Aplikasi Teori Rasch untuk Penyekalaan Vertikal Tes Catur Wulan

Kumaidi

Abstract: This article discusses a part of findings of a study performed by Kumaidi and associates. The issue discussed in this article mainly focused on the aplication of Rasch model in analyzing data across classrooms and demonstrated the worth of the analysis. Data used in this article were responses of SMUN students in answering tests of English as an end of first quarter (Cawu I) test. The results of the study sugget that Rasch approach might be implemented to plot items from different classes in one single metric scale. This single metric would be useful to monitor student's progress in learning. It is suggested that the testing procedures be implemented which give rise a multilevel testing for end quarter tests in the future.

Kata-kata kunci: Teori Rasch, penyekalaan vertikal, tes catur wulan.

Sistem pengujian yang dilaksanakan dalam persekolahan kita sebenarnya sangat potensial untuk dikembangkan menjadi model pengujian yang dapat dipakai untuk memantau perkembangan pendidikan peserta didik, di samping kepentingan lainnya. Di dunia persekolahan Indonesia saat ini, kita mengenal pengujian rutin setiap empat bulan sekolah, baik yang dikelola secara sekolah, daerah, maupun nasional. Sistem pertama, yaitu pengujian empat bulanan, ada yang dilaksanakan sekolah dan ada yang secara bersama, atas kesepakatan musyawarah Kepala-kepala Sekolah, disebut ujian Catur

---

Kumaidi adalah Ketua Tim dan Peneliti Utama, Spesialis Pengujian IKIP Padang. *Artikel ini diangkat dari hasil penelitian HB VI/2 yang didanai oleh Proyek Peningkatan Penelitian dan Pengabdian pada Masyarakat, Ditbinlitabmas, Kontrak Kerja No. 29/P2IPT/DPPM/98/PHB/VI/2/V/1998 tanggal 20 Mei 1998.*

Wulan (Cawu); yang kedua, yang dilakukan secara nasional, disebut Evaluasi Belajar Tahap Akhir Nasional (Ebtanas).

Ujian Cawu dilakukan untuk menguji prestasi belajar semua peserta didik di Sekolah Menengah, mulai dari kelas I sampai dengan kelas III, sedangkan ujian Ebtanas (yang mungkin dapat dipandang sebagai Cawu terakhir) hanya diberikan kepada siswa kelas III. Berbagai pemakaian hasil pengujian ini dapat disebutkan, namun yang paling mencolok adalah untuk pengisian angka rapor (fungsi sumatif) dan penentuan kenaikan kelas atau kelulusan siswa (Ebtanas dan/atau dua Cawu sebelumnya untuk siswa kelas III). Di samping itu, hasil Ebtanas dapat juga dipakai untuk seleksi siswa baru bagi sekolah di atasnya, kecuali perguruan tinggi. Tambahan pula, ada cita-cita untuk memakai hasil Ebtanas untuk menentukan standar atau mutu pendidikan nasional.

Satu hal penting dalam pemakaian hasil pengujian yang saat ini sering dilupakan pengelola dan pelaksana pendidikan kita adalah fungsi diagnostik dan formatif. Padahal, setidaknya menurut penulis, dua fungsi ini merupakan unsur penting dalam implementasi tekad peningkatan mutu pendidikan. Di samping itu, fungsi utama hasil pengujian sebagaimana yang diamanatkan oleh Undang-undang Nomor 2 Tahun 1989 adalah fungsi diagnostik ini. Pasal 43 UU No. 2/1989 (Depdikbud, 1989) ini menyebutkan bahwa terhadap kegiatan dan kemajuan belajar peserta didik dilakukan penilaian. Sedangkan kalimat pertama dari penjelasan dari pasal 43 ini berbunyi: "Penilaian kegiatan belajar-mengajar dilakukan untuk **membantu perkembangan** peserta didik dalam usaha mencapai tujuan pendidikannya" (Depdikbud, 1989: 70, huruf tebal dari penulis). Dari kutipan itu dapat diketahui pesan yang ingin disampaikan adalah memanfaatkan hasil pengujian terutama untuk memberi bantuan belajar agar (setiap) peserta didik dapat mencapai tujuan belajar (masing-masing).

Untuk membantu memantau perkembangan pendidikan peserta didik, khususnya dari waktu ke waktu, dari Cawu ke Cawu, atau dari tahun ke tahun, alangkah idealnya apabila sistem pengujian Cawu dan Ebtanas dapat diletakkan dalam satu ukuran kemajuan belajar. Dalam upaya mengembangkan model pengujian ini, akan sangat baik apabila sejak awal pengembangan sistem pengujian Cawu dan Ebtanas disadari dan didasari, untuk **suatu saat** nanti, Cawu dan Ebtanas dapat diintegrasikan ke dalam satu ukuran tunggal tersebut, meskipun tanpa harus mengubah pelaksanaan pengujiannya, atau pelaksanaan pengujian tetap dilaksanakan sebagaimana yang

ada selama ini. Tentunya, strategi pengembangan tes Cawu (dan/atau Ebtanas) dikembangkan dengan memasukkan unsur diagnostik tersebut.

Untuk menjajaki kemungkinan pelaksanaan model tersebut, peneliti dengan dukungan dana dari Ditbinlitabmas, Ditjen Dikti, dan bekerjasama dengan Kanwil Depdikbud Sumbar, melakukan pengembangan model (manajemen dan pengolahan data) pengujian yang diarahkan untuk itu. Namun, karena baru pada tahap awal (dua tahun pertama), berbagai kendala masih dijumpai sehingga hasilnya belum sebagaimana diharapkan. Salah satu analisis yang dilakukan adalah analisis penyekalaan vertikal *(vertical equating)* untuk menempatkan tes dengan tingkat kesukaran yang berbeda pada skala atau ukuran yang sama. Tes dengan tingkat kesukaran yang berbeda ini dimaksudkan tes kelas I (sebagai mudah), kelas II (sedang), dan kelas III (sukar), karena tes kelas atas berkemungkinan sulit bagi siswa kelas rendah. Tulisan ini mencoba menjelaskan strategi penyekalaan vertikal dan mendiskusikan hasil awal pengembangan model pengujian tersebut, dengan contoh diambil dari tes Bahasa Inggris kelas II dan III SMU.

Berbagai model dan strategi penyekalaan *(equating)* telah dibicarakan para ahli, misalnya Angoff (1971), Wright dan Stone (1979), Petersen dkk. (1989), dan Kolen dan Brennan (1995). Pada dasarnya, penyekalaan *(equating)* dapat dilaksanakan dengan pendekatan atau teori *linear equating*, *equipercentile*, ataupun *Item Response Theory* (IRT). Salah satu pendekatan IRT adalah pendekatan Rasch (yang menurut Hambleton dan Swaminathan, 1985 juga digolongkan sebagai IRT dengan satu parameter—butir soal).

Pendekatan Rasch merupakan bagian dari IRT, sering juga disebut sebagai teori modern (Crocker dan Algina, 1986). Teori ini mulai berkembang pesat pada dasa warsa 1970an, termasuk di dalamnya teori pengukuran yang ditemukan oleh George Rasch, sehingga dikenal dengan teori pengukuran Rasch atau lebih dikenal sebagai *Rasch model* (Crocker dan Algina, 1986; Wright dan Stone, 1979). Model Rasch ini dalam banyak literatur juga dikenalkan sebagai model IRT *(Item Response Theory)* berparameter (butir soal) tunggal. Model ini mencoba menghubungkan prilaku atau karakteristik sebuah butir soal ketika diujikan kepada sejumlah peserta ujian dengan berbagai tingkat kemampuan. Perilaku butir soal ini akan dapat dilukiskan sebagai suatu kurva karakteristik butir soal (KKBS) sehingga seorang perserta ujian dengan tingkat kemampuan tertentu akan memiliki peluang menjawab benar sebuah butir soal (dengan tingkat kesulitan) tertentu pula. Rumusan matematis untuk teori (model) Rasch ini dapat diungkapkan sebagai berikut (Wright dan Stone, 1979:15):

$$P\{X_{vi} = 1 \mid \beta_v, \delta_i\} = \exp(\beta_v - \delta_i)/[1 + \exp(\beta_v - \delta_i)]$$

Catatan: $P\{X_{vi} = 1 \mid \beta_v - \delta_i\}$ adalah peluang menjawab benar dari seorang peserta ujian (v) terhadap sebuah butir soal (i); β_v adalah parameter (kemampuan) seorang peserta ujian (v); δ_i adalah parameter (tingkat kesukaran) sebuah butir soal (i).

Dalam kaitannya dengan penyekalaan, hasil estimasi tingkat kesukaran butir soal dan/atau tingkat kemampuan peserta ujian untuk tes yang berbeda, posisi parameter peserta ujian dan butir soal, kemudian diletakkan secara relatif terhadap posisi netral yang sama *(common zero)*. Oleh karena itu, dalam estimasi seperti itu selalu dibutuhkan (salah satunya) butir soal yang dikerjakan dua kelompok yang berbeda atau dua set butir soal yang dikerjakan kelompok peserta yang sama. Pendekatan pertama disebut *common items*, yang terakhir disebut *common test takers*. Dari dua rancangan penyekalaan ini, pendekatan dengan *common items* dapat dinyatakan sebagai lebih baik (Wright dan Stone, 1979). Oleh karena itu, dalam penelitian ini dipakai pendekatan pertama, yaitu dengan memanfaatkan *common* atau *linking items* yang berasal dari butir soal kelas bawah, karena tes yang dikaji diberikan kepada kelompok siswa dari kelas yang berbeda (yaitu kelas I dan kelas II, atau kelas II dan kelas III). Dalam istilah pengukuran, penyekalaan tersebut dikenal sebagai penyekalaan vertikal *(vertical equating)*. Dalam bahasa teori Rasch, pendekatan yang dijelaskan ini juga didiskusikan sebagai konsep kalibrasi butir soal. Dampaknya bagi pengembangan tes adalah pengembang tes berpeluang membuat berbagai perangkat tes dari variasi butir soal yang berbeda, yang disesuaikan dengan kemampuan calon peserta ujian (dari sekolah tertentu), namun hasilnya dapat diinterpretasikan dalam metrik yang sama, sehingga pengujian menjadi lebih efektif namun pengungkapan tingkat "kemampuan" peserta ujian lebih akurat.

Rancangan set butir dalam perangkat-perangkat tes yang dikaji dalam makalah ini dapat diilustrasikan sebagai Gambar 1. Dari Gambar 1 dapat dipahami bahwa hasil estimasi *linking items* (butir soal penjangkar) akan dipakai untuk proses penempatan butir soal lain pada salah satu skala butir soal yang dipilih, dalam kasus penelitian ini, butir soal diletakkan ke dalam skala tes kelas II, sehingga yang ditetapkan sebagai acuan adalah hasil estimasi butir soal penjangkar dalam tes kelas II. Kemudian, (rerata) perbedaan hasil estimasi butir soal penjangkar kelas III dengan kelas II, dipakai untuk proses penyesuaian skala tes kelas III.

Gambar 1 Ilustrasi Rancangan Penyekalaan Vertikal antara Tes Bahasa Inggris Kelas II dan Kelas III (dimodifikasi dari Wright dan Stone, 1979)

Proses penyekalaan ini, pada dasarnya, dapat dijelaskan sebagai berikut (dalam pelaksanaannya diproses dengan program komputer *Bigsteps*). Berdasarkan hasil kalibrasi butir soal menurut data pengujian perkelas, dianalisis lebih lanjut hasil kalibrasi butir soal penjangkar *(linking items)* perkalibrasi. Karena butir soal penjangkar dikalibrasikan dua kali, maka akan terdapat perbedaan hasil kalibrasi sesuai dengan kelompok siswanya. Diharapkan dari hasil kalibrasi butir soal penjangkar akan diperoleh kenyataan bahwa **butir soal yang sama ketika diujikan kepada siswa kelas II (bawah) akan lebih sulit dibandingkan apabila diujikan kepada siswa kelas III (atas).** Apabila harapan ini dapat dipenuhi, butir soal penjangkar tersebut dapat dipakai untuk proses penyekalaan; sebaliknya, kalau terjadi butir soal yang sama diujikan kepada siswa kelas bawah ternyata lebih mudah dibandingkan apabila butir soal tersebut diujikan kepada siswa kelas atas, maka butir soal itu tidak dapat dipakai sebagai butir soal penjangkar (dibuang).

Kemudian, dari beda tingkat kesukaran (dalam *logits*) butir soal penjangkar untuk kelas bawah dan atas tersebut, dicari rerata dan simpangan baku bedanya. Rerata beda akan dipakai sebagai faktor penyekala, yang dalam proses ini disebut **BETA32** (artinya faktor penyekala dari skala

kelas 3 ke skala kelas 2, dengan mengambil skala kelas II sebagai skala baku atau *reference*). Simpangan baku beda skala dipakai untuk memperkirakan besarnya kesalahan penyekalaan. Dari proses **pemindahan** tersebut akan diperoleh tingkat kesukaran butir soal (dalam skala *logits*) butir soal kelas III **dilihat** atau **dipersepsikan** apabila butir soal ini (yang diujikan di kelas III) diujikan kepada siswa kelas II. Hasil penyekalaan ini adalah butir-butir soal yang memiliki skala atau metrik yang sama. Karena dalam rumusan matematis Rasch itu, skala butir soal dan skala kemampuan siswa telah diletakkan pada satu skala tunggal, maka apabila dilakukan pemontenan dengan memperhatikan tingkat kesukaran hasil penyekalaan ini, perkembangan pendidikan (dalam hal ini prestasi belajar Bahasa Inggris kelas III dari kelas II dapat dipahami secara lebih baik). Apabila kemajuan yang tercapai belum sebagaimana diharapkan, maka guru bersama pengelola pendidikan dapat memberi intervensi pembelajaran seperlunya.

METODE

Data pengujian yang dipakai dalam proses penyekalaan ini berasal dari data ujian Catur Wulan pertama (Cawu I), tahun pelajaran 1998/99, Sekolah Menengah Umum Negeri (SMUN) di Kotamadya Padang, Sumatera Barat. Siswa yang diikutkan dalam pengujian sebenarnya mulai dari kelas I, II, dan III, untuk lima SMUN peringkat terbaik di Padang dalam mata pelajaran Bahasa Inggris, Matematika, dan Fisika. Namun dalam tulisan ini, data yang ditampilkan dibatasi dari tes Bahasa Inggris kelas II dan III saja. Alasan yang dapat diberikan adalah hanya dua tes ini yang memiliki butir soal penjangkar yang cukup untuk menghasilkan penyekalaan yang konsisten.

Data Cawu I Bahasa Inggris kelas II dan III ini pun dibatasi untuk tes pilihan ganda, dengan pola pemontenan benar (1) dan salah (0). Data pengujian ini direkam dalam komputer melalui Lembar Jawaban Komputer Alat Baca Optik (LJK-ABO) yang proses perekamannya dilakukan secara "komputerisasi". Namun, untuk menghindari kesalahan pengisian informasi dan identifikasi siswa (nomor ujian, kode sekolah, dan kode mata pelajaran) dilakukan *editing* data terlebih dahulu. Data respon atau jawaban siswa tidak dilakukan *editing*. Analisis yang diterapkan memang dimulai dengan analisis butir soal, yang bertujuan untuk pembersihan data. Butir soal yang kurang konsisten perilakunya kemudian dibuang agar tidak mengganggu

proses estimasi parameter model (Rasch); demikian pula dengan peserta ujian yang kurang stabil, datanya juga dibuang dengan alasan yang sama. Berdasarkan data butir soal dan peserta ujian yang tersisa, proses penyekalaan dilanjutkan.

Analisis data dilaksanakan dengan program komputer Iteman (teori klasik), *Microcat 3.30*, dan *Bigsteps* (teori Rasch) versi 2.30 (Wright dan Linacre, 1992). Dari analisis tes klasik akan diungkapkan tingkat kesulitan dan indeks daya pembeda butir soal. Analisis item (klasik) ini dipakai untuk memperkirakan butir soal yang berkemungkinan bermasalah dalam analisis Rasch. Tingkat kesukaran akan didekati dengan proporsi jawaban benar, sedangkan indeks daya pembeda butir soal akan dipilih korelasi *point biserial* (yang juga merupakan korelasi *product moment* antara skor butir soal dan skor total tes). Apabila diketemukan butir soal dengan proporsi jawaban yang sangat tinggi/rendah dan *point biserial* rendah, maka butir soal ini mungkin dibuang dan tidak dianalisis dengan *Bigsteps*, karena mungkin akan mengganggu proses estimasi parameter (butir soal lainnya). Dari analisis *Bigsteps* akan diungkapkan estimasi parameter butir soal (δ_i) yang merupakan parameter indeks kesulitan butir soal dan estimasi parameter kemampuan siswa (β_v). Kemudian, parameter butir soal (δ_i) dari dua estimasi yang berbeda (kelas II dan III), diproses untuk meletakkan hasil estimasi parameter kelas III ke dalam skala (metrik) parameter butir soal kelas II. Proses inilah yang disebut sebagai proses penyekalaan vertikal (karena dari kelas III ke kelas II atau sebaliknya).

HASIL DAN PEMBAHASAN

Tabel 1 menunjukkan hasil kalibrasi atau estimasi tingkat kesukaran butir soal tes Bahasa Inggris kelas II dan kelas III, yang dilakukan secara terpisah. Beberapa butir soal terpaksa tidak dicantumkan, karena dibuang. Butir soal yang dibuang ini, tidak dapat dipertahankan karena berbagai penyebab, salah satunya adalah butir soal yang dibuang memiliki berbagai *flaw* (ketidaktepatan atau cacat penulisan butir soal). Hasil kalibrasi ini, di dalamnya termasuk butir soal penjangkar, menunjukkan tingkat kesukaran yang berbeda. Butir soal dengan (δ_i) negatif merupakan butir soal yang mudah untuk rata-rata peserta ujian, sebaliknya yang positif lebih sulit.

Tabel 1 Hasil Kalibrasi Butir Soal Bahasa Inggris Kelas II dan III secara Terpisah

No. butir	δ_i Kelas 2	No. butir	δ_i Kelas 3	No. butir	δ_i Kelas 2	No. butir	δ_i Kelas 3
1	-2,80					4	0,24
2	-2,00					5	-1,05
3	-1,27					6	-4,09
4	-1,73					7	-0,22
5	-2,89					8	0,58
6	0,68					9	1,35
7	0,87					10	-0,11
8	0,02					11	-2,51
9	-1,17					12	0,50
11	-2,27					13	0,41
13	-0,12					14	0,29
14	0,13					15	2,75
15	1,20					16	0,37
16	-1,14					17	-1,10
17	1,17					18	-1,54
18	-0,65					19	-0,82
19	1,09					20	0,80
20	1,20					21	-0,25
23	0,34					22	-0,58
24	0,08					23	0,33
25	0,98					25	0,91
26	-0,55					32	-0,99
27*	-1,23	26	-1,46			33	0,98
28*	-0,83	27	-1,25			34	1,51
29*	0,51	28	0,46			35	-0,63
30*	1,19	29	0,80			36	0,84
31*	0,52	31	0,38			37	0,41
32	1,76					38	0,28
33	-1,06					40	0,64
34	1,56					41	0,86
35	1,15					42	0,80
38	3,07					43	0,23
39	1,20					44	0,86
40	-1,74					45	3,04

Lanjutan Tabel 1

No. butir	δ_i Kelas 2	No. butir	δ_i Kelas 3	No. butir	δ_i Kelas 2	No. butir	δ_i Kelas 3
		1	-3,22			46	-0,05
		2	-2,96			47	0,46
		3	-0,23				

Catatan: *menunjukkan butir soal penjangkar *(linking items)*

Kemudian, Tabel 2 menunjukkan proses penyekalaan khusus untuk butir soal penjangkar, yang ternyata posisi butir soal tersebut dalam perangkat tes kelas II dan III tidak dalam posisi yang sama, meskipun urutannya relatif sama. Proses penghitungan faktor penyekalaan dicoba untuk dijelaskan dengan ilustrasi ditampilkan dalam tabel ini. Dalam praktik proses tersebut tidak tampak, karena diproses dengan program komputer.

Tabel 2 Hasil Kalibrasi Butir Soal Penjangkar dan Proses Penyekalaannya

No. butir soal tes kelas II	δ_i kelas II	No. butir soal tes kelasa III	δ_i kelas III	beda δ_i kelas II dan δ_i kelas III
27*	-1,23	26	-1,46	0,23
28*	-0,83	27	-1,25	0,42
29*	0,51	28	0,46	0,05
30*	1,19	29	0,80	0,39
31*	0,52	31	0,38	0,14
rerata	0,032		-0,214	0,246
simpangan baku	1,018		1,056	0,159

Butir soal nomor 29 pada tes kelas II atau nomor 28 kelas III merupakan butir soal penjangkar *(linking item)* terlemah dalam kelompok ini. Hal ini dapat dilihat dari simpangan antara parameter tingkat kesukaran butir soal ketika diujikan di kelas II dan kelas III paling kecil. Sebaliknya, butir soal penjangkar telah teruji sehingga perkiraan tingkat kesukaran (klasik dan Rasch) telah diketahui sebelumnya. Untuk itu, perlu pembakuan butir soal.

Tabel 3 menunjukkan hasil akhir penyekalaan setelah semua butir soal Bahasa Inggris Kelas III diletakkan pada skala butir soal Bahasa Inggris Kelas II. Dari kolom "Skala baru" dapat dilihat perubahan tingkat kesulitan butir soal kelas III (bandingkan dengan tingkat kesukaran *linking item* dalam skala baru ini dengan dalam Tabel 2).

Tabel 3 Hasil Akhir Proses Penyekalaan Butir Soal Bahasa Inggris Kelas III ke dalam Skala Tes Bahasa Inggris Kelas II

No. butir	δ_i Kelas 2	No. butir	δ_i Kelas 3	δ_i Skala baru (II/III)	No. Butir	δ_i Kelas 3	δ_i Skala baru (III)
1	-2,80			-2,80	4	0,24	0,486
2	-2,00			-2,00	5	-1,05	-0,804
3	-1,27			-1,27	6	-4,09	-3,844
4	-1,73			-1,73	7	-0,22	0,026
5	-2,89			-2,89	8	0,58	0,826
6	0,68			0,68	9	1,35	1,596
7	0,87			0,87	10	-0,11	0,136
8	0,02			0,02	11	-2,51	-2,264
9	-1,17			-1,17	12	0,50	0,746
11	-2,27			-2,27	13	0,41	0,656
13	-0,12			-0,12	14	0,29	0,536
14	0,13			0,13	15	2,75	2,996
15	1,20			1,20	16	0,37	0,616
16	-1,14			-1,14	17	-1,10	-0,854
17	1,17			1,17	18	-1,54	-1,294
18	-0,65			-0,65	19	-0,82	-0,574
19	1,09			1,09	20	0,80	1,046
20	1,20			1,20	21	-0,25	-0,004
23	0,34			0,34	22	-0,58	-0,334
24	0,08			0,08	23	0,33	0,576
25	0,98			0,98	25	0,91	1,156
26	-0,55			-0,55	32	-0,99	-0,744
27*	-1,23	26	-1,46	-1,214	33	0,98	1,226
28*	-0,83	27	-1,25	-1,004	34	1,51	1,756
29*	0,51	28	0,46	0,706	35	-0,63	-0,384
30*	1,19	29	0,80	1,046	36	0,84	1,086

Lanjutan Tabel 3

No. butir	δ_i Kelas 2	No. butir	δ_i Kelas 3	δ_i Skala baru (II/III)	No. Butir	δ_i Kelas 3	δ_i Skala baru (III)
31*	0,52	31	0,38	0,626	37	0,41	0,656
32	1,76			1,76	38	0,28	0,526
33	-1,06			-1,06	40	0,64	0,886
34	1,56			1,56	41	0,86	1,106
35	1,15			1,15	42	0,80	1,046
38	3,07			3,07	43	0,23	0,476
39	1,20			1,20	44	0,86	1,106
40	-1,74			-1,74	45	3,04	3,286
		1	-3,22	-2,974	46	-0,05	0,196
		2	-2,96	-2,714	47	0,46	0,706
		3	-0,23	0,016			

Catatan: *menunjukkan butir soal penjangkar *(linking items)*

Dari Tabel 3 terlihat bahwa skala baru untuk tes kelas II masih tetap sebagaimana tingkat kesukaran butir soal (δ_i) dengan skala lama, sedangkan untuk tes kelas III skala baru lebih tinggi angkanya (δ_i) dibandingkan dengan skala lamanya. Perbedaan atau perubahan angka skala (δ_i) ini pada dasarnya adalah kalibrasi butir soal tes (Bahasa Inggris) kelas III dilihat dalam posisi kelas II. Dengan kata lain, seandainya tes kelas III diberikan kepada siswa kelas II butir soalnya akan bertambah sulit. Hal ini ditandai dengan skala (δ_i) baru untuk kelas III lebih tinggi dari skala (δ_i) lama, sedangkan skala (δ_i) baru tes kelas II tetap sama dengan skala (δ_i) lamanya. Proses penempatan skala butir soal dari dua kelas yang berbeda inilah yang disebut dengan *vertical equating* (penyekalaan vertikal).

Apabila pendekatan ini dilaksanakan untuk semua tes yang dipakai dalam ujian Cawu, maka guru dan pengelola sekolah dapat dengan mudah memperkirakan kemungkinan sukses siswa sejak kelas I sampai nanti (seandainya) dia mencapai kelas III. Seandainya seseorang atau sekelompok murid diproyeksikan akan mengalami kegagalan atau berprestasi rendah, maka suatu program intervensi, seperti pemberian bimbingan belajar atau jam belajar tambahan dapat diberikan. Bimbingan belajar atau jam belajar tambahan inilah yang akan bermanfaat untuk peningkatan prestasi belajar peserta didik, yang akhirnya akan bermuara kepada peningkatan mutu pen-

didikan (lulusan) sebuah sekolah atau daerah. Antisipasi awal inilah yang perlu diketahui oleh pengelola dan pelaksana pendidikan.

Manfaat lain dari suatu penyekalaan vertikal ini adalah peluang pengelola dan pelaksana pendidikan untuk mengkaji perkembangan kemajuan belajar siswa orang perorang. Dengan demikian, guru dan guru pembimbing dapat memberikan orientasi belajar lanjutan sejak awal dengan memperhatikan kemampuan dan proyeksi prestasi belajarnya. Hal ini dapat dipakai untuk memotivasi siswa belajar lebih giat dan juga bermanfaat untuk mempersiapkan mereka secara fisik dan mental untuk menerjuni (peluang) profesi dan karier yang mungkin tidak pernah dibayangkan sebelumnya. Untuk yang terakhir ini memang diperlukan informasi lain, misalnya minat dan cita-cita siswa.

Oleh karena itu, sangat penting kiranya ujian Cawu dirancang sedemikian rupa sehingga data pengujian dapat dipakai untuk melakukan penyekalaan vertikal ini. Pengembangan tes model ini dalam dunia pengujian pendidikan disebut sebagai *multi level testing*. Perancangan tes Cawu yang mengakomodasikan "peluang" penyekalaan vertikal ini sebenarnya tidak terlalu sulit, dan dapat diintegrasikan dengan model ujian Cawu yang selama ini telah dikenal.

KESIMPULAN DAN SARAN

Kesimpulan

Dari uraian dan pembahasan yang diberikan di muka dapat disimpulkan bahwa teori Rasch, sebagaimana diimplementasikan dalam program komputer *Bigsteps*, sangat mungkin dan mudah dipakai untuk melakukan penyekalaan vertikal *(vertical equating)* untuk tes mata uji (pelajaran) yang sama tetapi diujikan untuk kelas yang berbeda. Dalam ilustrasi ditunjukkan bahwa tes kelas III sangat mudah diskalakan kepada tes kelas II. Penyekalaan seperti ini dimaksudkan sebagai upaya untuk memahami perilaku butir soal, seandainya tes kelas atas (kelas III) diberikan kepada siswa kelas rendah (kelas II). Hasil penyekalaan vertikal dalam diskusi di muka menunjukkan bahwa tes kelas III seandainya diberikan kepada siswa kelas II akan bertambah sulit. Ini mudah dipahami, karena siswa kelas III secara teoretis telah lebih lama belajar sehingga memiliki pemahaman ataua kemampuan yang lebih tinggi.

Ujian Cawu SMU sangat potensial untuk dikembangkan menjadi tes yang bersifat berkelanjutan, mulai dari kelas I sampai dengan kelas III,

sehingga dapat dikembangkan skala pengukuran atau penilaian dari kelas I ke kelas III. Tes semacam ini mungkin dikenal sebagai *multilevel test*, yang bermanfaat untuk memantau perkembangan mutu pendidikan dan juga sekaligus perkembangan kemajuan belajar siswa. Seandainya dalam suatu kelompok siswa telah diketahui bahwa siswa dengan prestasi tertentu akan menghasilkan lulusan yang kurang bermutu (penguasaan minimalnya rendah), dan dalam waktu yang bersamaan dapat dirunut ke kelas I tipe prestasi siswa ini pada waktu itu, maka apabila kasus yang sama di kemudian hari menimpa siswa lain, suatu program intervensi (perlakuan) khusus dapat dipersiapkan untuk mengubah jalan prestasi siswa tersebut.

Saran

Pengembangan model Cawu yang memungkinkan penilaian berkelanjutan (dalam format *multilevel test*) akan sangat bermanfaat untuk memantau perkembangan mutu pendidikan, baik untuk individu siswa maupun kelompok siswa. Pemanfaatan seperti itu jelas akan membantu pengelolan dan pelaksana pendidikan meningkatkan mutu lulusan (suatu sistem dan jenjang pendidikan). Kajian longitudinal seperti itu akan sangat bermanfaat untuk membantu mengarahkan karier dan pendidikan lanjut siswa. Dampaknya, efisiensi internal dan eksternal pendidikan mungkin dapat ditingkatkan.

Untuk mencapai hal ini, berbagai syarat pengembangan tes Cawu memang perlu dipenuhi. Salah satunya adalah rancangan (kisi-kisi) ujian atau tes Cawu perlu dipersiapkan secara cermat dengan memperhatikan topik bahasan atau konsep esensial mana yang layak dijadikan butir soal penjangkar *(linking items)*. Pendekatan pengujian seperti ini masih perlu juga disosialisasikan kepada pejabat Depdikbud, baik tingkat Kanwil maupun Kandep, Kepala Sekolah, dan guru. Pengembangan butir soal, khususnya dalam arah bank soal, perlu dilakukan di tingkat Kandep atau Kanwil Depdikbud. Pelatihan dan orientasi pemanfaatan model pengujian ini sehingga guru dan pengelola pendidikan lainnya tidak mengalami kesulitan dalam interpretasi hasil pengujian dan penilaiannya.

CATATAN

Penulis mengucapkan banyak terima kasih kepada Direktur Direktorat Pembinaan Penelitian dan Pengabdian pada Masyarakat, Ditjen Dikti, yang melalui proyek penelitian ilmu terapan berkenan memberi dana bagi terlaksananya penelitian yang menjadi landasan penulisan artikel ini. Peneliti juga menyampaikan terima kasih kepada rekan Nonny Swediati, Ph.D yang atas kritik, saran, dan bantuan analisis dalam penelitian dan publikasi ini.

DAFTAR RUJUKAN

Angoff, W.H. 1971. Scales, Norms, and Equivalent Scores. Dalam Thorndike, R.L. (Ed.). *Educational Measurement* (halaman 508—600). Washington, DC: American Council on Education.

Crocker, L. dan Algina, J. 1986. *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart and Winston.

Depdikbud. 1989. *Undang-undang Republik Indonesia Nomor 2 Tahun 1989 tentang Sistem Pendidikan Nasional beserta Penjelasan*. Jakarta: Balai Pustaka.

Hambleton, R.K. dan Swaminathan, H. 1985. *Item Response Theory: Principles and Applications*. Boston, MA: Kluwer Academic Publishers.

Kolen, M.J. dan Brennan, R.L. 1995. *Test Equating: Methods and Practices*. New York: Springer-Verlag.

Petersen, N.S., Kolen, M.J. dan Hoover, H.D. 1989. Scaling, Norming, and Equating. Dalam Linn, R.L. (Ed.). *Educational Measurement* (halaman 221—262). New York: American Council on Education-Macmillan Publishing.

Wright, B.D. dan Linacre, J.M. 1992. *A User's Guide to Bigsteps: Rasch-Model Computer Program*. Chicago, Il: Mesa Press.

Wright, B.D. dan Stone, M.H. 1979. *Best Test Design: Rasch Measurement*. Chicago, Il: Mesa Press.