KARAKTERISTIK INSTRUMEN

Oleh:
Samsul Hadi

A. Pendahuluan

Tes atau ujian digunakan untuk mengetahui pencapaian standar kompetensi lulusan atau turunannya berupa standar kompetensi, kompetensi dasar, atau indikator dari suatu mata pelajaran yang harus dikuasai oleh siswa. Supaya tes atau ujian dapat mengukur pencapaan kompetensi, maka soal tes atau ujian harus dibuat memenuhi validitas isi, yaitu butir-butir soal yang ada harus benar-benar diturunkan dari standar kompetensi lulusan, standar kompetensi, kompetensi dasar, atau indikator dari suatu mata pelajaran.

Supaya soal memenuhi validitas isi penyusunan butir soal diawali dengan mengkaji standar kompetensi lulusan, standar kompetensi, kompetensi dasar, atau indikator dari suatu mata pelajaran. Berdasarkan kisi-kisi tersebut kemudian dibuat butir-butir soal. Butir-butir soal yang sudah jadi kemudian diminta untuk ditelaah oleh pihak lain yang dianggap mampu dengan memperhatikan: materi, konstruksi, dan bahasa yang digunakan dalam butir soal.

Telaah materi bertujuan untuk melihat kesesuaian soal sudah sesuai dengan indikator, kesesuaian materi soal dengan tuntutan kompetensi (urgensi, relevansi, kontinyuitas, keterpakaian yang tinggi), dan kesesuaian materi dengan jenjang jenis sekolah atau tingkat kelas. Telaah konstruksi bertujuan untuk menjamin bahwa soal telah dirumuskan dengan singkat, bebas dari pernyaatn yang tidak relevan, bebas dari pernyataan negatif ganda, dan bebas dari pernyataan yang multi interpretasi. Telaah bahasa dilakukan agar soal komunikatif dan sesuai dengan jenjang pendidikan siswa serta menggunakan bahasa Indonesia yang baku.

Hasil telaah materi, konstruksi, dan bahasa dijadikan masukan perbaikan terhadap butir soal yang ada. Setelah perbaikan dilakukan berdasarkan telaah ketiga hal tersebut, soal siap diujicobakan kepada siswa. Data yang diperoleh dari hasil ujicoba perlu dianalisis untuk mengetahui karakteristik soal atau butir soal

secara empiris. Ada dua pendekatan untuk menganalisis data hasil ujicoba soal, yaitu menggunakan teori tes klasik dan menggunakan teori respons butir.

B. Teori Tes Klasik

Kualitas tes atau butir soal penyusun tes yang baik dapat dilihat dari karakteristikya. Karakteristik tes atau butir dapat diketahui dengan dua pendekatan teori. Kedua pendekatan tersebut yakni teori tes klasik dan teori respons butir. Teori tes klasik, atau disebut juga teori tes skor murni klasik, didasarkan pada model aditif, yaitu skor amatan merupakan penjumlahan dari skor sebenarnya dan skor kesalahan pengukuran (Allen & Yen, 1979: 57). Secara matematis pernyataan tersebut dapat dirumuskan sebagai berikut.

$$X = T + E$$

dengan keterangan X = skor amatan, T = skor murni, dan E = skor kesalahan pengukuran *(error score).*

Kesalahan pengukuran dalam teori tes klasik merupakan kesalahan yang tidak sistematis atau acak. Kesalahan pengukuran merupakan penyimpangan secara teoretis dari skor amatan yang diperoleh dengan skor amatan yang diharapkan. Kesalahan pengukuran yang sistematis dianggap bukan merupakan kesalahan pengukuran.

Asumsi-asumsi yang mendasari teori tes klasik tersebut dijadikan dasar untuk mengembangkan rumus-rumus matematis untuk mengestimasi validitas dan koefisien reliabilitas tes. Validitas dan koefisien reliabilitas pada perangkat tes digunakan untuk menilai kualitas tes. Kualitas tes dalam teori tes klasik juga dapat ditentukan dengan indeks kesukaran dan daya pembeda.

Tingkat kesukaran, disimbolkan dengan p, merupakan salah satu parameter butir soal yang sangat berguna dalam analisis soal. Tingkat kesukaran dapat dihitung dengan berbagai cara, yaitu (a) skala kesukaran linear, (b) skala bivariat, (c) indeks Davis, dan (d) proporsi menjawab benar (Bahrul Hayat, dkk., 1999). Secara matematis tingkat kesukaran yang dihitung dengan proporsi menjawab benar dirumuskan dengan:

$$p = \frac{\sum B}{N}$$

dengan keterangan B adalah banyak peserta tes yang menjawa benar, dan N jumlah peserta tes yang menjawab. Dengan rumus tersebut, maka dapat diketahui bahwa jika p mendekati 0, maka soal tersebut terlalu sukar, sedang jika p mendekati 1 maka soal tersebut terlalu mudah. Soal yang terlalu mudah atau terlalu sukar tidak dapat membedakan kemampuan peserta tes sehingga perlu dibuang.

Menurut Allen dan Yen (1979) tingkat kesukaran butir soal sebaiknya antara 0,3 – 0,7. Pada rentang tersebut informasi tentang kemampuan siswa akan diperoleh secara maksimal. Namun angka tersebut perlu disesuaikan dengan tujuan pengembangan soal. Soal untuk keperluan seleksi, remidi, atau ulangan umum seharusnya mempunyai p yang berbeda-beda untuk mencapai tujuan yang maksimal.

Daya beda merupakan parameter butir soal yang memberikan informasi tentang seberapa besar butir soal tersebut dapat membedakan peserta tes yang skornya tinggi dan peserta tes yang skornya rendah. Daya beda dapat dihitung dengan beberapa cara antara lain dengan menghitung koefisien korelasi *point biserial* dan koefisien korelasi *biserial*. Korelasi *point biserial* secara matematis dirumuskan sebagai berikut.

$$r_{pbis} = \frac{M_p - M_q}{S_t}\sqrt{pq}$$

dimana:

r_{pbis} : koefisien korelasi *point biserial*

M_p : *mean* skor pada tes dari peserta tes yang memiliki jawaban benar pada butir soal

M_q : *mean* skor pada tes dari peserta tes yang memiliki jawaban salah pada butir soal

p : proporsi peserta tes yang menjawab benar pada butir soal

q : $1 - p$

S_t : standar deviasi seluruh skor tes

Korelasi *biserial* secara matematis dinyatakan dengan rumus sebagai berikut.

$$r_{bis} = \frac{M_p - M_T}{S_T} . \frac{p}{y}$$

dengan keterangan r_{bis} adalah koefisien korelasi *biserial*, y adalah ordinat p dalam distribusi normal, sedangkan simbol lain sama dengan keterangan sebelumnya. Nilai korelasi *point biserial* selalu lebih rendah dibanding dengan nilai korelasi *biserial*. Hubungan antara keduanya dinyatakan dengan rumus:

$$r_{pbis} = r_{bis} . \frac{y}{\sqrt{p.q}}$$

Soal pilihan ganda perlu memiliki pengecoh, yaitu jawaban yang tidak bernilai benar. Pengecoh perlu dibuat sedemikian rupa sehingga menarik perhatian peserta tes yang belum memiliki konsep yang baik terhadap materi yang diujikan. Allen dan Yen (1979) menyatakan bahwa pengecoh yang baik minimum berindeks 0,1 yang berupa koefisien korelasi point biserial, bernilai positif untuk kunci jawaban dan bernilai negatif untuk pengecoh.

Kesalahan Pengukuran (*Standard Error of Measurement*, SEM) membantu penyusun tes dalam memahami kesalahan yang bersifat acak yang mempengaruhi skor peserta tes. Kesalahan pengukuran dihitung dengan rumus sebagai berikut (Bahrul Hayat, dkk., 1999):

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{XX'}}$$

dengan keterangan σ_X adalah standar deviasi dari skor total dan $\rho_{XX'}$ adalah koefisien reliabilitas tes.

Reliabilitas tes dapat diartikan sebagai keajegan atau konsistensi hasil pengukuran atau hasil tes yang dilakukan pada waktu yang berbeda pada subjek yang sama. Allen dan Yen (1979) menyatakan bahwa tes disebut reliabel jika skor amatan mempunyai korelasi yang tinggi dengan skor yang sebenarnya. Mereka juga menyatakan bahwa reliabilitas merupakan koefisien korelasi antara dua skor amatan yang diproleh dari hasil pengukuran menggunakan tes yang paralel.

Reliabilitas suatu tes dapat dihitung dengan beberapa cara dan formula. Cara atau formula belah dua, alfa (α) Cronbach, Guttman, dan paralel dapat digunakan. Nilai hasil perhitungan dari formula tersebut sering dikatakan sebagai koefisien reliabilitas. Mehrens dan Lehmann (1973) menyatakan bahwa meskipun

tidak ada ketentuan umum, tetapi secara luas dapat diterima bahwa untuk tes yang digunakan untuk membuat keputusan secara perorangan harus memiliki koefisien reliabilitas minimal 0,85.

Keterbatasan pada teori tes klasik adalah adanya sifat *group dependent* dan *item dependent* (Hambleton, Swaminathan, & Rogers, 1991: 2-5), juga indeks daya pembeda, tingkat kesulitan, dan koefisien reliabilitas tes juga tergantung kepada peserta tes yang mengerjakan tes tersebut.

Untuk mengatasi kelemahan-kelemahan yang ada pada teori tes klasik, para ahli pengukuran mencari model alternatif. Hambleton, Swaminathan, & Rogers (1991: 2-5) serta Hulin, Drasgow, & Parsons (1983), menyatakan seharusnya model alternatif ini memiliki sifat : (a) statistik butir tidak tergantung pada kelompok subjek, (b) skor tes dapat menggambarkan kemampuan subjek, (c) model dinyatakan dalam tingkatan butir, tidak dalam tingkatan tes, d) model tidak memerlukan tes paralel untuk menghitung koefisien reliabilitas, dan e) model menyediakan ukuran yang tepat untuk setiap skor kemampuan. Model alternatif ini adalah model pengukuran yang disebut dengan teori respons butir (*Item Response Theory*).

C. Teori Respons Butir

1. Butir Dikotomus

Hambleton, Swaminathan, & Rogers (1991: 2-5) menyatakan bahwa teori respons butir didasarkan pada dua buah postulat, yaitu : (a) prestasi subjek pada suatu butir soal dapat diprediksikan dengan seperangkat faktor yang disebut kemampuan laten (*latent traits*), dan (b) hubungan antara prestasi subjek pada suatu butir soal dan perangkat kemampuan yang mendasarinya sesuai dengan grafik fungsi naik monoton tertentu, yang disebut kurva karakteristik butir (*item characteristic curve, ICC*). Kurva karakteristik butir ini menggambarkan bahwa semakin tinggi level kemampuan peserta tes, semakin meningkat pula peluang menjawab benar suatu butir.

Ada tiga model logistik dalam teori respons butir, yaitu model logistik satu parameter (1 PL), model logistik dua parameter (2 PL), dan model logistik tiga

parameter (3 PL). Perbedaan dari ketiga model tersebut terletak pada banyaknya parameter yang digunakan dalam menggambarkan karakteristik butir dalam model yang digunakan. Parameter-parameter yang digunakan tersebut adalah indeks kesukaran, indeks daya beda butir dan indeks tebakan semu *(pseudoguessing*).

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).

$$P_i(\theta) = c_i + \frac{(1-c_i)e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}}$$

Keterangan :

θ : tingkat kemampuan (*ability*) peserta tes

$P_i(\theta)$: 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

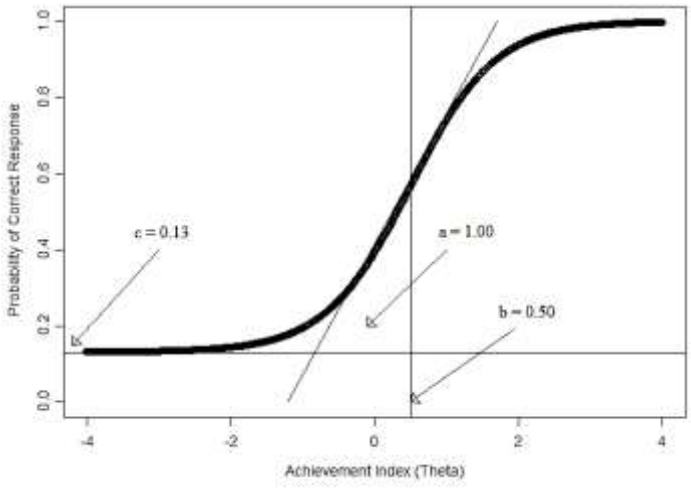
c_i : indeks tebakan semu butir ke-i

e : bilangan natural yang nilainya mendekati 2,718

D : faktor penskalaan yang harganya 1,7.

Kurva karakteristik butir soal yang dianalisis dengan model 3 parameter logistik yang memiliki a = 1,00; b = 0,50; dan c = 0,13 ditunjukkan pada Gambar 1. Gambar ini menunjukkan bahwa probabilitas menjawab benar tidak berawal dari 0, tetapi berawal dari 0,13. Jadi jawaban yang sifatnya tebakan mempunyai kemungkinan benar 13,0%. Daya beda pada kurva karakteristik butir ditunjukkan dengan kemiringan grafik yang ada. Semakin vertikal kurva karakteristik suatu

butir soal, berarti butir soal tersebut semakin bisa membedakan peserta pandai atau kurang pandai.



Gambar 1. Kurva Karakteristik Butir Soal dengan a = 1,00; b = 0,50; dan c = 0,13

Model 2 parameter dan 1 parameter merupakan bagian dari model 3 parameter. Model 2 parameter merupakan kasus khusus dari model 3 parameter, yakni ketika c = 0. Model 1 parameter merupakan kasus khusus dari model 2 parameter, yakni ketika a = 1 atau a merupakan tetapan untuk keseluruhan butir tes. Model 2 parameter logistik secara matematika dapat dirumuskan sebagai berikut:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1+e^{Da_i(\theta - b_i)}};$$

sedagkan model 1 parameter logistik rumus matematikanya adalah sebagai berikut:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

Estimasi parameter dapat dilakukan dengan menggunakan bantuan program komputer. Nilai-nilai indeks parameter butir dan kemampuan peserta merupakan hasil estimasi. Karena merupakan hasil estimasi, maka kebenarannya bersifat probabilistik dan tidak terlepaskan dengan kesalahan pengukuran. Dalam teori respons butir, kesalahan pengukuran standar *(Standard Error of Measurement, SE)* berkaitan erat dengan fungsi informasi.

Fungsi informasi dengan *SE* mempunyai hubungan yang berbanding terbalik kuadratik, semakin besar fungsi informasi maka *SE* semakin kecil atau sebaliknya (Hambleton, Swaminathan, & Rogers, 1991, 94). Jika nilai fungsi informasi dinyatakan dengan $I_i(\theta)$, nilai estimasi SE dinyatakan dengan$SE(\theta)$, dan *N* adalah jumlah butir yang ada, hubungan keduanya menurut Hambleton, Swaminathan, & Rogers (1991 : 94) dan Baker (2001, 119) dinyatakan dengan

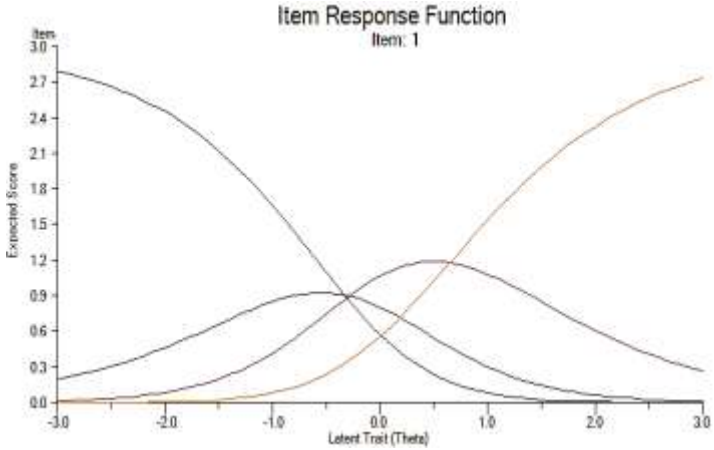
$$SE(\theta) = \frac{1}{\sqrt{\sum_{i=1}^{N} I_i(\theta)}}$$

2. Butir Politomus

Dalam butir dikotomus respons jawaban hanya ada dua kemungkinan, yaitu nol atau satu. Kadang-kadang respons lebih dari dua kemungkinan tersebut, misalnya pada angket skala Likert, atau soal dengan jawaban bergradasi sesuai dengan tingkat kebenarannya. Butir yang respons jawabannya lebih dari dua kemungkinan disebut butir politomus.

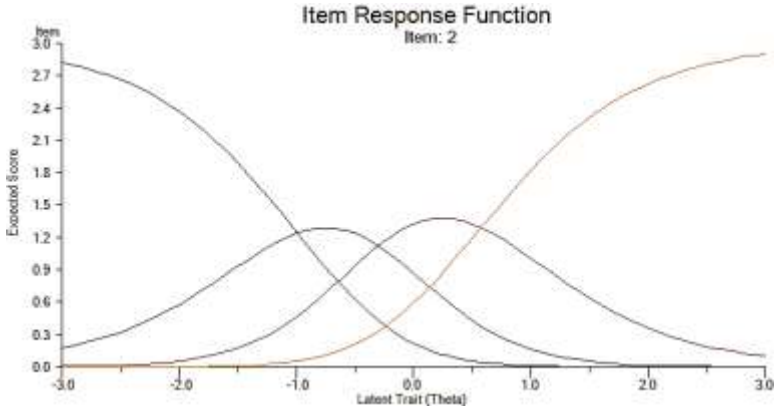
Ada beberapa model dalam analisis butir politomus, antara lain adalah politomus satu parameter (b) yang disebut *Partial Credit Model* (PCM), dua parameter (a dan b) yang disebut *Gegeralized Partial Credit Model* (GPCM), dan dua parameter (a dan b) yang disebut *Graded Respons Model* (GRM). Model matematis ketiganya dapat dipelajari di teori respons butir lanjut.

Contoh karakteristik butir PCM secara grafik ditunjukkan pada Gambar 2, GPCM ditunjukkan pada Gambar 3, dan GRM ditunjukkan pada Gambar 4.



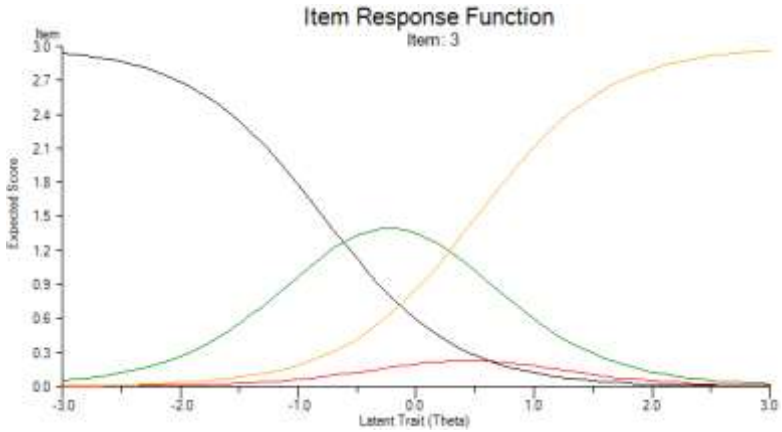
b = -0.331,-0.281,0.651

Gambar 2. Karakteristik Butir PCM



a = 1.396
b = -0.991,-0.301,0.575

Gambar 3. Karakteristik Butir GPCM



a = 1.779
b = -0.789,0.344,0.515

Gambar 4. Karakteristik Butir GRM

DAFTAR PUSTAKA

Allen, M. J & Yen, W. M. (1979). *Introduction to measurement theory*. Belmont: Wadsworth.

Bahrul Hayat, Sumarno S. Pranata, dan Herwindo Haribowo. (1999). *Manual item and test analysis (ITEMAN)*. Jakarta: Pusbang Sisjian Depdikbud.

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

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

Han, K. T. and Hambleton, R. K. (2007). User's Manual for WinGen: Windows Software that Generates IRT Model Parameters and Item Responses. Massachusetts: University of Massachusetts Amherst.

Hulin, C.L., Drasgow, F. & Parsons, C.K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

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