PENGEMBANGAN TES PENGETAHUAN PRAKTIKUM BIOLOGI BERDASARKAN *GRADED RESPONSE* DAN *GENERALIZED PARTIAL CREDIT*

Saiful Ridlo
Pend. Biologi, FMIPA, UNNES
sridlo@yahoo.co.id

Abstrak

Penelitian ini bertujuan untuk menghasilkan model tes yang cocok dengan data. Pengembangan item pada penelitian menggunakan pendekatan teori respons butir politomus (TRBP). Subjek ujicoba diambil dari siswa lima SMP kelas VII akhir mewakili peringkat SMP di Kota Yogyakarta sebanyak 1030 siswa. Hasil Model TRBP yang cocok dipilih berdasarkan hasil parametrisasi menggunakan PARSCALE dan deskripsi hubungan fungsional antara respons peserta tes dengan tingkat kemampuannya yang dinyatakan dalam *test information curves* (TIC). Penelitian ini menghasilkan 16 butir untuk bank soal dengan karakteristik masing-masing butir memiliki nilai daya beda yang tidak rendah (>0,25 skala logit) dan nilai kesulitan butir pada selang -3 sampai +3 skala logit. Berdasarkan informasi yang dihasilkan, kedua macam model penskoran GRM dan GPCM cocok memodelkan penskoran TPPB yang diadministrasikan. GPCM mungkin lebih merefleksikan realitas bagaimana data dihasilkan sehingga dari TIC tampak lebih akurat menaksir kemampuan dibanding GRM.

Kata Kunci: *tes pengetahuan praktikum biologi, GRM, GPCM*

DEVELOPMENT OF A TEST OF BIOLOGY PRACTICUM KNOWLEDGE WITH GRADED RESPONSE AND GENERALIZED PARTIAL CREDIT MODELS

Saiful Ridlo
Pend. Biologi, FMIPA, UNNES
sridlo@yahoo.co.id

Abstract

This study aims to generate information to define the polytomous item response models which are more suitable with the data. The items were developed by the polytomous item response theory approach. The tryout participants were 1030 Year VII students selected from five junior high schools in Yogyakarta City. A suitable model was selected based on the result of PARSCALE parameterization and a description of the functional relationship between the testees' responses and their ability levels indicated by the test information curves (TIC). The study yields 16 items for the item bank in which the discrimination index of each item is > 0.25 logit scale and the difficulty index ranges from -3 to +3 logit scale. The information shows that GRM and GPCM models of are suitable for scoring the administered TBPK. GPCM possibly reflects reality more regarding how the data are yielded so that on the basis of TIC it seems more accurate to estimate students' ability than GRM.

Keywords: *a test of biology practicum knowledge (TBPK), GRM, GPCM*

Pendahuluan

Penggunaan asesmen autentik terutama dengan asesmen kinerja sebagaimana dilaksanakan di kelas 9 SMP/MTs dan 12 SMA/MA membutuhkan kesiapan lebih sehingga ujian praktikum tidak dilaksanakan untuk kelas 7, 8, 10, dan 11. Bahkan, dengan beragamnya kondisi sekolah, beberapa sekolah tidak menyelenggarakannya. Oleh karena itu diperlukan alat asesmen yang dapat menjembatani kondisi tersebut. Model *paper and pencil task* yang dikembangkan Reynold, Doran, Allers, et al. (1996) merupakan salah satu model tes untuk mengukur pengetahuan praktikum. Asesmen keterampilan yang dilakukan melalui *paper and pencil test* berarti mengukur *knowledge of performance* (pengetahuan kinerja).

Sebuah butir soal dapat diskor dengan dua skor kategori (0 dan 1) seperti butir pilihan ganda, B-S, dan menjodohkan. Butir soal yang harus dijawab siswa bukan sekedar memilih tetapi harus dikonstruksi siswa biasanya diskor lebih dari dua skor kategori. Butir demikian disebut butir politomus. Tes dengan butir politomus sering memberi lebih banyak informasi tentang kecakapan peserta tes (Nandakumar, Feng Yu, Hsin-Hung Li, et al., 1998). Data politomus yang kemudian didikotomisasikan mungkin akan mengorbankan informasi yang dikandung oleh data asli, seperti akan menghasilkan model yang sama sekali tidak sesuai dengan model aslinya (Bastari, 1998). Berdasarkan hal tersebut, apabila dipandang suatu topik yang diajarkan membutuhkan balikan dari siswa bukan sekedar memilih jawaban maka seharusnya dikembangkan butir-butir politomus.

Pengembangan tes dapat dilakukan sesuai dengan teori tes klasik (TTK) atau teori respons butir (TRB). Penggunaan TRB memiliki keuntungan jika dibandingkan dengan TTK. Pertama, statistik butir yang independen terhadap sampel peserta tes yang diambil dalam suatu tes. Ke dua, statistik kemampuan peserta independen terhadap instrumen yang digunakan. Ke tiga, akurasi indek estimasi kemampuan individu memadahi (De Ayala, 1993:172). Berdasarkan hal-hal tersebut maka TRB dipandang memiliki keunggulan dibandingkan dengan TTK dan dapat memperbaiki keterbatasan yang ada dalam TTK. Keberhasilan penggunaan TRB pada

analisis dan interpretasi hasil tes hanya jika asumsi unidimensi dan independensi lokal terpenuhi.

Nandakumar, Feng Yu, Hsin-Hung Li, et al. (1998) mengelompokkan ragam data tes yang dihasilkan butir-butir politomus menjadi dua: (1) tes-tes dengan semua butir memiliki banyaknya kategori respons yang sama, dan (2) tes-tes dengan butir-butir memiliki banyaknya kategori respons yang beraneka. Beberapa peneliti menyebutkan penggunaan 7 skala akan memaksimalkan reliabilitas konsistensi internal, tetapi ada yang menyebutkan 4 skala dan yang lain 3 skala (Lei Chang, 1994:205). Linn dan Gronlund (Boughton, Klinger & Gierl, 2001:2) merekomendasikan penggunaan skala 3-7 kategori skor.

Tes pengetahuan praktikum pada disertasi ini memiliki banyaknya kategori respons yang beraneka untuk tiap butir. Artinya, dalam satu set tes terdapat butir yang diskor 3, 5, dan 7 kategori. Terkait dengan parameter diskriminan, asumsi model Rasch (dalam hal ini *partial credit model*, PCM) di mana antar butir memiliki nilai diskriminasi yang sama pada data empiris biasanya dilanggar (Ware, Bjorner & Kosinski, 2000). RSM (*rating scale model*) tidak dapat digunakan jika kategori skor butir beragam antar butir dan PCM adalah model dengan nilai faktor diskriminan tetap untuk semua butir, maka dua model yang dapat digunakan adalah GRM dan GPCM.

Pada GPCM, Tang (1996) menjelaskan probabilitas peserta ujian yang memilih kategori *k* dijelaskan dengan perbedaan pada probablitas untuk orang yang memiliki skor lebih dari atau sama dengan *k* dan memiliki skor lebih dari atau sama dengan *k+1*. Jika terdapat sebuah item yang diskor politomus memiliki *m* kategori skor, De Ayala (1993) menjelaskan bahwa kategori skor yang lebih tinggi (2 > 1 dan 0; dan 1 > 0) mencerminkan kemampuan yang lebih tinggi tetapi tidak harus tingkat kesulitan pada tahap ke-2 lebih tinggi dari pada tahap pertama dan sebaliknya. Tingkat kesulitan tidak harus urut. Berdasarkan GPCM, item tersebut memiliki sebuah parameter diskriminasi atau parameter *a*, sebuah parameter lokasi atau parameter *b*-global, dan satu set *m-1* parameter tingkat kesulitan/parameter *b*. Sebuah parameter lokasi dan *m-1* parameter tingkat kesulitan dapat dikombinasikan menjadi satu set *m-1* parameter *step*. Parameter diskriminasi item menjelaskan seberapa besar item dapat

membedakan antara individu dengan kemampuan berbeda. Parameter lokasi menjelaskan kesulitan item. Parameter *b* diinterpretasikan sebagai kesulitan relatif dari sebuah *step* dibandingkan dengan *step* lain dalam satu item. GPCM diformulasikan berdasarkan asumsi bahwa probabilitas untuk memilih kategori ke *k* lebih dari kategori *k-1* dipengaruhi oleh model respons dikotomus (Muraki, 1992:163). Dengan kata lain, probabilitas memperoleh skor *k* melebihi probabilitas untuk memperoleh skor *k-1.*

Pada GRM dari Samejima, masing-masing item mempunyai sebuah parameter diskriminasi dan satu set *m-1* parameter tingkat kesulitan. Parameter diskriminasi diinterpretasikan sama seperti pada GPCM. Masing-masing parameter tingkat kesulitan *m-1* membedakan probabilitas dari penskoran kurang dari kategori skor *k* dan lebih dari atau sama dengan kategori skor *k* (Tang, 1996). Childs & Wen-Hung Chen (1999), menjelaskan bahwa fungsi respons kategori $P_{jk}(\theta)$ adalah probabilitas peserta tes memberikan respons dalam kategori *k* pada item *j.* Probabilitas dihitung dengan mengurangkan probabilitas merespons pada suatu kategori *given* (cenderung dipilih) atau yang lebih tinggi dari probabilitas merespons pada kategori yang berbatasan atau lebih rendah.

De Ayala (1993) mencontohkan sebuah butir matematika: (6/3) + 2 = ? Agar dapat menyelesaikan butir tersebut seorang harus mampu menghitung dahulu 6/3 sebagai tahap pertama, baru pada tahap ke-2 menghitung hasil perhitungan tahap pertama + 2. Dalam hal ini terdapat dua anggapan, yaitu (a) seorang tidak dapat menyelesaikan tahap ke-2 dengan benar jika tahap pertama tidak diselesaikan dengan benar, dan (b) skor 0 bagi seorang yang menjawab persamaan tersebut salah. Sehingga ada 3 kemungkinan skor (x_i) yaitu 0, 1 atau 2. Penyekoran seperti dicontohkan pada butir matematika tersebut dapat didekati dengan *graded response* (GR). GRM merupakan ekstensi dari metode Thurstone yang muncul pada 1928. GRM tepat digunakan ketika respons peserta tes terhadap butir digolongkan sebagai respons kategori yang berurutan dan tingkat penyelesaiannya cenderung meningkat seperti yang ada pada skala Likert. Nilai tingkat kesulitan relatif katageori 1 > 2 > ...> n atau urut.

Dengan membangun sintaks tertentu, kalibrasi data respons butir politomus dengan PARSCALE (Muraki & Bock, 1997) dapat dihasilkan

statistik item yang *fit*, taksiran parameter θ, a, *b-global* dan *d* serta berbagai grafik. Parameter b_{jk} diperoleh dengan mengurangkan nilai parameter $b\text{-}global_j$ dengan nilai parameter d_{jk}. Grafik yang dapat dihasilkan antara lain *item characteristic curve* (ICC), *item information curve* (IIC), TIC dan histogram distribusi kemampuan/parameter θ masing-masing pada selang -3 sampai 3 skala logit. Fungsi informasi butir dapat digunakan untuk membandingkan beberapa perangkat tes. Dengan fungsi informasi butir diketahui butir yang mana yang cocok dengan model sehingga membantu dalam seleksi butir tes. Demikian juga dengan fungsi informasi tes. Semakin tinggi nilai fungsi informasi dan semakin kecil kesalahan bakunya berarti semakin akurat model penskoran tersebut dalam menaksir kemampuan.

Di Indonesia, sepanjang pengetahuan penulis, masih langka dijumpai penelitian pengembangan tes pengetahuan praktikum. Apalagi pengembangannya mendasarkan pada analisis hasil kalibrasi model TRBP (GRM dan GPCM). Penelitian untuk membandingkan kinerja dua model teori respon butir politomus, GRM dan GPCM, pada data tes dengan banyaknya kategori respons berbeda dalam satu perangkat dalam rangka pengembangan tes sangatlah penting. Penelitian penting untuk dilakukan mengingat adanya kecenderungan semakin meningkatnya penggunaan butir politomus dalam berbagai tes di dunia pendidikan, terutama pendidikan IPA.

Tujuan penelitian ini adalah menghasilkan model tes pengetahuan praktikum biologi menggunakan pendekatan teori respons butir politomus (GRM atau kah GPCM) yang cocok dengan data tes. Pembandingan kedua macam model TRBP tersebut dilakukan dengan menganalisis hasil parametrisasi respons subjek ujicoba dan TIC *output* PARSCALE.

Metode Penelitian

Pengembangan item menggunakan pendekatan teori respons butir politomus (TRBP) sesuai Stark, Chemyschenko, Chuah, et al. (2001). Persyaratan penggunaan TRBP sangatlah ketat. Mula-mula telah dikembangkan /dibongkar pasangkan 24 item setelah dilakukan konsultasi, *review* ahli dan diskusi. Selanjutnya, dilakukan ujicoba dengan 80 dan 177

subjek (ujicoba skala kecil) dan penskoran oleh dua orang *rater*. Analisis item dengan pendekatan teori klasik. Item yang dikehendaki didasarkan pada hasil perhitungan reliabilitas dengan □-Cronbach >0,6. Uji korelasi skor antara *rater* 1 dan 2 terbukti erat dan nilai *t* berpasangan nonsignifikans yang membuktikan persepsi dua orang *rater* terhadap rubrik sama. Nilai koefisien generalizability >0,75 agar desain setiap peserta tes dinilai oleh seorang *rater* dan *rater* tersebut menilai seluruh peserta tes pada *decision study* layak dilaksanakan. Item terbukti valid pada uji *goodness of fit* (GOF) pada *first order* dan *second order*. Karakteristik item yang dikehendaki dengan selang indeks kesulitan p = 0,3 – 0,8 dan indeks diskriminasi D minimal 0,2 sesuai kriteria Nina Deng & Hambleton (2008:5). Hal-hal tersebut dilakukan untuk memenuhi validitas, reliabilitas, dan menemukan karakteristik item yang dikehendaki.

Setelah mendapatkan perangkat soal sesuai kriteria tersebut di atas, selanjutnya diaplikasikan. Subjek ujicoba pada uji model item yang telah dikembangkan diambil dari siswa lima SMP kelas VII akhir mewakili peringkat SMP di Kota Yogyakarta sebanyak 1030 siswa. Sebelum dikalibrasi menggunakan TRBP, telah dilakukan uji asumsi unidimensi dengan *exploratory factor analysis* (EFA). Data hasil penelitian diperoleh dari *output* PARSCALE dan deskripsi hubungan fungsional antara respons peserta tes dengan tingkat kemampuannya yang dinyatakan dalam TIC.

Kinerja GRM dan GPCM dilihat dari hasil kalibrasi menggunakan PARSCALE. Hasil kalibrasi dilihat untuk mengetahui karakteristik item dan orang/kemampuan siswa. Hasil penaksiran parameter butir soal berupa daya beda butir (parameter *a* atau *slope*) dan tingkat kesukaran butir (parameter *b-global*) dapat dibaca dari *output* fase 2 atau file *.PAR. Pada file *.PAR juga dapat dibaca parameter *d* atau parameter kategori. Hubungan parameter *b-global* dan parameter *d* digunakan untuk menghitung parameter kesukaran relatif atau parameter *b*. Sedangkan hasil penaksiran parameter kemampuan dapat dibaca pada keluaran fase 3 atau file *.SCO. Item yang memenuhi kriteria, yaitu parameter *a* nilainya > 0,25 pada skala logit dan parameter *b-global* dan *b* nilainya pada selang -3 sampai 3 skala logit. Item yang terpilih dimasukkan dalam bank soal. Selain menggunakan kinerja hasil parametrisasi, pembandingan kesesuaian model pengembangan item

dilihat berdasarkan TIC. Model yang menghasilkan TIC pada selang -3 sampai 3 skala logit dengan akurasi yang lebih baik dianggap sebagai model yang lebih cocok dengan data. Model tes terbaik juga didasarkan pada kurva karakteristik total (*total characteristic curve/TCC*). Model terbaik dipilih jika memiliki TCC yang mengindikasikan diskriminasi antar peserta tes yang lebih baik.

Hasil Penelitian dan Pembahasan

Setelah melalui ujicoba skala kecil maka didapat hasil sebagai berikut. Satu item dari 24 item tidak valid dan secara umum pengembangan model konseptual didukung oleh data empirik (P-value=0,564 dan RMSEA=0,000). Reliabilitas tes menunjukkan harga Cronbach's Alpha = 0,760 yang berarti perangkat tes memiliki reliabilitas yang cukup tinggi. Uji korelasi berpasangan (*rater* 1 dan *rater* 2) pada semua item memiliki nilai korelasi yang signifikans (p<0,05). Artinya korelasi antara skor yang diberikan *rater* 1 dan 2 adalah sangat erat dan benar-benar berhubungan secara nyata. Hasil uji *t* menunjukkan ada 23 butir soal yang memiliki nilai *t* yang tidak signifikans (p>0,05) yang berarti pada 23 item tersebut *rater* memberikan persepsi yang tidak berbeda terhadap rubrik. Taksiran koefisien *generalizability* $\hat{\rho}^2_{I*}$=0,79 yang berarti setiap peserta tes dinilai oleh seorang *rater* dan *rater* tersebut menilai seluruh peserta tes pada *D-study* menaikkan nilai reliabilitas relatif secara klasik maka layak apabila tes dengan kondisi yang sama ujicoba skala kecil diskor oleh seorang *rater*. Didapatkan 17 item yang dikehendaki dengan selang p = 0,3 – 0,8 dan D minimal 0,2.

Hasil penelitian pada subjek uji target menunjukkan bahwa 17 item yang digunakan layak untuk dianalisis dengan TRBP karena memenuhi asumsi unidimensi sesuai pendapat Hattie (1985) dan unidimensionalitas esensial dari Stout (Kyong Hee Chon, Won-Chan Lee & Ansley, 2007). Besarnya variasi yang diterangkan oleh komponen pertama = 26,245% yang berarti tidak kurang dari 15%, perbandingan total nilai eigen 1 dan 2 = 1: 2,6129 dan perbandingan nilai selisih nilai eigen pertama dan ke-2 banding ke-3 dan ke-4 = (4,199 – 1,607) / (1,186 – 1,064) = 21,24.

Meskipun dari total nilai eigen 1 tidak ada 3 kali total nilai eigen 2 tetapi dari besarnya variasi yang diterangkan oleh komponen pertama lebih dari 15% dan perbandingan nilai selisih nilai eigen diketahui nilainya lebih dari 10. Berdasarkan *scree plo-t*nya terlihat ada satu dimensi dominan dan 3 dimensi minor.

Dari ke 17 item yang dikalibrasi dengan GRM dan GPCM terpilih 16 item dan sebuah item disisihkan karena memiliki nilai daya beda rendah pada selang 0 – 0,25 dan nilai kesulitan item berada di luar selang -3 sampai +3. Ke 16 item yang tersisa selanjutnya dikalibrasi kembali dengan kedua model tersebut di atas. Reparametrisasi 16 item memiliki karakteristik parameter daya beda pada selang 0,25 – 1,5 dan parameter kesulitan item berada pada selang -3 sampai +3. Selengkapnya pada Tabel 1.

Tabel 1.Nilai Parameter *a, b-global,* b_1, b_2, b_3, b_4, b_5 dan b_6
Hasil Penelitian[1)]

No	Kalib	Σ Kat	Parameter item		Parameter kesulitan kategori (b_{jk})[2)]					
			a_j	$b\text{-}glob_j$	b_1	b_2	b_3	b_4	b_5	b_6
2	GPCM	7	0,173	-1,168	1,924	-3,032	-2,053	-0,025	1,311	-5,554
3	GPCM	5	0,205	-1,550	2,239	-4,142	0,847	-5,465		
5	GPCM	3	0,453	-0,655	-0,349	-0,962				
6	GPCM	3	0,187	1,436	2,217	0,627				
9	GPCM	3	0,378	-0,335	1,354	-2,075				
10	GPCM	5	0,331	-0,900	-1,936	-1,937	0,090	0,078		
11	GPCM	7	0,155	-0,457	0,485	-2,935	-0,052	0,408	0,655	-1,495
12	GPCM	5	0,220	-0,499	2,490	-2,463	0,997	-2,882		
14	GPCM	7	0,320	-0,342	-0,900	-1,026	-0,876	-0,543	2,155	-1,105
16	GPCM	7	0,462	-1,212	-0.575	-2,652	-1,437	-2,770	4,034	-2,857
18	GPCM	3	1,365	-0,723	-1,794	0,391				
19	GPCM	5	0,218	-2,173	0,882	-3,346	-4,099	0,637		
20	GPCM	3	0,305	-0,979	-0,260	-1,565				
22	GPCM	7	0,166	-1,053	2,209	-2,094	-3,394	0,095	-2,186	1,113

No	Kalib	Σ Kat	Parameter item		Parameter kesulitan kategori (b_{jk})[2)]					
			a_j	$b\text{-}glob_j$	b_1	b_2	b_3	b_4	b_5	b_6
23	GPCM	3	0,567	-0,305	-1,091	0,591				
24	GPCM	5	0,270	-0,218	0,886	-1,542	1,420	-1,531		
2	GRM	7	0,671	-1,238	-2,482	-2,086	-1,500	-0,810	-0,255	0,123
3	GRM	5	0,648	-1,630	-2,363	-2,016	-1,146	-0,676		
5	GRM	3	0,735	-0,655	-1,170	-0,140				
6	GRM	3	0,293	1,421	0,216	2,656				
9	GRM	3	0,713	-0,360	-0,599	-0,071				
10	GRM	5	0,743	-0,926	-2,611	-1,551	-0,276	0,840		
11	GRM	7	0,567	-0,489	-2,496	-1,743	-0,764	0,005	0,722	1,532
12	GRM	5	0,615	-0,464	-1,384	-0,954	-0,114	0,455		
14	GRM	7	0,949	-0,382	-1,865	-1,249	-0,721	-0,114	0,747	1,151
16	GRM	7	0,964	-1,042	-2,780	-2,479	-1,905	-1,390	0,591	0,691
18	GRM	3	1,350	-0,701	-1,844	0,399				
19	GRM	5	0,421	-1,481	-4,051	-3,356	-2,134	0,845		
20	GRM	3	0,435	-0,912	-1,744	-0,214				
22	GRM	7	0,529	-0,709	-2,839	-2,351	-1,788	-0,825	-0,076	1,557
23	GRM	3	0,720	-0,250	-1,295	0,684				
24	GRM	5	0,653	-0,191	-1,342	-0,715	0,288	0,895		

[1)] diolah berdasarkan *output* PARSCALE dimana [2)] $b_{jk} = b\text{-}global_j - d_{jk}$

Butir terbaik sesuai hasil kalibrasi kedua model adalah butir 11 dengan kode TPPB 18 karena memiliki nilai parameter a paling baik. Butir tersebut memiliki banyak kategori 3. Butir TPPB 18 digunakan untuk mengukur konstruk mengklasifikasikan makhluk hidup berdasarkan ciri-ciri yang dimiliki dengan indikator siswa mampu menuliskan nama ilmiah dengan benar. Batang soal dan rubrik butir tersebut adalah sebagai berikut. Catatan: pada perangkat tes lengkap telah disediakan gambar berbagai spesies *Triangulum* dan kunci dikotomisnya.

Batang soal:

18. Apakah nama jenis/spesies pada gambar berikut 

.....

Butir 18 : menuliskan nama spesies ..... total 2 poin

2 poin jika siswa memberi nama dengan benar (*Triangulum oddcilius* atau Triangulum oddcilius atau *T. oddcilius* atau T. oddcilius)

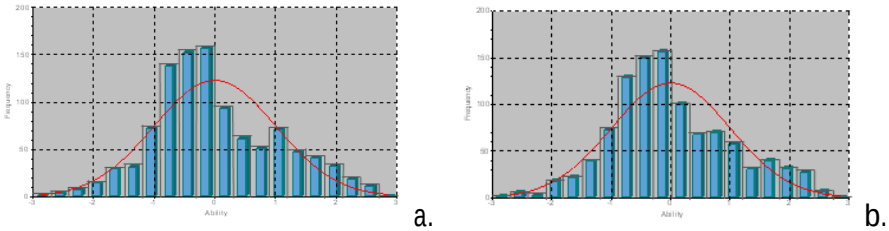
1 poin jika siswa memberi nama tetapi menuliskannya tidak seperti salah satu cara menuliskan nama ilmiah di atas

0 poin jika siswa memberi nama lain atau tidak menjawab

Ketentuan untuk teori respons butir dengan respons politomus adalah nilai a_j pada selang 0 – 0,25 dikatakan memiliki daya beda rendah, nilai a_j pada selang 0,25 – 1,5 merupakan selang yang khas (*typical range*), dan lebih dari 1,5 merupakan nilai a_j yang tinggi (Wells, Hambleton dan Purwono, 2008). Jika menggunakan kriteria ini maka hasil kalibrasi dengan GRM lebih sedikit butir ditemukan memiliki daya beda rendah (1 butir atau 6%) dibanding hasil kalibrasi dengan GPCM (8 butir atau 47%). Berdasarkan hal ini, maka sesuai pernyataan Dodeen (2004) mungkin yang menyebabkan lebih sedikitnya item yang *fit* pada kalibrasi dengan GPCM adalah akibat rendahnya parameter *a* atau daya beda item-item.

Sesuai kriteria parameter a_j dan b_j - global yang diinginkan, maka sampai parameterisasi 16 butir sudah tidak lagi ditemukan butir yang tidak dikehendaki. Parameterisasi dengan GRM menghasilkan nilai parameter *a* pada selang 0,293 – 1,350 dengan rerata 0,688 dan parameter *b*-global pada selang -2,174 – 1,436 dengan rerata -0,696. Parametrisasi dengan GPCM menghasilkan nilai parameter *a* pada selang 0,155 – 1,365 dengan rerata 0,361 dan parameter *b*-global pada selang -1,481 – 1,422 dengan rerata -0,626. Parameterisasi dengan GRM tidak dijumpai butir memiliki daya beda rendah (<0,25), tetapi parameterisasi dengan GPCM masih ditemukan 7 butir (44%) yang memiliki daya beda rendah. Berdasarkan

hasil proses pemilihan dan parameterisasi item maka dihasilkan item sebanyak 16 butir soal dalam bank item



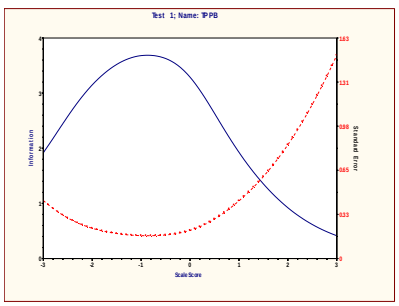
a. b.

Gambar 1. Histogram Distribusi Skor Kemampuan Peserta pada Kalibrasi 16 Butir dengan (a). GRM dan (b). GPCM

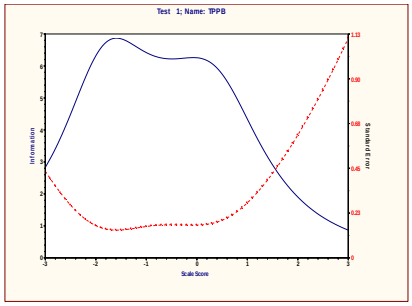
Hasil penaksiran kemampuan, antara hasil kalibrasi dengan GRM dan GPCM menunjukkan lebar selang yang tidak terlalu berbeda. Pada GRM antara 2,6305 sampai -2,5450, dan GPCM dari 2,6103 sampai -2,7909. Dari histogramnya ternyata lebih banyak siswa ditaksir berkemampuan tinggi pada GRM daripada GPCM atau dapat dikatakan GRM menaksir parameter kemampuan lebih tinggi dibandingkan GPCM. Hasil penelitian Thissen, Nelson & Rosa, *et al* (2001:155) juga menunjukkan hal yang sama. Histogram distribusi parameter θ estimate dapat dilihat pada Gambar 1. Besarnya nilai kemampuan dapat dibaca pada sumbu X dan banyaknya/ frekuensi siswa yang memiliki kemampuan tertentu dapat dibaca pada sumbu Y.

Kurva informasi total (TIC) untuk GRM menghasilkan tes cukup akurat untuk menaksir peserta dengan kemampuan -3 sampai 1,4 bahkan siswa yang memiliki kemampuan di bawah -3. Tes memberikan informasi paling tinggi pada nilai theta sekitar -0,7. Pada kalibrasi dengan GPCM perangkat tes secara total memberikan informasi cukup akurat untuk menaksir peserta dengan kemampuan -3 sampai 1,6 (SE $\approx 0{,}45$) atau lebih lebar 0,2257 skala. Tes paling akurat untuk menaksir peserta dengan kemampuan sekitar -1,7. Dari kedua kurva dapat dijelaskan bahwa tes yang diadministrasikan akurat untuk mengukur peserta tes dengan kemampuan lebih rendah sehingga dapat dikatakan sebenarnya tes yang diadministrasikan merupakan tes yang mudah. GPCM memberikan

informasi lebih baik terhadap theta yang dikehendaki. Pada tingkat kemampuan yang sama, informasi yang diberikan oleh item yang dikalibrasi dengan GPCM lebih tinggi dibandingkan dengan GRM. Maka jika harus memilih model, akan dipilih GPCM yang memberikan akurasi penaksiran yang lebih tinggi. Selengkapnya dapat dilihat pada Gambar 2. Garis tanpa putus menunjukkan besarnya nilai informasi dan garis putus-putus menunjukkan besarnya kesalahan baku.

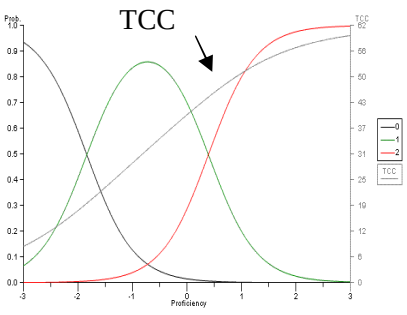


a.

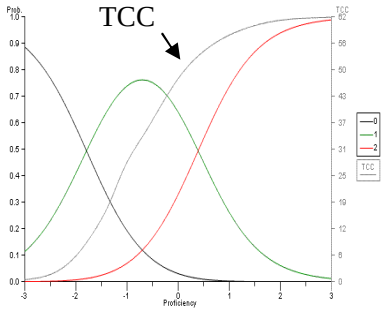


b.

Gambar 2. TIC untuk Tes Pengetahuan Praktikum Biologi Kalibrasi dengan (a). GRM dan (b). GPCM



a.



b.

Gambar 3. TCC untuk Item Total TPPB dan ICC untuk Item PPB 18 Hasil Parameterisasi (a). GRM dan (b). GPCM dengan PARSCALE

TCC kedua model yang digambar bersamaan keluaran ICC nomor 11 (TPPB 18) dapat dilihat pada Gambar 3. Fokus pada TCC-nya, kenaikan yang relatif tajam dapat dilihat pada kalibrasi dengan GPCM, yaitu antara theta sama dengan -2,0 – 0,5. Hal tersebut mengindikasikan adanya diskriminasi antar peserta tes yang lebih baik pada selang tersebut (Stark, Chemyschenko, Chuah, et al., 2001: 8). Berdasarkan hasil tersebut maka berarti model GPC lebih baik dari pada model GR. Jika tes dianalisis berdasarkan model GPC maka item-item dalam perangkat tes secara umum memiliki daya beda yang lebih baik sehingga lebih dapat membedakan kemampuan peserta tes.

Kesimpulan

Penelitian ini menghasilkan 16 butir untuk bank soal dengan karakteristik masing-masing butir memiliki nilai daya beda yang tidak rendah (>0,25 skala logit) dan nilai kesulitan butir pada selang -3 sampai 3 skala logit. Berdasarkan informasi yang dihasilkan, kedua macam model penskoran GRM dan GPCM cocok memodelkan penskoran Tes Pengetahuan Praktikum Biologi yang diadministrasikan. GPCM mungkin lebih merefleksikan realitas bagaimana data dihasilkan sehingga dari TIC dan TCC tampak lebih akurat menaksir kemampuan dan diindikasikan lebih dapat membedakan kemampuan peserta tes dibanding GRM.

Saran

Saran dan rekomendasi yang dapat diberikan adalah sebagai berikut. Pertama, guru dapat menggunakan model penskoran GPCM atau GRM ketika mengadministrasikan Tes Pengetahuan Praktikum Biologi sebagaimana kondisi tes pada penelitian ini. Ke dua, diperlukan penelitian dan analisis psikometri lebih lanjut menggunakan hasil-hasil pengadministrasian tes. Rekomendasinya, lakukan penelitian akurasi model berbagai macam kondisi tes melalui studi simulasi. Gunakan nilai parameter butir hasil pengadministrasian tes sebagai parameter pembangkit.

Daftar Pustaka

Bastari. (December 1998). *Comparison of IRT models that handle dichotomous and polytomous response data simultaneusly*. Makalah disajikan di University of Massachusetts.

Boughton, K.A., Klinger, D.A. & Gierl, M.J. (April 2001). *Effect of random rater error on parameter recovery of the generalized partial credit model and graded response model*. Paper presented at the annual meeting of the national council on measurement in education, Seattle, WA.

Childs, R.A., & Wen-Hung Chen. (1999). Software note: Obtaining comparable item parameter estimates in MULTILOG and PARSCALE for two polytomous IRT models [Versi elektronik]. *Applied Psychological Measurement,* 23, 4, 371-379.

De Ayala, R.J. (1993). An introduction to polytomous item response theory models. *Measurement and Evaluation in Conseling and Development,* 25, 172-189.

De Mars, C.E. (April, 2002). *Recovery of graded response and partial credit parameters in MULTILOG and PARSCALE*. 28p. Paper presented at the Annual Meeting of the American Education Research Association, Chicago.Dodd, B.G., De Ayala, R.J. & Koch, W.R. (1995). Computerized adaptive testing with polytomous items. [Versi elektronik]. *Apllied Psychological Measurement, 19, 5-23.*

Dodeen, H. (2004). The relationship between item parameters and item fit. *Journal of Educational Measurement.* Fall 2004, Vol.41, No.3, pp.261-270.

Hattie, J. (1985). Methodology Review: Assessing unidimensionality of tests and items. [Versi elektronik]. *Applied Psychological Measurement*, vol 9 (3): 139-164.

Kyong Hee Chon, Won-Chan Lee & Ansley, T.N. (November 2007). *Assessing IRT model-data fit for mixed format tests*, CASMA Report

Number 26, Center for Advanced Studies in Measurement and Assessment.

Lei Chang. (1994). A Psychometric evaluation of 4-poin and 6-point Likert-type scales in relatiopn to reliability and validity. [Versi elektronik]. *Apllied Psychological Measurement,* 18, 3, 205-215

Muraki, E., & Bock, R.D. (1997). *PARSCALE: IRT item analysis and test scoring for rating scale data.* Chicago: Scientific Software International.

Nandakumar, R., Feng Yu, Hsin-Hung Li, et al. (1998). Assessing unidimensionality of polytomous data [Versi elektronik]. *Applied Psycholohical Measurement,* 22, 2, 99-115.

Nina Deng & Hambleton, R.K. (February 4, 2008). *Psychometric analyses of the 2006 MCAS high school introductory physics test*. Center for Educatinal Assessment Research Report No. 647. Amherrst, MA: Center for Educational Assessment, University of Massachusetts

Ostini, R. & Nering, M.L. (2006). *Polytomous item response theory models, Series: Quantitative application in the social sciences; no. 07-144*. Thousand Oaks, CA: Sage.

Reynolds, D.S., Doran, R.L., Allers, R.H. et al. (1996). *Alternative assessment in science: A teacher's guide*. New York: New York State Education Departmen University of Buffalo.

Stark, S., Chemyschenko, S., Chuah, D., et al. (2001). Selecting a polytomous IRT model. *IRT Modelling Lab*. Diambil pada 12 Oktober 2006, dari University of Illinois IRT Laboratoty.htm http://work.psych.uiuc.edu/irt

Tang, K.L. (1996). Polytomous item response theory (IRT) models and their aplications in large-scale testing program: Review of literature. *Educational Testing Science.* Princeton, NJ. RM-96-8 TOEFL Monograph Series.

Thissen, D., Nelson, L., Rosa, K., et al. (2001). Item response theory for items scored in more than two catagories. Dalam D. Thissen & H.

Wainer (Eds.), *Test scoring* (pp.141-186). Mahwah, NJ: Lawrence Erlbaum Associates.

Ware Jr., J.E., Bjorner, J.B. & Kosinski, M. (2000). Practical implications of item response theory and compterized adaptive testing, A brief summary of ongoing studies of widely used headache impact scales. [Versi elektronik]. *Medical Care*, 38, 9, II.73-II.82 .

Wells, C.S., Hambleton, R.K. & Urip Purwono. (Juni 2008). *Item response theory: Polytomous respons IRT models and application.* Handout disampaikan dalam Pelatihan Asesmen Pendidikan dan Psikologi (Psikometri), di Universitas Negeri Yogyakarta