ANALISIS DAYA BEDA SOAL, TARAF KESUKARAN, VALIDITAS BUTIR TES, INTERPRETASI HASIL TES DAN VALIDITAS RAMALAN DALAM EVALUASI PENDIDIKAN

Mujianto Solichin
mujiantosolichin@fai.unipdu.ac.id
Universitas Pesantren Tinggi Darul Ulum (Unipdu) Jombang

Abstrak: Daya beda soal adalah kemampuan suatu soal untuk membedakan antara siswa yang pandai (berkemampuan tinggi) dengan siswa yang kurang pintar (berkemampuan rendah). Angka yang menunjukkan besarnya daya beda disebut *indeks diskriminasi*, yang mana berkisar antara 0,00 sampai 1,00. Bilangan yang menunjukkan sukar dan mudahnya sesuatu soal disebut *indeks kesukaran* (*difficulty index*). Besarnya indeks kesukaran antara 0,00 sampai dengan 1,0. Indeks kesukaran ini menunjukkan taraf kesukaran soal. Soal dengan indeks kesukaran 0,0 menunjukkan bahwa soal itu terlalu sukar, sebaliknya indeks 1,0 menunjukkan bahwa soalnya terlalu mudah. Validitas ramalan *(predictive validity)* merupakan suatu tes yang diharapkan mampu meramalkan keberhasilan studi para calon mahasiswa dalam mengikuti program pendidikan di suatu perguruan tinggi pada masa-masa yang akan datang. Adapun yang menjadi permasalahan di sini adalah bagaimana cara yang dapat ditempuh agar kita dapat sampai pada kesimpulan bahwa suatu tes telah memiliki validitas ramalan? Analisis soal sesungguhnya bertujuan untuk mengadakan identifikasi soal-soal yang baik, kurang baik, dan soal yang jelek. Dengan analisis soal dapat diperoleh informasi tentang kejelekan sebuah soal dan "petunjuk" untuk mengadakan perbaikan dalam pembelajaran. Kualitas tes dan butir soal sangat ditentukan oleh: (1) validitas, (2) reliabilitas, (3) objektivitas, (4) praktikabilitas, (5) daya pembeda, (6) taraf atau derajad kesukaran, (7) efektivitas pilihan, dan (8) efisiensi.
Kata kunci: daya beda, taraf kesukaran, tes, validitas.

Abstract: What is meant here is the ability of the question to distinguish between clever students and not-clever students. Figures indicating the magnitude of discriminating ability are called discriminative indices, which range from 0.00 to 1.00. Numbers that indicate difficult and easy questions are called difficulty indexes. The magnitude of the index of difficulty between 0.00 to 1.0. This difficult index indicates the difficulty level of the question. Questions with a .07 difficult index suggest that the question is too difficult, otherwise index 1.0 indicates that the question is too easy. Predictive validity is a test that is expected to predict the success of the study of students in following the education program in college in the future. As for the problem here is how do we get to the conclusion that a test has had the validity of the prediction? Question analysis aims to identify good and bad questions. The

analysis of the questions obtained informations about the badness of questions and "guidance" to make improvements in learning. The quality of tests and questions is largely determined by: (1) validity, (2) reliability, (3) objectivity, (4) practice, (5) distinguishing ability, (6) level or degree of difficulty, (7) effective option (8) efficiency.
Keywords: distinguishing ability, levels of difficulty, test, validity.

Pendahuluan

Kegiatan evaluasi dalam dunia pendidikan merupakan komponen integral dalam program pembelajaran di samping rencana pembelajaran (kurikulum), tujuan pembelajaran, bentuk pembelajaran, cara pembelajaran (metode), dan alat pembelajaran (media), serta metode pembelajaran.[1] Tujuan utama dalam pelaksanaan evaluasi pembelajaran adalah untuk mendapatkan informasi yang akurat mengenai tingkat pencapaian tujuan pembelajaran oleh siswa sehingga dapat diupayakan tindak lanjutnya.[2]

Evaluasi dalam proses pendidikan menurut H.A.R. Tilaar, berkaitan dengan kegiatan mengontrol sejauh mana hasil yang telah dicapai sesuai dengan program yang telah ditetapkan dalam kurikulum pendidikan.[3] Kegiatan evaluasi ini perlu terutama untuk menciptakan kesempatan bagi para siswa untuk memperlihatkan prestasi mereka dalam kaitannya dengan tujuan yang telah ditentukan dalam kurikulum tersebut. Sehingga evaluasi merupakan alat pemicu pengantar prestasi belajar siswa secara merata.

Evaluasi tes yang diadakan pada tiap-tiap mata pelajaran, akhir semester, menjadi sangat penting *(urgent)* kedudukan dan fungsinya dalam mengukur tingkat kemampuan dan pemahaman siswa. Aktivitas evaluasi sebenarnya harus selalu dilakukan pada saat akhir pelajaran, gunanya untuk menilai sampai seberapa besar tingkat penguasaan ilmu pengetahuan yang diberikan dan diserap siswa. Dalam hal ini, proses persiapan, pembuatan soal, pelaksanaan tes, observasi dan penilaian tes, hendaknya direncanakan secara teratur dan kontinyu sehingga guru dapat benar-benar mengevaluasi dan membimbing perkembangan siswa secara positif[4] sesuai dengan

[1] Secara umum, ada 2 (dua) macam evaluasi yang kita kenal, yakni evaluasi hasil belajar dan evaluasi proses pembelajaran. Evaluasi hasil pembelajaran disebut juga evaluasi substantif, atau populer dengan sebutan tes dan pengukuran hasil belajar. Sedang Evaluasi proses pembelajaran, yang oleh beberapa ahli ada pula yang menyebutkan sebagai evaluasi diagnostik atau juga evaluasi manajerial. Lihat www.depdiknas.go.id/evaluasi-proses-pembelajaran-sebagai-kontrol-kualitas-di-lembaga-pendidikan-yang-otonom.html. Diakses pada 21 Oktober 2015 Pukul 06.03 WIB.

[2] Daryanto, *Evaluasi Pendidikan* (Jakarta: Rineka Cipta, 1999), 11 dan 19.

[3] H. A. R. Tilaar, *Menuju Pendidikan Nasional: Kajian Pendidikan Masa Depan* (Bandung: PT. Remaja Rosdakarya, 1994), 45.

[4] Ngalim Purwanto, *Prinsip-prinsip dan Teknik Evaluasi Pengajaran* (Bandung: Remaja Rosdakarya, 2003), 25.

penerapan Kurikulum Tingkat Satuan Pendidikan (KTSP) 2006,[5] 2010 maupun Kurikulum 2013.

Proses pembelajaran merupakan sistem yang terdiri atas beberapa komponen. Salah satu komponen yang terpenting dalam proses belajar mengajar adalah evaluasi. Evaluasi dalam bahasa Inggris dikenal dengan istilah evaluation adalah suatu proses sistematis untuk menentukan atau membuat keputusan sampai sejauh mana tujuan program telah dicapai.[6] Dalam kaitannya dengan pendidikan, Nurkancana dan Sunartana mengemukakan bahwa evaluasi pendidikan merupakan suatu tindakan atau suatu proses untuk menentukan nilai terhadap segala sesuatu yang berkaitan dalam dunia pendidikan.[7]

Dalam kegiatan evaluasi diperlukan alat atau teknik penilaian, sehingga pelaksanaannya akan lebih terarah. Alat evaluasi dalam pendidikan yang digunakan untuk mengumpulkan data dapat berupa tes atau nontes.[8] Berkaitan dengan pemahaman kita terhadap tes, Nurkancana dan Sunartana menyatakan tes adalah suatu cara untuk mengadakan penilaian yang berbentuk suatu tugas atau serangkaian tugas yang harus dikerjakan oleh siswa atau sekelompok siswa sehingga menghasilkan nilai tentang tingkah laku atau prestasi siswa sebagai peserta didik.[9]

Evaluasi belajar secara teratur bukan hanya ditunjukkan untuk mengetahui tingkat daya serap dan kemampuan siswa, tetapi yang terpenting adalah memanfaatkan hasilnya untuk memperbaiki dan menyempurnakan proses pembelajaran. Sistem evaluasi harus mampu memberikan umpan balik kepada guru untuk terus menerus meningkatkan kemampuan peserta didik.[10]

Suatu tes evaluasi yang baik memiliki ciri dan sifat yang merupakan persyaratan-persyaratan yang harus dipenuhi, yaitu tes tersebut harus valid atau memiliki tingkat validitas yang absah/baik. Sebuah tes evaluasi dikatakan valid apabila tes tersebut secara tepat dan benar dapat mengukur

[5] Secara umum tujuan Kurikulum Tingkat Satuan Pendidikan (KTSP) adalah untuk memandirikan dan memberdayakan satuan pendidikan melalui pemberian kewenangan (otonomi) kepada lembaga pendidikan dan mendorong sekolah untuk melakukan pengambilan keputusan secara partisipatif dalam pengembangan kurikulum. Kurikulum ini merupakan penyempurnaan dari kurikulum berbasis kompetensi (KBK) 2004. Enco Mulyasa, *Kurikulum Tingkat Satuan Pendidikan (KTSP)* (Bandung: PT. Remaja Rosdakarya, 2007), 22.

[6] Djaali dan Pudji Muljono, *Pengukuran dalam Bidang Pendidikan* (Jakarta: Grasindo, 2008), 1.

[7] Wayan Nurkancana dan P.P.N. Sunartana, *Evaluasi Pendidikan* (Cetakan ke-4). (Surabaya: Usaha Nasional, 1986), 1.

[8] Purwanto, *Evaluasi Hasil Belajar* (Yogyakarta: Pustaka Pelajar, 2011), 56.

[9] Nurkancana dan Sunartana, *Evaluasi Pendidikan*, 25.

[10] Enco Mulyasa, *KBK: Konsep, Karakteristik dan Implementasinya* (Bandung: Remaja Rosdakarya, 2003), 64.

apa yang hendak diukur.[11] Validitas di sini, dapat berupa validitas isi, prediktif atau ramalan dan validitas konstruksi, kemudian tes tersebut harus realiabel, obyektif, praktis dan ekonomis.[12]

Dalam evaluasi pendidikan baik tes maupun nontes, keduanya merupakan instrumen atau alat bantu pengumpulan dan pengolahan data tentang variabel-variabel yang diteliti. Ciri-ciri/karakteristik instrumen yang baik sebagai alat evaluasi adalah memenuhi persyaratan validitas dan reliabilitas. Inilah alasan mengapa alat evaluasi yang baik dapat dilihat dari beberapa segi antara lain: (1) validitas, (2) reliabilitas, (3) objektivitas, (4) praktikabilitas, (5) daya pembeda, (6) taraf atau derajat kesukaran, (7) efektivitas option, (8) efisiensi.[13]

Makalah ini secara khusus mengurai persoalan kualitas tes dan butir soal. Tes sendiri merupakan alat atau prosedur yang dipergunakan dalam rangka pengukuran dan penilaian dalam pembelajaran. Diakhir pembahasan kami suguhkan contoh analisis kualitas tes dan butir soal sebagai gambaran teknis implementatif instrument pengukuran sebuah tes.

Analisis Kualitas Tes dan Butir Soal

Sebagaimana pemaparan kami di atas tentang karakteristik instrumen yang baik sebagai alat evaluasi adalah memenuhi persyaratan validitas dan reliabilitas (keterhandalan). Baik buruknya suatu tes atau alat evaluasi dapat ditinjau dari validitas, reliabilitas, tingkat kesukaran dan daya beda.[14] Sebuah tes disebut valid atau memiliki validitas apabila tes itu dapat tepat mengukur apa yang hendak diukur. Validitas butir perlu dicari untuk mengetahui butir-butir tes manakah yang menyebabkan soal secara keseluruhan jelek karena memiliki validitas rendah. Lebih lanjut Butir soal dikatakan valid apabila memiliki dukungan besar terhadap skor total. Skor pada butir soal menyebabkan skor total menjadi tinggi atau rendah. Dengan kata lain dapat dikatakan bahwa butir soal memiliki validitas yang tinggi jika skor pada butir soal memiliki kesejajaran dengan skor total. Kesejajaran ini diartikan dengan korelasi, dilakukan analisis validitas butir soal dengan menggunakan rumus korelasi *product moment* angka kasar.

Berikut kami uraikan pembahasan terkait tentang analisis butir soal tes taraf kesukaran, daya beda, pola jawaban soal dan analisis pengecoh *(distractor)*. Namun sebelumnya, ada baiknya kita membahas terlebih dahulu proses melakukan penilaian pada tes standar buatan sendiri (guru).

---

[11] Suharsimi Arikunto, *Dasar-Dasar Evaluasi Pendidikan* (Jakarta: Bumi Aksara, 1996), 56.
[12] Anas Sudijono, *Pengantar Evaluasi Pendidikan* (Jakarta: Raja Grafindo, 2003), 163.
[13] M. Subana dan Sudrajat, *Dasar-dasar Penelitian Ilmiah* (Bandung: Pustaka Setia, 2005), 128.
[14] Nurkancana dan Sunartana, *Evaluasi Pendidikan*, 127.

Menilai Tes Buatan Sendiri

Terdapat 4 (empat) cara untuk menilai tes buatan sendiri, sebagaimana di bawah.

a. Meneliti secara jujur soal-soal yang sudah disusun, kadang-kadang dapat diperoleh jawaban tentang ketidakjelasan perintah atau bahasa, taraf kesukaran, dan lain-lain keadaan soal tersebut.
b. Mengadakan analisis soal *(item analysis)*. Analisis soal adalah suatu prosedur yang sistematis, yang akan memberikan informasi-informasi yang sangat khusus terhadap butir tes yang kita susun.
c. Mengadakan *checking validitas*. Validitas yang paling penting dari tes buatan guru adalah validitas kurikuler *(content validity)*. Untuk mengadakan *checking validitas* kurikuler, kita harus merumuskan tujuan setiap bagian pelajaran secara khusus dan jelas sehingga setiap soal dapat kita jodohkan dengan setiap tujuan khusus tersebut.
d. Mengadakan *checking realibilita*. Salah satu indikator untuk tes yang mempunyai reliabilitas yang tinggi adalah bahwa kebanyakan dari soal-soal tes itu mempuyai daya pembeda yang tinggi.[15]

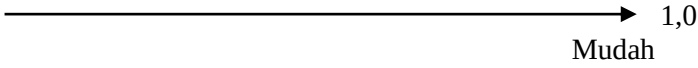
Analisis Butir Soal Tes Taraf Kesukaran

Analisis soal sesungguhnya bertujuan untuk mengadakan identifikasi soal-soal yang baik, kurang baik, dan soal yang jelek. Dengan analisis soal dapat diperoleh informasi tentang kejelekan sebuah soal dan "petunjuk" untuk mengadakan perbaikan. Dalam analisis soal terdapat tiga komponen yang saling berkaitan, yaitu taraf kesukaran, daya pembeda, dan pola jawaban soal.[16]

a. Taraf kesukaran

Soal yang baik adalah soal yang tidak terlalu mudah atau tidak terlalu sukar. Soal yang terlalu mudah tidak merangsang siswa untuk mempertinggi usaha memecahkannya. Sebaliknya soal yang terlalu sukar akan menyebabkan siswa menjadi putus asa dan tidak mempuyai semangat untuk mencoba lagi karena di luar jangkauannya.[17]

Adapun bilangan yang menunjukkan sukar dan mudahnya sesuatu soal disebut *indeks kesukaran (difficulty index)*. Besarnya indeks kesukaran antara 0,00 sampai dengan 1,0. Indeks kesukaran ini menunjukkan taraf kesukaran soal. Soal dengan indeks kesukaran 0,0 menunjukkan bahwa soal itu terlalu sukar, sebaliknya indeks 1,0 menunjukkan bahwa soalnya terlalu mudah.



[15] Arikunto, *Dasar-dasar Evaluasi Pendidikan*, 205-206.
[16] Ibid., 207.
[17] Ibid., 207.

Dalam istilah evaluasi, indeks kesukaran ini diberi simbol P (p besar), singkatan dari kata "proporsi". Dengan demikian maka soal dengan P = 0,70 lebih mudah jika dibandingkan dengan P = 0,20. Sebaliknya soal dengan P = 0,30 lebih sukar daripada soal dengan P = 0,80.

Menurut ketentuan yang sering diikuti, indeks kesukaran sering diklasifikasikan sebagai berikut:

1) soal dengan P = kurang dari 0,30 adalah soal terlalu sukar;
2) soal dengan ~~- = 0,30 s/d 0,70 adalah soal cukup (sedang);
3) soal dengan P = lebih dari 0,70 adalah soal terlalu mudah.[18]

Adapun rumus untuk mencari P (proporsi) adalah:

$$P = \frac{B}{JS}$$

P = Indeks kesukaran.

B = Banyaknya siswa yang menjawab soal itu dengan betul.

JS = Jumlah seluruh siswa peserta tes. [19]

b. Daya pembeda

Daya pembeda soal adalah kemampuan sesuatu soal untuk membedakan antara siswa yang pandai (berkemampuan tinggi) dengan siswa yang kurang pintar (berkemampuan rendah).[20] Angka yang menunjukkan besarnya daya pembeda disebut indeks diskriminasi, disingkat D (d besar). Indeks diskriminasi berkisar antara 0,00 sampai 1,00. Hanya bedanya, indeks kesukaran tidak mengenal tanda negatif (-), tetapi pada indeks diskriminasi ada tanda negatif. Tanda negatif pada indeks diskriminasi digunakan jika sesuatu soal "terbalik" menunjukkan kualitas *testee*.[21]

Angka yang menunjukkan besarnya daya beda disebut *Indeks Diskriminasi* berkisar antara 0,00 sampai 1,00. Akan tetapi pada indeks diskriminasi ini mengenal/ ada tanda negatif (-) yakni -1,0 ----------0,0----------1,0 (semakin ke kanan soal semakin baik, semakin ke kiri maka soal semakin jelek, sebab semakin ke kanan siswa yang pandai semakin sulit/tidak bisa menjawab dan sebaliknya siswa yang kurang pintar (kiri) bisa menjawab dengan asal-asalan).[22]

Rumus:

---

[18] Robert L. Thorndike dan Elizabeth, "Measurement and Evaluation in Psychology and Education," dalam Sudijono, *Pengantar Evaluasi Pendidikan*, 372.

[19] Daryanto, *Evaluasi Pendidikan*, 180-182. Arikunto, *Prosedur*, 212-214. Subana dan Sudrajat, *Dasar-Dasar Penelitian Ilmiah*, 133-135.

[20] Arikunto, *Dasar-Dasar Evaluasi Pendidikan*, 211.

[21] Ibid., 211.

[22] Ibid., 212

$$D = \frac{BA-BB}{JA-JB} = PA - PB \qquad PA = \frac{BA}{JA}, PB = \frac{BB}{JB}$$

D = Indek diskriminasi (daya beda)
JA = Banyaknya peserta kelompok atas
JB = Banyaknya peserta kelompok bawah
BA = Banyaknya peserta kelompok atas yang menjawab soal dengan benar
BB = Banyaknya peserta kelompok bawah yang menjawab soal dengan benar
PA = Proporsi peserta kelompok atas yang menjawab benar
PB = Proporsi peserta kelompok bawah yang menjawab benar
D = 0,00 ---------- 0,20 → Jelek *(poor)*
D = 0,21 ---------- 0,40 → Cukup *(satisfactory)*
D = 0,41 ---------- 0,70 → Baik *(good)*
D = 0,71 ---------- 1,00 → Baik Sekali *(exellent)*
D = Negatif, semuanya → tidak baik, jadi semua butir soal yang mempunyai nilai D negatif sebaiknya dibuang saja.[23]

Dalam kegiatan analisis kualitas tes dan butir soal terdapat manfaat daya pembeda butir soal sebagaimana kami kutip berdasarkan pendapat Karjono Natar berikut ini.

1) Untuk meningkatkan mutu setiap butir soal melalui data empiriknya. Berdasarkan indeks daya pembeda, setiap butir soal dapat diketahui apakah butir soal itu baik, direvisi, atau ditolak.
2) Untuk mengetahui seberapa jauh setiap butir soal dapat mendeteksi/membedakan kemampuan siswa, yaitu siswa yang telah memahami atau belum memahami materi yang diajarkan guru. Apabila suatu butir soal tidak dapat membedakan kedua kemampuan siswa itu, maka butir soal itu dapat dicurigai "kemungkinannya" seperti berikut ini: (a) kunci jawaban butir soal itu tidak tepat; (b) butir soal itu memiliki 2 (dua) atau lebih kunci jawaban yang benar; (c) kompetensi yang diukur tidak jelas; (d) pengecoh tidak berfungsi; (e) materi yang ditanyakan terlalu sulit, sehingga banyak siswa yang menebak; (f) sebagian besar siswa yang memahami materi yang ditanyakan berpikir ada yang salah informasi dalam butir soalnya.[24]

Butir soal tes yang baik juga harus dapat menunjukan daya pembedanya. Sebagaimana penuturan Arikunto di atas, "daya beda adalah kemampuan suatu soal untuk membedakan antara siswa yang pandai (berkemampuan tinggi) dengan siswa yang kurang (berkemampuan rendah)."[25] Menurut Anastasi dan Urbina dalam Purwanto, daya beda

[23] Ibid., 213-217.
[24] Karjono Natar, *Panduan Analisis Butir Soal* (Lampung: UNILA Press, 2011), 12.
[25] Arikunto, *Dasar-Dasar Evaluasi Pendidikan*, 211.

berhubungan dengan derajad kemampuan butir membedakan dengan baik prilaku pengambil tes dalam tes yang dikembangkan. Soal dapat dikatakan mempunyai daya pembeda jika soal tersebut dapat dijawab oleh siswa berkemampuan tinggi dan tidak dapat dijawab oleh siswa berkemampuan rendah. Jika suatu soal dapat dijawab oleh siswa pintar maupun kurang, berarti soal tersebut tidak mempunyai daya beda, demikian juga jika soal tersebut tidak dapat dijawab oleh siswa pintar dan siswa kurang, berarti soal tersebut tidak baik sebab tidak mempunyai daya pembeda.[26]

c. Pola jawaban soal dan analisis pengecoh *(distractor)*

Pola jawaban soal adalah distribusi *testee (responden yang sedang mengerjakan tes)* dalam hal menentukan pilihan jawaban pada soal bentuk pilihan ganda. Pola jawaban soal diperoleh dengan menghitung banyaknya *testee* yang memilih pilihan jawaban a, b, c, atau d atau yang tidak memilih pilihan manapun *(blangko)*. Dalam istilah evaluasi disebut *Omit (tidak menjawab)*, disingkat *O*. Menganalisis fungsi pengecoh *(distractor)* dikenal dengan istilah menganalisis pola penyebaran jawaban butir soal pada soal bentuk pilihan ganda.Dari pola penyebaran jawaban butir soal dapat ditentukan apakah pengecoh berfungsi dengan baik atau tidak. Suatu pengecoh dapat dikatakan berfungsi dengan baik jika paling sedikit dipilih oleh 5% pengikut tes.[27]

Berikut beberapa pertimbangan terhadap analisis pengecoh:

1) diterima, karena sudah baik;
2) ditolak, karena tidak baik;
3) ditulis kembali, karena kurang baik;
4) Sebuah pengecoh dikatakan berfungsi baik jika paling sedikit dipilih oleh 5% pengikut tes.[28]

Untuk tes pilihan ganda dengan 5 alternatif jawaban dan P = 0,8, dilihat dari segi *omitted* (O), sebuah butir soal dikatakan baik jika persentase O-nya ≤ 10%.[29]

Contoh:

Pilihan jawaban	A	B	C*	D	E	O	Jumlah
Kelompok atas	5	7	15	3	3	0	33
Kelompok bawah	8	8	6	5	7	3	37
Jumlah	13	15	21	8	10	3	70

O = Omitted (tidak menjawab)

C* = kunci jawaban

Pengecoh A = 13/70 x 100% > 5%, berfungsi

B = 15/70 x 100% > 5%, berfungsi

[26] Purwanto, *Evaluasi Hasil Belajar* (Yogyakarta: Pustaka Pelajar, 2011), 102.
[27] Arikunto, *Dasar-Dasar Evaluasi Pendidikan*, 219.
[28] Ibid.
[29] Ibid.

D = 8/70 x 100% > 5%, berfungsi
E = 10/70 x 100% > 5%. Berfungsi

Option biasanya berjumlah tiga atau lima buah, dan dari kemungkinan- kemungkinan jawaban yang terpasang pada setiap butir soal itu salah satunya adalah jawaban betul (kunci jawaban) sedangkan sisanya merupakan jawaban salah. Jawaban salah itulah yang biasa dikenal dengan istilah *"distractor"* (*distractor*: pegecoh). Tujuan pemasangan distraktor pada setiap butir item adalah agar dari sekian banyak siswa mengikuti tes ada yang tertarik memilihnya, sebab mereka menyangka bahwa distraktor yang mereka pilih merupakan jawaban betul. Semakin banyak siswa terkecoh, maka distraktor makin dapat menjalankan fungsinya sebaik-baiknya. Sebaliknya, jika distraktor yang dipasang tidak ada yang memilih, maka distraktor tidak dapat menjalankan fungsinya dengan baik. Menurut Sudijono "distraktor dinyatakan telah berfungsi dengan baik apabila distraktor tersebut sekurang-kurangnya sudah dipilih 5% dari seluruh peserta tes."[30]

Istilah Validitas

Istilah validitas dipakai dalam tiga hal: (1) validitas penelitian, (2) validitas soal, dan (3) validitas alat ukur.[31] Validitas penelitian adalah derajat kesesuaian hasil penelitian dengan keadaan sebenarnya, atau sejauh mana hasil penelitian menggambarkan keadaan yang sebenarnya. Sedangkan validitas internal penelitian adalah kesesuaian data hasil penelitian dengan keadaan sebenarnya. Untuk mencapai validitas internal penelitian maka instrumen penelitian harus memenuhi syarat tertentu. Adapun validitas eksternal suatu penelitian adalah derajat kesesuaian antara generalisasi hasil penelitian dengan keadaan yang sebenarnya, atau sejauh mana generalisasi hasil penelitian sesuai dengan keadaan yang sebenarnya. Untuk menjamin validitas eksternal, permasalahan sampling harus menjadi perhatian serius peneliti.

Detail istilah validitas bisa dijelaskan sebagaimana berikut.

1. Validitas item bukan validitas tes. Validitas item adalah derajat kesesuaian antara suatu item dengan perangkat item-item yang lain dari alat ukur yang sama. Ukuran validitas item adalah korelasi antara skor dari suatu item dengan skor pada perangkat item (*item total correlation*) yang biasanya dihitung dengan korelasi *point-biserial* (rpbis = $(X_b - X_s/ S_t)$ (pq/ y)) atau korelasi *product moment*. Isi validitas soal adalah daya pembeda soal *(item discriminating power)* dan bukan validitas tes/alat ukur. Apabila masing-masing soal atau item semuanya berkorelasi tinggi dengan perangkat soal atau perangkat item berarti perangkat soal

[30] Sudijono, *Pengantar Evaluasi*, 411.
[31] Sumadi Suryabrata, *Metodologi Penelitian* (Jakarta: RajaGrafindo Persada, 2004), 40.

dalam suatu tes bersama-sama mengukur sesuatu yang sama.[32] Tapi apakah sesuatu itu adalah persoalan validitas tes?

2. Validitas tes atau alat ukur. Secara umum tes atau alat ukur dipandang valid apabila ia mampu mengukur apa yang hendak diukurnya, atau sejauh mana tes itu mengukur apa yang dimaksudkan untuk diukur. Secara konvensional orang mengkaji validitas alat ukur berdasar tiga arah, yaitu (1) dari arah isi yang diukur, (2) dari arah rekaan teoritis atribut yang diukur, dan (3) dari arah kriterium yang diukur.

Oleh karenanya macam-macam validitas didasarkan pada tiga arah tersebut, yaitu (1) *content validity*, (2) *construct validity*, dan (3) *criterion related validity*.[33]

1. *Content validity* adalah validitas yang diestimasi melalui pengujian terhadap isi tes dengan analisis rasional. Valid-tidaknya suatu tes adalah sampai sejauhmana item-itemnya dapat mencakup seluruh kawasan variabel yang hendak diukur. Estimasi terhadap validitas isi ini tidak perlu menggunakan perhitungan-perhitungan statistik apapun, tapi hanya melalui analisis rasional.

   Ada dua macam *content validity*, yaitu *face validity* dan *logical validity*. Pertama, *Face validity* (validitas tampang) adalah suatu tes dipandang valid apabila item-item tes telah tampak sesuai dengan variabel yang hendak diukur. Dipilihnya validitas tampang ini biasanya karena alasan praktis seperti halnya membuat soal ujian. Kedua, *logical validity* (validitas logik) atau validitas sampling. Valid-tidaknya suatu tes atau alat ukur tergantung pada sejauhmana item-item tes mencerminkan (merepresentasikan) aspek-aspek yang akan diukur. Dengan demikian diharapkan item-item yang dibuat tidak menyimpang dari aspek-aspek variabel yang hendak diukur. Validitas logik mempunyai peranan penting dalam tes prestasi, dengan memberikan kisi-kisi (*blue-print*) yang mencakup isi dan kompetensi yang hendak diukur.[34]
2. *Construct validity* adalah jenis validitas yang menunjukkan sampai sejauh mana suatu tes mampu mengukur suatu trait atau konstruk teoritis (biasa juga disebut sebagai *latent variable*) yang hendak diukur. Atau: validitas konstruk adalah sejauh mana skor-skor hasil pengukuran dari suatu instrumen merefleksikan konstruksi teoritis yang mendasari penyusunan instrumen tersebut.[35]

   Validitas konstruk diestimasi melalui indikator-indikatornya (biasa juga disebut sebagai *observed-variable*) dengan analisis statistik yang cukup

[32] Sudijono, *Pengantar Evaluasi Pendidikan*, 191.
[33] Purwanto, *Evaluasi Hasil Belajar*, 120.
[34] Purwanto, *Prinsip-Prinsip dan Teknik Evaluasi*, 178-180.
[35] Purwanto, *Evaluasi Hasil Belajar*, 121.

rumit (analisis faktor (gunakan SPSS atau *structure equation modeling)* atau validitas konvergen dan diskriminan).[36]

Ada dua metode untuk menguji validitas konstruk yaitu: (a) dengan metode statistik analisis faktor (*confirmatory*); (b) dasar fikiran validasi konvergen dan diskriminan adalah: suatu tes harus berkorelasi tinggi dengan variabel-variabel yang secara teori memang harus berkorelasi tinggi (validasi konvergen) dan sekaligus tes itu tidak berkorelasi dengan variabel-variabel lain yang secara teori memang tak berkorelasi (validasi diskriminan).[37]

3. *Criterion related validity*. Suatu tes dipandang valid apabila skor tes tersebut memiliki korelasi dengan skor dari suatu kriterium (tes lain yang mengungkap hal yang sama) yang berada di luar tes. Untuk mengetahui apakah antara skor tes dengan skor kriterium memiliki korelasi digunakan analisis statistik.[38]

   Berdasar atas kapan skor kriterianya diperoleh, maka *criterion related validity* ini ada dua macam, yaitu *predictive validity* dan *concurrent validity*.[39]

   Pertama, *predictive validity* adalah jenis validitas yang menggunakan kriterium berupa skor performansi subyek diwaktu mendatang. Oleh sebab itu validitas ini sangat penting artinya apabila suatu tes dimaksudkan sebagai prediktor (untuk memprediksi atau meramalkan) performansi subyek diwaktu mendatang. Misalnya skor tes masuk yang diperoleh calon mahasiswa digunakan untuk memprediksi Indeks Prestasi Kumulatif (IPK) mahasiswa tersebut setelah ia menempuh kuliah. Jadi tes masuk suatu Perguruan Tinggi baru bisa diuji validitasnya setelah diperoleh IPK mahasiswa. Cara pengujiannya atau proses validasinya adalah dengan mengkorelasikan skor tes masuk dengan skor IPK yang diperoleh dengan menggunakan teknik korelasi *product moment*.[40]

   Kedua, *concurrent validity* adalah jenis validitas yang skor kriteriumnya diperoleh dalam waktu yang sama dengan skor suatu tes/alat ukur lain. Dengan sendirinya alat ukur yang dipakai sebagai kriterium haruslah mengungkap hal yang sama dengan alat ukur yang akan diestimasi validitasnya. Suatu alat ukur secara konkuren dipandang valid apabila antara skor alat ukur tersebut berkorelasi dengan skor kriteriumnya.[41]

---

[36] Ibid., 121-122.
[37] Ibid., 122-123.
[38] Nurkancana dan Sunartana, *Evaluasi Pendidikan*, 128.
[39] Ibid.
[40] Ibid.
[41] Purwanto, *Evaluasi Hasil Belajar*, 121-125; Nurkancana dan Sunartana, *Evaluasi Pendidikan*, 129.

Dari 3 (tiga) jenis validitas di atas yang proses validasinya dengan menggunakan teknik statistik korelasi adalah *Criterion Related Validity*. Caranya adalah dengan mengkorelasikan antara skor tes dengan skor kriterium sekelompok subyek dengan menggunakan teknik korelasi *product moment*. Koefisien korelasi antara dua perangkat skor (tes) itu disebut *koefisien validitas*. Karena koefisien validitas diperoleh dengan cara korelasi maka orang melakukan uji signifikansi untuk menafsirkan koefisien validitas tersebut. Ini tidak benar. Koefisien validitas harus ditafsirkan dari koefisien determinasi, yaitu angka yang menunjukkan proporsi (persentase) varians suatu variabel yang dapat dijelaskan dari variabel lainnya. Makin tinggi angka ini berarti kecermatan prediksinya makin tinggi pula. Cara meningkatkan koefisien determinasi adalah dengan menambah prediktornya.[42]

Sekalipun untuk *content validity* tidak menuntut perhitungan statistik bagi proses validasinya, namun umumnya orang mencari daya beda item-item dalam suatu alat ukur (yang juga menggunakan teknik korelasi), yang secara tidak langsung juga mengindikasikan validitasnya. Teknik statistik yang digunakan tergantung pada jenis data variabelnya; tapi umumnya adalah teknik korelasi *product moment*, diikuti koreksi *part-whole*, atau teknik korelasi *point-biserial*.[43]

Berikut ini adalah kesimpulan tentang jenis validitas dan cara estimasinya.

Jenis Validitas	Estimasi Melalui	Analisis	Keterangan
Validitas isi: (1) Validitas tampang (2) Validitas logis	(1) Pengujian isi tes dengan analisis rasional (2) Butir tes nampak sesuai dengan variabel (3) Item mencerminkan aspek dari variabel	Konsistensi internal, biasanya dipakai teknik korelasi	(1) Kebutuhan praktis (2) Variabel dibatasi secara jelas, tegas dan konkrit
Validitas konstruk	Konstruk teori (*latent variable*) melalui indikatornya (*observed-variable*)	Analisis faktor atau menggunakan *structure equation model*	
Validitas berkorelasi dengan kriteria: (1) Validitas prediktif (2) Validitas konkuren	(1) Korelasi dengan suatu kriterium (2) Korelasi dengan kriterium performansi subjek mendatang (3) Korelasi dengan kriterium (tes lain) pada waktu yang sama	Teknik korelasi *product moment*	Penting untuk prediksi

[42] Purwanto, *Evaluasi Hasil Belajar* , 123-125.
[43] Ibid.

Analisis Validitas Butir Soal

Validitas adalah salah satu ciri yang menandai tes hasil belajar yang baik. Untuk dapat menentukan apakah suatu tes hasil belajar telah memiliki suatu validitas atau daya ketepatan mengukur, kita dapat melakukannya dari dua segi, yaitu: (1) dari segi tes itu sendiri sebagai suatu totalitas, dan (2) dari segi itemnya sebagai bagian tak terpisahkan dari tes tersebut.[44] Sementara dalam menentukan reliabilitas tes hasil belajar, hal yang paling pokok adalah bagaimana cara kita menentukan reliabilitas tes hasi belajar bentuk uraian dan bagaimana pula menentukan reliabilitas tes hasil belajar bentuk obyektif.

Pengujian Validitas Tes

Penganalisaan terhadap tes hasil belajar sebagai suatu totalitas dapat dilakukan dengan dua cara. Pertama, penganalisisan yang dilakukan dengan jalan berfikir secara rasional atau penganalisisan dengan menggunakan logika *(logical analysis)*. Kedua, penganalisisan yang dilakukan dengan mendasarkan diri kepada kenyataan empiris, di mana penganalisisan dilakukan dengan menggunakan *empirical analysis.*[45]

a. Pengujian validitas tes secara rasional

Validitas rasional dapat juga kita pahami sebagai penganalisisan tes hasil belajar secara rasional yang teryata memiliki daya ketepatan mengukur *(logical validity)*. Ketika kita menentukan apakah tes hasil belajar sudah memiliki validitas rasional ataukah belum, maka perlu dilakukan suatau penelusuran dari dua segi, yaitu dari segi isinya *(content)* dan dari segi susunan atau konstruksinya *(construct).*[46]

1) Validitas isi *(content validity)*. Validitas isi dari suatu tes hasil belajar adalah validitas yang diperoleh setelah dilakukan penganalisisan, penelusuran atau pengujian terhadap isi yang terkandung dalam tes hasil belajar tersebut. Validitas isi adalah validitas yang ditilik dari segi isi tes itu sendiri sebagai alat pengukur hasil belajar, artinya bahwa sejauh manakah tes hasil belajar sebagai alat pengukur hasil belajar siswa/i isinya telah dapat mewakili secara representatif terhadap keseluruhan materi atau bahan pelajaran yang seharusnya diujikan atau diteskan. Karenanya, validitas isi sebenarnya identik dengan populasi dan sampel.[47]

   Dalam prakteknya validitas isi dari suatu tes hasil belajar dapat diketahui dengan cara membandingkan antara isi yang terkandung dalam tes hasil belajar dengan Kompetensi Dasar (KD) pada masing-

[44] Sudijono, *Pengantar Evaluasi Pendidikan*, 163-164.
[45] Ibid.
[46] Ibid,168.
[47] Thoha, *Teknik Evaluasi Pendidikan*, 111.

masing mata pelajaran yang ada. Upaya lain yang dapat digunakan dalam rangka mengetahui validitas isi adalah dengan jalan menyelenggarakan diskusi panel.[48]

2) Validitas konstruksi *(construct validity)*. Tes hasil belajar disebut sebagai validitas konstruksi apabila tes hasil belajar tersebut telah memiliki validitas susunan butir-butir soal atau item yang membangun tes tersebut secara tepat dapat mengukur aspek-aspek berfikir (aspek kognitif, afektif dan psikomotorik) dalam Kompetensi Dasar (KD). Validitas konstruksi ini juga diketahui dengan cara menggelar diskusi panel.[49]

b. Pengujian validitas tes secara empirik

Validitas empirik adalah ketepatan mengukur yang didasarkan pada hasil analisis yang bersifat empirik (validitas yang bersumber atas dasar pengamatan di lapangan). Dalam menentukan validitas empirik ini dapat dilakukan melalui dua langkah, pertama dari segi daya ketepatan meramalnya *(predictive validity)* dan daya ketepatan bandingannya *(concurrent validity)*.[50]

1) Validitas ramalan *(predictive validity)*. Validitas ramalan dari suatu tes adalah suatu kondisi yang menunjukkan seberapa jauhkan sebuah tes telah dapat dengan secara tepat menunjukkan kemampuannya untuk meramalkan apa yang bakal terjadi pada masa mendatang. Misalnya tes seleksi penerimaan calon Mahasiswa baru pada sebuah perguruan tinggi merupakan tes yang diharapkan mampu meramalkan keberhasilan studi para calon mahasiswa dalam mengikuti program pendidikan di perguruan tinggi tersebut pada masa-masa yang akan datang.[51]

   Kemudian untuk mengetahui apakah suatu tes hasil belajar dapat dinyatakan sebagai tes yang memiliki validitas ramalan ataukah belum maka perlu mencari korelasi antara tes hasil belajar yang sedang diuji validitas ramalannya dengan kriterium yang ada. Jika diantara kedua variabel tersebut terdapat korelasi positif yang signfikan, maka tes hasil belajar yang sedang diuji validitas ramalannya itu, dapat dinyatakan sebagai tes hasil belajar yang telah memiliki daya ramal yang tepat (benar-benar terjadi secara nyata dalam praktek).[52]

   Cara sederhana yang paling sering digunakan adalah dengan menerapkan teknik analisis korelasional *product moment* dari Karl Pearson, yaitu: (a) hipotesis nihil (H0) yang akan diuji; (b) tidak terdapat korelasi positif yang signifikan antara tes hasil belajar yang

[48] Ibid.
[49] Sudijono, *Pengantar Evaluasi Pendidikan*, 166.
[50] Ibid.
[51] Arikunto, *Dasar-Dasar Evaluasi Pendidikan*, 66.
[52] Ibid, 67.

sedang diuji validitas ramalannya (variabel X); (c) kriteria yang telah ditentukan (variabel Y).[53]

2) Validitas bandingan (*predictive validity*). Selama ini tes dijadikan sebagai alat pengukur dapat dikatakan telah memiliki validitas bandingan apabila tes tersebut dalam kurun waktu yang sama dengan secara tepat telah mampu menunjukkan adanya hubungan yang searah, antara tes tes pertama dan terakhir. Validitas bandingan disebut juga dengan istilah validitas pengalaman yang ada pada saat sekarang ini.[54] Seperti halnya validitas ramalan, maka untuk mengetahui ada tidaknya hubungan yang searah antara tes yang pertama dengan tes berikutnya, dapat digunakan teknik analisis korelasional *product moment* dari Karl Pearson, yaitu: jika korelasi antara variabel X (tes pertama) dengan variabel Y (tes berikutnya) adalah positif dan signifikan, maka tes tersebut dapat dinyatakan sebagai tes yang telah memiliki validitas bandingan.[55]

c. Pengujian validitas item tes hasil belajar

Validitas item dari suatu tes adalah ketepatan mengukur yang dimiliki oleh sebutir item (yang merupakan bagian tak terpisahkan dari tes sebagai suatu totalitas), dalam mengukur apa yang seharusnya diukur lewat butir item tersebut. Karenanya validnya suatu tes akan sangat tergantung pada validitas yang dimiliki oleh masing-masing butir item yang membangun tes tersebut.[56]

Sebutir item dapat dikatakan telah memiliki validitas yang tinggi atau dapat dinyatakan valid, jika skor-skor pada butir item yang bersangkutan memiliki kesesuaian atau kesejajaran arah dengan skor totalnya. Dalam bahasa "statistik" dapat pula dinyatakan: "ada korelasi positif yang signifikan antara skor item (variabel bebas atau *independent variable*) dengan skor totalnya (variabel terikat atau *dependent variable*)."[57]

Pengujian Reliabilitas Tes

Selain berfungsi sebagai alat pengukur hasil belajar, tes hasil belajar dapat dibedakan menjadi dua golongan, yaitu: tes hasil belajar bentuk uraian yang lebih dikenal dengan istilah *essay test* atau *subjektive test*, dan tes hasil belajar bentuk objektif yang dikenal dengan istilah *objektive test* atau *new type test.* (1) Pengujian reliabilitas tes hasil belajar bentuk uraian. Dalam teknik pengujian reliabilitas tes hasil belajar bentuk uraian umumnya

[53] Ibid, 68.
[54] Nana Sudjana, *Penilaian Hasil Proses Belajar Mengajar* (Bandung: PT. Remaja Rosda Karya, 2005), 15-16.
[55] Sudijono, *Evaluasi Pendidikan*, 170.
[56] Ibid, 182.
[57] Ibid, 183.

digunakan rumus Alpha.[58] (2) Pengujian reliabilitas tes hasil belajar bentuk obyektif.

Pada penentuan reliabilitas tes dapat dilakukan dengan menggunkan tiga macam pendekatan, yaitu: (1) pendekatan *single test-single trial method*, (2) pendekatan *single test-double trial method*, dan (3) pendekatan *double test-double test method*.[59]

1. Pengujian reliabilitas tes hasil belajar bentuk obyektif dengan menggunakan pendekatan *single test-single trial method*. Pendekatan *single test-single trial* memungkinkan tinggi rendahnya reliabilitas test hasil belajar bentuk obyektif dapat diketahui dengan melihat besar kecilnya koefisien reliabilitas tes, yang pada tes uraian dilambangkan dengan: r_{11} atau r_{tt} (koefisien reliabilitas tes secara total). Adapun untuk mencari atau menghitung r_{11} atau r_{tt} dapat digunakan lima jenis formula, yaitu: (1) formula Spearman-Brown, (2) formula Flanagan, (3) formula Rulon, (4) formula Kuder-Richardson dan (5) formula C. Hoyt.[60]
2. Pengujian reliabilitas tes hasil belajar bentuk obyektif dengan menggunakan pendekatan pendekatan *single test-double trial method*.[61] Pada pendekatan *single test-double trial* atau pendekatan *test-retest* ini, sering juga dikenal dengan sebutan istilah pendekatan bentuk ulangan, maka penentuan reliabilitas tes dilakukan dengan menggunakan teknik ulangan, di mana tester hanya menggunakan satu seri tes, tetapi percobaannya dilakukan sebanyak dua kali. Karenanya pendekatan ini sering dikenal dengan istilah *single test-double trial method.*[62]
3. Pengujian reliabilitas tes hasil belajar bentuk obyektif dengan menggunakan pendekatan pendekatan *alternate form* (*double test-double test method*).[63]

Penentuan reliabilitas tes dengan menggunakan pendekatan *alternate form* sering juga dikenal dengan istilah pendekatan bentuk paralel. Pendekatan jenis ketiga ini dipandang lebih baik daripada dua jenis pendekatan yang tersebut di atas dengan alasan bahwa: (a) karena butir-butir item dibuat sejenis tetapi tidak sama, maka tes hasil belajar (yang akan diujikan reliabilitasnya) dapat terhindar dari kemungkinan timbulnya pengaruh yang datang dari *testee* (latihan atau menghafal); (b) karena kedua tes itu dilaksanakan secara berbareng (paralel), maka dapat dihindarkan timbulnya perbedaan-perbedaan situasi dan kondisi ynag

[58] Ibid, 252.
[59] Ibid, 124-130.
[60] Chabib Thoha, *Teknik Evaluasi Pendidikan* (Jakarta: Rajawali, 1991), 124.
[61] Sudijono, *Evaluasi Pendidikan*,269.
[62] Ibid, 270.
[63] Ibid, 271.

diperkirakan akan dapat mempengaruhi pelaksanaan tes, baik yang bersifat sosial maupun yang bersifat alami.[64]

Reliabilitas Alat Ukur

Reliabilitas alat ukur menunjukkan sejauh mana hasil pengukuran dengan alat tersebut dapat dipercaya. Hal ini dapat ditunjukkan oleh taraf keajegan (konsistensi) skor yang diperoleh subjek yang diukur dengan alat ukur yang sama pada kondisi yang berbeda.[65] Dalam arti yang paling luas, reliabilitas alat ukur menunjuk pada sejauh mana perbedaan-perbedaan skor perolehan itu mencerminkan perbedaan-perbedaan atribut yang sebenarnya. Jadi reliabilitas tes adalah proporsi varians skor perolehan yang merupakan varians skor murni. Karena reliabilitas alat ukur berkenaan dengan derajat konsistensi dua perangkat skor, maka dia dinyatakan dalam bentuk koefisien korelasi, tapi tidak perlu diuji signifikansinya.[66]

Estimasi Reliabilitas

Reliabilitas alat ukur tak dapat ditentukan dengan pasti, melainkan hanya dapat diestimasi. Ada tiga pendekatan untuk mengestimasi reliabilitas alat ukur yaitu (1) pendekatan tes ulang (cocok untuk keterampilan fisik), (2) pendekatan dengan tes paralel (sulit menyusun tes paralel), dan (3) pendekatan satu kali pengukuran (menghasilkan informasi mengenai keajegan/ konsistensi internal alat ukur).[67]

a. *Test-retest approach*. Pendekatan *test-retest* mengenakan satu test dua kali pada sekelompok subjek dengan jarak waktu. Dua kelompok skor yang diperoleh dari dua kali pengetesan kemudian dikorelasikan (PM). Kelemahan *test-retet* adalah (a) kemungkinan terjadinya *carry over effect*, (b) *rejection* dari subjek, (c) hanya cocok untuk mengukur aspek fisik atau aspek psikologis yang relatif stabil.[68]
b. *Paralel form* (*alternate form*) *approach.* Pendekatan *paralel form*: dua tes yang sama tujuan ukurnya dan setara kualitas dan kuantitas isi itemnya, diberikan kepada kelompok subjek yang sama pada waktu yang bersamaan. Dua kelompok skor yang diperoleh dari dua tes paralel tersebut kemudian dikorelasikan (PM) untuk mengestimasi reliabilitasnya.[69]

   Kelemahan: (1) sukarnya menyusun dua tes yang paralel (spesifikasinya harus sama: indikatornya, jumlah item, format item, taraf kesukaran

[64] Ibid.
[65] Suryabrata, *Metodologi Penelitian*, 28.
[66] Ibid.
[67] Arikunto, *Dasar-Dasar Evaluasi Pendidikan*, 220-222; Saifuddin Azwar, *Metode Penelitian* (Yogyakarta: Pustaka Pelajar, 1998), 36- 43.
[68] Ibid.
[69] Arikunto, *Dasar-Dasar Evaluasi Pendidikan*, 220-222.

item); (2) pengalaman subjek mengerjakan tes yang pertama dapat meningkatkan kinerja subjek pada tes yang kedua karena adanya faktor belajar.[70]

Untuk mengatasi ini dikembangkan pendekatan ganjil-genap (*odd-even splits*) (ini termasuk kelompok konsistensi internal), artinya dua tes paralel tersebut digabung menjadi satu tes saja dengan memberikan nomor urut item pada tes pertama dengan nomor ganjil dan nomor genap untuk item-item dari tes yang kedua. Setelah data diperoleh, maka data dipisah menjadi dua kelompok berdasar nomor ganjil dan nomor genap tadi, selanjutnya kelompok skor item ganjil dikorelasikan dengan kelompok skor item genap untuk mengestimasi reliabilitasnya.[71]

c. *Internal consistency approach.* Satu tes diberikan sekali pada sekelompok subjek. Keuntungan pendekatan ini adalah praktis dan efisien karena hanya dilaksanakan satu kali pengetesan. Item-item dari tes tersebut dibelah menjadi dua, tiga, atau empat belahan, bahkan banyaknya belahan bisa sebanyak jumlah item yang terdapat dalam tes tersebut (misal Anava Hoyt). Estimasi reliabilitasnya adalah dengan melihat konsistensi antar item atau kelompok-kelompok item dalam tes itu sendiri Cara pengelompokan item (disebut belahan tes) tergantung pada jenis tes (*speed* atau *power test*), homogenitas, dan taraf kesukaran item. Belahan tes tersebut diusahakan setara agar tidak terjadi *under estimate* atau *over estimate*. Cara pembelahan tes menjadi dua: (a) secara random, (b) ganjil-genap, (c) *matched random subsets*.[72]

Berikut ini adalah macam-macam pendekatan belah dua yang perlu diketahui dan diperhatikan dengan saksama.[73]

a. Formula Spearman-Brown[74]

Pendekatan belah dua yang populer adalah formula Spearman-Brown. Formula ini mengasumsikan kedua belahan tes adalah paralel (*mean* dan varian setara). Keuntungan formula ini dapat dipakai baik untuk skor dikotomi maupun nondikotomi.

Subjek	Belahan 1	Belahan 2
1 2 . 10		

[70] Ibid.
[71] Ibid.
[72] Soetarlinah Sukadji, *Menyusun dan Mengevaluasi Laporan Penelitian* (Jakarta: UI-Press, 2000), 31-32; Azwar, *Metode Penelitian,* 45-53.
[73] Sukadji, *Menyusun dan Mengevaluasi Laporan Penelitian*, 31-32; Azwar, *Metode Penelitian,* 45-53.
[74] Thoha, *Teknik Evaluasi Pendidikan,* 119-141.

r_{12} = (rumus PM) -------→ *under estimate* karena panjang tes hanya setengah (20 menjadi 10), oleh karena itu perlu dikoreksi $r_{xx'} = 2\ r_{12}/(1 + r_{12})$ -----→ Spearman Brown (Semakin panjang tes maka koefisien reliabilitasnya juga semakin meningkat).

b. Formula Rulon[75]

Formula Rulon tanpa harus berasumsi bahwa kedua belahan homogen variansnya.

Rumus estimasi reliabilitasnya: $r_{xx'} = 1 - Vd/ Vx$

dimana: Vd = varians beda skor kedua belahan

Vx = varians skor total

Contoh:

Subjek	Belahan 1	Belahan 2	Beda skor (d)
1 2 . 10			

c. Formula Flanagan

rtt = 2 (1- (V1 + V2/ Vt))

V1 = varians belahan 1

V2 = varians belahan 2

Vt = varians total

d. Koefisien Alpha[76]

Untuk dua belahan, pendekatan reliabilitas dengan menggunakan koefisien alpha tidak harus memenuhi asumsi bahwa kedua belahan adalah setara mean dan variansnya, namun syaratnya r harus ekuivalen; kalau tidak akan terjadi *under etimate* terhadap koefisien alpha yang diperoleh.

Rumus koefisien alpha: $\alpha = 2\ (1 - (V_1 + V_2/ V_t))$

dimana: V_1 = varians skor belahan 1

V_2 = varians skor belahan 2

V_t = varians skor total

Untuk belahan lebih dari dua, syarat antar belahan paralel harus dipenuhi atau paling tidak memenuhi r (tau) ekuivalen.

Rumusnya: $\alpha = (k/k\text{-}1)(1\text{-}\Sigma V_j/V_t)$

dimana: k = banyaknya belahan

V_j = varians belahan j

V_t = varians skor total

e. Formula Kuder-Richardson[77]

Formula K-R merupakan modifikasi dari formula Alpha. Digunakan apabila skornya dikhotomi dan jumlah itemnya sedikit, banyaknya belahan

[75] Ibid.
[76] Ibid.
[77] Ibid.

sebanyak jumlah itemnya. Rumusnya: K-20 = (k/k-1) (1-(Σp (1-p)/ V_t)) atau (k/ k-1) (1 (V_t + Σpq/ V_t)) di mana:

k = banyaknya item

V_t = varians skor total

p = proporsi subjek yang mendapat skor 1 pada suatu item

f. Formula Kristof[78]

Banyaknya item yang tidak genap menyebabkan tes tidak dapat dibelah menjadi dua dengan mempertahankan asumsi paralel. Untuk mengatasinya dapat dipakai formula Kristof, dengan membelah tes menjadi tiga, dan tidak menuntut masing-masing belahan jumlah itemnya sama. Rumusnya:

$s_m^2 = s_{12}s_{13}/ s_{23} + s_{12}s_{23}/ s_{13} + s_{13}s_{23}/ s_{12} + 2 (s_{12} + s_{13} + s_{23})$

dimana: s_m^2 = estimasi terhadap varians skor murni

s_{12} = kovarians belahan 1 dan 2

s_{13} = kovarians belahan 1 dan 3

s_{23} = kovarians belahan 2 dan 3

$rtt = s_m^2/ v_x$

Contoh:

Subjek	Nomor Item	Belahan	X
	1 2 3 4 5 6 7 8 9 10 11 12	1 2 3	
A	2 3 3 1 2 4 4 3 3 4 2 3	9 13 12	34
B			41
C			30
dst.			

g. Menggunakan analisis variansi dari Hoyt[79]

Analisis untuk mengestimasi reliabilitasnya analog dengan rancangan eksperimen faktorial dua jalur tanpa replikasi (*treatment by subject design*), di sini disebut *item by subject design*.

Rumusnya: $r_{xx'} = 1 – MK_{is}/ MK_s$

dimana: MKis = Mean Kuadrat interaksi item dan subjek

MKs = Mean Kuadrat antar subjek

Makna koefisien reliabilitas.[80] Skor yang diperoleh subjek dalam mengerjakan tes umumnya bukanlah skor yang sebenarnya (skor murni) karena adanya kesalahan pengukuran. Melalui koefisien reliabilitas dapat diestimasi letak skor murni dalam rentang wilayah (interval) tertentu. Banyak orang mmberi makna atau tafsiran koefisien reliabilitas dengan melalui uji signifikansi. Ini tidak benar, karena koefisien reliabilitas harus ditafsirkan melalui *standard error of measurement* dengan rumus:

SEm = SD√1-rtt

(SEm kecil berarti reliabilitas tinggi)

[78] Ibid.
[79] Ibid.
[80] Arikunto, *Dasar-Dasar Evaluasi Pendidikan*, 220-222. Azwar, *Metode Penelitian*, 36- 43.

Misal WAIS: Mean = 100; SD = 15; dan rtt = 0,96

SEm = 15√1-0,96

Tidak ada harga mati untuk koefisien reliabilitas. Tinggi rendahnya koefisien reliabilitas yang diperlukan bergantung pada tujuan penerapan tes; kalau untuk membuat keputusan bagi seseorang maka direkomendasikan paling tidak 0,90.

Berikut kesimpulan tentang tabel jenis reliabilitas dan cara estimasinya.

Jenis Reliabilitas	Teknik	Banyaknya Belahan (Kelompok Skor)	Keterangan
Test-retest	*Product-moment*	Dua	*Carry over effect, reject,* aspek stabil
Paralel/alternate form	*Product-moment*	Dua	Sukar menyusun dua belahan paralel, *learning effect*
Konsistensi Internal: (a) Spearman-Brown (b) Rulon (c) Alpha (d) K-R (e) Kristof (f) Hoyt	Konsistensi item Korelasi Korelasi Koefisien Alpha KR-20; KR-21 Korelasi *Anava*	Dua atau lebih Dua Dua Dua atau lebih Sebanyak jumlah item Tiga Sebanyak jumlah item	Praktis, efisien Dua belahan paralel, dapat untuk skor dikhotomi Varians tidak harus homogen Belahan tidak harus setara Dikhotomi, item sedikit Atasi jumlah item ganjil Belahan sebanyak item

Catatan Akhir

Berdasarkan serangkaian pemaparan tulisan ini tersebut di atas kiranya kami dapat memberikan kesimpulan konkret mengenai kajian *Analisis Kualitas Tes dan Butir Soal*. Si dalamnya juga disajikan analisis validitas butir soal dan analisis reliabilitas sebuah tes. Di samping itu, guna melengkapi pemahaman kita tentang seputar analisis kualitas tes, di bagian lampiran dilengkapi contoh-contoh soal terkait masalah ini sehingga diharapkan mengantarkan pembaca pada pemahaman yang paripurna.

Berikut kesimpulan yang dapat kami berikan berdasarkan ulasan singkat pada diskusi kita hari ini:

1. Analisis soal sesungguhnya bertujuan untuk mengadakan identifikasi soal-soal yang baik, kurang baik, dan soal yang jelek. Dengan analisis soal dapat diperoleh informasi tentang kejelekan sebuah soal dan "petunjuk" untuk mengadakan perbaikan dalam pembelajaran.

2. Kualitas tes dan butir soal sangat ditentukan oleh: (1) validitas; (2) reliabilitas; (3) objektivitas; (4) praktikabilitas; (5) daya pembeda; (6) taraf atau derajad kesukaran; (7) efektivitas *option*; dan (8) efisiensi.[]

Daftar Rujukan

Arikunto, Suharsimi. *Dasar-dasar Evaluasi Pendidikan.* Jakarta: Bumi Aksara, 1996.

Daryanto. *Evaluasi Pendidikan.* Jakarta: Rineka Cipta, 1999.

Djaali dan Pudji Muljono. *Pengukuran dalam Bidang Pendidikan*. Jakarta: Grasindo, 2008.

Mulyasa, Enco. *KBK: Konsep, Karakteristik dan Implementasinya.* Bandung: Remaja Rosdakarya, 2003.

Mulyasa, Enco. *Kurikulum Tingkat Satuan Pendidikan (KTSP).* Bandung: PT. Remaja Rosdakarya, 2007.

Natar, Karjono. *Panduan Analisis Butir Soal.* Lampung: UNILA Press, 2011.

Nurkancana, Wayan dan P.P.N. Sunartana. *Evaluasi Pendidikan* (Cetakan ke-4). Surabaya: Usaha Nasional, 1986.

Purwanto. *Evaluasi Hasil Belajar.* Yogyakarta: Pustaka Pelajar, 2011.

Purwanto, Ngalim. *Prinsip-prinsip dan Teknik Evaluasi Pengajaran.* Bandung: Remaja Rosdakarya, 2003.

Subana, M. dan Sudrajat. *Dasar-dasar Penelitian Ilmiah.* Bandung: Pustaka Setia, 2005.

Sudijono, Anas. *Pengantar Evaluasi Pendidikan.* Jakarta: RajaGrafindo, 2003.

Sudjana, Nana. *Penilaian Hasil Proses Belajar Mengajar*. Bandung: PT. Remaja Rosda Karya, 2005.

Sukadji, Soetarlinah. *Menyusun dan Mengevaluasi Laporan Penelitian.* Jakarta: UI-Press, 2000.

Suryabrata, Sumadi. *Metodologi Penelitian.* Jakarta: RajaGrafindo Persada, 2004.

Thoha, Chabib. *Teknik Evaluasi Pendidikan.* Jakarta: Rajawali, 1991.

Thorndike, Robert L. dan Elizabeth. "Measurement and Evaluation in Psychology and Education." dalam Anas Sudijono. *Pengantar Evaluasi Pendidikan.* Jakarta: RajaGrafindo, 2003.

Tilaar, H. A. R. *Menuju Pendidikan* Nasional*: Kajian Pendidikan Masa Depan.* Bandung: PT. Remaja Rosdakarya, 1994.

www.depdiknas.go.id.