

An illustration of a classroom scene. A female teacher with short dark hair, wearing a dark button-down shirt, stands and holds a piece of paper. She is looking towards a male student on the left who is also holding a paper. In the foreground, the back of a student's head is visible as they sit at a desk. To the right, another student is partially visible, sitting at a desk. A dark rectangular shape, possibly a chalkboard, is in the background on the left.

BAB 3

VALIDITAS DAN RELIABILITAS

Wainer & Braun (1998:20) memberikan tiga syarat suatu instrumen penilaian yang bermutu, yaitu *valid*, *reliabel*, dan *usabel*. Validitas penilaian terkait dengan kemampuan instrumen penilaian itu menyajikan informasi yang tepat tentang kondisi anak yang dinilai. Dalam bahasa lain, seberapa akurat (*appropriateness*), bermakna (*meaningfulness*) dan berguna (*usefulness*) kesimpulan yang diambil dari hasil penilaian. Informasi ini tentu sangat bermanfaat untuk menangani anak yang bersangkutan. Bila suatu penilaian tidak valid, maka informasi yang diperoleh akan menyesatkan. Dalam istilah kedokteran kondisi yang demikian biasanya akan menyebabkan “malpraktek”.

**Tiga syarat
suatu instrumen
penilaian yang
bermutu, yaitu
valid, reliabel,
dan usable.**

50

tembakannya: (a) mengumpul di tempat tertentu, akan tetapi jauh dari titik sasaran yang telah ditetapkan; (b) menyebar pada bidang sasaran, tidak fokus pada titik yang telah ditetapkan; (c) mendekati atau mengumpul pada titik yang telah ditetapkan. Ilustrasi hasil tembakan tampak pada Gambar 3.1.



Gambar 3.1: Ilustrasi Sasaran

Dalam konteks penilaian, penilaian yang reliabel sangat diperlukan namun belum cukup untuk mendapatkan hasil yang valid. Artinya, masih diperlukan syarat lain yang harus dipenuhi agar penilaian tersebut valid. Sebaliknya, penilaian yang valid dapat dipastikan akan memberikan hasil yang reliabel.

Usabel memiliki makna penilaian yang digunakan praktis prosedurnya. Artinya, penilaian tidak menyulitkan bagi siapa pun yang melaksanakannya. Petunjuk harus dibuat sedemikian rupa hingga mudah dipahami, singkat, dan jelas. Dengan kata lain, penilaian tidak membingungkan bagi siapa pun yang terlibat, mulai dari siswa, guru yang melaksanakan, maupun orang yang mengolah nilai. Tujuannya, apabila penilaian dilaksanakan oleh guru atau orang lain, penilaian itu mudah dilakukan. Tidak memengaruhi hasil dan tidak menimbulkan banyak penafsiran sehingga substansi yang diinginkan tidak meleset.

A. Validitas Penilaian

Validitas memiliki beberapa karakteristik. Pertama, validitas merujuk pada *ketepatan interpretasi hasil* suatu penilaian yang dikenakan terhadap siswa, bukan merujuk pada instrumen penilaian tersebut. Untuk memudahkan, memang sering kali orang menyebutnya dengan istilah "validitas penilaian." Namun, lebih tepat validitas merujuk pada kesimpulan yang dibuat dari hasil penilaian.

Kedua, validitas berkaitan dengan pengategorian derajat (*degree*) tertentu, seperti validitasnya tinggi, sedang, atau rendah. Jadi, harus dihindari pemikiran tentang hasil penilaiannya valid dan tidak valid.

Ketiga, validitas senantiasa berkaitan dengan kondisi khusus. Artinya, tidak ada penilaian yang valid untuk semua tujuan. Sebagai contoh, hasil penilaian aritmetika mungkin memiliki validitas yang tinggi dalam hal kemampuan berhitung. Namun, hasil penilaian tersebut memiliki validitas yang rendah dalam hal penalaran aritmetika dan sedang untuk memprediksi keberhasilan mengikuti pelajaran matematika di masa depannya.

Sangat mungkin penilaian tersebut tidak valid untuk memprediksi keberhasilan anak dalam bidang musik atau seni. Jadi, ketika mendeskripsikan validitas, penting untuk mempertimbangkan penafsiran secara khusus tentang hasil penilaian. Hasil penilaian tidak pernah dikatakan valid saja, akan tetapi hasil penilaian tersebut memiliki derajat validitas berbeda pada setiap penafsiran yang dibuat.

1. Berbagai Faktor yang Memengaruhi Validitas

Terdapat sejumlah faktor yang bisa memengaruhi validitas penilaian. Pertama, karakteristik dan kondisi siswa. Unsur-unsur fisik dan psikologis sangat memengaruhi siswa pada saat dinilai

sehingga memengaruhi validitas hasil penilaian. Sebagai contoh, jika seorang siswa pada waktu diuji memiliki tingkat kecemasan tinggi, tentu hal itu akan memengaruhi hasil penilaian dibandingkan dalam kondisi normal. Anak yang sedang sakit dan lapar, bila dinilai juga akan berbeda hasilnya bila mereka dalam kondisi sehat dan afiat. Demikian pula, siswa yang motivasinya rendah juga akan memberikan hasil berbeda. Artinya, hasil penilaian yang didapatkan dari anak dalam kondisi kurang normal, tidak akan mampu memotret kondisi sebenarnya dari anak yang bersangkutan.

Kedua, pelaksanaan penilaian dan prosedur penskoran. Pelaksanaan dan prosedur penskoran yang menyimpang dari ketentuan akan memengaruhi validitas penilaian. Artinya, bila penilaian tidak mampu memberikan petunjuk yang jelas dan tepat sehingga menimbulkan banyak penafsiran di kalangan siswa yang mengikutinya, maka akan mengurangi validitas penilaian. Waktu pelaksanaan yang tidak adil, misalnya ada anak yang diberikan waktu berbeda, juga dapat memengaruhi validitas penilaian. Terkait prosedur penskoran, penskoran yang bias dan tidak reliabel dapat mengurangi validitas. Untuk itulah, perlu dibuatkan pedoman penskoran atau rubrik penilaian.

Ketiga, proses pembelajaran. Cara guru menyampaikan materi dapat memengaruhi validitas. Jika selama pembelajaran, siswa banyak diajar dan dibimbing menyelesaikan soal-soal yang akan diujikan (khususnya penilaian dalam bentuk tes), maka validitas penilaian tersebut dipertaruhkan. Jadi, masalah utama dalam hal ini adalah ketika guru bukan mengajar materi, tetapi mengajar menyelesaikan soal yang akan diujikan.

2. Berbagai Pendekatan dalam Menentukan Validitas Penilaian

Terdapat beberapa macam pendekatan dalam mengklasifikasikan validitas penilaian. Salah satunya membagi validitas penilaian ke dalam validitas terkait isi (*content-related validity*), validitas terkait kriteria (*criterion-related validity*), dan validitas terkait konstruk (*construct-related validity*). Ketiga pendekatan ini dideskripsikan pada Tabel 3.1.

Tabel 3.1
Deskripsi dari Validitas Penilaian

Jenis	Maksud	Prosedur
Validitas terkait isi	Kemampuan instrumen penilaian mewakili domain yang hendak diukur secara tepat	Membandingkan instrumen penilaian dengan kisi-kisi
Validitas terkait kriteria	Kemampuan instrumen penilaian memprediksi potensi siswa di masa mendatang secara baik	Membandingkan skor hasil penilaian dengan skor yang didapatkan pada penilaian lain di masa yang akan datang atau dengan skor pada penilaian lain sekaligus.
Validitas terkait konstruk	Kemampuan instrumen penilaian mengukur pengertian-pengertian yang terkandung pada materi yang diukur. Tujuannya mendapatkan bukti tentang sejauh mana hasil pengukuran memerikan konstruk variabel yang diukur.	Mengkaji teori-teori yang berkaitan dengan konstruk yang relevan diukur alat penilaian yang dikembangkan. Berdasarkan telaah teori-teori tersebut diturunkanlah butir-butir penilaian.

a. Validitas Terkait Isi (*Content-Related Validity*)

Validitas ini berkaitan derajat kemampuan instrumen penilaian mengukur cakupan substansi yang ingin diukur. Dua aspek penting, yaitu valid isi dan valid teknik samplingnya. Valid isi mencakup

Agar validitas ini dapat dicapai, maka selama pengkonstruksian atau pengembangan instrumen penilaian disesuaikan dengan kisi-kisi. Artinya, perlu adanya keselarasan antara item yang ada pada instrumen penilaian dengan kisi-kisi. Oleh karena itu, penting kiranya mengembangkan kisi-kisi yang cermat sehingga cakupan isi yang dibidik benar-benar terwujud.

Validitas terkait kriteria dapat dimaknai sebagai kemampuan instrumen penilaian memprediksi kemampuan anak di masa mendatang. Caranya, hasil penilaian tersebut dibandingkan dengan alat ukur yang lain (disebut kriteria). Dari pengertian ini, maka instrumen pengukuran kedua (yang disebut kriteria) bisa diperoleh di masa mendatang (ketika tertarik memprediksi kemampuan di masa yang akan datang). Bisa saja menggunakan data hasil penilaian lain yang dilakukan secara hampir bersamaan (bila tertarik mengestimasi kemampuan saat ini). Untuk lebih mudah, perhatikan contoh berikut.

55

akibat dikenai perlakuan (*treatment*) tertentu. Jadi, berdasarkan teori yang mendasari kemampuan yang diukur, dapat diprediksi bahwa skor hasil penilaian tertentu akan berubah (tetap) dalam berbagai macam kondisi. Jika prediksi itu diuji kebenarannya, hasilnya akan mendukung lebih lanjut validitas konstruk.

B. Reliabilitas Penilaian

Reliabilitas merujuk pada konsistensi suatu pengukuran. Artinya, bagaimana hasil penilaian konsisten dari pengukuran yang satu ke lainnya. Misalkan, Bu Retno hari ini memberikan ulangan Bahasa Indonesia kepada murid-muridnya. Apakah hasil ulangan tadi bila diujikan minggu depan atau dua bulan lagi pada siswa yang sama, akan memberikan skor yang serupa? Pertanyaan ini merupakan pertanyaan yang berkaitan dengan reliabilitas.

Berdasarkan pengertian dan uraian contoh di atas, maka reliabilitas memiliki karakteristik berikut. Pertama, reliabilitas merujuk pada hasil yang diperoleh suatu instrumen penilaian, bukan merujuk kepada instrumennya. Suatu instrumen tertentu mungkin memiliki reliabilitas berbeda, tergantung pada kelompok yang terlibat dan situasi dalam instrumen penilaian itu. Jadi, lebih tepat menyebutnya dengan reliabilitas “hasil penilaian”.

Kedua, reliabilitas merupakan syarat perlu, tetapi belum cukup untuk syarat validitas. Suatu instrumen penilaian yang memberikan hasil tidak konsisten mungkin dapat memberikan informasi yang valid berkaitan dengan kemampuan yang diukur. Di sisi lain, hasil penilaian yang sangat konsisten bisa saja mengukur sesuatu yang salah atau digunakan dengan cara yang tidak tepat. Jadi, reliabilitas yang rendah dapat membatasi tingkat validitas yang didapat, tetapi reliabilitas yang tinggi

1. Beberapa Metode Mengestimasi Reliabilitas

Belah Dua (<i>Split half</i>)	Mengukur konsistensi internal	Satu format tes	Sekali diujikan	Memberikan tes sekali. Tes dibagi menjadi dua (misalkan butir tes genap dan ganjil), skor dari masing-masing paruhan dikorelasikan dengan rumus Spearman-Brown
Koefisien alpha atau Kuder Richardson	Mengukur konsistensi internal	Satu format tes	Sekali diujikan	Memberikan tes sekali. Skor total tes dikorelaskan dengan rumus Kuder-Richardson
Inter-rater	Mengukur konsistensi penskoran, bila sebuah tes diskor dua orang atau lebih	Satu format tes	Sekali diujikan	Menyelenggarakan tes pada sekelompok siswa dalam sekali waktu. Dua atau lebih <i>rater</i> atau penilai menyekor tes secara independen

b. Metode Bentuk Ekuivalen

Dalam tes prestasi terdapat ribuan pertanyaan atau soal yang bisa dimunculkan pada suatu tes tertentu. Namun, karena keterbatasan waktu dan beberapa faktor lain, hanya beberapa soal yang digunakan. Oleh karena itu, penyusun tes harus mampu mengambil sampel dari cakupan materi yang ingin diuji.

c. Metode Tes-Retes dengan Bentuk Ekuivalen

62

ekuivalensi. Bentuk ini merupakan cara terbaik dari menentukan reliabilitas karena menunjukkan kemantapan karakteristik anak yang diukur dan mampu merepresentasikan sampel dari materi yang diujikan.

Oleh karena itu, metode ini dianggap sebagai prosedur yang baik dalam mengestimasi reliabilitas. Seperti halnya pada metode tes-retes, koefisien reliabilitas harus diinterpretasikan secara jelas dalam interval waktu antar-kedua bentuk tes. Bila jeda waktu antar-kedua tes lama, maka koefisien reliabilitas yang diperoleh akan semakin kecil.

Metode ini memiliki kelebihan karena mampu mengurangi pengaruh ingatan yang dibawa anak pada saat mereka mengerjakan tes pertama. Namun demikian, pengaruh ingatan anak saat mengerjakan tes kedua tidak sepenuhnya bisa dihilangkan. Kelemahan metode ini adalah hanya sedikit tes (baik tes standar ataupun tes yang dibuat oleh guru) yang memiliki bentuk ekuivalen. Mengembangkan tes yang benar-benar ekuivalen juga membutuhkan waktu lama, di samping sulit. Namun, bentuk ini penting untuk dipertimbangkan.

d. Metode Belah Dua (*Split Half*)

Dilakukan dengan cara mengujikan seperangkat tes, kemudian tes itu dibagi atau dibelah menjadi dua bagian yang ekuivalen. Selanjutnya, masing-masing diskor secara terpisah (independen). Hasil dari belahan pertama dikorelasikan dengan hasil belahan kedua, dihitung dengan menggunakan korelasi produk moment Pearson.

Banyak cara dapat dilakukan untuk membagi atau membelah tes menjadi dua. Salah satunya dengan menggunakan metode belah nomor ganjil dan genap. Butir tes bernomor ganjil dikumpulkan ke dalam satu kelompok dan butir tes bernomor genap juga

Untuk menghitung reliabilitas tes, digunakan rumus Spearman Brown berikut:

Misalnya, korelasi antara butir tes genap dan ganjil pada suatu tes tertentu adalah 0,74. Dengan demikian, nilai r bila digunakan rumus Spearman Brown adalah:

e. Metode Kuder Richardson atau Koefisien Alpha

Mengestimasi reliabilitas dengan metode ini sensitif terhadap kesalahan pengukuran yang disebabkan oleh kesalahan sampel isi. Metode ini juga sensitif terhadap heterogenitas isi tes. Bila merujuk pada heterogenitas isi, hal ini berkaitan dengan derajat butir tes mengukur karakteristik yang berkaitan.

64

Rumus yang dapat digunakan adalah KR-20 maupun KR-21.
Rumus untuk KR-20 adalah:

di mana: k = banyaknya butir tes
 SD^2 = varian skor tes total
 p_i = proporsi jawaban benar pada sebuah butir tes
 q_i = proporsi jawaban salah pada sebuah butir tes

	Baris 1	Baris 2	Baris 3	Baris 4	Baris 5	
Siswa 1	1	0	1	1	1	4
Siswa 2	1	1	1	1	1	5
Siswa 3	1	0	1	0	0	2
Siswa 4	0	0	0	1	0	1
Siswa 5	1	1	1	1	1	5
Siswa 6	1	1	0	1	1	4
p_i	0,83	0,5	0,67	0,83	0,67	$SD^2 =$
q_i	0,17	0,5	0,33	0,17	0,33	2,25
$p_i \times q_i$	0,14	0,25	0,22	0,14	0,22	

$$KR - 20 = \frac{5}{4} \left(\frac{2,25 - 0,97}{2,25} \right)$$

$$= \left(\frac{1,278}{2,25} \right)$$

$$= 1,25 \text{ (0,568)}$$

$$= 0,71$$

Formula ini paling banyak digunakan untuk menghitung reliabilitas, karena sangat mudah diterapkan. Namun, ada beberapa keterbatasan yang dimilikinya, yakni kurang cocok untuk tes kecepatan (*speed test*). Tes kecepatan adalah tes yang dibatasi waktu dalam pengerjaannya. Untuk tes yang menuntut kecepatan, penghitungan reliabilitas disarankan menggunakan metode tes-retes dan bentuk ekuivalen.

f. Metode Inter-Rater

Pada penskoran instrumen yang melibatkan beberapa penyekor atau *rater*, perlu dihitung tingkat persetujuan masing-masing *rater*. Hal ini disebut sebagai reliabilitas *inter-rater*. Mengestimasi reliabilitas *inter-rater* menjadikan proses penskoran menjadi lebih adil. Pada metode ini, tes dilaksanakan satu kali terhadap sejumlah peserta tes dengan menggunakan dua orang *rater*. Agar tidak saling memengaruhi, maka masing-masing *rater* bekerja secara terpisah.

Pada kondisi tertentu, reliabilitas *inter-rater* tidak terlalu penting, seperti soal pilihan ganda atau Benar-Salah. Untuk soal uraian, reliabilitas *inter-rater* menjadi sangat penting. Ketika reliabilitas *inter-rater* rendah, maka peserta tes merasakan adanya perbedaan atau kesenjangan yang cukup tinggi dari skor yang diperoleh dari dua penyekor yang berbeda.

C. LATIHAN

1. Carilah dokumen penilaian yang dikembangkan guru. Cermati validitas penilaian dari dokumen penilaian tersebut. Selanjutnya, uraikan secara ringkas bagaimana validitas penilaian itu. Apa yang dapat Anda lakukan untuk membantu guru agar penilaian tersebut memiliki validitas yang lebih baik.
2. Cermati setiap pernyataan di bawah ini dan nyatakan apakah pernyataan tersebut benar atau salah. Jelaskan!
 - a) Suatu tes penalaran verbal dikenakan pada siswa kelas VII. Skor tes ini dikorelasikan dengan nilai Bahasa Inggris yang diberikan pada kelas IX, X, XI, dan XII. Korelasi antara skor tes penalaran verbal dengan nilai Bahasa Inggris di kelas XII adalah paling rendah di antara ke-4 korelasi kelas lainnya.
 - b) Suatu tes prediktif memiliki reliabilitas sempurna (dengan koefisien korelasi sebesar 1). Hal ini berarti, tes tersebut memiliki korelasi yang tinggi dengan sebarang alat ukur lainnya.
 - c) Suatu tes prediktif lainnya memiliki reliabilitas sebesar 0. Artinya, tes tersebut memiliki korelasi yang sangat rendah dengan sebarang alat ukur lainnya.
3. Pernyataan-pernyataan di bawah ini merupakan pertanyaan seorang guru terkait prosedur penilaian yang digunakan. Lakukan identifikasi terhadap setiap pernyataan, masuk dalam kategori jenis validitas apa? Jelaskan jawaban Anda!
 - a) Dapatkah skor siswa dari tes membaca, saya gunakan untuk menilai mata pelajaran lain?
 - b) Apakah penilaian *performance* dalam matematika benar-benar mencerminkan kemampuan matematika siswa tersebut?

