Aplikasi Pemodelan Rasch pada Asesmen Pendidikan: Implementasi Penilaian Formatif (*assessment for learning*) [1]

Oleh: Bambang Sumintono[2]

Pendahuluan

Penilaian pendidikan adalah proses yang tidak terpisahkan dari kegiatan pendidikan. Proses belajar mengajar yang terjadi di sekolah selalu melibatkan penilaian pendidikan sebagai hal yang sangat penting dilakukan. Sebab, tanpa hal tersebut dilakukan, sulit diketahui secara pasti apakah kemajuan belajar yang dilakukan apakah sudah tercapai atau tidak.

Hampir semua ujian yang dilakukan di sekolah umumnya menggunakan pendekatan skor untuk menjelaskan pencapaian prestasi siswa. Pada saat yang sama terdapat kelemahan yang tidak terhindarkan dengan pendekatan ini yang biasanya tidak bisa mendukung umpan balik yang efektif. Penggunaan pendekatan pemodelan rasch bisa digunakan untuk memberikan perspektif berbeda dari data yang sama. Makalah ini akan menjelaskan ruang lingkup penilaian, khususnya tes formatif dan bagaimana menggunakannya sesuai dengan perspektif *assessment for learning* melalui pendekatan rasch model.

Penilaian Pendidikan

Definisi penilaian pendidikan sangat beragam, namun biasanya hal itu menyebutkan bahwa penilaian adalah cara untuk menempatkan pembelajar dalam konteks yang dapat menyatakan apa yang dia ketahui dan mampu lakukan (juga menjelaskan apa yang dia belum tahu dan belum mampu dia lakukan). Definisi penilaian pendidikan seperti ini memang sangat luas yang mengindikasikan bahwa untuk mengetahui kemajuan belajar seseorang bisa dilakukan baik secara formal maupun informal, kapan saja dan dalam waktu jangka waktu yang tidak harus dibatasi (Musial et al., 2009).

Bentuk penilaian pendidikan yang banyak dikenal adalah ujian. **Ujian** atau tes adalah prosedur evaluasi yang biasa dilakukan oleh seorang guru terhadap pengetahuan dan keterampilan siswa

[1] Makalah dipresentasikan dalam Kuliah Umum pada Jurusan Statistika, Institut Teknologi Sepuluh November, Surabaya, 17 Maret 2016.

[2] dosen pada Institute of Educational Leadership, Universiti Malaya, Kuala Lumpur, Malaysia
email: bambang@um.edu.my dan deceng@gmail.com
Blog: http://deceng2.wordpress.com (pengalaman dosen) dan http://deceng3.wordpress.com (rasch model)

untuk mengetahui kinerjanya dengan menggunakan instrumen tertentu. Sedangkan, yang disebut **instrumen** pun beragam, yaitu bisa berbentuk set soal yang harus dikerjakan maupun suatu tugas menghasilkan suatu produk tertentu. Ujian bisa dilakukan dalam berbagai bentuk, dimaksudkan untuk memberikan pengukuran yang objektif dari kegiatan pembelajaran yang telah dilakukan. Bentuk ujian atau tes yang paling umum dipakai oleh guru dalam menguji siswanya di kelas adalah tes tertulis. Namun bentuk ujian lain juga bisa dilakukan seperti ujian pada mata pelajaran olahraga, ujian keterampilan dan lainnya.

Dalam aktivitas kegiatan belajar mengajar di sekolah, yang lebih dikenal secara luas dalam konteks penilaian pendidikan disebut sebagai **penilaian formatif** dan **penilaian sumatif**. Penilaian formatif adalah kegiatan penilaian oleh guru terhadap siswa dimana tujuannya lebih kepada memberikan informasi yang bermanfaat sehingga pembelajaran berikutnya kualitasnya lebih baik lagi. Hal ini berimplikasi bahwa pada penilaian formatif guru mengumpulkan informasi dan melakukan interpretasi dari bukti hasil belajar yang ada, tentang apa yang perlu diketahui lebih lanjut oleh siswa, serta melakukan adaptasi pengajarannya sesuai dengan kebutuhan siswa. Dalam bahasa yang populer ini juga disebut sebagai ***assessment for learning***.

Penilaian sumatif adalah penilaian yang dilakukan untuk mengetahui apa yang sudah diketahui pelajar atau yang bisa dia lakukan, pada periode akhir masa belajar yang ditetapkan. Tujuannya memang untuk memberikan informasi, prestasi apa yang telah dicapai; dalam istilah populernya disebut ***assessment of learning***. Pada jenis penilaian ini, siswa selalunya berada dalam situasi dimana mereka harus menampilkan segala yang telah dikuasai selama waktu tertentu yang menunjukkan prestasi belajarnya, misalnya dalam Ujian Akhir Nasional (UAN).

Hasil kegiatan ujian yang dilakukan pada siswa biasanya digunakan dalam berbagai cara. Nilai atau skor yang didapat oleh siswa dalam satu ujian bisa menunjukkan seberapa bagus prestasinya dibanding temannya di kelas, ataupun dibanding prestasi yang telah dia raih sebelumnya di kelas yang sama. Secara lebih lengkapnya, hasil ujian ini dapat digunakan oleh guru untuk: (a) menentukan abilitas siswa relatif terhadap siswa lain dalam tes yang sama; (b) menunjukkan perkembangan kemampuan siswa dalam suatu jangka waktu dalam pengetahuan dan ketrampilan tertentu; (c) menunjukkan bukti pemahaman akan satu materi pelajaran, pengetahuan atau ide tertentu; dan (d) hal itu dapat meramalkan kinerja siswa di masa hadapan. Supaya hasil tes bisa dipercaya dan tepat untuk digunakan maka aspek validitas dan reliabilitas instrumen adalah hal esensial yang harus dipenuhi.

Analisis Hasil Ujian

Analisi hasil ujian dimulai dari proses mendapatkan informasi mengenai abilitas siswa dari hasil ujian yang dilakukan terhadap siswa, yang disebut juga skor ujian. Terdapat berbagai cara untuk mendapatkan skor yang menunjukkan kemampuan siswa. Cara yang umum dilakukan adalah menjumlahkan banyaknya jawaban yang benar. Skor ini menunjukkan kemampuan siswa. Analisis lebih lanjut adalah dengan melakukan prosedur statistik sederhana untuk bisa menjelaskan lebih jauh tentang kualitas soal, kualitas siswa maupun perbandingan atribut yang diukur.

Pendekatan yang banyak dipakai saat ini dalam analisis hasil ujian adalah pendekatan teori tes klasik (*clasical test theory* atau CTT). Teori tes klasik bisa digunakan untuk melakukan prediksi tentang hasil dari suatu ujian (tes). Prediksi ini dilakukan dengan mempertimbangkan beberapa parameter seperti kemampuan siswa dan tingkat kesulitan soal. Charles Spearman mengemukakan teori tes klasik ini pada tahun 1904 dan banyak diaplikasikan dalam bidang pendidikan khususnya penilaian pendidikan. Asumsi dasar yang dipunyai oleh teori tes klasik ini adalah, skor yang didapat dilambangkan dengan X, tidak lain adalah terdiri dari skor murni (T) dan eror pengukuran (E), sehingga persamaannya: $X = T + E$

Artinya di dalam skor hasil ujian yang didapat satu siswa misalnya, didalamnya terkandung skor murni dan eror pengukuran. Hal yang perlu dicatat adalah, skor tampak (X) bersifat nyata (muncul di dalam data secara langsung) sedangkan skor murni (T) dan eror pengukuran (E) bersifat tersembunyi (*latent*) atau tidak bisa diamati secara langsung. Keduanya muncul di dalam data setelah melalui proses estimasi. Asumsi lain yang perlu diketahui adalah eror pengukuran (E) dalam CTT bersifat acak dan tidak berkorelasi dengan X maupun T, dan korelasi yang diharapkan muncul adalah 0 (nol). Teori tes klasik (CTT) hanya menekankan pada skor tampak dari satu ujian, yang biasanya disimpulkan sebagai kemampuan (abilitas) seseorang dari ujian yang diikuti.

Dari skor mentah ini maka berbagai analisis dan interpretasi bisa dihasilkan sesuai dengan keperluan yang dilakukan, diantaranya adalah: a) Statistik deskriptif, yaitu tendensi sentral (misalnya rata-rata), ukuran keragaman (misalnya varians) dan tabel frekuensi. Ketiganya akan memberikan informasi secara langsung butir soal mana yang berguna dan mana yang tidak. Misalnya, keragaman skor antar siswa yang rendah menunjukkan rendahnya kualitas soal-soal di dalam tes; b) tingkat kesulitan butir. Tingkat kesulitan menunjukkan proporsi siswa yang dapat

mengerjakan soal secara benar dari satu ujian. Titik terendah sebesar 1,0, artinya semua siswa dapat menjawab dengan betul soal tes dan titik tertinggi tingkat kesulitan adalah 0,0, menunjukkan tidak ada satupun (0%) individu yang bisa menjawab dengan benar. Butir soal yang memiliki titik ekstrim (0% atai 100%) seperti kedua contoh di muka tidak banyak berguna karena tidak bisa membedakan kemampuan individu, dengan kata lain hal itu soal yang tidak bagus kualitasnya; c) Daya diskriminasi butir menunjukkan seberapa jauh sebuah soal mampu membedakan individu yang memiliki kemampuan yang tinggi dan rendah. Sederhananya, jika siswa berkemampuan tinggi dan rendah dapat mengatasi soal nomor 10, maka soal ini memiliki daya diskriminasi butir yang rendah. Sebaliknya, jika siswa berkemampuan tinggi dapat mengatasi soal nomor 10 sedangkan yang berkemampuan rendah tidak dapat mengatasi, maka butir nomor 10 memiliki daya diskriminasi yang tinggi; d) Pembobotan butir soal, umumnya dalam konteks CTT, skor untuk tiap butir soal diberikan sama (misal 1 untuk jawaban betul; dan 0 untuk jawaban salah), pembobotan skor diberlakukan bila satu soal yang diberikan mempunyai bobot yang berbeda untuk menghasilkan total skor mentah. Terdapat banyak cara untuk memberikan pembobotan, misal melalui reliabilitas soal, dimana soal dengan reliabilitas tinggi memiliki bobot lebih besar.

Catatan Mengenai Teori Skor Klasik

Teori skor klasik bukan satu-satunya pendekatan dalam penilaian pendidikan dan psikometri. Ada beberapa pendekatan lain yang merupakan alternatif dari pendekatan teori klasik. Pada dasarnya penggunaan skor mentah/*raw score* sebagai ukuran prestasi memiliki beberapa beberapa kelemahannya, diantaranya adalah (Alagumalai et al., 2005). :

a. **Skor mentah pada dasarnya bukanlah hasil pengukuran**. Lebih tepatnya skor mentah adalah jumlah jawaban benar dari soal yang dikerjakan siswa
b. **Skor mentah adalah informasi awal**. *S*kor mentah juga biasanya dinyatakan dalam persentase (%) yang tidak lain hanyalah ringkasan data berupa angka, tetapi tidak memberikan data dari suatu pengukuran
c. **Skor mentah memiliki makna kuantitatif yang lemah.** Makna kuantitatif dari skor mentah yang didapat akan berbeda, tergantung banyaknya soal, sedangkan persentase jawaban betul selalu tergantung pada tingkat kesulitan soal
d. **Skor mentah tidak menunjukkan kemampuan seseorang terhadap tugas tertentu.** Skor mentah juga tidak bisa banyak menjelaskan tingkat kesulitan soalnya; dan terakhir,

e. **Skor mentah dan persentase jawaban benar tidak selalu bersifat linier**. Dalam sebuah tes yang bersifat linier, siswa yang memiliki skor 15 (skala 0 hingga 100) selalu memiliki kemampuan lebih tinggi dibanding yang memiliki skor 10. Namun secara empirik terkadang keduanya memungkinkan memiliki kemampuan yang sama.

Oleh karena itu melihat alternatif lain dalam melakukan analisis hasil ujian sangat diperlukan, khususnya dengan berbagai kelemahan teori tes klasik di atas. Kekurangan CTT kemudian diperbaiki dengan teori respon butir (*item response theory* atau IRT) dengan berbagai variasi parameter logistiknya (PL), salah satunya adalah 1PL yang dikembangkan menjadi model rasch. Tidak seperti CTT yang selalu bergantung pada skor, IRT tidak tergantung pada sampel soal/pernyataan tertentu dan abilitas orang yang terlibat dalam ujian /survey. Pada bagian selanjutnya akan dijelaskan secara singkat tentang model rasch, pengukuran yang objektif, serta aplikasi model rasch dalam penilaian pendidikan dengan penggunaan perangkat lunak (software) yang dirancang untuk aplikasi rasch model (Bond and Fox, 2015).

Model Rasch

Georg Rasch mengembangkan satu model analisis dari teori respon butir (atau *Item Response Theory*, IRT) pada tahun 1960-an biasa disebut 1PL (satu parameter logistic) (Olsen, 2003). Model matematika ini kemudian dipopulerkan oleh Benjamin Wright (Linacre, 2011). Dengan data mentah berupa data dikotomi (berbentuk benar dan salah) yang mengindikasikan kemampuan siswa, Rasch memformulasikan hal ini menjadi satu model yang menghubungan antara siswa dan aitem (Sumintono & Widhiarso, 2015).

Sebagai ilustrasi, seorang siswa yang mampu mengerjakan 80% soal dengan benar tentu mempunyai abilitas yang lebih baik dari siswa lain yang hanya bisa mengerjakan 65% soal. Data tersebut (persentase) menunjukkan bahwa data mentah yang diperoleh tidak lain adalah jenis data ordinal yang menunjukkan peringkat dan tidak linier (Linacre, 1999). Oleh karena data ordinal tidak mempunyai interval yang sama, maka data tersebut perlu diubah menjadi data rasio untuk keperluan analisis statistik. Sehingga bila seseorang mendapat skor 80%, maka nilai *odds ratio*-nya adalah 80:20 (bermakna: 80 skor benar dibandingkan 20 skor salah), yang tidak lain adalah data perbandingan frekuensi/rasio yang lebih tepat untuk tujuan pengukuran. Melalui data rasio ini Rasch mengembangkan model pengukuran yang menentukan hubungan antara tingkat kemampuan siswa (*person ability*) dan tingkat kesulitan aitem (*item difficulty*) dengan menggunakan fungsi logaritma untuk menghasilkan pengukuran dengan interval yang sama.

Hasilnya adalah satuan baru yang disebut logit (*log odds unit*) yang menunjukkan abilitas siswa dan kesulitan aitem; sehingga nantinya dari nilai logit yg didapat, disimpulkan bahwa tingkat kesuksesan siswa dalam mengerjakan soal sangat tergantung dari tingkat abilitasnya dan tingkat kesulitan soal (Englehard, 2013).

Untuk data yang berbentuk dikotomi, pemodelan Rasch menggabungkan suatu algoritma yang menyatakan hasil ekspektasi probabilistik dari aitem 'i' dan responden 'n', yang secara matematis dinyatakan sebagai (Bond dan Fox, 2007):

$$P_{ni}(x_{ni}=1/\beta_n, \delta_i) = \frac{e^{(\beta n - \delta i)}}{1 + e^{(\beta n - \delta i)}}$$

dimana: $P_{ni}(x_{ni}=1 /\beta_n , \delta_i)$ adalah probilitas dari responden n dalam aitem i untuk menghasilkan jawaban betul (x = 1); dengan kemampuan responden, βn, dan tingkat kesulitan aitem δi.

Persamaan di atas dapat lebih disederhanakan dengan memasukkan fungsi logaritma dan menjadikannya:

$$\log (Pni (Xni = 1 / \beta n, \delta i)) = \beta n - \delta i$$

Sehingga probabilitas akan satu keberhasilan dapat dituliskan sebagai:

probabilitas untuk berhasil	=	kemampuan responden	–	tingkat kesulitan aitem

Racsh Model dan Pengukuran Objektif

Dalam lingkup penilaian pendidikan, maka mendapatkan data berupa angka yang merupakan skor data mentah dari ujian yang dikerjakan oleh siswa misalnya berasal dari soal ujian/instrumen yang diberikan. Instrumen tersebut dirancang dari variabel yang sudah didefinisikan secara memuaskan (misalnya kemampuan kuantitatif), kemudian diidentifikasi konstruk-konstruk yang relevan (yaitu yang dapat diukur melalui tes hitungan, deret bilangan, komparasi kuantitatif); dari sana lah aitem-aitem dibuat dan dikembangkan untuk bisa mengukur variabel yang dimaksud. Pada saat yang sama pilihan jawaban yang disediakan umumnya kemudian mengikuti pola penskoran yang dianut oleh teori test klasik (CTT). Dalam konteks

model rasch, pola penskoran yang ‘menetap’ ini tidak lain adalah pengukuran yang hasilnya bergantung pada siapa yang diukur (*test dependent scoring*); sedangkan yang harus dilakukan dalam riset kuantitatif dalam penilaian pendidikan adalah pengukuran yang objektif (*objective measurement*).

Konsep pengukuran yang objektif dalam ilmu-ilmu sosial dan penilaian pendidikan menurut Mok dan Wright (2004) harus mempunyai lima kriteria, yaitu:

1. Memberikan ukuran yang linear dengan interval yang sama;
2. Melakukan proses estimasi yang tepat;
3. Menemukan aitem yang tidak tepat (*misfits*) atau tidak umum (*outliers*);
4. Mengatasi data yang hilang;
5. Menghasilkan pengukuran yang *replicable* (independen dari parameter yang diteliti)

Dari kelima syarat tadi, sejauh ini hanya rasch model lah yang bisa memenuhi kelima syarat tersebut. Dengan kata lain kualitas pengukuran dalam penilaian pendidikan yang dilakukan dengan rasch model akan mempunyai kualitas yang sama seperti halnya pengukuran yang dilakukan dalam dimensi fisik dalam bidang fisika (misal mengukur panjang dengan mistar centimeter, mengukur berat dengan neraca kilogram dll).

Bila dilihat lebih lanjut, skala logit (*log odds unit*) yang dihasilkan dalam model rasch adalah skala dengan interval yang sama dan bersifat linear yang berasal dari data ratio (*odds ratio*) dan bukannya data mentah skor yang didapat (1). Oleh karena itu proses estimasi abilitas seseorang ataupun tingkat kesulitan soal akan mempunyai nilai estimasi yang lebih tepat dan bisa saling dibandingkan karena mempunyai satuan yang sama (logit) (2). Berhubung algoritma yang digunakan akan melakukan pengurutan secara terstruktur antara responden dari abilitas tinggi ke rendah, yang secara bersamaan juga mengurutkan soal dari yang mudah ke yang sulit, maka adanya ketidaktepatan/konsistensi jawaban dari responden (*misfit*) ataupun pola yang diluar kebiasaan (*outlier*) akan mudah dideteksi; demikian juga untuk pola respon yang diterima satu soal tertentu (3). Pengurutan abilitas responden dan kesulitan soal secara terstruktur juga membuat model rasch dapat melakukan prediksi bila terdapat data yang hilang (4). Skala logit yang dihasilkan akan memunculkan nilai yang tergantung dari pola respon yang diberikan, bukannya pada skor awal yang ditentukan, sehingga rasch model akan selalu menghasilkan pengukuran yang independen (5).

Analisis dengan model Rasch menghasilkan analisis statistik kesesuaian (*fit statistics*) yang memberikan informasi pada peneliti apakah data yang didapatkan memang secara ideal menggambarkan bahwa orang yang mempunyai abilitas tinggi memberikan pola jawaban terhadap aitem sesuai dengan tingkat kesulitannya. Parameter yang digunakan adalah *infit* dan *outfit* dari kuadrat tengah (*mean square*) dan nilai terstandarkan (*standardized values*). Menurut Sumintono dan Widhiarso (2014), *infit* (*inlier sensitive* atau *information weighted fit*) adalah kesensitifan pola respon terhadap aitem sasaran pada responden (*person*) atau sebaliknya; sedangkan *outfit* (*outlier sensitive fit*) mengukur kesensitifan pola respon terhadap aitem dengan tingkat kesulitan tertentu pada responden atau sebaliknya.

Riset kuantitatif dalam ilmu sosial dan penilaian pendidikan selalu menghadapi kritik yang mendasar dalam hal pengujian instrumen risetnya. Uji kuantitatif instrument yang biasa dilakukan dalam CTT adalah indeks reliabilitas (alpha Cronbach) yang hanya mengukur interaksi antara aitem dan person; bagaimana kualitas individual aitem tidak pernah bisa dilakukan karena tiadanya indeks pengukuran yang bisa dilakukan; saat yang sama untuk mendeteksi jawaban responden yang tidak konsisten pun tidak tersedia. Hal yang berbeda dengan teori test klasik, dalam rasch model analisis aitem dilakukan ke tingkat masing-masing aitem. Selain terhadap aitem, rasch model juga secara bersamaan menguji person (responden), dimana akan terlihat pola jawaban responden yang konsisten, yang cenderung untuk menyetujui (dalam instrument sikap) maupun mengidentifikasi jawaban yang asal saja. Uji untuk instrument riset pun bisa dilakukan dalam bentuk uji dimensionalitas, maupun deteksi adanya bias dari aitem yang diujikan. Kesemua itu bisa dilakukan karena pada dasarnya model rasch memenuhi semua syarat pengukuran objektif.

Aplikasi Model Rasch dengan Winstep untuk Penilaian Formatif

1. Peta Konstruk Ukur

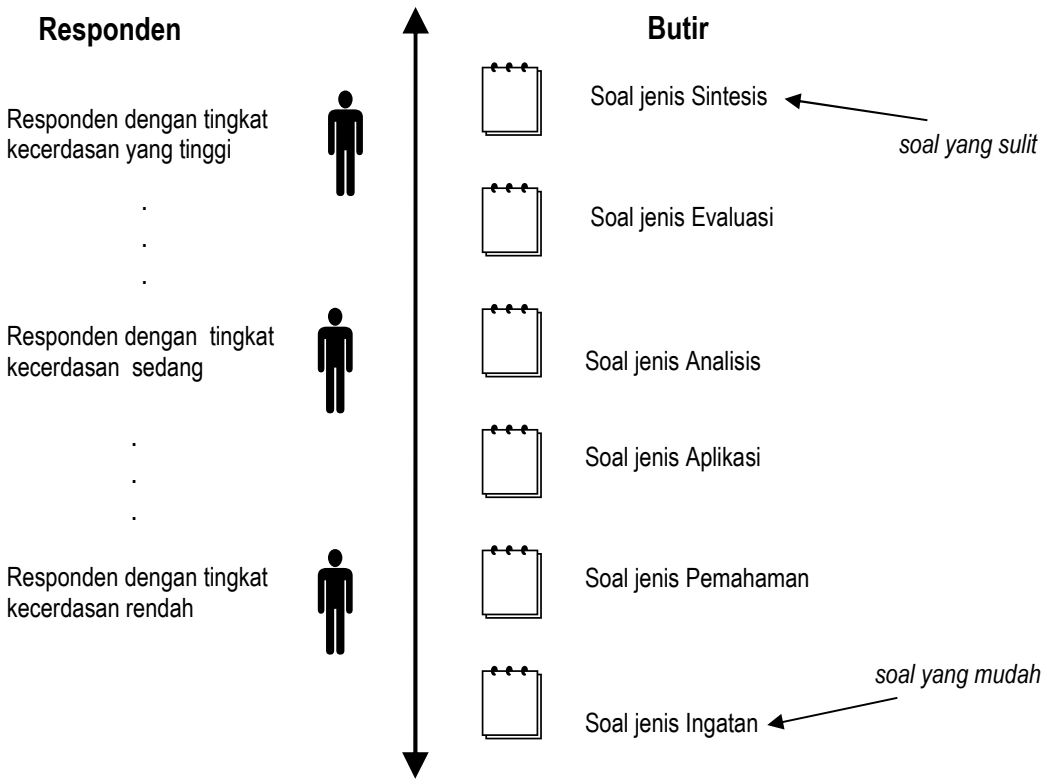
Peta konstruk pada dasarnya adalah representasi visual untuk dapat mengetahui ada dimana tepatnya lokasi dari butir soal sekaligus responden dalam hal dimensi-dimensi yang diukur, Dengan kata lain,

1) Peta konstruk memberikan konteks yang menyeluruh dan substantif pada isi konstruk-konstruk yang diukur;

2) Peta konstruk adalah suatu ide yang menggambarkan kontinum yang didapat dari pengukuran yang isinya terdapat dua hal, yaitu:
   a) peta konstruk responden yang menyatakan kualitas responden, diperingkatkan dari yang terendah ke tertinggi (misalnya abilitas dalam TPA);

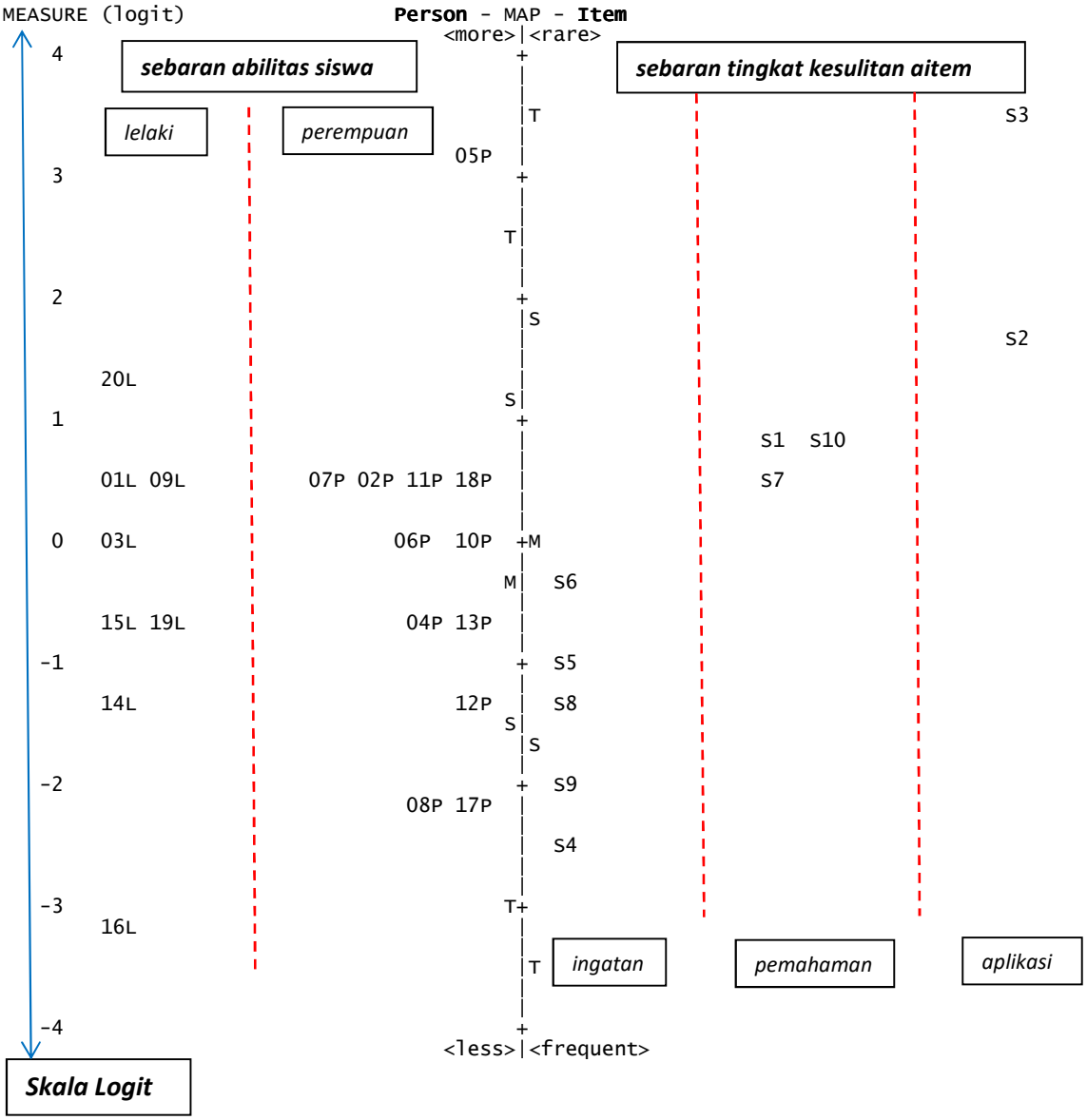
   b) peta konstruk butir yang menunjukan pola jawaban yang diberikan oleh responden terhadap butir yang kemudian diurutkan dari soal yang mudah ke soal yang sangat susah;

Secara teoritik, contoh kontinum tingkat kesulitan butir dapat mengikuti Taksonomi Bloom. Pada tahun 1950-an Benjamin Bloom mengajukan sebuah taksonomi proses kognitif. Taksonomi ini begitu berpengaruh dalam pendidikan, dan telah mengalami berbagai revisi. Menurut Bloom, butir soal yang menekankan pada masalah ingatan atau pengetahuan termasuk pada proses kognitif di level terendah. Oleh karena itu butir-butir yang mengukur proses ini cenderung memiliki tingkat kesulitan rendah. Semakin tinggi level proses kognitif yang dilakukan maka semakin tinggi tingkat kesulitan soal yang mengukurnya. Level proses kognitif yang dikembangkan oleh Bloom bergerak mulai dari ingatan, aplikasi, analisis, evaluasi dan berakhir pada sintesis. Artinya soal-soal jenis sintesis secara teoritik adalah soal yang sulit untuk dikerjakan dengan benar oleh siswa.



Gambar 1. Ilustrasi peta konstruk taksonomi Bloom

Bila kita ilustrasikan taksonomi Bloom ini (Gambar 1), maka individu yang mampu mengerjakan dengan benar soal-soal yang mengukur kemampuan melakukan sintesis dapat dipastikan dia mempunyai abilitas yang lebih tinggi dibanding individu lain yang tidak dapat mengerjakan soal tersebut dengan benar. Dengan menggunakan peta konstruk ini juga kita dapat menggambarkan bahwa siswa yang hanya dapat menjawab soal-soal jenis ingatan, adalah siswa yang memang abilitasnya berbeda dengan siswa yang dapat menjawab dengan benar sampai ke tingkatan analisis seperti terlihat pada gambar di bawah. Pada program Winsteps peta konstruk bisa dihasilkan dengan menampilkan fungsi **Variable Map** (Gambar 2).



Gambar 2. Ilustrasi peta konstruk dengan Winsteps

Terlihat bahwa identifikai peserta test atau *person* ada di bagian kiri, yang menunjukkan abilitas dari setiap peserta tes dalam skala logit; sedangkan pada sebelah kanan adalah kontinum tingkat kesulitan item. Untuk membuat peta lebih informatif maka modifikasi berdasar jenis kelamin (sebelah kiri), dan jenis pertanyaan dari soal yang diberikan (kanan) untuk mengetahui secara lebih tepat informasi tentang peserta test dan butir soal yang diberikan.

2. Pengembangan Instrumen Pengukuran

Pemodelan Rasch menjadi alternatif pengembangan instrumen pengukuran pada penilaian pendidikan selain menggunakan teori klasik. Beberapa tahap yang biasanya dilalui dalam prosedur pengembangan instrumen pengukuran adalah:

a) Verifikasi asumsi unidimensionalitas dan independensi lokal pengukuran
b) Pengujian ketepatan butir-individu dengan model. Butir yang memiliki nilai ketepatan rendah dikeluarkan dari analisis. Analisis diulang lagi hingga semua butir memiliki ketepatan dengan model.
c) Jika jumlah butir yang tersisa masih melebihi jumlah butir yang ditargetkan, maka kita dapat menyeleksi butir dengan berbagai pertimbangan, misalnya : (a) butir yang tidak overlap lokasinya dengan butir lain, (b) butir yang dapat meningkatkan reliabilitas pengukuran, butir yang opsi-opsi responsnya sesuai dengan urutannya (menelaah grafik karakteristik butir) atau (d) butir yang memberikan informasi yang sesuai dengan fungsi pengukuran (menelaah grafik fungsi informasi).

Proses evaluasi terhadap instrumen pengukuran merupakan proses analisis yang bersifat iteratif, yang dilakukan berulang-ulang hingga peneliti menemukan komposisi yang optimal, dimana semua kriteria dapat terpenuhi. Pada program Winsteps, unidimensionalitas terdapat pada fungsi **Item : dimensionality** dan ketepatan butir dengan model (*infit-outfit*) dan lokasinya (*measure*) dapat dilihat pada **Item: fit order**.

TABLE 10.1 C:\Users\user\Desktop\dikotomi.prn ZOU280WS.TXT Jul 1 9:23 2015
INPUT: 20 Person 10 Item REPORTED: 20 Person 10 Item 2 CATS MINISTEP 3.75.0

Person: REAL SEP.: 1.03 REL.: .52 ... Item: REAL SEP.: 2.17 REL.: .82

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
9	16	20	-2.05	.63	1.12	.4	2.28	1.4	A .22	.42	85.0	82.4	S9
4	17	20	-2.49	.70	1.60	1.3	1.90	1.0	B .00	.39	80.0	86.3	S4
3	1	20	3.46	1.12	1.39	.7	.97	.5	C .14	.31	95.0	94.9	S3
5	13	20	-1.06	.54	1.11	.5	1.21	.6	D .39	.47	70.0	74.8	S5
6	10	20	-.25	.51	1.09	.5	.96	.0	E .43	.47	65.0	70.9	S6
7	7	20	.55	.53	.87	-.6	.71	-.5	e .56	.46	75.0	71.2	S7
2	4	20	1.52	.63	.86	-.3	.81	.0	d .50	.43	85.0	83.2	S2
1	6	20	.84	.55	.83	-.6	.69	-.5	c .56	.45	80.0	74.9	S1
10	6	20	.84	.55	.76	-1.0	.57	-.7	b .61	.45	80.0	74.9	S10
8	14	20	-1.35	.56	.55	-1.8	.42	-1.0	a .74	.46	85.0	76.9	S8
MEAN	9.4	20.0	.00	.63	1.02	-.1	1.05	.1			80.0	79.0	
S.D.	5.1	.0	1.71	.17	.29	.9	.56	.7			8.1	7.2	

Tiga kolom ini: **OUTFIT MNSQ**, **OUTFIT ZSTD** & **PT-MEASURE CORR** adalah kriteria menilai kesesuaian butir soal (*item outliers atau misfit*)

Menurut Boone, Staver dan Yale (2014), kriteria yang digunakan untuk memeriksa kesesuaian butir soal yang tidak sesuai (*outliers* atau *misfits*) adalah:

a. Nilai Outfit mean square (MNSQ) yg diterima: 0,5 < MNSQ < 1,5
b. Nilai Outfit Z-standard (ZSTD) yg diterima: -2,0 < ZSTD < +2,0
c. Nilai Point Measure Correlation (Pt Mean Corr): 0,4 < Pt Measure Corr < 0,85

Soal yang berada di baris teratas (soal no 9, S9) menunjukkan pola respon yang tidak fit, demikian juga S4, yang bila dibandingkan dengan butir lain, kedua soal ini mempunyai masalah.

3. Deteksi Bias Pengukuran

Butir maupun instrumen pengukuran dapat bersifat bias, yaitu ketika sebuah butir lebih memihak pada salah satu individu dengan karakteristik tertentu. Sementara itu individu dengan karakteristik oposisinya justru dirugikan. Misalnya, butir sebuah tes kecerdasan anak melibatkan gambar berupa salju untuk dikenali kejanggalannya. Bagi anak-anak yang pernah berinteraksi

dengan salju, soal ini cukup mudah dipahami. Sebaliknya bagi anak-anak yang tidak berinteraksi dengan salju, soal ini sulit dipahami. Butir ini cenderung bias dalam mengukur, yang dalam psikometri dinamakan dengan butir yang terjangkit keberfungsian butir diferensial (DIF/*differential item functioning*). Pemodelan Rasch menyediakan menu untuk memfasilitasi peneliti yang hendak mendeteksi adanya butir-butir yang yang terjangkit DIF.

Pada paket program Winsteps, informasi mengenai bias butir ini dapat dilihat melalui **Item: DIF, between/within**. Butir-butir yang memiliki nilai P (PROB.) di bawah 0,05 menunjukkan bahwa butir tersebut terjangkit DIF pada tabel tersebut akan muncul nilai selisih tingkat kesulitan butir ditinjau dari dua sampel yang diuji. Misalnya jender, latar belakang budaya atau lokasi demografis; bahkan Winsteps juga bisa mendeteksi kombinasi data demografis misalnya jender dan lokasi demografis.

DIF class specification is: DIF=$S3W1

Person CLASSES	SUMMARY DIF CHI-SQUARE	D.F.	PROB.	BETWEEN-CLASS MEAN-SQUARE	t=ZSTD	Item Number	Name
2	.6894	1	.4064	.3675	-.1305	1	S1
2	.5313	1	.4661	.2785	-.2645	2	S2
1	.0000	0	1.0000	.0000	.0000	3	S3
2	.2156	1	.6424	.1174	-.6111	4	S4
2	4.0188	1	.0450	2.4373	1.2049	5	S5
2	.0245	1	.8756	.0128	-1.1540	6	S6
2	1.8816	1	.1702	1.0593	.5126	7	S7
2	.2470	1	.6192	.1282	-.5803	8	S8
2	.0961	1	.7566	.0501	-.8678	9	S9
2	.0596	1	.8071	.0298	-.9922	10	S10

Butir soal S5 mempunyai nilai probabilitas 0,045 yang menunjukkan nilainya kurang dari 5%. Hal ini menunjukkan butir soal ini perlu diperbaiki supaya tidak merugikan kelompok jender tertentu.

4. Analisis Abilitas Individu

Selain mengukur kemampuan abilitas individu secara lebih tepat, dengan rasch model juga bisa diketahui ketepatan abilitas dengan pola respon yang diberikan. Dalam aspek ini guru atau dosen bisa mengetahui lebih awal informasi dari hasil tes yang dilakukan, dimana tes formatif ini akan memberikan informasi yang berharga untuk perbaikan pengajaran maupun membantu siswa

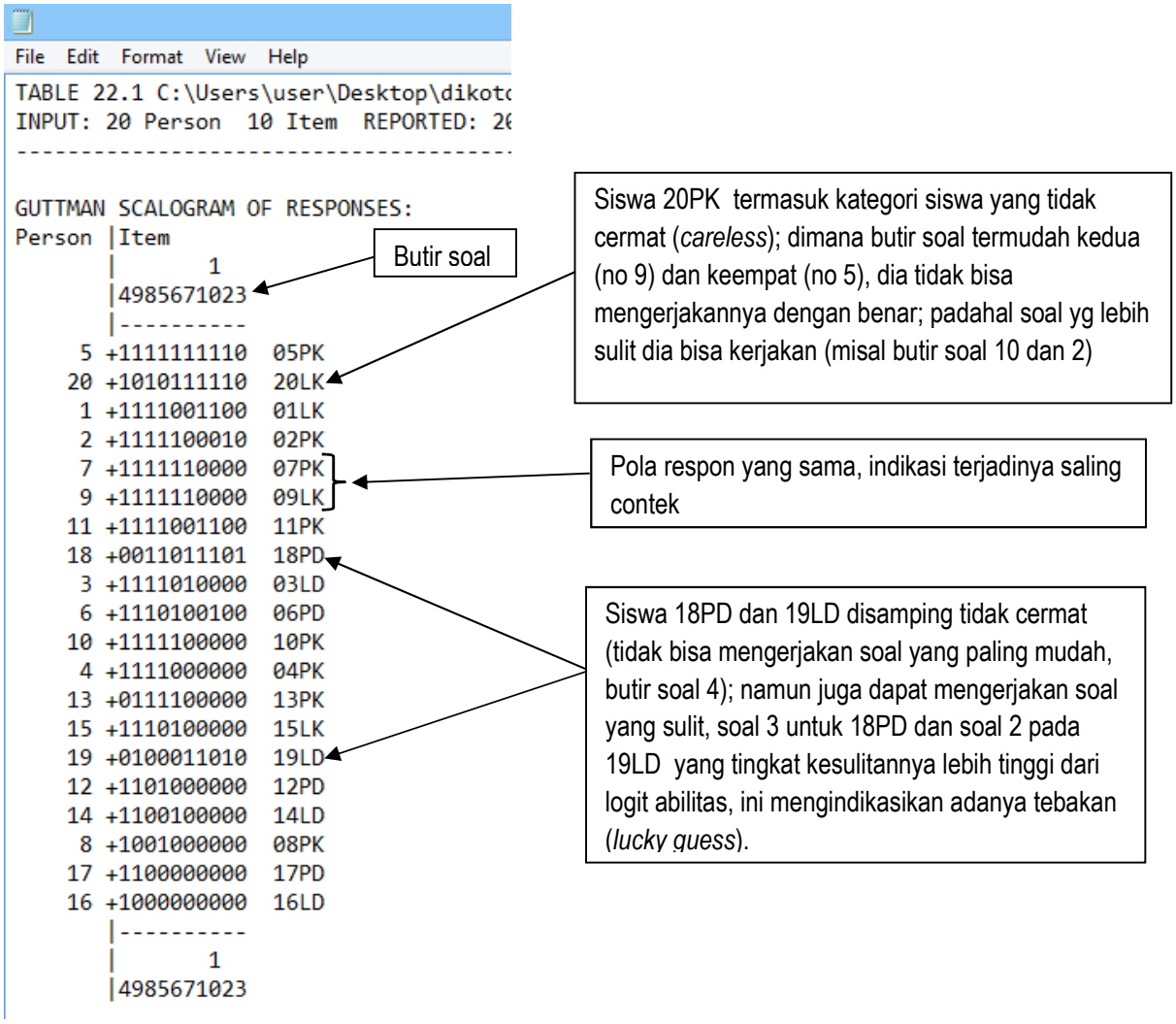
secara lebih tepat. Deteksi yang bisa dilakukan berupa identifikasi miskonsepsi siswa pada pokok bahasan tertentu, yang bisa diketahui dari informasi fit statistik-nya maupun pola respon yang diluar kebiasaan.

Para meter yang digunakan seperti halnya untuk mengetahui ketepan item yaitu pada outfit mean-square, outfit-zstd dan point measure correlation (Boone, Staver & Yale, 2014). Nilai yang di luar batas ketepatan statistik menunjukkan pola responnya yang perlu diketahui lebih jauh. Pada program winstep, tabel informasi ini bisa dimunculkan dengan fungsi person fir order (Tabel 6) yang mengurutkan dari yang tidak fit yang berada di urutan atas.

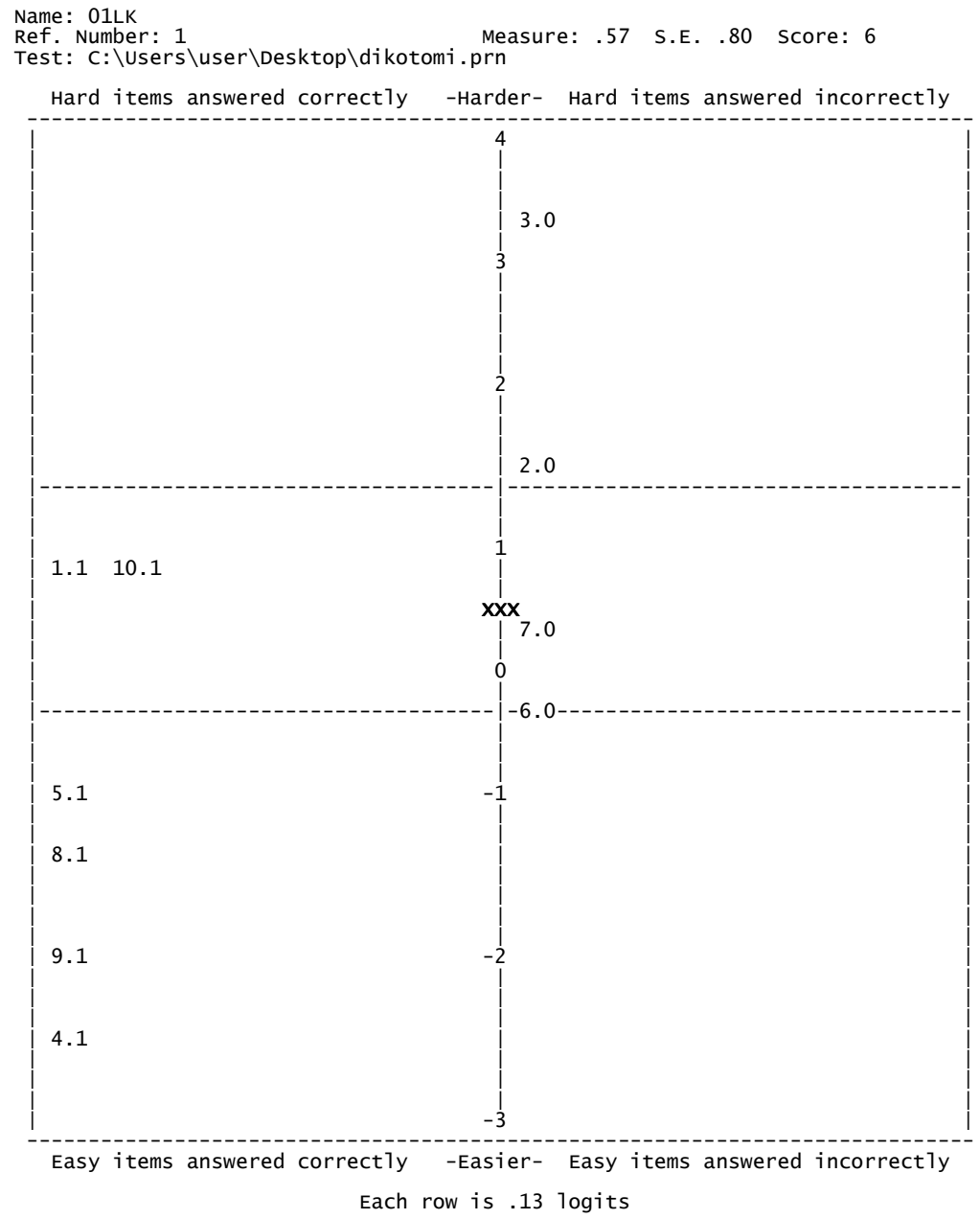
```
TABLE 6.1 C:\Users\user\Desktop\dikotomi.prn     ZOU280WS.TXT  Jul  1  9:23 2015
INPUT: 20 Person  10 Item  REPORTED: 20 Person  10 Item  2 CATS  MINISTEP 3.75.0
-------------------------------------------------------------------------------
Person: REAL SEP.: 1.03  REL.: .52 ... Item: REAL SEP.: 2.17  REL.: .82

         Person STATISTICS:  MISFIT ORDER

-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL          MODEL|   INFIT  |  OUTFIT  |PT-MEASURE |EXACT MATCH|       |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Person|
|------------------------------------+----------+----------+-----------+-----------+-------|
|    18      6     10     .57     .80|2.67   3.3|5.95   3.5|A-.39   .59| 40.0  76.0| 18PD  |
|    20      7     10    1.24     .85|1.84   1.8|4.02   2.2|B .05   .57| 70.0  79.3| 20LK  |
|    19      4     10    -.68     .80|2.46   2.7|2.76   1.6|C-.10   .57| 40.0  78.1| 19LD  |
|    13      4     10    -.68     .80| .99    .1| .99    .3|D .56   .57| 80.0  78.1| 13PK  |
|     1      6     10     .57     .80| .97    .0| .68   -.2|E .63   .59| 60.0  76.0| 01LK  |
|    11      6     10     .57     .80| .97    .0| .68   -.2|F .63   .59| 60.0  76.0| 11PK  |
|     6      5     10    -.05     .79| .97    .0| .77   -.1|G .62   .59| 80.0  77.3| 06PD  |
|     8      2     10   -2.08     .91| .89   -.1| .53    .1|H .52   .45| 90.0  81.8| 08PK  |
|    14      3     10   -1.33     .83| .83   -.3| .60    .0|I .61   .52| 80.0  78.5| 14LD  |
|     2      6     10     .57     .80| .83   -.4| .61   -.3|J .68   .59| 80.0  76.0| 02PK  |
|    16      1     10   -3.08    1.13| .67   -.3| .26   -.3|j .48   .33| 90.0  89.8| 16LD  |
|     3      5     10    -.05     .79| .66   -.9| .50   -.6|i .75   .59| 80.0  77.3| 03LD  |
|    15      4     10    -.68     .80| .65   -.8| .48   -.4|h .73   .57| 80.0  78.1| 15LK  |
|    12      3     10   -1.33     .83| .59  -1.0| .39   -.2|g .71   .52| 80.0  78.5| 12PD  |
|     7      6     10     .57     .80| .54  -1.4| .39   -.8|f .79   .59|100.0  76.0| 07PK  |
|     9      6     10     .57     .80| .54  -1.4| .39   -.8|e .79   .59|100.0  76.0| 09LK  |
|    17      2     10   -2.08     .91| .51  -1.2| .29   -.2|d .66   .45| 90.0  81.8| 17PD  |
|    10      5     10    -.05     .79| .42  -1.8| .33  -1.0|c .84   .59|100.0  77.3| 10PK  |
|     4      4     10    -.68     .80| .40  -1.8| .31   -.8|b .83   .57|100.0  78.1| 04PK  |
|     5      9     10    3.21    1.23| .36   -.8| .13   -.6|a .67   .45|100.0  91.1| 05PK  |
|------------------------------------+----------+----------+-----------+-----------+-------|
| MEAN     4.7   10.0    -.25     .85| .94   -.2|1.05    .0|           | 80.0  79.0|       |
| S.D.     1.9     .0    1.34     .12| .63   1.3|1.45   1.1|           | 17.9   4.2|       |
-------------------------------------------------------------------------------------------
```

Pada tabel di atas terlihat bawah peserta tes 18PD (paling atas) mempunyai pola respon yang sangat tidak fit dibanding lainnya, demikian juga untuk 20 LK dan 19LD.

Alat analisis lain yang dipergunakan adalah skalogram yang menunjukkan pola respon secara sistematik antara peserta tes (diurutkan dari yang abilitas tinggi ke yang rendah secara vertical, atas ke bawah) dan butir soal (diurutkan dari yang mudah ke yang susah secara horizontal, dari kiri ke kanan).



Analisis di tingkat individu bisa dilakukan dengan *person diagnostic*, seperti di bawah ini:



Terlihat bahwa peserta ujian no 1 dengan nilai logit 0,57 (lambang XXX), dimana dua garis horizontal adalah batas *standard error measurement* abilitasnya. Kotak di sebelah kiri adalah soal yang dijawab dengan benar; dan di sebelah kanan adalah butir soal yang dijawab dengan salah. Terlihat bahwa soal yang tingkat kesulitannya berada di bawah abilitasnya yang ada di sebelah kiri dapat dijawab dengan benar (butir soal 4, 9, 8 dan 5); demikian juga untuk soal yang tingkat kesulitan agak sedikit dia atas abilitasnya (soal no 1 dan 10).

Namun pada sisi sebelah kanan, dua soal yang seharusnya dia jawab ternyata tidak dapat diselesaikan dengan benar, yaitu butir soal 6 dan 7. Hal ini bermakna untuk seperta test no 1 ini, perlu remedial untuk materi soal no 6 dan 7. Sedangkan dua soal sulit yang tidak bisa dia kerjakan yaitu butir soal nomor 2 dan 3, alternatifnya adalah mengajarkannya kembali atau membuat ulasan lebih mendalam supaya bisa memahami lebih baik.

Kesimpulan

Pengujian instrumen dan penentuan abilitas siswa dalam penilaian pendidikan adalah hal yang esensial. Analisis yang bisa menghasilkan pengukuran yang lebih tepat (karena bersifat *equal-interval*) akan menentukan kualitas hasil analisis dan upaya perbaikan proses pendidikan untuk bisa membantu kesulitan belajar siswa. Model Rasch dapat banyak membantu guru, dosen dan peneliti penilaian pendidikan dalam meningkatkan kualitas analisis yang dilakukan, karena prinsip dasar yang tepat dan model pengolahan data yang sesuai untuk analisis hasil ujian khususnya dalam pengolahan data ordinal. Hal ini karena model Rasch sesuai dengan lima persyaratan pengukuran yang objektif.

Aplikasi pemodelan rasch dalam ujian formatif siswa dengan rasch mempunyai banyak kelebihan karena memanfaatkan ketepatan pengukuran. Hal ini bisa untuk deteksi kualitas soal, maupun pada deteksi abilitas individu dan identifikasi bantuan pada kebutuhan belajarnya.

Daftar Pustaka

Alagumalai, S., Curtis, D.D. and Hungi, N. (editors) (2005). *Applied Rasch Measurement: book of exemplars*. papers in honour of John P. Keeves. Dordrecth: Springer.

Bond, T.G., & Fox, C. (2015). *Applying the Rasch Model. Fundamental measurement in the Human Sciences*. 3rd edition. New York: Routledge.

Boone, W. J., Staver, J.R. and Yale, M.S. (2014). *Rasch Analysis in the Human Sciences.* Dordrecht: Springer.

Englehard, G. (2013). *Invariant Measurement, using rasch models in the social, behavioral and health sciences.* New York: Routledge.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*. 3(2), 103-122.

Linacre, J.M. (2011). *A User's guide to WINSTEPS Ministeps*; Rasch-model Computer Program. Program Manual 3.73.

Mok, M. and Wright, B. (2004). Overview of Rasch Model Families. **In** *Introduction to Rasch Measurement: Theory, Models and Applications* (hal 1-24). Minnesota: Jam Press.

Musial, D., Nieminen, G., Thomas, J. dan Burke, K. (2009). *Foundations of Meaningful Education Assessment.* Boston: McGraw-Hill Higher Education.

Olsen, L. W. (2003). Essays on Georg Rasch and his contributions to statistics. Unpublished PhD thesis at Institute Of Economics University of Copenhagen.

Sumintono, B dan Widhiarso, W. (2014). *Aplikasi Model Rasch untuk Penelitian Ilmu-ilmu Sosial (edisi revisi).* Cimahi: Trim Komunikata Publishing House.

Sumintono, B dan Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan.* Cimahi: Trim Komunikata Publishing House.