

# Reliabilitas Tes Hasil Belajar dan Aplikasinya Dalam Evaluasi Pembelajaran

Oleh: Gito Supriadi

## ABSTRAK

Reliabilitas adalah tingkat keajegan atau kemantapan hasil dari hasil dua pengukuran terhadap hal yang sama. Hasil pengukuran itu diharapkan akan sama apabila pengukuran diulangi pada waktu yang berbeda.

Reliabilitas atau keajegan suatu skor merupakan hal yang sangat penting dalam menentukan apakah tes telah menyajikan pengukuran yang baik atau belum. Pengukuran yang baik tentunya akan ajeg membedakan kemampuan peserta didik antara yang pandai dan yang kurang pandai.

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.

**Kata Kunci :** Reliabilitas

### A. Pendahuluan

Evaluasi pendidikan melibatkan banyak kegiatan teknis dalam menentukan metode dan format penilaian yang dapat digunakan untuk mendapatkan informasi yang dibutuhkan. Informasi tersebut diperlukan dalam menafsir dan menetapkan keputusan untuk kepentingan pendidikan. Penilai

membutuhkan keterampilan dalam mengidentifikasi dan memahami berbagai macam perspektif penilaian, baik penilaian kontekstual dan proses maupun penilaian hasil. Karena penilaian merupakan pusat kontrol keberhasilan program pendidikan, maka terdapat dua syarat utama yang harus dipenuhi oleh suatu instrumen

penilaian, yaitu validitas dan reliabilitas.

Validitas mengacu pada keberartian, kebenaran, kemanfaatan, dan kesesuaian skor tes. Validitas merupakan karakteristik suatu tes ketika diujikan pada suatu kelompok peserta tes. Validasi suatu instrumen mencakup pengumpulan data empiris dan argumentasi logis untuk menunjukkan bahwa kesimpulan tertentu adalah tepat. Sedangkan reliabilitas yang berarti konsistensi adalah ciri umum dari suatu instrumen pengukuran dan penilaian pendidikan. Konsistensi tinggi skor instrumen dari suatu pengukuran ke pengukuran berikutnya merupakan ciri terpenting dari instrumen yang berkualitas tinggi.

Dalam tulisan ini penulis akan menekankan pada pembahasan reliabilitas yang merupakan sebuah kajian teoretis tentang apa dan bagaimana reliabilitas itu apabila dikaitkan dengan kualitas instrumen dan penerapannya dalam penilaian asil suatu program pembelajaran. Selanjutnya tulisan ini diharapkan mengundang wacana bagi pembaca tentang: bagaimana mengestimasi reliabilitas? Berapa koefisien reliabilitas suatu instrumen penilaian yang baik? Faktor-faktor apa yang mempengaruhi reliabilitas suatu

instrumen penilaian?.

## **B. Konsep Reliabilitas**

Reliabilitas telah didefinisikan dengan cara yang berbeda oleh pengarang yang berbeda. Cara yang terbaik untuk membahas reliabilitas adalah sejauhmana hasil pengukuran dari suatu instrumen mewakili karakteristik yang diukur. Sebagai contoh, reliabilitas didefinisikan seberapa besar konsistensi skor tes yang dicapai peserta tes pada pengujian ulang. Definisi ini akan memuaskan jika skor tes dapat menggambarkan kemampuan peserta tes; jika tidak maka skor tes tidak sistematis, tidak dapat diulangi atau tidak terikat. Reliabilitas juga diartikan sebagai indikator ketidakhadiran kesalahan acak. Jika kesalahan acak dapat diperkecil maka skor tes akan lebih konsisten dari suatu pengujian ke pengujian berikutnya.

Menurut Nunally (1970), Allen dan Yen (1979), dan Anastasi (1986) bahwa reliabilitas adalah kestabilan skor yang diperoleh orang yang sama ketika diuji ulang dengan tes yang sama pada situasi yang berbeda atau dari satu pengukuran ke pengukuran yang lainnya. Jadi reliabilitas dapat dinyatakan sebagai tingkat keajegan atau kemantapan hasil dari

hasil dua pengukuran terhadap hal yang sama. Hasil pengukuran itu diharapkan akan sama apabila pengukuran diulangi.

Definisi teoretis dari reliabilitas adalah proporsi keragaman skor tes yang disebabkan oleh keragaman sistematis dalam populasi peserta tes. Jika terdapat keragaman sistematis yang lebih besar dalam suatu populasi dibandingkan dengan populasi lainnya, seperti dalam semua siswa sekolah negeri dibandingkan hanya dengan kelas tertentu, tes akan mempunyai reliabilitas lebih besar untuk populasi yang lebih bervariasi. Reliabilitas adalah karakteristik bersama antara tes dan kelompok peserta tes. Reliabilitas tes bervariasi dari suatu kelompok dengan kelompok lainnya.

Para profesional pengukuran menganggap reliabilitas sebagai persyaratan utama suatu instrumen penilaian. Dalam teori tes diakui bahwa skor tes akan valid (benar) jika skor tes tersebut reliabel (Mehrens & Lehmann, 1991). Asumsi ini didasarkan pada suatu model matematika teori tes dimana skor perolehan terdiri atas skor tulen dan skor galat ( $\text{obtained score} = \text{true score} + \text{error score}$ ). Semakin sedikit kesalahan dalam suatu tes (yaitu semakin reliabel)

semakin valid skor tes. Karenanya, suatu penilaian yang tidak reliabel secara otomatis tidak valid. Hal ini sejalan dengan pendapat Anderson dalam Arikunto (1999: 87) menyatakan bahwa persyaratan bagi tes adalah validitas dan reliabilitas. Dalam hal ini validitas lebih penting, dan reliabilitas perlu karena menyokong terbentuknya validitas. Sebuah tes mungkin reliabel tetapi tidak valid. Sebaliknya sebuah tes yang valid biasanya reliabel.

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 memperoleh 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.

### C. Model Umum Reliabilitas

Pengukuran yang baik tentunya akan ajeg membedakan kemampuan peserta didik. Misalnya saja Melati lebih pintar daripada Mawar, skor yang diperoleh Melati mestinya lebih tinggi dibandingkan dengan Mawar. Dalam kenyataannya, pengukuran kita tidak pernah sempurna. Teori reliabilitas dikembangkan untuk menentukan ketidakajegan tersebut.

Setiap pengukuran, baik dalam sains maupun dalam penilaian kelas selalu mengandung kesalahan pengukuran. Pengukuran materan dengan skala mm misalnya memiliki kesalahan pengukuran sebesar 0,01mm. Pengukuran dengan skala cm memiliki kesalahan sebesar 0,1cm. Konsep dasar reliabilitas juga menggunakan kesalahan pengukuran tersebut. Setiap skor yang diperoleh seorang peserta tes terdiri atas tiga hal: skor amatan yang sering disebut skor perolehan (*observed test score*), skor yang sebenarnya (*true score*), yaitu skor yang sesuai dengan kemampuan peserta tes yang sebenarnya, dan kesalahan pengukuran, yaitu faktor-faktor yang mempengaruhi ketidakajegan suatu pengukuran adalah situasi yang mempengaruhi perolehan skor (*error of measure-*

*ment*). Secara umum konsep dasar tersebut dinyatakan dalam persamaan matematis sebagai berikut :

$$X = T + e$$

$$X = \text{observed test score}$$

$$T = \text{true score}$$

$$e = \text{error of measurement}$$

Kesalahan pengukuran merupakan selisih antara skor amatan (perolehan) dan skor sebenarnya.

$$e = X - T$$

Berbagai hasil telah diturunkan dari teori ini yang pada akhirnya mengandung beberapa implikasi terhadap pengukuran (penilaian berbasis kelas khususnya). Teori reliabilitas ditentukan berdasarkan varian antara skor perolehan yang merupakan penjumlahan dari varian skor sebenarnya dengan varian kesalahan pengukuran, yaitu :

$$\sigma_x^2 = \sigma_T^2 + \sigma_e^2$$

Persamaan ini menunjukkan bahwa skor tes ditentukan oleh dua hal, yaitu variabilitas skor sebenarnya dan variabilitas kesalahan pengukuran. Jika kesalahan pengukuran itu memberikan kontribusi yang signifikan, maka tes menjadi tidak ajeg. Dengan kata lain apabila varian kesalahan pengukuran relatif tidak berarti maka pengukuran menjadi ajeg.

Koefisien reliabilitas ( $r_{xx}$ ) menyajikan indeks relatif pengaruh skor sebenarnya dan skor

kesalahan pengukuran terhadap skor yang diperoleh.

#### **D. Metode Sederhana Mengestimasi Reliabilitas**

Tujuan utama mengestimasi reliabilitas adalah untuk menentukan seberapa besar variabilitas yang terjadi akibat adanya kesalahan pengukuran dan seberapa besar variabilitas skor tes sebenarnya.

Reliabilitas memiliki dua keajaiban, pertama adalah keajaiban internal yakni tingkat sejauhmana tingkat butir soal itu homogen baik dari segi tingkat kesukaran maupun bentuk soalnya. Keajaiban kedua adalah keajaiban 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 consistency* (Surapranata, 2004:90)

Dalam pembahasan ini penulis tidak akan menguraikan secara keseluruhan dari metode-metode tersebut, akan tetapi akan menyajikan salah satu metode yang sederhana yaitu *test-retest method* (stabilitas).

Metode *test-retest method* atau metode tes ulang sering pula dinamakan metode stabilitas merupakan pendekatan yang paling tua digunakan untuk mengestimasi reliabilitas. Pendekatan stabilitas sering pula dinamakan *single-test-double-trial method*. Pelaksanaan metode ini adalah sebagai berikut :

1. Misalkan pada bulan Mei disajikan bentuk tes Pendidikan Agama Islam kepada peserta didik kelas XII SMA.
2. Setelah beberapa waktu, misalnya pada bulan Juni disajikan kembali tes Pendidikan Agama Islam tersebut kepada peserta didik kelas XII SMA yang sama.
3. Skor perolehan kedua kali tes tersebut dikorelasikan. Korelasi skor tes yang pertama dengan skor tes yang kedua digunakan untuk mengestimasi reliabilitas tes. Korelasi yang digunakan adalah korelasi product moment. Jika koefisien korelasinya tinggi, maka reliabilitas tesnya juga tinggi. Tabel berikut menunjukkan skor yang diperoleh 36 orang peserta didik yang mengerjakan dua kali tes Pendidikan Agama Islam kelas XII SMA.

**Tabel 1.**  
**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.	Kenanga	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.	Gita	3	4	9	16	12
36.	Rahmati	4	5	16	25	20
	$\Sigma$	566	747	11136	20017	14101

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

$$r_{x_1 x_2} = \frac{N \Sigma X_1 X_2 - (\Sigma X_1)(\Sigma X_2)}{\sqrt{((N \Sigma X_1^2) - (\Sigma X_1)^2)(N \Sigma X_2^2 - (\Sigma X_2)^2)}}$$

$$r_{x_1 x_2} = \frac{(36 \times 14101) - (566)(747)}{\sqrt{(36 \times 11136 - (566)^2)(36 \times 20017 - (747)^2)}}$$

$$r_{x_1 x_2} = 0.7413$$

Reliabilitas hasil perhitungan dengan tes kedua cukup ajeg, adalah  $r_{x_1 x_2} = 0.7413$ . Angka ini sebagaimana tabel makna koefisien menunjukkan bahwa tes pertama korelasi sebagai berikut :

**Tabel 2.**  
**Makna Angka Korelasi**

Angka korelasi	Makna
0.800 – 1.000	Sangat tinggi/sangat baik
0.600 – 0.800	Tinggi/baik
0.400 – 0.600	Cukup
0.200 – 0.400	Rendah
0.000 – 0.200	Sangat rendah

Metode tes ulang sangat berguna untuk melihat kestabilan pengukuran. Oleh karena itu, metode ini biasa juga disebut sebagai metode kestabilan tes. Metode test-retest sangat peka terhadap masalah *carry-over effect* diantara kedua penyajian. Skor subjek pada penyajian kedua misalnya sangat mungkin dipengaruhi oleh penyajian test yang pertama. Misalnya, bila subjek



masih ingat jawaban yang diberikannya pada penyajian pertama ia sekedar mengulangi saja jawaban tersebut. Hal ini akan meningkatkan korelasi antara dua penyajian dan dapat menyebabkan overestimasi terhadap korelasi  $xx'$ . Kemungkinan pula terjadi peningkatan hasil test karena subjek berlatih dan belajar sesuatu dalam tenggang waktu diantara kedua penyajian. Perubahan sikap subjek juga dapat mempengaruhi estimasi reliabilitas, kalau penyajian yang kedua subjek bersikap negatif, menolak bekerja dengan bersungguh-sungguh, atau dengan sengaja memberi jawaban sekennanya, akan mengakibatkan pula korelasi yang rendah (Mardapi, 2004:50).

### **E. Seberapa Tinggi Koefisien Reliabilitas**

Reliabilitas tes adalah proporsi varians tulen (*true variance*) dalam skor tes (Guilford, 1982:492). Penilaian kecukupan koefisien reliabilitas tes dapat diacuh dari pendapat Aiken (1988:138) bahwa jika tes akan digunakan untuk menentukan signifikansi perbedaan rerata skor dua kelompok siswa maka koefisien reliabilitas sebesar 0,65 dianggap memuaskan. Dan jika tes akan digunakan untuk membandingkan

siswa yang satu dengan yang lainnya maka paling tidak diperlukan koefisien reliabilitas sebesar 0,85. Untuk menjelaskan keberartian koefisien reliabilitas dapat pula diacuh dari galat baku pengukuran, yang dihitung dengan menggunakan rumus: dimana  $S_m$  = galat baku pengukuran;  $S_x$  = simpangan baku skor tes; dan  $r_x$  = koefisien reliabilitas tes.

Misalnya, dari hasil perhitungan koefisien reliabilitas instrumen dengan menggunakan rumus Alpha Cronbach diperoleh 0,93 dengan galat baku pengukuran 6,88. Hal ini berarti bahwa tes tersebut sangat terandalkan karena dapat mengukur 93 persen keragaman skor yang sebenarnya, dan bila dalam jangka waktu tertentu dan dalam kondisi yang sama para responden merespon kembali tes tersebut maka rentangan penyimpangan skor total yang dicapai masing-masing responden berkisar antara  $+6,88$ ; jadi bila pada tes pertama seseorang siswa memperoleh skor total 450 maka kemungkinan rentangan skor total yang dicapai pada tes berikutnya adalah  $-450+6,88$  atau paling rendah 443,12 dan paling tinggi 456,88.

Jika tes yang diadministrasikan memiliki konsekuensi tinggi,



seperti tes yang digunakan untuk penempatan dalam pendidikan, misalnya ujian akhir SMA, dan sertifikasi profesional, maka diperlukan reliabilitas konsistensi internal yang tinggi paling sedikit di atas 0,90, dan paling baik jika di atas 0,95. Kesalahan klasifikasi yang disebabkan oleh kesalahan pengukuran harus diperkecil. Tetapi perlu dicatat bahwa tidak satu pun tes dengan sendirinya dapat digunakan untuk membuat suatu keputusan penting bagi seseorang.

Tes di kelas tidak selalu membutuhkan koefisien reliabilitas tinggi. Ketika para siswa lebih menguasai materi yang diujikan, variabilitas tes akan menurun, sehingga reliabilitas tes juga akan menurun. Para guru mengawasi siswa mereka sepanjang hari dan mempunyai peluang untuk mengumpulkan masukan dari berbagai sumber informasi. Jika pengetahuan dan pertimbangan guru digunakan bersama dengan informasi yang diperoleh dari tes, maka akan dapat menyediakan informasi yang lebih lengkap. Jika suatu tes tidak reliabel atau tidak akurat untuk siswa secara perorangan, maka guru perlu membuat koreksi penyesuaian. Suatu koefisien reliabilitas sebesar 0.50 atau 0.60 mungkin cukup

untuk tes di kelas.

Selanjutnya, reliabilitas adalah karakteristik bersama antara tes dan kelompok peserta tes. Reliabilitas juga perlu dievaluasi dalam kaitan dengan kelompok peserta tes. Suatu tes dengan koefisien reliabilitas 0.92 ketika diujikan pada siswa dalam beberapa kelas maka koefisien reliabilitas yang diperoleh tidak akan sama jika tes tersebut hanya diujikan pada satu kelas saja.

Reliabilitas berhubungan dengan konsistensi hasil pengukuran. Reliabilitas dipengaruhi oleh cakupan instrumen penilaian. Misalnya, suatu instrumen tes tertentu yang mencakup sasaran belajar dan butir yang terbatas memiliki reliabilitas yang lebih rendah dibanding dengan tes yang mencakup sasaran belajar yang lebih luas dengan jumlah butir yang lebih banyak.

Instrumen yang representatif dengan kesalahan pengukuran yang relatif kecil akan memiliki reliabilitas tinggi. Kesalahan pengukuran dapat diperkecil melalui penulisan butir instrumen yang jelas, petunjuk yang mudah dipahami, administrasi instrumen yang sesuai, dan penskoran yang konsisten. Suatu instrumen tes adalah suatu sampel perilaku dari keterampilan yang diinginkan, tes

lebih panjang dengan sampel yang lebih besar, memungkinkan untuk lebih reliabel. Hasil ujian akhir dari suatu unit pembelajaran dengan waktu satu jam akan lebih reliabel ketimbang hasil ujian harian dengan jangkauan materi dan waktu yang terbatas.

### **F. Ancaman Terhadap Reliabilitas**

Semua jenis instrumen tes atau nontes tidak terlepas kesalahan. Hal ini berlaku untuk instrumen tes dalam ilmu-ilmu eksakta dan dalam ilmu-ilmu psikologi dan pendidikan. Misalnya, dalam mengukur panjang dengan suatu penggaris, mungkin ada kesalahan sistematis berhubungan dengan di mana titik nol dicetak pada penggaris dan kesalahan acak berhubungan dengan kemampuan mata dalam membaca tanda-tanda dan memperhitungkan tanda-tanda tersebut. Juga memungkinkan bahwa panjang obyek dapat berubah dari waktu ke waktu dan pada lingkungan yang berbeda (misalnya perubahan temperatur). Salah satu tujuan penilaian adalah untuk mengurangi kesalahan tersebut hingga ke tingkatan yang sesuai dengan tujuan tes. Tes yang beresiko tinggi (*high-stakes test*), seperti ujian untuk mendapatkan SIM, harus mempunyai kesalahan

yang sangat kecil. Tes di kelas dapat mentolerir kesalahan yang lebih tinggi secara wajar kesalahan tersebut mudah dikoreksi sepanjang proses pengujian. Reliabilitas hanya mengacu pada derajat tingkat kesalahan yang tidak sistematis, yang disebut kesalahan acak.

Ada tiga sumber kesalahan utama dalam menentukan reliabilitas, yaitu: faktor dalam tes itu sendiri, faktor siswa yang dites, dan faktor penskoran. Umumnya tes berisi suatu koleksi butir yang mengukur keterampilan tertentu. Adakalanya guru secara khas menggeneralisasikan masing-masing butir tes ke semua materi yang diukur oleh tes itu. Sebagai contoh, jika seorang siswa dapat memecahkan beberapa permasalahan seperti  $7 \times 8$ , maka mungkin akan disamaratakan kemampuannya dalam mengalikan angka tunggal bilangan bulat. Juga mungkin akan menyamaratakan suatu kumpulan materi kepada suatu domein yang lebih luas. Jika siswa dapat menyelesaikan penjumlahan, pengurangan, perkalian, dan pembagian, maka mungkin akan disimpulkan bahwa siswa tersebut mampu menyelesaikan operasi pecahan. Kesalahan dapat pula disebabkan oleh pemilihan butir untuk mengukur

*domain* dan keterampilan tertentu. Materi yang tercakup dalam tes berbeda menurut format masing-masing tes, kesalahan pensampelan, pembatasan butir tes, dan karena menyamaratakan ke data yang tidak diamati, yakni, kemampuan siswa terhadap keseluruhan butir yang mungkin terdapat dalam tes. Ketika keterampilan dan domain yang diukur menjadi lebih rumit, mungkin akan terjadi lebih banyak kesalahan yang disebabkan oleh pensampelan materi. Sumber lain kesalahan tes adalah ketidakefektifan pengecoh dalam tes pilihan ganda, seperti jawaban benar yang lebih banyak, dan tingkat kesukaran butir tes.

Sebagai manusia, para siswa tidaklah selalu konsisten dan juga tidak terlepas dari kesalahan dalam menyelesaikan tes. Apakah tes itu dimaksudkan untuk mengukur kemampuan khusus atau kemampuan siswa secara optimal, perubahan dalam berbagai hal seperti sikap siswa, kesehatan, dan rasa kantuk dapat mempengaruhi kualitas usaha dan konsistensi siswa dalam menyelesaikan tes. Sebagai contoh, peserta tes mungkin membuat kesalahan karena teledor, salah menafsirkan petunjuk tes, melupakan instruksi tes, melupakan beberapa butir tes, atau salah baca butir tes.

Kesalahan penskoran merupakan sumber sepertiga dari kesalahan potensial. Pada bentuk tes objektif, penskoran bersifat mekanik, dan kesalahan penskoran harus diperkecil. Pada tes uraian, sumber kesalahan meliputi ketidakjelasan rubrik penskoran, ketidakjelasan apa yang diharapkan dari siswa, dan beberapa kesalahan yang bersumber dari penilai. Para penilai tidaklah selalu konsisten, kadang-kadang merubah ukuran-ukuran mereka selagi menskor, dan terkadang terpengaruh oleh hal-hal yang tidak berhubungan dengan skor tes seperti *efek halo*, latar belakang siswa, perbedaan persepsi, kebaikan hati atau kepelikan, dan kesalahan dalam penskalaan (Rudner, 1992).

### **G. Kesimpulan**

Reliabilitas adalah kestabilan skor yang diperoleh orang yang sama ketika diuji ulang dengan tes yang sama pada situasi yang berbeda atau dari satu pengukuran ke pengukuran yang lainnya. Jadi reliabilitas dapat dinyatakan sebagai tingkat keajegan atau kemantapan hasil dari hasil dua pengukuran terhadap hal yang sama. Hasil pengukuran itu diharapkan akan sama apabila pengukuran diulangi.

Reliabilitas memiliki dua ke-  
ajegan, 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.

Penilaian koefisien reliabilitas  
tes dianggap cukup adalah jika tes  
akan digunakan untuk menen-

takan signifikansi perbedaan  
rerata skor dua kelompok siswa  
maka koefisien reliabilitas sebesar  
0,65. Dan jika tes akan digunakan  
untuk membandingkan siswa yang  
satu dengan yang lainnya maka  
paling tidak diperlukan koefisien  
reliabilitas sebesar 0,85.

Ada tiga sumber kesalahan  
utama dalam menentukan  
koefisien reliabilitas, yaitu : faktor  
dalam tes itu sendiri, faktor siswa  
yang dites, dan faktor penskoran.

### DAFTAR PUSTAKA

- Aiken, Lewis R. *Psychological Testing and Assessment*. Massachusetts: Allyn and Bacon Inc. 1988.
- Allen, M.J., & Yen.W.M., *Introduction to Measurement Theory*. California: Brooks/Cole Publishing Company, 1979.
- Anas Sudijono, *Pengantar Evaluasi Pendidikan*, Jakarta, PT. Raja Grafindo Persada, 2003.
- Anastasi, A, *Psychology Testing*, (6<sup>th</sup>ed), New York: MacMillan Publishing Company, 1988.
- Djemari Mardapi, *Penyusunan Tes Hasil Belajar*, Program Pascasarjana, Universitas Negeri Yogyakarta, Tanpa Penerbit, 2004.
- Guilford, J.P. *Psychometric Methods*. New Delhi: Tata McGraw-Hill Publishing Co. Ltd., 1982.
- Mehrens, W. A., & Lehmann, I. J. *Measurement and Evaluation in Education and Psychology*. Fourth Edition. Fort Worth, TX: Holt, Rinehart, and Winston, Inc., 1991.

*Nunnally, Jum C. Jr. Introduction to Psychological Measurement. New York: McGraw-Hill Book Company, 1970.*

Sumarna Surapranata, *Analisis, Validitas, reliabilitas dan Interpretasi Hasil Tes Implementasi Kurikulum 2004*, Bandung, PT. Remaja Rosdakarya, 2005.

Suharsimi Arikunto, *Dasar-Dasar Evaluasi Pendidikan*, Jakarta, Bumi Aksara, 1999.