# Analisis Butir Soal
# dengan Pendekatan Teori Respons Butir

Untuk mendapatkan instrumen berkualitas tinggi, selain dilakukan analisis secara teori (telaah butir berdasarkan aspek isi, konstruksi, dan bahasa) perlu juga dilakukan analisis butir secara empirik. Secara garis besar, analisis butir secara empirik ini dapat dibedakan menjadi dua, yaitu dengan pendekatan teori tes klasik dan teori respons butir (*Item Response Theory, IRT*).

## A. Pendahuluan Teori Tes Klasik *(Classical Test Theory)*

Teori tes klasik atau disebut teori skor murni klasik (Allen & Yen, 1979:57) didasarkan pada suatu model aditif, yakni skor amatan merupakan penjumlahan dari skor sebenarnya dan skor kesalahan pengukuran. Jika dituliskan dengan pernyataan matematis, maka kalimat tersebut menjadi

$$X = T + E \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

dengan :

X : skor amatan,

T : skor sebenarnya,

E : skor kesalahan pengukuran *(error score).*

Kesalahan pengukuran yang dimaksudkan dalam teori ini merupakan kesalahan yang tidak sistematis atau acak. Kesalahan ini merupakan penyimpangan secara teoritis dari skor amatan yang diperoleh dengan skor amatan yang diharapkan. Kesalahan pengukuran yang sistematis dianggap bukan merupakan kesalahan pengukuran.

Ada beberapa asumsi dalam teori tes klasik. Skor kesalahan pengukuran tidak berinteraksi dengan skor sebenarnya, merupakan asumsi yang pertama. Asumsi yang kedua adalah skor kesalahan tidak berkorelasi dengan skor sebenarnya dan skor-skor kesalahan pada tes-tes yang lain untuk peserta tes *(testee)* yang sama. Ketiga, rata-rata dari skor kesalahan ini sama dengan nol.

Asumsi-asumsi pada teori tes klasik ini dijadikan dasar untuk mengembangkan formula-formula dalam menentukan validitas dan reliabilitas tes.

Validitas dan reliabilitas pada perangkat tes digunakan untuk menentukan kualitas tes. Kriteria lain yang dapat digunakan untuk menentukan kualitas tes adalah indeks kesukaran dan daya pembeda.

**1). Reliabilitas**

Mehrens & Lehmann (1973: 102) menyatakan bahwa reliabilitas merupakan derajat keajegan *(consistency)* di antara dua buah hasil pengukuran pada objek yang sama. Definisi ini dapat diilustrasikan dengan seseorang yang diukur tinggi badannya akan diperoleh hasil yang tidak berubah walaupun menggunakan alat pengukur yang berbeda dan skala yang berbeda. Dalam kaitannya dengan dunia pendidikan, prestasi atau kemampuan seorang siswa dikatakan reliabel jika dilakukan pengukuran, hasil pengukuran akan sama informasinya, walaupun penguji berbeda, korektornya berbeda atau butir soal yang berbeda tetapi memiliki karakteristik yang sama.

Allen & Yen (1979: 62) menyatakan bahwa tes dikatakan reliabel jika skor amatan mempunyai korelasi yang tinggi dengan skor yang sebenarnya. Selanjutnya dinyatakan bahwa reliabilitas merupakan koefisien korelasi antara dua skor amatan yang diperoleh dari hasil pengukuran menggunakan tes yang paralel. Dengan demikian, pengertian yang dapat diperoleh dari pernyatan tersebut adalah suatu tes itu reliabel jika hasil pengukuran mendekati keadaan peserta tes yang sebenarnya.

Dalam pendidikan, pengukuran tidak dapat langsung dilakukan pada ciri atau karakter yang akan diukur. Ciri atau karakter ini bersifat abstrak. Hal ini menyebabkan sulitnya memperoleh alat ukur yang stabil untuk mengukur karakteristik seseorang (Mehrens & Lehmann, 1973: 103).

Berdasarkan uraian di atas, maka dalam pembuatan alat ukur dalam dunia pendidikan harus dilakukan secermat mungkin dan disesuaikan dengan kaidah-kaidah yang telah ditentukan oleh ahli-ahli pengukuran di bidang pendidikan. Untuk melihat reliabilitas suatu alat ukur, yang berupa suatu indeks reliabilitas,

dapat dilakukan penelaahan secara statistik. Nlai ini biasa dinamakan dengan koefisien reliabilitas *(reliability coefficient*).

Untuk menentukan nilai reliabilitas suatu tes (butir soal berbentuk pilihan ganda *(multiple choice)*) dapat digunakan formula sebagai berikut .

$$\hat{\sim} = \frac{R}{R-1}\left(1-\frac{\sum \pm_i^2}{\pm_x^2}\right) \text{.....(2)}$$

dengan :

R : banyaknya butir soal,

$\sigma^2$ : varians.

Mehrens & Lehmann (1973: 104) menyatakan bahwa meskipun tidak ada perjanjian secara umum, tetapi secara luas dapat diterima bahwa untuk tes yang digunakan untuk membuat keputusan pada siswa secara perorangan harus memiliki koefisien reliabilitas minimal sebesar 0,85. Dengan demikian, pada penelitian ini, tes seleksi digunakan untuk menentukan keputusan pada siswa secara perorangan, sehingga indeks koefisien reliabilitasnya diharapkan minimal sebesar 0,85.

**2). Validitas**

Validitas suatu perangkat tes dapat diartikan merupakan kemampuan suatu tes untuk mengukur apa yang seharusnya diukur (Allen & Yen, 1979: 97; Syaifudin Azwar, 2000: 45; Kerlinger, 1986). Ada tiga tipe validitas, yaitu validitas isi, validitas konstruk dan validitas kriteria (Allen & Yen, 1979: 97; Syaifudin Azwar, 2000: 45 ; Kerlinger, 1986 : 731).

Ada dua macam validitas isi , yaitu validitas kenampakan dan validitas logika (Syaifudin Azwar, 2000: 45-47). Validitas isi berarti sejauh mana suatu perangkat tes mencerminkan keseluruhan kemampuan yang hendak diukur (Syaifudin Azwar, 2000: 45), yang berupa analisis rasional terhadap domain yang hendak diukur. Validitas kenampakan didasarkan pada pertanyaan apakah suatu butir-butir dalam perangkat tes mengukur aspek yang relevan dengan domainnya. Validitas logika berkaitan dengan keseksamaan batasan pada domain yang hendak

diukur, dan merupakan jawaban apakah keseluruhan butir merupakan sampel representatif dari keseluruhan butir yang mungkin dibuat.

Validitas kriteria, disebut juga validitas prediktif, merupakan kesahihan suatu perangkat tes dalam membuat prediksi, dapat meramalkan keberhasilan siswa pada masa yang akan datang. Validitas prediktif suatu perangkat tes dapat diketahui dari korelasi antara perangkat tes dengan kriteria tertentu yang dikehendaki, yang disebut dengan variabel kriteria (Allen & Yen, 1979 : 97; Syaifudin Azwar, 2000: 51).

**3). Tingkat Kesukaran**

Tingkat kesukaran suatu butir soal, yang disimbolkan dendan $p_i$, merupakan salah satu parameter butir soal yang sangat berguna dalam penganalisian suatu tes. Hal ini disebabkan karena dengan melihat parameter butir ini, akan diketahui seberapa baiknya kualitas suatu butir soal. Jika $p_i$ mendekati 0, maka soal tersebut terlalu sukar, sedangkan jika $p_i$ mendekati 1, maka soal tersebut terlalu mudah, sehingga perlu dibuang. Hal ini disebabkan karena butir tersebut tidak dapat membedakan kemampuan seorang siswa dengan siswa lainnya.

Allen dan Yen (1979 : 122) menyatakan bahwa secara umum indeks kesukaran suatu butir sebaiknya terletak pada interval 0,3 – 0,7. Pada interval ini, informasi tentang kemampuan siswa akan diperoleh secara maksimal. Dalam merancang indeks kesukaran suatu perangkat tes, perlu dipertimbangkan tujuan penyusunan perangkat tes tersebut. Untuk menentukan indeks kesukaran dari suatu butir pada perangkat tes pilihan ganda, digunakan persamaan sebagai berikut :

$$p_i = \frac{\sum B}{N} \quad \text{.....(3)}$$

dengan :

p = proporsi menjawab benar pada butir soal tertentu.

ΣB = banyaknya peserta tes yang menjawab benar.

N = jumlah peserta tes yang menjawab.

**4). Daya Pembeda**

Untuk menentukan daya pembeda, dapat digunakan indeks diskriminasi, indeks korelasi biserial, indeks korelasi *point biserial*, dan indeks keselarasan. Pada analisis butir dalam penelitian ini, hanya digunakan indeks korelasi *point biserial*. Koefisien korelasinya untuk suatu butir tes ditentukan dengan rumus:

$$\mathbf{r_{pbis}} = \left[\frac{\overline{X}_1 - \overline{X}}{s_X}\right]\sqrt{\frac{p_1}{1-p_1}} \quad \text{.....(4)}$$

dengan $r_{pbis}$ = koefisien korelasi point biserial, $X_i$ merupakan variabel kontinu, $\overline{X}_1$ merupakan rerata skor $X$ untuk peserta tes yang menjawab benar butir tersebut, $\overline{X}$ merupakan rerata skor $X$, $s_X$ merupakan standar deviasi dari skor $X$, dan $p_1$ merupakan proporsi peserta tes yang menjawab benar butir tersebut.

Pada suatu butir soal, indeks daya beda dikatakan baik jika lebih besar atau sama dengan 0,3. Indeks daya pembeda suatu butir yang kecil nilainya akan menyebabkan butir tersebut tidak dapat membedakan siswa yang kemampuannya tinggi dan siswa yang kemampuannya rendah. Pada analisis tes dengan *Content-Referenced Measures*, indeks daya pembeda butir tidak terlalu perlu menjadi perhatian, asalkan tidak negatif (Ebel & Frisbie, 1986; Frisbie, 2005). Jika nilainya kecil, menunjukkan bahwa kemencengan distribusi skor dari populasi, yang juga mengakibatkan validitas tes menjadi rendah.

**5). Kesalahan Pengukuran**

Kesalahan Baku Pengukuran *(Standard Error of Measurement, SEM)* dapat digunakan untuk mamahami kesalahan yang bersifat acak/random yang mempengaruhi skor peserta tes dalam pelaksanaan tes. Kesalahan pengukuran, yang disimbulkan dengan $\sigma_E$ , dapat dihitung dengan rumus pada persaman 5, yang diturunkan dari rumus reliabilitas (Allen & Yen, 1979 : 73).

$$\sigma_E = \sigma_x \sqrt{1 - \rho_{xx'}} \quad \text{.....(5)}$$

dengan $\sigma_x$ merupakan simpangan baku dari skor total dan $\rho_{xx'}$ merupakan koefisien reliabilitas.

Teori tes klasik memiliki beberapa kelemahan mendasar. Kebanyakan statistik yang digunakan dalam model tes klasik seperti tingkat kesukaran dan daya pembeda soal sangat tergantung pada sampel yang dipergunakan dalam analisis. Rerata tingkat kemampuan, rentang, dan sebaran kemampuan siswa yang dijadikan sampel dalam analisis sangat mempengaruhi nilai statistik yang diperoleh. Sebagai contoh, tingkat kesukaran soal akan tinggi apabila sampel yang akan digunakan mempunyai kemampuan lebih tinggi dari rerata kemampuan siswa dalam poulasinya. Daya pemeda soal akan tinggi apabila tingkat kemampuan sampel bervariasi atau mempunyai rentang kemampuan yang besar. Demikian pula dengan reliabilitas tes.

Kelemahan kedua yakni skor siswa yang diperoleh dari suatu tes sangat terbatas pada tes yang digunakan. Kesimpulan hasil tes tidak dapat digeneralisasikan di luar tes yang digunakan. Skor perolehan seseorang sangat tergantung pada pemilihan tes yang digunakan bukan pada kemampuan peserta tes tersebut. Karena keterbatasan penggunaan skor tes, teori tes klasikal tidak mempunyai dasar untuk mempelajari perkembangan kemampuan siswa dari waktu ke waktu, kecuali jika siswa tersebut menempuh tes yang sama dari waktu ke waktu.

Ketiga, konsep keajegan/reliabilitas tes dalam konteks teori tes klasik didasarkan pada kesejajaran perangkat tes sangat sukar untuk dipenuhi. pada praktiknya, sulit sekali memperoleh dua perangkat tes yang benar-benar sejajar. Jika prosedur tes retes digunakan, sampel yang diambil sangat tidak mungkin berperilaku sama pada saat tes dikerjakan untuk yang kedua kalinya.

Keempat, teori tes klasik tidak memberikan landasan untuk menentukan bagaimana respons seseorang peserta tes apabila diberikan butir tertentu. Tidak adanya informasi ini tidak memungkinkan melakukan desain tes yang bervariasi sesuai dengan kemampuan peserta tes (*adaptive or tailored testing*).

Kelima, indeks kesalahan baku pengukuran dipraasumsikan sama untuk setiap peserta tes. Padahal seseorang peserta tes mungkin berperilaku lebih konsisten dalam menjawab soal dibandingkan peserta tes lainnya. Demikian pula sebaliknya, banyak sekali kesalahan individual. Kesalahan pengukuran sebenarnya merupakan perilaku peserta tes yang bersifat perorangan dan bukan perilaku tes.

Terakhir, prosedur-prosedur yang berkaitan dengan teori tes klasik seperti pengujian bias butir soal dan penyetaraan tes tidak bersifat praktis dan sukar untuk dilakukan. Demikian pula halnya dengan penyetaraan yang sifatnya vertikal. Untuk mengatasi hal itu, digunakanlah pendekatan teori lain yang disebut denga teori respons butir.

## B. Teori Respons Butir

Dalam evaluasi yang dilaksanakan dalam pendidikan, siswa menjawab butir soal suatu tes yang berbentuk pilihan ganda dengan benar, biasanya diberi skor 1 dan 0 jika menjawab salah. Pada penyekoran dengan pendekatan teori tes klasik, kemampuan siswa dinyatakan dengan skor total yang diperolehnya. Prosedur ini kurang memperhatikan interaksi antara setiap orang siswa dengan butir.

Pendekatan teori respons butir merupakan pendekatan alternatif yang dapat digunakan dalam menganalisis suatu tes. Ada dua prinsip yang digunakan pada pendekatan ini, yakni prinsip relativitas dan prinsip probabilitas. Pada prinsip relativitas, unit dasar dari pengukuran bukanlah siswa atau butir, tetapi lebih kepada kemampuan siswa relatif terhadap butir. Jika $\beta_n$ merupakan indeks dari kemampuan siswa ke n pada trait yang diukur, dan $\delta_i$ merupakan indeks dari tingkat kesulitan dari butir ke-i relative yang terkait dengan kemampuan yang diukur, maka bukan $\beta_n$ atau $\delta_i$ yang merupakan unit pengukuran, tetapi lebih kepada perbedaan antara kemampuan dan dari siswa relatif terhadap tingkat kesulitan butir atau $(\beta_n - \delta_i)$ perlu dipertimbangkan. Sebagai alternatifnya perbandingan antara kemampuan terhadap tingkat kesulitan dapat digunakan. Jika kemampuan dari siswa melampaui tingkat kesulitan butir, maka respons siswa diharapkan benar, dan jika kemampuan siswa kurang dari tingkat kesulitan butir, maka respons siswa diharapkan salah (Keeves dan Alagumalai, 1999:24).

Pada teori respons butir, prinsip probabilitas menjadi perhatian. Misalkan kemampuan siswa ke n dinyatakan dengan $\theta_n$ dan tingkat kesulitan dari butir dinyatakan dengan $\Delta_i$ maka sesuai dengan prinsip relativitas, jika $\theta_n > \Delta_i$ siswa diharapkan menjawab dengan benar, dan $\theta_n < \Delta_i$ siswa diharapkan menjawab salah. Probabilitas respons menjawab benar berada pada rentang 0 sampai dengan 1.0 dan hal ini menghalangi data dinyatakan sebagai skala interval. Skor mentah yang dihasilkan dari cara ini sulit dinyatakan sebagai skala. Untuk mengatasi permasalahan ini, dapat digunakan transformasi logistik, sehingga hubungan antara tingkat kesulitan butir dan peluang menjawab benar bukan hubungan linear.

Dalam teori respons butir, model matematisnya mempunyai makna bahwa probabilitas subjek untuk menjawab butir dengan benar tergantung pada kemampuan subjek dan karakteristik butir. Ini berarti bahwa peserta tes dengan kemampuan tinggi akan mempunyai probabilitas menjawab benar lebih besar jika dibandingkan dengan peserta yang mempunyai kemampuan rendah. Hambleton & Swaminathan (1985: 16) dan Hambleton, Swaminathan, & Rogers (1991: 9) menyatakan bahwa ada tiga asumsi yang mendasari teori respon butir, yaitu unidimensi, independensi lokal dan invariansi parameter. Ketiga asumsi dapat dijelaskan sebagai berikut.

Unidimensi, artinya setiap butir tes hanya mengukur satu kemampuan. Contohnya, pada tes prestasi belajar bidang studi matematika, butir-butir yang termuat di dalamnya hanya mengukur kemampuan siswa dalam bidang studi matematika saja, bukan bidang yang lainnya. Pada praktiknya, asumsi unidimensi tidak dapat dipenuhi secara ketat karena adanya faktor-faktor kognitif, kepribadian dan faktor-faktor pelaksanaan tes, seperti kecemasan, motivasi, dan tendensi untuk menebak. Oleh karena itu, asumsi unidimensi dapat ditunjukkan hanya jika tes mengandung satu saja komponen dominan yang mengukur prestasi subjek.

Pada teori respons butir, hubungan antara kemampuan peserta dan skor tes yang dicapai dinyatakan dengan kurva yang tidak linear. Pada Gambar 2 disajikan ilustrasi suatu distribusi kondisional di suatu bagian level kemampuan pada subpopulasi peserta tes. Di sepanjang garis regresi, terdapat sebaran skor tes. Variabilitas kesalahan pengukuran skor tes mungkin terjadi. Jika distribusi

bervariasi lintas beberapa subpopulasi, maka tes tidak hanya mengukur kemampuan tunggal saja (Hambleton & Swaminathan, 1985).

Jika faktor-faktor yang mempengaruhi prestasi konstan, maka respons subjek terhadap pasangan butir yang manapun akan independen secara statistik satu sama lain. Kondisi ini disebut dengan independensi lokal. Asumsi independensi lokal ini akan terpenuhi apabila jawaban peserta terhadap suatu butir soal tidak mempengaruhi jawaban peserta terhadap terhadap butir soal yang lain. Tes untuk memenuhi asumsi independensi lokal dapat dilakukan dengan membuktikan bahwa peluang dari pola jawaban setiap peserta tes sama dengan hasil kali peluang jawaban peserta tes pada setiap butir soal.

Menurut Hambleton, Swaminathan, & Rogers (1991: 10), independensi lokal secara matematis dinyatakan sebagai:

$$P(u_1,u_2, \ldots,u_n|\theta) = P(u_1|\theta).P(u_2|\theta)\ldots.P(u_n|\theta)$$

$$= \prod_{i=1}^{n} P(u_i|\theta) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

Keterangan :

i : 1, 2, 3, …n

n : banyaknya butir tes

$P(u_i|\theta)$ : probabilitas peserta tes yang memiliki kemampuan $\theta$ dapat menjawab butir ke-i dengan benar.

$P(u_1,u_2, \ldots,u_n|\theta)$ : probabilitas peserta tes yang memiliki kemampuan $\theta$ dapat menjawab butir ke-1 sampai ke-n dengan benar

Invariansi parameter artinya karakteristik butir soal tidak tergantung pada distribusi parameter kemampuan peserta tes dan parameter yang menjadi ciri peserta tes tidak bergantung dari ciri butir soal. Kemampuan seseorang tidak akan berubah hanya karena mengerjakan tes yang berbeda tingkat kesulitannya dan parameter butir tes tidak akan berubah hanya karena diujikan pada kelompok peserta tes yang berbeda tingkat kemampuannya.

Menurut Hambleton, Swaminathan, & Rogers (1991: 18), invariansi parameter kemampuan dapat diselidiki dengan mengajukan dua perangkat tes atau lebih yang memiliki tingkat kesukaran yang berbeda pada sekelompok peserta tes.

Invariansi parameter kemampuan akan terbukti jika hasil estimasi kemampuan peserta tes tidak berbeda walaupun tes yang dikerjakan berbeda tingkat kesulitannya. Invariansi parameter butir dapat diselidiki dengan mengujikan tes pada kelompok peserta yang berbeda. Invariansi parameter butir terbukti jika hasil estimasi parameter butir tidak berbeda walaupun diujikan pada kelompok peserta yang berbeda tingkat kemampuannya.

Dalam teori respons butir, selain asumsi-asumsi yang telah diuraikan sebelumnya, hal penting yang perlu diperhatikan adalah pemilihan model yang tepat. Pemilihan model yang tepat akan mengungkap keadaan yang sesungguhnya dari data tes sebagai hasil pengukuran. Ada 3 model hubungan antara kemampuan dengan parameter batir, yaitu model 1 parameter (model Rasch), model 2 parameter, dan model 3 parameter.

Model Rasch dituliskan sebagai berikut :

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$ , dengan i : 1,2,3, …,n ……. (7)

$P_i(\theta)$ : probabilitas peserta tes yang memiliki kemampuan θ dipilih secara acak dapat menjawab butir i dengan benar

θ : tingkat kemampuan subyek (sebagai variabel bebas)

$b_i$ : indeks kesukaran butir ke-i

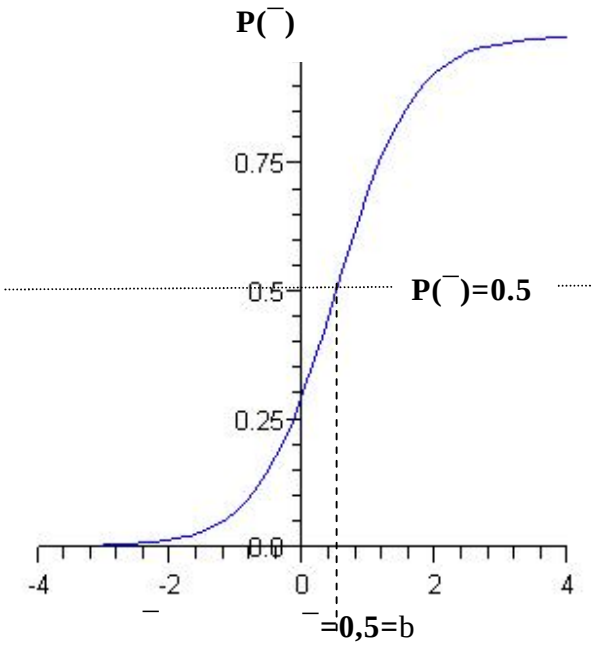
e : bilangan natural yang nilainya mendekati 2,718

n : banyaknya butir dalam tes

Parameter $b_i$ merupakan suatu titik pada skala kemampuan agar peluang menjawab benar sebesar 50%. Misalkan suatu butir tes mempunyai parameter $b_i$ = 0,3, artinya diperlukan kemampuan minimal 0,3 pada skala untuk dapat menjawab benar dengan peluang 50%. Semakin besar nilai parameter $b_i$ , maka semakin besar kemampuan yang diperlukan untuk menjawab benar dengan peluang 50%. Dengan kata lain, semakin besar nilai parameter $b_i$, maka makin sulit butir soal tersebut.

Hubungan peluang menjawab benar $P_i$ (θ) dengan tingkat kemampuan peserta (θ) dapat digambarkan sebagai kurva karakteristik butir (*item characteristic curve, ICC*). Gambar 1 berikut merupakan ilustrasi kurva karakteristik butir untuk model Rasch (1 parameter, 1P), dengan butir 1 (b=-0,5), butir 2 (b=0) dan butir 3(b=0,5).

Gambar 1 berikut merupakan ilustrasi kurva karakteristik butir untuk model Rasch dengan tingkat kesulitan b=0,5.




Gambar 1
Kurva Karakteristik Butir untuk Model 1P, dengan b=0,5

Pada model logistik dua parameter, probabilitas peserta tes untuk dapat menjawab benar suatu butir soal ditentukan oleh dua karakteristik butir, yaitu indeks kesukaran butir ($b_i$) dan indeks daya beda butir ($a_i$). Parameter $a_i$ merupakan indeks daya pembeda yang dimiliki butir ke-i. Pada kurva karakteristik, $a_i$ proporsional terhadap koefisien arah garis singgung *(slope)* pada titik θ = b. Butir soal yang memiliki daya pembeda yang besar mempunyai kurva

yang sangat menanjak, sedangkan butir soal yang mempunyai daya pembeda kecil mempunyai kurva yang sangat landai. Secara teoretis, nilai $a_i$ ini terletak antara $-\infty$ dan $+\infty$. Pada pada butir yang baik nilai ini mempunyai hubungan positif dengan performen pada butir dengan kemampuan yang diukur, dan $a_i$ terletak antara 0 dan 2 (Hambleton & Swaminathan, 1985: 37 ).

Menurut Hambleton, Swaminathan, & Rogers (1991: 15), secara matematis model logistik dua parameter dapat dituliskan sebagai berikut.

$$P_i(\bar{\ }) = \frac{e^{Da_i(\bar{\ }-b_i)}}{1+e^{Da_i(\bar{\ }-b_i)}} \quad \text{dengan i : 1,2,3, …,n} \quad \ldots\ldots\ldots\ldots\ldots\ldots (8)$$

Keterangan :

$\bar{\ }$ : tingkat kemampuan peserta tes

$P_i(\bar{\ })$ : probabilitas peserta tes yang memiliki kemampuan θ dapat menjawab butir i dengan benar

$a_i$ : indeks daya pembeda

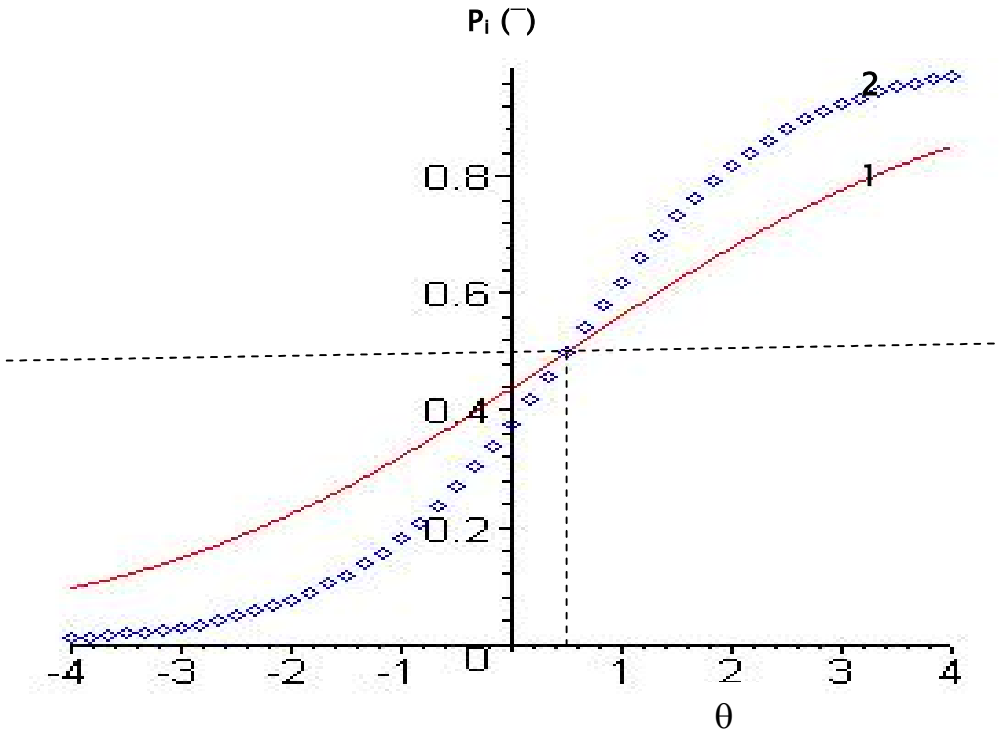
$b_i$ : indeks kesukaran butir ke-i

$e$ : bilangan natural yang nilainya mendekati 2,718

n : banyaknya butir dalam tes

$D$ : faktor penskalaan yang harganya 1,7.

Pada gambar 2 disajikan kurva karakteristik butir 1 (a=0,5; b=0,5) dan butir 2 (a=1; b=0,5). Berdasarkan gambar tersebut, jika indeks daya pembeda butir 1 lebih rendah dibandingkan butir 2, maka akan nampak bahwa kurva karakteristik butir 1 lebih landai dibandingkan butir 2.




Gambar 2
Kurva Karakteristik Butir Model 2P, dengan Butir 1 (a=0,5; b=0,5)
dan Butir 2 (a=1; b=0,5)

Sesuai dengan namanya, model logistik tiga parameter ditentukan oleh tiga karakteristik butir yaitu indeks kesukaran butir soal, indeks daya beda butir, dan indeks tebakan semu (*pseudoguessing*). Dengan adanya indeks tebakan semu pada model logistik tiga parameter, memungkinkan subjek yang memiliki kemampuan rendah mempunyai peluang untuk menjawab butir soal dengan benar. Secara matematis, model logistik tiga parameter dapat dinyatakan sebagai berikut (Hambleton, & Swaminathan, 1985 : 49; Hambleton, Swaminathan, & Rogers, 1991: 17).

$$P_i(\bar{\ }) = c_i + (1-c_i)\ \frac{e^{Da_i(\bar{\ }-b_i)}}{1+e^{Da_i(\bar{\ }-b_i)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (9)$$

Keterangan :

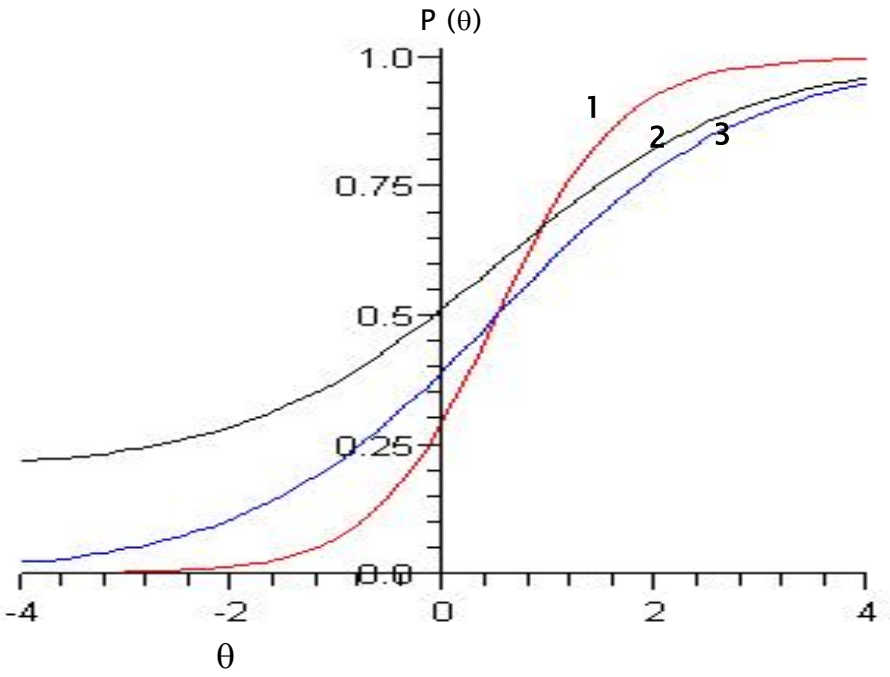
¯ : tingkat kemampuan peserta tes

$P_i(\bar{\ })$ : probabilitas peserta tes yang memiliki kemampuan θ dapat menjawab butir i dengan benar

$a_i$ : indeks daya pembeda
$b_i$ : indeks kesukaran butir ke-i
$c_i$ : indeks tebakan semu butir ke-i
$e$ : bilangan natural yang nilainya mendekati 2,718
n : banyaknya butir dalam tes
$D$ : faktor penskalaan yang harganya 1,7.

Nilai kemampuan peserta (θ) terletak di antara –3 dan +3, sesuai dengan daerah asal distribusi normal. Pernyataan ini merupakan asumsi yang mendasari besar nilai $b_i$. Secara teoretis, nilai $b_i$ terletak di $-\infty$ dan $+\infty$. Suatu butir dikatakan baik jika nilai ini berkisar antara –2 dan +2 (Hambleton & Swaminathan, 1985: 107). Jika nilai $b_i$ mendekati –2, maka indeks kesukaran butir sangat rendah, sedangkan jika nilai $b_i$ mendekati +2 maka indeks kesukaran butir sangat tinggi untuk suatu kelompok peserta tes.

Peluang menjawab benar pada saat kemampuan peserta tes sangat rendah dilambangkan dengan $c_i$, yang disebut dengan tebakan semu *(pseudoguessing)*. Parameter ini merupakan suatu kemungkinan asimtot bawah yang tidak nol *(nonzero lower asymtote)* pada kurva karakteristik butir (ICC). Parameter ini menggambarkan probabilitas peserta dengan kemampuan rendah menjawab dengan benar pada suatu butir yang mempunyai indeks kesukaran yang tidak sesuai dengan kemampuan peserta tersebut. Besarnya harga $c_i$ diasumsikan lebih kecil daripada nilai yang akan dihasilkan jika peserta tes menebak secara acak jawaban pada suatu butir. Gambar 3 menyajikan kurva karakteristik butir 1 (a=1, b=0,5, c=0), butir 2(a=0,5, b=0,5, c=0) dan butir 3 (a=0,5, b=0,5, c=0,2).




Gambar 3
Kurva Karakteristik Butir Model 3P, dengan Butir 1 (a=1, b=0,5, c=0), Butir 2 (a=0,5, b=0,5, c=0) dan Butir 3 (a=0,5, b=0,5, c=0,2)

Mencermati gambar tersebut, nampak bahwa pada skala kemampuan peserta tes yang sangat rendah (θ = -4), peluang menjawab benar butir 3 sebesar 0,2, sedangkan pada butir 1 dan butir 2 mendekati 0.

Pada suatu butir tes, nilai $c_i$ ini berkisar antara 0 dan 1. Suatu butir dikatakan baik jika nilai $c_i$ tidak lebih dari 1/k, dengan k banyaknya pilihan (Hullin, 1983: 36). Jadi misalkan pada suatu perangkat tes pilihan ganda dengan 4 pilihan untuk setiap butir tesnya, butir ini dikatakan baik jika nilai $c_i$ tidak lebih dari 0,25.

Fungsi informasi butir *(Item Information Functions)* merupakan suatu metode untuk menjelaskan kekuatan suatu butir pada perangkat tes, pemilihan butir tes, dan pembandingan beberapa perangkat tes. Fungsi informasi butir menyatakan kekuatan atau sumbangan butir tes dalam mengungkap latent trait yang diukur dengan tes tersebut. Dengan fungsi informasi butir diketahui butir

yang mana yang cocok dengan model sehingga membantu dalam seleksi butir tes. Secara matematis, fungsi informasi butir memenuhi persaman sebagai berikut.

$$I_i(\theta) = \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)Q_i(\theta)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (10)$$

keterangan :

i : 1,2,3,…,n

$I_i(\theta)$ : fungsi informasi butir ke-i

$P_i(\theta)$ : peluang peserta dengan kemampuan θ menjawab benar butir i

$P'_i(\theta)$ : turunan fungsi $P_i(\theta)$ terhadap θ

$Q_i(\theta)$ : peluang peserta dengan kemampuan θ menjawab benar butir i

Fungsi informasi tes merupakan jumlah dari fungsi informasi butir penyusun tes tersebut (Hambleton dan Swaminathan, 1985: 94). Berhubungan dengan hal ini, fungsi informasi perangkat tes akan tinggi jika butir tes mempunyai fungsi informasi yang tinggi pula. Fungsi informasi perangkat tes secara matematis dapat dituliskan sebagai berikut.

$$I_i(\theta) = \sum_{i=1}^{n} I_i(\theta) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (10)$$

Nilai-nilai indeks parameter butir dan kemampuan peserta merupakan hasil estimasi. Karena merupakan hasil estimasi, maka kebenarannya bersifat probabilitas dan tidak terlepaskan dengan kesalahan pengukuran. Dalam teori respon butir, kesalahan baku pengukuran *(Standard Error of Measurement, SEM)* berkaitan erat dengan fungsi informasi. Fungsi informasi dengan *SEM* mempunyai hubungan yang berbanding terbalik kuadratik, semakin besar fungsi informasi maka *SEM* semakin kecil atau sebaliknya (Hambleton, Swaminathan dan Rogers, 1991, 94). Jika nilai fungsi informasi dinyatakan dengan $I_i(\theta)$ dan nilai estimasi SEM dinyatakan dengan *SEM* ( $\hat{\theta}$ ), maka hubungan keduanya, menurut Hambleton, Swaminathan, dan Rogers (1991 : 94) dinyatakan dengan

$$\text{SEM}\ (\hat{\bar{\ }}) = \frac{1}{\sqrt{I(\bar{\ })}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (11)$$

Pada model Rasch untuk menganalisis jawaban siswa, yang perlu menjadi perhatian adalah pengestimasian parameter butir dan parameter kemampuan peserta. Dalam pengestimasian ini, dikenal fungsi likelihood. Fungsi likelihood untuk kasus dengan N siswa dan n butir dapat dinyatakan dengan

$$L(\theta, b;u) = \prod_i \prod_j P_i(\bar{\ }_j;b_i)^{u_{ij}} \left[P_i(\bar{\ }_j;b_i)\right]^{1-u_{ij}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (12)$$

Selanjutnya diestimasi nilai-nilai yang memaksimumkan fungsi ini. Prosedur yang dapat dipilih yakni prosedur likelihood maksimum gabungan (*joint maximum likelihood, JML*) atau prosedur likehood maksimum marginal (*marginal maximum likelihood, MML*) atau juga dengan pendekatan Bayes.

Untuk mengestimasi parameter-parameter butir pada model logistik Rasch, ada beberapa perangkat lunak yang dapat digunakan, diantaranya Bigsteps, Rascal, Ascal, Bilog, Xalibrate dan Multilog. Keluaran (*output*) dari program-program ini juga menyediakan hasil pengestimasian parameter peserta tes.

Langkah selanjutnya adalah mengetahui kecocokan model dari data yang dianalisis. Uji statistik untuk kecocokan model salah satunya uji perbandingan likelihood (*likelihood ratio test*). Uji ini digunakan untuk mengecek apakah estimasi parameter butir dalam grup skor yang berbeda bernilai sama pada kesalahan penyampelan dari estimasi. Secara teoritis, responden yang berukuran N dapat dibuat menjadi interval-interval pada skala kontinum untuk θ, yang merupakan dasar untuk mengestimasi nilai θ. Statistik Khi-kuadrat dari perbandingan *likelihood* digunakan untuk membandingkan frekuensi menjawab benar dan tidak benar dari respons pada interval yang diharapkan dari model yang cocok pada rata-rata interval $\hat{\bar{\ }}_h$, dengan persamaan :

$$G_j^2 = 2\sum_{h=1}^{n_g}\left[ r_{hj} \ln \frac{r_{hj}}{N_h P_j(\bar{\ }_h)} + (N_A - r_{hj}) \ln \frac{N_h - r_{hj}}{N_h[1 - P(\bar{\ }_h)]} \right] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (13)$$

dengan $n_g$ merupakan banyaknya interval, $r_{hj}$ merupakan frekuensi respons yang benar untuk butir pada interval h, $N_h$ merupakan banyaknya anggota sampel yang berada dalam interval, dan $P_j(\bar{\theta}_h)$ merupakan nilai dari fungsi respons sesuai model untuk butir j pada $\bar{\theta}_h$, yang merupakan kemampuan rata-rata responden pada interval *h* (Mislevy dan Bock, 1990).

Pada program Bilog, untuk menentukan banyaknya interval yang dibuat pada skala kontinu untuk θ, mula-mula dibuat maksimum 20 interval. Setiap responden disarangkan pada interval tersebut termasuk estimasi EAP (*expected a posteriori),* berdasarkan tipe prior yang dispesifikasikan pemakai dari skor yang diperoleh responden. Pada setiap butir tes, probablilitas harapan dari respons yang sesuai dengan estimasi rata-rata EAP untuk kemampuan dari kasus yang berada dalam interval digunakan sebagai proporsi harapan untuk interval tersebut.

Khi-kuadrat perbandingan kemungkinan dihitung setelah mengkombinasikan interval-interval yang ekstrim, hingga frekuensi harapan pada gabungan interval-interval tersebut lebih dari lima. Derajat kebebasan dari khi-kuadrat perbandingan kemungkinan sama dengan banyaknya interval-interval yang telah dikombinasikan (Mislevy dan Bock, 1990).

**Referensi :**


Allen, M. J. & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole Publishing Company.

Anastasi, A. & Urbina,S. (1997). *Psychological testing*. Upper Saddle River, NJ : Prentice Hall.

Ebel, R.L. & Frisbie, D.A. (1986). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Gronlund, N.E. (1976). *Measurement and evaluation in teaching.* New York : Macmillan Publishing Co.

Hambleton, R.K., Swaminathan, H & Rogers, H.J. (1991). *Fundamental of item response theory*. Newbury Park, CA : Sage Publication Inc.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory*. Boston, MA : Kluwer Inc.

Heri Retnawati. (2003). Keberfungsian butir diferensial pada perangkat tes seleksi masuk SMP. *Tesis*. Universitas Negeri Yogyakarta, tidak dipublikasikan.

Hosmer, D.W. dan Lemeshow,S. (1989). Applied Logistic Regressions. New York : John Willwy and Sons.

Hullin, C. L., et al. (1983). *Item response theory : Application to psichologycal measurement.* Homewood, IL : Dow Jones-Irwin.

Kerlinger, F.N. (1986). *Asas-asas penelitian behavioral* (Terjemahan L.R. Simatupang). Yogyakarta: Gajahmada University Press.

Keeves, J.P. dan Alagumalai,S. (1999). New appoaches to measurement. Dalam Masters, G.N. dan Keeves, J.P.(Eds). Advances in measurement in educational research and assesment. Amsterdam : Pergamon.

Mehrens, W.A. & Lehmann, I.J. (1973). *Measurement and evaluation in education and psychology*. New York : Hold, Rinehart and Wiston,Inc.

Mislevy, R.J. & Bock,R.D. (1990). *BILOG 3 : Item analysis & test scoring with binary logistic models*. Moorseville : Scientific Sofware Inc.

Van der Linden, W.J. dan Hambleton, R.K. (1997). Item response theory:brief history, common models and extentions. Dalam Van der Linden, W.J. dan Hambleton, R.K. (Eds). Handbook of item response theory. New York : Springer.

Syaifudin Azwar. (2000). *Reliabilitas dan validitas* (Edisi 4). Yogyakarta: Pustaka Pelajar.