Analisis Butir Soal Tes Prestasi Hasil Belajar

Oleh: Gito Supriadi

ABSTRAK

Untuk mendapatkan soal yang berkualitas baik maka perlu dilakukan analisis butir soal. Secara garis besar ada dua cara menganalisis soal, yaitu analisis soal secara kualitatif dan analisis soal secara kuantitatif.

Analisis soal secara kualitatif dilakukan sebelum diadakan ujicoba, yakni dengan cara mencermati butir-butir soal yang telah disusun dilihat dari kesesuaian dengan kemampuan dasar dan indikator yang diukur serta pemenuhan persyaratan baik dari aspek materi, kontruksi, dan bahasa. Sedangkan analisis soal secara kuantitatif menekankan pada karakteristik internal tes melalui data yang diperoleh secara empiris. Karakteristik internal secara kuantitatif dimaksudkan meliputi parameter tingkat kesukaran, daya pembeda, fungsi distraktor, dan reliabilitas.

Indeks Kesukaran soal yang ideal berkisar antara 0.30 – 0.70, indeks daya pembeda yang ideal adalah mendekati angka 1; distraktor berfungsi dengan baik apabila dipilih lebih banyak oleh kelompok rendah, dan reliabilitas tes yang baik apabila memiliki indeks reliabilitas minimum 0.70.

Kata-kata kunci : Indeks kesukaran, daya pembeda, dan reliabilitas

A. Pendahuluan

Sebagai alat ukur, suatu tes baru dapat dikatakan berhasil menjalankan fungsi ukurnya apabila mampu memberikan hasil ukur yang cermat dan akurat. Tes yang hasil ukurnya tidak cermat atau tidak dapat menunjukkan perbedaan-perbedaan kecil yang ada pada objek ukurnya tidaklah banyak memberikan informasi yang berguna. Tidak ada arti sebuah tes prestasi apabila ia tidak mampu menunjukkan perbedaan

antara siswa yang mempunyai sedikit kemampuan dan yang mempunyai lebih banyak kecakapan. Tidak berguna sebuah tes prestasi yang sedemikian mudahnya sehingga semua siswa dapat menjawab semua soal dengan benar dan kemudian penguji tidak dapat membedakan antara mereka yang benar-benar menguasai pelajaran dan mereka yang menjawab benar semata-mata karena soal itu terlalu mudah. Manfaat apakah yang dapat diambil dari sebuah tes prestasi yang demikian sukarnya sehingga tidak seorang pun yang mampu menjawab satu soalpun dengan benar?.

Sebuah tes yang berisi soal-soal berkualitas tinggi walaupun dalam jumlah yang sedikit akan jauh lebih berguna daripada sebuah tes yang berisi puluhan soal berkualitas rendah. Soal-soal yang berkualitas rendah tidak saja akan menurunkan fungsi tes akan tetapi akan memberikan hasil pengukuran yang menyesatkan.

Oleh karena itu setiap tes yang telah selesai ditulis, masih harus diuji kualitasnya secara empirik. Soal-soalnya masih harus diuji dengan menggunakan data yang diperoleh melalui suatu prosedur *try-out* atau dari hasil pengenaan tes di kelas yang sesungguhnya *(field tested)*. Dari data hasil pengenaan tes ini akan diperoleh bukti mengenai kualitas soal-soal tes yang bersangkutan. Kemudian dari hasil analisis terhadap data empirik ini pula diperoleh dasar untuk melakukan perbaikan-perbaikan yang diperlukan. Prosedur kerja dalam melakukan pengujian seluruh soal tes yang didasarkan pada data empirik tersebut dinamai prosedur analisis butir soal.

Dalam tulisan ini penulis akan memaparkan sebuah prosedur analisis butir soal dengan memusatkan pada teori tes klasik, dengan tujuan dapat memberikan sebuah wawasan bagi para penyusun tes prestasi, khususnya tes bentuk objektif, sehingga tes yang disusun akan menghasilkan butir-butir soal yang memiliki kualitas baik dilihat dari segi indeks kesukaran soal dan daya diskriminasi.

B. Analisis Butir Soal

Analisis butir soal dilakukan untuk mengetahui berfungsi tidaknya sebuah soal. Analisis pada umumnya dilakukan melalui dua cara, yaitu analisis soal secara teoritik atau kualitatif dan analisis soal secara empiris atau analisis soal secara kuantitatif.

Analisis soal secara teoritik atau analisis kualitatif dilakukan sebelum diadakan ujicoba, yakni dengan cara mencermati butir-

butir soal yang telah disusun dilihat dari kesesuaian dengan kemampuan dasar dan indikator yang diukur serta pemenuhan persyaratan baik dari aspek materi, kontruksi, dan bahasa (Mardapi, 2004: 130). Sedangkan analisis soal secara kuantitatif menekankan pada karakteristik internal tes melalui data yang diperoleh secara empiris. Karakteristik internal secara kuantitatif dimaksudkan meliputi parameter soal tingkat kesukaran, daya pembeda, distribusi jawaban, dan reliabilitas (Surapranata, 2005:10)

Pada pembahasan berikut penulis tidak akan membicarakan analisis soal secara kualitatif, akan tetapi difokuskan pada analisis soal secara kuantitatif yang meliputi parameter tingkat kesukaran soal, daya pembeda, fungsi distraktor, dan reliabilitas.

1. Indeks Kesukaran Soal

Sangatlah penting untuk melihat tingkat kesukaran soal dalam rangka menyediakan berbagai macam alat diagnostik kesulitan belajar peserta didik ataupun dalam rangka meningkatkan penilaian berbasis kelas. Baik buruknya butir tes juga ditentukan oleh tingkat kesukaran butir tersebut, yang diperoleh dari analisis soal. Secara umum, menurut teori klasik, tingkat kesukaran dapat dinyatakan melalui beberapa cara diantaranya : (1) proporsi menjawab benar, (2) skala kesukaran linear, (3) indeks Davis, dan (4) skala bivariat. Proporsi jawaban benar (p), yaitu jumlah peserta tes yang menjawab benar pada butir soal yang dianalisis dibandingkan dengan jumlah peserta tes seluruhnya merupakan tingkat kesukaran yang paling umum digunakan (Surapranata, 2005:12). Indeks kesukaran suatu soal dinyatakan oleh suatu indeks yang dinamakan indeks kesukaran soal dan disimbolkan oleh huruf p. Indeks kesukaran soal merupakan rasio antara penjawab soal dengan benar dan banyaknya penjawab soal. Secara teoretik dikatakan bahwa p sebenarnya merupakan probabilitas empirik untuk lulus soal tertentu bagi kelompok siswa tertentu. Formulasi indeks kesukaran soal adalah:

$$p = n_i / N$$

Keterangan

p = proporsi menjawab benar atau tingkat kesukaran

n_i = banyaknya siswa yang menjawab soal dengan benar

N = jumlah peserta tes

Besarnya indeks kesukaran berkisar antara 0,00 sampai dengan 1,00. Suatu soal yang mempunyai p = 0, artinya soal itu terlalu sukar karena tidak ada peserta tes yang menjawab benar, sedangkan butir yang mempunyai harga p = 1, artinya soal itu terlalu mudah karena setiap peserta tes dapat menjawab dengan benar. Dari penjelasan diatas dapat disimpulkan bahwa semakin tinggi harga p, butir soal tersebut semakin mudah. Hal demikian secara logis sebetulnya dikatakan tingkat kemudahan butir soal (Allen & Yen, 1979:120).

Tingkat kesukaran biasanya dibedakan menjadi tiga kategori seperti nampak pada tabel 1 :

Tabel 1 :
Kategori Tingkat Kesukaran

Nilai p	Kategori
$p < 0,3$	Sukar
$0,3 \leq p \leq 0,7$	Sedang
$p > 0,7$	Mudah

Sebagai contoh, dari 80 orang siswa yang mengikuti tes ternyata soal nomor 1 dapat dijawab dengan benar oleh 60 orang siswa, sedangkan soal nomor 2 dijawab dengan benar oleh 25 orang siswa. Maka untuk soal nomor 1 $n_i = 60$, dan $p = 60/80 = 0,75$, sedangkan untuk soal nomor 2, $n_i = 25$ dan $p = 25/80 = 0,31$.

Dalam contoh di atas soal nomor 1 adalah lebih mudah daripada soal nomor 2 dikarenakan soal nomor 1 dapat dijawab oleh lebih banyak siswa (60 orang), sedangkan soal nomor 2 hanya dapat dijawab oleh 25 orang. Akan tetapi, p untuk soal nomor 1 angkanya lebih besar daripada p untuk soal nomor 2. Hal itu menunjukkan bahwa semakin besar angka p berarti soal yang bersangkutan semakin mudah dan sebaliknya semakin kecil p berarti soal yang bersangkutan semakin sukar.

Berapakah besar p yang ideal? Walaupun tidak selalu benar, namun umumnya p yang berada disekitar 0,50 dianggap yang terbaik. Kadang-kadang dikehendaki harga p yang lebih kecil daripada 0,50 (yaitu soalnya lebih sulit) (Azwar, 2005:137). Menurut Allen & Yen (1979:122) indek kesukaran sekitar 0,30 – 0,70 merupakan indeks kesukaran yang baik.

2. Indeks Daya Pembeda (Diskriminasi) Soal

Daya beda soal adalah kemampuan suatu soal untuk membedakan antara siswa yang pandai (siswa yang mempunyai kemam-

puan tinggi) dengan siswa yang kurang pandai (siswa yang mempunyai kemampuan rendah). Fungsi dari daya beda itu adalah mendeteksi perbedaan individual yang sekecil-kecilnya di antara para subyek tes, sejalan dengan fungsi dan tujuan tes itu sendiri. Butir yang demikian dikatakan valid atau cermat (Azwar, 2005: 137).

Indeks daya pembeda dihitung atas dasar pembagian kelompok menjadi dua bagian, yaitu kelompok atas yang merupakan kelompok peserta tes yang berkemampuan tinggi dengan kelompok bawah yang merupakan kelompok peserta tes yang berkemampuan rendah. Kemampuan tinggi ditunjukkan dengan perolehan skor yang tinggi dan kemampuan rendah ditunjukkan dengan dengan perolehan skor yang rendah. Indeks daya pembeda didefinisikan sebagai selisih antara proporsi jawaban benar pada kelompok atas dengan proporsi jawaban benar pada kelompok bawah (Crocker & Algina, (1986). Pembagian kelompok menurut Kelley (1939), Crocker & Algina (1986) dalam Surapranata (2005:24), yang paling stabil dan sensitif serta paling banyak digunakan adalah dengan menentukan 27% kelompok atas dan 27% kelompok bawah. Menurut Ebel (1979) kriteria indeks daya beda adalah sebagai berikut:

Tabel 2:
Indeks Daya Pembeda Soal

Nilai D	Kategori	Keterangan
$D \geq 0{,}40$	Sangat baik	Diterima
$0{,}30 \leq D \leq 0{,}39$	Baik	Perlu peningkatan
$0{,}20 \leq D \leq 0{,}29$	Cukup	Perlu perbaikan
$D \leq 0{,}19$	Tidak baik	Dibuang

Sebagai contoh misalkan sebuah tes berjumlah 40 soal yang diikuti oleh 36 peserta tes, selanjutnya perolehan skor diurutkan dari skor tertinggi sampai skor terendah. Teknik pembagian kelompok atas dan kelompok bawah dapat dilihat pada tabel 2 berikut:

Tabel 3 : Pembagian kelompok 27 % - 27 %.

No	Peserta Tes	Butir Soal/item										Skor Total
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	
1.	Ahmad	1	1	1	1	1	1	1	0	1	1	9
2.	Azizah	1	1	1	1	1	1	1	0	1	1	9
3.	Andi	1	1	1	1	1	1	1	0	1	1	9
4.	Asmawati	1	1	1	1	0	1	1	0	1	1	8
5.	Asyifa	0	1	1	1	1	1	1	0	1	1	8
6.	Aminah	1	1	0	1	1	1	1	0	1	0	7
7.	Aini	1	0	1	1	1	1	1	0	1	0	7
8.	Bahriah	1	1	0	1	1	0	1	0	1	1	7
9.	Bambang	0	1	1	0	1	1	1	0	1	1	7
10.	Budi	1	0	0	1	1	1	1	0	1	1	7
11.	Cinta	1	0	0	1	1	1	1	1	0	0	6
12.	Choiriyah	1	0	0	1	1	1	1	1	0	0	6
13.	Endang	1	0	0	1	1	1	1	1	0	0	6
14.	Erna	1	0	0	0	0	1	0	1	1	1	6
15.	Eniwati	1	1	0	1	1	0	1	1	0	0	6
16.	Farida	1	1	1	0	0	0	1	1	1	0	6
17.	Fitria	1	0	0	1	1	1	0	0	0	1	5
18.	Farhan	1	0	0	1	1	1	0	0	0	1	5
19.	Fiqrah	1	0	0	1	1	1	0	0	0	1	5
20.	Hani	1	0	0	1	1	1	0	0	0	1	5
21	Kemuning	1	0	0	1	1	1	0	0	0	1	5
22	Mardiana	1	0	0	1	1	1	0	0	0	1	5
23	Hesty	0	0	0	1	1	0	0	0	0	1	4
24	Hamidah	0	1	1	0	1	0	1	0	0	0	4
25	Jamilah	0	0	1	0	1	0	1	1	0	0	4
26	Kenanga	0	1	0	0	1	0	0	1	1	0	4
27	Mawar	1	0	1	0	0	0	0	1	0	1	4
28	Melati	0	1	0	1	0	0	1	1	0	0	4
29	Mukminah	0	1	1	0	0	1	0	0	0	0	3
30	Maskanah	0	0	0	0	1	1	0	1	1	0	3
31	Murdiano	1	0	0	0	1	0	0	1	0	0	3
32	Joko	0	0	1	0	0	0	0	1	0	1	3
33	Apriani	0	1	0	1	0	0	0	1	0	0	3
34	Basuki	0	1	0	1	0	0	0	1	0	0	3
35	Imas	1	1	0	0	0	1	0	0	0	0	3
36	Rahmati	0	1	0	0	0	0	1	1	0	0	3
	ΣX	23	18	13	23	25	22	19	16	14	18	
Jumlah Peserta Tes		36	36	36	36	36	36	36	36	36	36	
Tingkat kesukaran (p)		0.64	0.50	0.36	0.64	0.69	0.61	0.53	0.44	0.39	0.50	

Formulasi yang digunakan untuk mencari indeks daya pembeda adalah:

$d = n_{iT} / N_T - n_{iR} / N_R$

Keterangan:

n_{iT} = Banyaknya penjawab soal dengan benar dari kelompok atas

N_T = Banyaknya penjawab dari kelompok tinggi

niR = Banyaknya penjawab soal dengan benar dari kelompok rendah

N_R = Banyaknya penjawab dari kelompok rendah

Dari tabel 2 di atas dapat dibagi dua kelompok, yaitu 10 peserta tes dari kelompok atas (27%) nomor peserta 1 sampai dengan 10, dan 27% kelompok bawah berjumlah 10 orang yaitu nomor 27 sampai dengan 36. dengan berdasarkan rumus indeks daya pembeda soal di atas, maka diperoleh indeks daya pembeda soal nomor 1 sampai 10 sebagai berikut:

Tabel 4: Kategori Tingkat Kesukaran dan Daya Pembeda

Soal	Kelompok atas	Kelompok bawah	Daya Pembeda
1	0.80	0.30	0.50
2	0.80	0.60	0.20
3	0.70	0.70	0.00
4	0.90	0.70	0.20
5	0.90	0.20	0.70
6	0.90	0.30	0.60
7	1.00	0.20	0.80
8	0.00	0.80	-0.80
9	1.00	0.10	0.90
10	0.80	0.20	0.60

Kembali ke tingkat kesukaran seperti ditunjukkan pada tabel 4, dapat dilihat bahwa soal nomor 3 merupakan soal yang mudah bagi kelompok bawah maupun bagi kelompok atas. Perhitungan indeks daya pembeda pada soal nomor 3 diperoleh sebesar 0.00, hal ini dapat memberikan informasi bahwa soal nomor 3 tidak dapat membedakan antara kelompok atas dan kelompok bawah. Sedangkan soal nomor 8 merupakan soal yang sangat mudah bagi kelompok bawah, tetapi sangat sukar bagi kelompok atas. Jika dilihat indeks daya pembeda sebesar -0.80 maka soal nomor 8 memiliki indeks daya

beda yang sangat baik, tetapi terbalik. Tanda negatif pada soal nomor 8 menunjukkan bahwa peserta tes yang kemampuannya tinggi (kelompok atas) tidak dapat menjawab soal dengan benar, tetapi peserta tes kelompok bawah dapat menjawab dengan benar. Dengan demikian data tersebut menunjukkan bahwa soal 3 dan nomor 8 merupakan soal yang tidak baik. Data statistik menunjukkan bahwa soal nomor 1, 5, 6, 7, 9 dan nomor 10 merupakan soal yang memiliki indeks daya beda yang sangat baik, ditinjau dari segi daya pembeda soal, sedangkan soal nomor 2 dan nomor 4 merupakan soal yang cukup, akan tetapi perlu perbaikan.

Daya pembeda soal maksimal tercapai apabila seluruh peserta tes kelompok tinggi dapat menjawab dengan benar ($n_{iT} = N_T$), sedangkan seluruh subjek kelompok rendah tidak seorang pun dapat menjawab dengan benar ($n_{iR} = 0$).

Dalam hal ini harga d = 1 – 0 = 1. Indeks daya pembeda soal sebesar 0 akan terjadi apabila proporsi penjawab benar dari kelompok tinggi dan dari kelompok rendah sama besarnya, yaitu ketika indeks kesukaran bagi kelompok tinggi sama besar dengan indeks kesukaran bagi kelompok rendah.

Secara matematis, indeks daya pembeda soal (d) besarnya akan berkisar mulai dari -1 sampai dengan +1, namun hanya harga d yang positif sajalah yang memiliki arti dalam analisis butir soal. Harga d yang berada di sekitar 0 menunjukkan bahwa soal yang bersangkutan mempunyai daya pembeda yang rendah sedangkan harga d yang negatif menunjukkan bahwa soal yang bersangkutan tidak ada gunanya, bahwa memberikan informasi yang menyesatkan.

Indeks daya pembeda yang ideal adalah yang sebesar mungkin mendekati angka 1, semakin besar indeks daya pembeda berarti soal tersebut semakin mampu membedakan antara mereka yang menguasai bahan yang diujikan dan mereka yang tidak menguasai bahan. Semakin kecil indeks daya pembeda (mendekati 0) berarti semakin tidak jelaslah fungsi soal yang bersangkutan dalam membedakan mana subjek yang menguasai bahan pelajaran dan mana subjek yang tidak tahu apa-apa.

3. Fungsi Distraktor

Apabila dilihat strukturnya tes bentuk pilihan ganda terdiri atas dua bagian yaitu pokok soal atau stem yang berisi permasalahan yang akan ditanyakan dan sejumlah kemungkinan jawaban atau

option. Kemungkinan jawaban itu dibagi dua yaitu kunci jawaban dan pengecoh. Dari sekian banyak alternatif jawaban hanya terdapat satu yang paling benar yang dinamakan kunci jawaban, sedangkan kemungkinan jawaban yang tidak benar dinamakan pengecoh atau distraktor (Surapranata, 2005: 43)

Pengecoh berfungsi sebagai pengidentifikasi peserta tes yang berkemampuan tinggi. Pengecoh dikatakan berfungsi efektif apabila banyak dipilih oleh peserta tes yang berasal dari kelompok rendah, sebaliknya apabila pengecoh banyak dipilih oleh peserta tes yang berasal dari kelompok atas, maka pengecoh itu tidak berfungsi sebagaimana mestinya.

Menurut Azwar (2005: 141) efektivitas distraktor dapat dilihat dari dua kriteria, yaitu ; (a) distraktor dipilih oleh peserta tes dari kelompok rendah, dan (b) pemilih distraktor tersebar relatif proporsional pada masing-masing distraktor yang ada. Lebih lanjut Surapranata (2005: 43) dan Sudijono (2005: 411) suatu pengecoh dapat dikatakan berfungsi baik jika paling sedikit dipilih oleh 5 % dari peserta tes. Apabila pengecoh dipilih secara merata, maka termasuk pengecoh yang sangat baik. Dan apabila pengecoh lebih banyak dipilih oleh peserta tes dari kelompok atas dibandingkan dengan kelompok bawah, maka termasuk pengecoh yang menyesatkan.

Berikut ini dikemukakan sebuah contoh bagaimana cara menganalisis fungsi distraktor. Misalnya sebuah tes diikuti oleh 50 orang peserta tes, bentuk soal pilihan ganda sebanyak 40 item, dimana setiap item dilengkapi dengan lima alternatif jawaban, yaitu A, B, C, D dan E. Dari 40 butir item tersebut khusus untuk butir item nomor 1, 2, dan 3 diperoleh pola penyebaran jawaban item sebagai berikut:

No. Soal	Altenatif jawaban					Keterangan
	A	B	C	D	E	
1	4	6	5	**30***	5	* Kunci jawaban
2	1	**44***	2	1	2	
3	1	1	**10***	1	37	

Dengan pola penyebaran jawaban item sebagaimana tergambar pada tabel di atas, maka dengan mudah dapat diketahui berapa persen peserta tes yang telah "terkecoh" untuk memilih distraktor yang dipasangkan pada item 1, 2, dan 3, yaitu :

a. untuk item nomor 1, kunci jawabannya D, sedangkan pengecohnya adalah A, B, C dan E. Pengecoh A dipilih oleh 4 orang, berarti 4/50x100% = 8%. Jadi pengecoh A sudah dapat berfungsi dengan baik, sebab angka persentasenya lebih dari 5%. Pengecoh B dipilih oleh 6 orang, berarti 6/50x100% = 12% (telah berfungsi dengan baik). Pengecoh C dipilih oleh 5 orang, berarti 5/50x100% = 10 % (telah berfungsi dengan baik). Pengecoh E dipilih oleh 5 orang = 10% (telah berfungsi dengan baik). Jadi keempat pengecoh yang dipasangkan pada item nomor 1 sudah dapat menjalankan fungsinya dengan baik.
b. Untuk item nomor 2 kunci jawabannya adalah B, sebagai pengecohnya adalah : A, C, D, dan E. Pengecoh A dipilih 1 orang, berarti 1/50x100% = 2% (belum berfungsi), pengecoh C dipilih 2 orang berarti 2/50x100% = 4% (belum berfungsi), pengecoh D dipilih 1 orang = 2% (belum berfungsi), dan pengecoh E dipilih 2 orang yang berarti juga 4% (belum berfungsi). Jadi keempat pengecoh yang dipasangkan di item nomor 2 belum dapat menjalankan fungsinya seperti yang diharapkan.
c. Untuk item nomor 3, kuncinya adalah C, sebagai pengecoh adalah; A, B, D dan E. Pengecoh A, B, dan D masing-masing dipilih oleh 1 orang (=2%) berarti ketiga pengecoh itu belum berfungsi. Adapun pengecoh E dipilih oleh 37 orang, berarti 37/50x100% = 74% (telah berfungsi dengan baik). Jadi soal nomor tiga hanya 1 buah pengecoh saja yang sudah dapat menjalankan fungsinya dengan baik.

4. Reliabilitas

Penekanan utama dalam mengumpulkan data untuk menentukan reliabilitas tes adalah pada konsistensi dihubungkan dengan reliabilitas skor atau reliabilitas penilai. Reliabilitas skor berarti bahwa jika suatu tes telah diadministrasikan pada penempuh ujian untuk kedua kalinya, maka penempuh ujian akan tetap mem-peroleh skor yang sama dengan pengadministrasian yang pertama. Salah

satu cara para spesialis pengukuran dalam menentukan reliabilitas skor tes adalah melalui tes standar. Jika penempuh ujian diuji kembali, mereka harus melengkapi tugas yang sama persis dalam kondisi yang juga persis sama. Hal ini akan membantu dalam pencapaian hasil tes yang konsisten.

Indeks reliabilitas soal dikatakan baik adalah minimum 0.70 (Mardapi, 2004: 119). Reliabilitas memiliki dua keajegan, pertama adalah keajegan internal yakni tingkat sejauhmana tingkat butir soal itu homogen baik dari segi tingkat kesukaran maupun bentuk soalnya. Keajegan kedua adalah keajegan eksternal yakni tingkat sejauhmana skor dihasilkan tetap sama sepanjang kemampuan orang yang diukur belum berubah.

Untuk dapat mengestimasi reliabilitas terdapat beberapa metode reliabilitas yaitu (1), *test-retest* atau stabilitas (2) *pararel* atau ekuivalen, (3) *split-half* atau belah dua, (4) *interval consintency* (Surapranata, 2005: 90). Pada saat sekarang sejalan dengan kecanggihan teknologi dengan bantuan komputer program *Iteman* dari *MicroCAT*, akan dengan mudah dan cepat untuk menghitung indeks reliabilitas tes hasil belajar. Berikut adalah hasil dari penghitungan sebuah tes (data tidak disampaikan disini) ujicoba tes hasil belajar mata pelajaran fiqih kelas VII MTs di Yogyakarta guna mencari reliabilitas tes.

Tabel 5: Mencari Reliabilitas dengan Komputer

```
        MicroCAT (tm) Testing System
   Copyright (c) 1982, 1984, 1986, 1988 by
        Assessment Systems Corporation

Item and Test Analysis Program -- ITEMAN (tm)
                Version 3.00

  Item analysis for data from file FIQH.TXT
                  Page  6


There were 24 examinees in the data file.

Scale Statistics
----------------

  Scale:            0
               -------
N of Items          30
N of Examinees      24
Mean            20.458
Variance        13.165
Std. Dev.        3.628
Skew            -0.544
Kurtosis        -0.952
Minimum         13.000
Maximum         25.000
Median          22.000
Alpha            0.668
SEM              2.091
Mean P           0.682
Mean Item-Tot.   0.327
Mean Biserial    0.473
```

Dari hasil analisis program *Iteman* dapat dilihat bahwa koefisien alpha sebesar 0.668. hal ini menunjukkan bahwa tes tersebut secara keseluruhan belum reliabel, sebab koefisien alpha kurang dari 0.70.

Secara manual berikut ini penulis sajikan salah satu teknik menghitung reliabilitas tes dengan menggunakan persamaan *test-retest*.

Tabel 6: Perhitungan reliabilitas dengan test-retest methods

No	Peserta	Tes Pertama	Tes Kedua	X_1^2	X_2^2	X_1X_2
		X_1	X_2			
1.	Ahmad	31	36	961	1296	1116
2.	Azizah	30	35	900	1225	1050
3.	Andi	30	34	900	1156	1020
4.	Asmawati	30	35	900	1225	1050
5.	Asyifa	31	33	961	1089	1023
6.	Aminah	29	35	841	1225	1015
7.	Aini	30	36	900	1296	1080
8.	Bahriah	16	40	256	1600	640
9.	Bambang	14	32	196	1024	448
10.	Budi	16	33	256	1089	528
11.	Cinta	18	31	324	961	558
12.	Choiriyah	12	36	144	1296	432
13.	Endang	13	21	169	441	273
14.	Erna	15	26	225	676	390
15.	Eniwati	11	25	121	625	275
16.	Farida	13	27	169	729	351
17.	Fitria	12	15	144	225	180
18.	Farhan	9	14	81	196	126
19.	Fiqrah	11	16	121	256	176
20.	Hani	13	18	169	324	234
21.	Kemuning	12	15	144	225	180
22.	Mardiana	21	18	441	324	378
23.	Hesty	15	9	225	81	135
24.	Hamidah	15	7	225	49	105
25.	Jamilah	9	12	81	144	108
26.	Kenangan	10	8	100	64	80
27.	Mawar	10	8	100	64	80
28.	Melati	16	11	256	121	176
29.	Mukminah	13	11	169	121	143
30.	Maskanah	11	11	121	121	121
31.	Murdiano	13	16	169	256	208
32.	Joko	15	18	225	324	270
33.	Apriani	9	8	81	64	72
34.	Basuki	6	8	36	64	48
35.	Badrun	3	4	9	16	12
36.	Rahmati	4	5	16	25	20
	Σ	566	747	11136	20017	14101

Dari tabel di atas diperoleh jumlah skor masing-masing tes : $\Sigma X_1 = 566$; $\Sigma X_2 = 747$; $\Sigma X_1^2 = 11136$; $\Sigma X_2^2 = 20017$ dan $\Sigma X_1X_2 = 14101$, selanjutnya menentukan korelasi antara tes I dan tes II sebagai berikut:

$$rx_1x_2 = \frac{N\sum X_1X_2 - (\sum X_1)(\sum X_2)}{\sqrt{((N\sum X_1^2) - (\sum X_1)^2)(N\sum X_2^2 - (X_2)^2)}}$$

$$rx_1x_2 = \frac{(36x14101) - (566)(747)}{\sqrt{(36x11136 - (566)^2)(36x20017 - (747)^2)}}$$

$$rx_1x_2 = 0.7413$$

Reliabilitas hasil perhitungan adalah $rx_1x_2 = 0.7413$. Angka ini menunjukkan bahwa tes pertama dengan tes kedua telah menunjukkan reliabilitas yang cukup baik, karena lebih dari 0.70.

C. Kesimpulan

Dari uraian tentang analisis butir soal di atas, dapat penulis simpulkan bahwa:

1. Indeks kesukaran butir soal yang ideal berkisar antara 0.30 – 0.70, indeks daya pembeda soal yang ideal adalah yang sebesar mungkin mendekati angka 1, semakin besar indeks daya pembeda berarti soal tersebut semakin mampu membedakan antara mereka yang menguasai bahan yang diujikan dan mereka yang tidak menguasai bahan;
2. Suatu pengecoh dapat dikatakan berfungsi baik jika paling sedikit dipilih oleh 5 % dari peserta tes. Apabila pengecoh dipilih secara merata, maka termasuk pengecoh yang sangat baik. Kemudian apabila pengecoh lebih banyak dipilih oleh peserta tes dari kelompok atas dibandingkan dengan kelompok bawah, maka termasuk pengecoh yang menyesatkan.
3. Reliabilitas memiliki dua keajegan, pertama adalah keajegan internal yakni tingkat sejauhmana tingkat butir soal itu homogen baik dari segi tingkat kesukaran maupun bentuk soalnya. Keajegan kedua adalah keajegan eksternal yakni tingkat sejauhmana skor dihasilkan tetap sama sepanjang kemampuan orang yang diukur belum berubah. Reliabilitas soal dikatakan baik apabila memiliki indeks minimum 0.70.

DAFTAR PUSTAKA

Allen, M.J. and Yen, W. *Introduction to Measurement Theory.* Monterey: Brooks/Cole Publishing Company. 1979.

Azwar, Saifuddin. *Tes Prestasi : Fungsi dan Pengembangan Pengukuran Prestasi Belajar.* Yogyakarta: Pustaka Pelajar Offset, 2005.

Ebel, R. L. *Essentials of Educational Measurement. (2nd ed.).* Englewood Cliff, New Jersey: Prentice-Hall, Inc. 1979.

Mardapi, Djemari, *Penyusunan Tes Hasil Belajar.* Yogyakarta: Program Pascasarjana UNY, 2004.

Sudijono, Anas., *Pengantar Evaluasi Pendidikan,* Jakarta: Raja Grafinfo Persada, 2005.

Surapranata, Sumarna., *Analisis, Validitas, Reliabilitas dan Interpretasi Hasil Tes. Implementasi kurikulum 2004.* Bandung: Remaja Rosdakarya Offset, 2005.