, dan PCM III terangkum dalam Tabel 2 di bawah ini. Berdasar hasil analisis korelasi di bawah ini, penskoran PCM I, II, dan III menghasilkan estimasi kemampuan yang relatif sama untuk 410 responden. Hal ini ditunjukkan oleh besarnya nilai koefisien korelasi antar ketiga estimasi kemampuan yang dihasilkannya, yakni sekitar 0,70 hingga 0,89. Penskoran dikotomus menghasilkan estimasi kemampuan yang berbeda dan cenderung berlawanan dibandingkan penskoran model *partial credit*. Hal ini terlihat dari koefisien korelasinya yang berkisar antara -0,15 hingga -0,21.

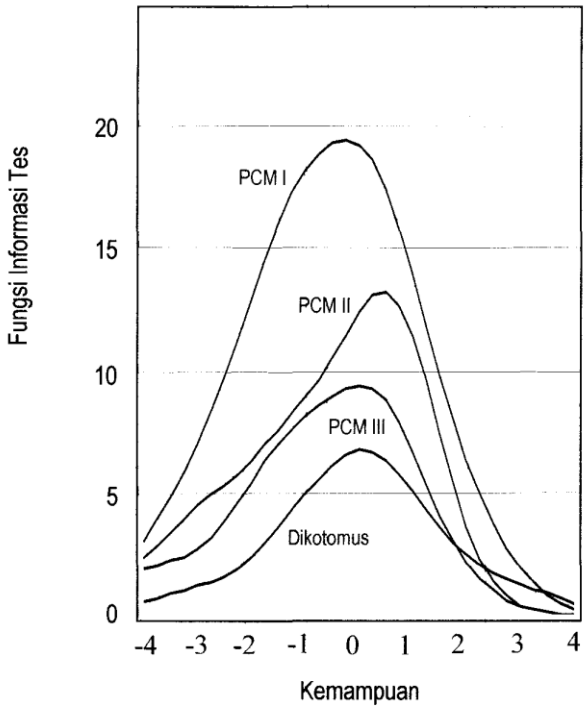
Tabel 2. Korelasi Estimasi Kemampuan Hasil Penskoran Dikotomus, PCM I, II, dan III

		Dikotomus	**PCM I**	**PCM II**	**PCM III**
	Korelasi Parsial	1	-0,149	-0,193	-0,209
	Sig. (2-tailed)	0,000	0,003	0,000	0,000
	N	410	410	410	410
PCM I	Korelasi Parsial	-0,149	1	0,890	0,699
	Sig. (2-tailed)	0,003	0,000	0,000	0,000
	N	410	410	410	410
PCM II	Korelasi Parsial	-0,193	0,890	1	0,847
	Sig. (2-tailed)	0,000	0,000	0,000	0,000
	N	410	410	410	410
PCM III	Korelasi Parsial	-0,209	0,699	0,847	1
	Sig. (2-tailed)	0,000	0,000	0,000	0,000
	N	410	410	410	410

Hal yang menarik, responden yang memiliki estimasi kemampuan tertinggi menurut keempat model penskoran di atas adalah orang yang sama, yaitu responden nomor 193. Responden yang memiliki estimasi kemampuan terendah menurut empat model penskoran tersebut, tidak dapat dinyatakan secara pasti sebagai orang yang sama. Estimasi kemampuan terendah pada penskoran PCM I sebesar -0,86 (dimiliki 1 orang, responden nomor 102); pada penskoran PCM II sebesar -1,22 (dimiliki 2 orang, responden nomor 59 dan 102); pada penskoran PCM III sebesar -2,90 (dimiliki 4 orang, responden nomor 115, 136, 345, dan 367); dan pada penskoran dikotomus sebesar -3,41 (dimiliki 151 orang, termasuk semua responden yang memiliki estimasi kemampuan terendah pada penskoran PCM I, PCM II, dan PCM III). Hasil analisis tersebut menunjukkan bahwa: (i) respons yang berasal dari responden dengan estimasi kemampuan tinggi memiliki tingkat kebenaran respons tinggi, sehingga diskor dengan model penskoran apapun memberikan posisi estimasi yang relatif tetap; (ii) respons yang berasal dari responden dengan estimasi kemampuan rendah memiliki tingkat kebenaran respons juga rendah, sehingga ketika diskor dengan model penskoran berbeda memberikan posisi estimasi yang tidak stabil.

Penskoran PCM I memiliki jumlah kategori 5-10, menghasilkan 1 estimasi kemampuan terendah. Penskoran PCM II memiliki jumlah kategori 5, menghasilkan 2 estimasi kemampuan terendah. Penskoran PCM III memiliki jumlah kategori 3-5, menghasilkan 4 estimasi kemampuan terendah. Penskoran dikotomus memiliki jumlah kategori 2, menghasilkan 151 estimasi kemampuan terendah. Ternyata, penskoran PCM I menghasilkan estimasi kemampuan terendah lebih pasti dibandingkan penskoran PCM II, PCM III, dan dikotomus. Berdasarkan hasil analisis di atas, semakin banyak jumlah kategori yang digunakan dalam suatu penskoran, semakin akurat penskoran tersebut mengestimasi kemampuan responden. Hasil ini menguatkan hasil penelitian Baker dkk. (2000), Tognolini & Davidson (2003), dan Wu (2003) yang menyatakan bahwa penskoran dengan banyak kategori dapat mengukur kemampuan peserta tes lebih baik dibandingkan penskoran dikotomus, yang hanya memiliki 2 kategori.

Untuk mendapatkan bukti yang lebih mantap, di bawah ini disajikan hasil analisis fungsi informasi dan kesalahan baku estimasi pada penskoran dikotomus, PCM I, PCM II, dan PCM III. Gambar 6 menunjukkan fungsi informasi tes yang dihasilkan oleh penskoran dikotomus, PCM I, PCM II, dan PCM III.



Gambar 6. Fungsi Informasi Tes Pada Penskoran Dikotomus, PCM I, PCM II, Dan PCM III

Gambar 6 menunjukkan bahwa fungsi informasi tes maksimum paling rendah diperoleh ketika respons peserta diskor secara dikotomus (2 kategori). Fungsi informasi tes menjadi meningkat ketika diskor dengan PCM III (3-5 kategori), PCM II (5 kategori), dan berlipat sekitar tiga kali

ketika diskor secara PCM I (5-10 kategori). Analisis sejenis dilakukan oleh Donoghue (2005) pada penskoran ujian membaca, hasilnya menunjukkan penskoran politomus menghasilkan rata-rata fungsi informasi 2,1 hingga 3,1 kali rata-rata fungsi informasi penskoran dikotomus. Semakin besar fungsi informasi, semakin akurat estimasi yang dihasilkan (Hambleton dkk, 1991; Lin, 2008).

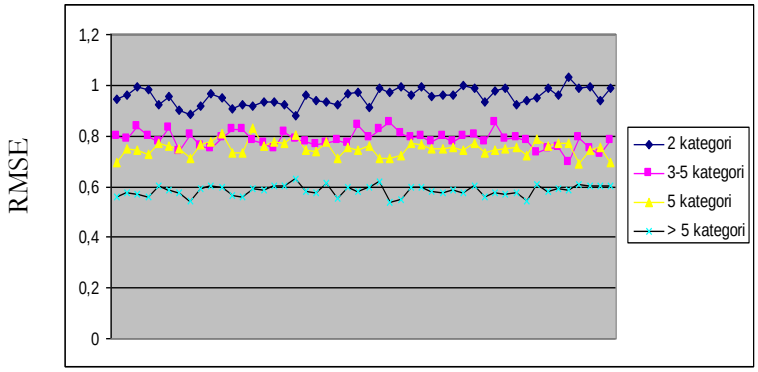
Bila fungsi informasi butir dan tes menunjukkan peningkatan ketika penskoran digeser dari dikotomus ke politomus, bagaimana dengan kesalahan baku estimasinya?

Tabel 3. Kesalahan Baku Estimasi Kemampuan θ = -0,89, -0,52, dan 0,18 pada Penskoran Dikotomus, PCM I, PCM II, dan PCM III

No	Estimasi Kemampuan	Kesalahan Baku Estimasi (diambil dua desimal)			
		Dikotomus	PCM III	PCM II	PCM I
1	-0,89	0,44	0,35	0,33	0,23
2	-0,52	0,40	0,34	0,31	0,23
3	0,18	0,38	0,32	0,28	0,23

Tabel 3 menunjukkan kesalahan baku estimasi hasil penskoran dikotomus paling besar dan menjadi berkurang ketika respons diskor dengan PCM III, PCM II, dan PCM I. Penskoran PCM I menghasilkan kesalahan baku estimasi yang paling kecil.

Untuk memantapkan hasil penelitian empiris, juga dilakukan penelitian simulasi. Berdasarkan parameter tingkat kesukaran butir atau *threshold* hasil penelitian empiris dibangkitkan 500 respons berdistribusi normal dan diperoleh RMSE untuk 48 iterasi sebagaimana ditunjukkan Gambar 7.



Gambar 7. RMSE 48 Iterasi Untuk Jumlah Kategori Yang Berbeda, Berdasar Parameter Butir Hasil Penelitian Empiris

Rata-rata RMSE dan simpangan baku yang dihasilkan ditunjukkan Tabel 4 di bawah ini.

Tabel 4. Rata-rata RMSE dan Simpangan Baku untuk Jumlah Kategori Berbeda Berdasarkan Parameter Empiris

Jumlah Kategori	Rata-rata RMSE	Simpangan Baku
2 (Dikotomus)	0,954	0,033
3-5 (PCM III)	0,786	0,032
5 (PCM II)	0,750	0,029
> 5 (PCM I)	0,584	0,021

Hasil uji ANOVA untuk keempat variasi jumlah kategori di atas menunjukkan pada taraf kepercayaan 95% keempat variasi rata-rata RMSE di atas adalah berbeda, dan dapat disimpulkan bahwa semakin banyak jumlah kategori yang digunakan dalam penskoran, rata-rata RMSE-nya semakin kecil.

Simpulan

Berdasarkan uraian hasil dan pembahasan di atas, dapat dirumuskan beberapa simpulan: (i) Model penskoran *partial credit* pada butir *multiple true-false* bidang fisika yang mampu menghasilkan estimasi kemampuan lebih akurat dibandingkan model penskoran lain adalah model penskoran *partial credit* dengan pengategorian berdasarkan pembobotan atau kompleksitas setiap *option*-nya; (ii) Semakin banyak jumlah kategori dalam penskoran *partial credit*, semakin akurat estimasi kemampuan yang dihasilkan. Hal ini ditunjukkan oleh hasil penelitian empiris maupun simulasi. Semakin banyak kategori yang digunakan, dihasilkan fungsi informasi tes yang lebih tinggi dan kesalahan baku estimasi yang lebih kecil. Semakin banyak kategori yang digunakan, semakin kecil nilai RMSE yang diperoleh. Fungsi informasi tes semakin tinggi, kesalahan baku estimasi semakin kecil, dan RMSE semakin kecil menunjukkan bahwa estimasi kemampuan yang dihasilkan semakin akurat. Menurut Bond & Fox (2007: 221) penambahan jumlah kategori dalam penskoran akan meningkatkan reliabilitas pengukuran, bila penambahan tersebut tidak dilakukan secara sembarangan. Kategori baru yang ditambahkan harus menunjukkan perbedaan tingkat kesukaran yang signifikan dengan kategori sebelumnya, sehingga tidak tumpang tindih atau saling meniadakan. Muraki (1998: 54) menyatakan jumlah kategori respons yang efektif untuk menaksir kemampuan adalah ≤15.

Daftar Pustaka

Adams, R. J., & Khoo, S. T. (1996). *Quest* (program komputer). *The interactive test analysis system*. Victoria: ACER.

Baker, J. G., Rounds, J. B., & Zeron, M. A. (2000). A comparison of graded response and rasch partial credit models with subjective well-being. *Journal of Educational and Behavioral Statistic,* 25(3), 253-270.

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

Dittendik, Ditjendikdasmen, Depdiknas. (2003). *Sistem penilaian kelas SD, SMP, SMA, dan SMK*. Jakarta: Dittendik, Ditjendikdasmen, Depdiknas.

Donoghue, J. R. (2005). An empirical examination of the IRT information of polythomously scored reading butirs under the generalized PCM. *Journal of Educational Measurement,* 31(4), 295-311.

Hambleton, R. K., & Jones, R. W. (tt). Comparison of classical test theory and butir response theory and their applications to test development. *BUTIRS* (*Instructional Topics in Educational Measurement)*. Diambil pada tanggal 27 November 2008 dari www.ncme.org/ pubs/butirs/24.pdf.

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

Kumaidi. (1987). *An exploratory study of the internal characteristics of the Indonesian public university entrance exam 'SIPENMARU': Implications for future test development*. PhD thesis, tidak diterbitkan. The University of Iowa, Iowa City, USA.

Kumaidi. (Desember 1988). *Studi analitik terhadap karakteristik internal ujian tulis seleksi masuk perguruan tinggi*. Makalah disajikan dalam Seminar Nasional Pengkajian Ujian Masuk Perguruan Tinggi Negeri, di Jakarta, 21-24 Desember 1988.

Lin, C. J. (2008). Comparisons between classical test theory and butir response theory in automated assembly of parallel test form. *The Journal of Technology, Learning, and Assessment*. 6(8), 1-42.

Muraki, E., & Bock, R. D. (1998). *Parscale. IRT butir analysis and test scoring for rating-scale data.* Chicago: Scientific Software International.

Oosterhof, A. (2003). *Developing and using classroom assessments* (*3th ed.*). Upper Saddle River: Merrill Prentice Hall.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice butirs: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, Summer, 3-13.

SAS Institute (1999). *SAS macro language: Reference version 8*. Cary, N. C.: SAS Institute, Inc.

Thiaragajan, S., Semmel, D. S., & Semmel, M. L. (1974). *Instructional Development for Training Teachers of Exceptional Children*. Minnesota: Indiana University.

Tognolini, J., & Davidson, M. (Juli 2003). *How do we operationalise what we value? Some technical chalenges in assessing higher order thinking skills.* Makalah disajikan dalam the Natinaonal Roundtable on Assessment Conference pada bulan Juli 2003 di Darwin, Australia.

Wu, B. C. (2003). Scoring multiple true-false butirs: A comparison of summed scores and response pattern scores at butir and test level. Research report. Lanham, Maryland: Educational Resources International Center (ERIC).