

Pengukuran, Penilaian, dan Tes

1. Pengukuran

Pengukuran adalah suatu kegiatan yang ditujukan untuk mengidentifikasi besar kecilnya obyek atau gejala (Hadi, 1995). Pengukuran dapat dilakukan dengan dua cara; 1) menggunakan alat-alat yang standar, 2) menggunakan alat-alat yang tidak standar.

Suryabrata (1984) mendefinisikan secara sederhana bahwa pengukuran terdiri atas aturan-aturan untuk mengenakan bilangan-bilangan kepada sesuatu obyek untuk mempresentasikan kuantitas atribut pada obyek tersebut. Cronbach yang dikutip oleh Mehren (1973) mendefinisikan pengukuran sebagai suatu prosedur yang sistematis untuk mengamati perilaku seseorang dan menggambarkannya dengan bantuan skala numerik atau sistem pengkategorian. Hamalik (1989), menyatakan bahwa kualitas dan kuantitas hasil pengukuran itu banyak bergantung pada jenis dan mutu alat ukur yang digunakan.

Menurut Umar (1991) pengukuran adalah suatu kegiatan untuk mendapatkan informasi data secara kuantitatif. Hasil dari pengukuran dapat berupa informasi-informasi atau data yang dinyatakan dalam bentuk angka ataupun uraian yang sangat berguna dalam pengambilan keputusan, oleh karena itu mutu informasi haruslah akurat.

Berdasarkan pendapat-pendapat diatas, dapat disimpulkan bahwa pengukuran adalah suatu prosedur yang sistematis untuk memperoleh informasi data kuantitatif baik data yang dinyatakan dalam bentuk angka maupun uraian yang akurat, relevan, dan dapat dipercaya terhadap atribut yang diukur dengan alat ukur yang baik dan prosedur pengukuran yang jelas dan benar.

2. Penilaian

Tahap berikutnya setelah pengukuran adalah penilaian yang bertujuan untuk mengambil suatu keputusan baik atau buruk. Mehren dan Lehmann (1973) mengatakan bahwa penilaian adalah proses penggambaran untuk memperoleh dan memberikan informasi yang berguna sebagai alternatif pengambilan keputusan.

Hasil pengukuran merupakan landasan yang terpenting dalam penilaian pendidikan, dan hanya data dari hasil pengukuran saja yang dapat dipercaya dan dapat dijadikan landasan kuat bagi pengambilan keputusan (Umar, 1991). Grondlund (1985) berpendapat bahwa penilaian merupakan serangkaian proses mulai dari pengumpulan data, analisis data, interpretasi hasil, serta pengambilan keputusan berkenaan dengan pencapaian tujuan belajar.

3. Tes

Tes pada dasarnya merupakan suatu pengukuran yang obyektif dan standar terhadap sampel perilaku (Anastasi, 1976). Brown (1976) mengatakan bahwa tes adalah prosedur yang sistematis guna mengukur sampel perilaku seseorang. Menurut Cronbach (1970) prosedur sistematis adalah untuk meneliti perilaku seseorang dan mengembarkannya dengan tujuan dari skala numerikal atau sistem kategori.

TIPOLOGI TES, FUNGSI, DAN BATASAN

Ada delapan tipe tes yang digunakan oleh dunia psikologi sampai saat ini. Lima tipe diantaranya adalah tes berbentuk tes proyektif dan sisanya bukan. Delapan tes berdasarkan kepada tiga dikotomi tipologi tes. Dengan tiga dikotomi ini dapat menghasilkan delapan tipe tes.

1. Subyektif vs Obyektif.

Dalam tes subyektif (*voluntary*) responden diminta untuk memahami bahwa apapun jawabannya adalah benar, dan tidak ada kriteria eksternal untuk membenarkan atau menyalahkan jawabannya. Dalam tes ini biasanya dicantumkan pernyataan “tes bukan menguji kecakapan anda” atau “tidak ada jawaban benar atau salah” atau “jawablah sesuai dengan yang benar-benar anda rasakan”. Tes ini adalah tes yang mengukur performansi tipikal. Tes tipikal bertujuan untuk mengungkap kecenderungan reaksi atau perilaku individu ketika berada dalam situasi-situasi tertentu. Tujuannya bukan untuk mengetahui apa yang mampu dilakukan seseorang tapi apa yang cenderung ia lakukan. Tes ini misalnya tes kepribadian dan tes motivasi. Sebaliknya dalam tes obyektif, akan selalu ada jawaban salah atau benar. Konsep ketepatan atau kesalahan akan selalu diperhitungkan oleh responden. Konsep benar-salah ini diperhitungkan karena ada patokan atau kriteria eksternal yang membenarkan atau menyalahkan jawaban dan kriteria ini biasa disebut kunci

jawaban. Tes ini adalah tes yang mengukur performansi maksimal. Tes ini bertujuan mengungkap apa yang mampu dilakukan oleh seseorang dan seberapa baik ia mampu melakukannya, tes ini misalnya tes belajar dan tes prestasi.

2. Langsung vs Tidak Langsung

Dalam tes langsung, responden paham dengan tes yang sedang dijalani. Jadi responden sangat paham akan apa yang ingin psikolog ungkapkan terhadap dirinya. Sebelum mengerjakan tes responden terlebih dulu membaca laporan mengenai tujuan, hasil tes yang pernah dilakukan, reliabilitas dan validitas tes, sehingga tidak ada yang akan mengejutkan dirinya ketika menghadapi tes. Tes tipe ini bisa diterapkan dalam tes prestasi, wawancara diagnostik, atau tes-tes minat dan inventori adjusmen.

Dalam tes tidak langsung, psikolog menginterpretasi jawaban-jawaban responden dalam dimensi dan kategori yang berbeda dengan apa yang dipikirkan responden. Jika seorang menceritakan cerita-cerita yang dia lihat di sebuah gambar dengan keyakinan bahwa kreativitasnya dalam bercerita dinilai kemudian psikolog menginterpretasikan ceritanya sebagai proyeksi mendalam, maka tes ini adalah tes tidak langsung. Jika seseorang mengekspresikan rasa suka dan tidak sukanya terhadap serangkaian gambar, dan hasilnya diinterpretasikan sebagai rasa pesimisme oral, maka tes ini adalah tes tidak langsung. Umumnya, jika jawaban-jawaban responden diambil sebagai simptom-simptom, daripada sebagai informasi literal maka tes itu masuk katagori tes tidak langsung.

Karakter tes tidak langsung adalah penyamaran. Artinya tes ini menyembunyikan tujuan sebenarnya, dan membuat sebuah alasan-alasan tertentu yang masuk akal sehingga responden mau mengikuti tes.

3. Terstruktur vs Tidak Terstruktur.

Dikotomi ini banyak terjadi dalam prosedur tes kepribadian dan pengukuran sikap. Karakter tes proyektif adalah jawabannya yang terbuka (open-ended), bebas, tidak terstruktur dan membebaskan responden untuk memproyeksikan dirinya terhadap sesuatu atau materi tes. Jawaban bebas memiliki ciri yaitu tidak menawarkan alternatif jawaban dan tidak membatasi jumlah jawaban.

Tes yang terstruktur jawaban-jawaban yang akan diberikan oleh responden telah disediakan pilihan-pilihannya. Tes tipe ini lebih banyak dilakukan untuk tes-tes obyektif meskipun banyak juga yang diterapkan dalam tes-tes subyektif seperti tes kepribadian atau tes-tes proyektif.

4. Delapan Tipe Tes

1. *Subyektif, tidak langsung, tidak terstruktur.* Tes ini termasuk tes yang menggunakan teknik-teknik proyektif klasik, misalnya asosiasi bebas, Rorschach, Thematic Apperception Test, gambar, dan pertanyaan-pertanyaan proyektif seperti “Apa yang paling kamu kagumi dari orang?” Atau “Apa yang kamu anggap paling memalukan?”

2. *Subyektif, tidak langsung, terstruktur.* Dalam kategori ini bisa dimasukkan pilihan ganda Rorschach dan pilihan ganda tes asosiasi. Selain itu tes-tes lain yang masuk kategori ini adalah: kuesioner tidak langsung (misalnya pengukuran skala F terhadap kecenderungan kepribadian otoriterian), Differential Semantic Osgood digunakan untuk sikap tidak langsung terhadap orang tua, dan figur-figur lain. Juga Pendekatan Q-Short untuk identifikasi ketidak sadaran. Tes humor dan inventori kejengkelan yang digunakan untuk tujuan diagnostik tidak langsung juga termasuk kategori ini.

3. *Subyektif, langsung, tidak terstruktur.* Masuk dalam kategori ini adalah tes-tes melengkapi-kalimat, kuesioner essay, wawancara open-ended.

4. *Subyektif, langsung, terstruktur.* Kategori ini termasuk analisis kuantitatif untuk mengukur adjusmen, kepribadian, minat, dan bakat. Tes kategori ini adalah Woodworth inventori, tes sikap Thurstone dan Likert, inventori minat Strong dan Kuder, Minnesota Multiphasic Personality Inventori. Tes-tes ini disekor dengan analisis empiris atau statistik atau jawaban dipilih dengan format pilihan tertentu.

5. *Obyektif, tidak langsung, tidak terstruktur.* Tes ini adalah tes proyektif yang menggunakan tampilan tes obyektif, yang berusaha mengarahkan responden kepada kriteria eksternal tapi dipersilahkan mengisi jawaban dengan pola tidak terstruktur. Misalnya dalam tes “Verbal Summator”. Dalam tes ini responden diinstruksikan untuk mendengarkan sebuah rekaman dari sebuah percakapan. Responden diberitahu

bahwa percakapan ini tidak tersusun dengan rapi, tapi jika didengarkan dengan cermat responden bisa menangkap ide dalam percakapan itu.

6. *Obyektif, tidak langsung, terstruktur*. Kategori ini adalah kategori lima yang distrukturkan. Misalnya penilaian terhadap sebuah gambar bisa menggunakan jawaban terstruktur atau tidak terstruktur.

7. *Obyektif, langsung, tidak terstruktur*.

8. *Obyektif, langsung, dan terstruktur*. Dua kategori ini adalah tes yang digunakan untuk mengukur kemampuan maksimal, yaitu tes kecakapan atau tes prestasi. Tes yang terstruktur diantaranya adalah tes tipe pilihan ganda, tipe pasangan, dan benar atau salah. Sedangkan tes dalam format tidak terstruktur adalah tes essay.

Jenis Tes

Tes Intelijensi: mengukur kecakapan umum individu misalnya verbal comprehension, pengaturan persepsi, atau penalaran. Tes ini membantu menentukan potensi akademik atau prestasi dalam bidang tertentu.

Tes Bakat: mengukur kemampuan individu dalam kemampuan atau ketrampilan khusus. Tes ini termasuk dalam tes kecakapan (ability).

Tes Prestasi: mengukur tingkat pembelajaran, keberhasilan, atau kemampuan menyelesaikan suatu tugas atau pekerjaan individu.

Tes Kreativitas: menilai kemampuan berfikir individu yang tidak biasa atau kemurnian berfikir atau kemampuannya memberi solusi yang tidak biasa atau tidak terduga, khususnya masalah-masalah yang membingungkan.

Tes Kepribadian: mengukur perilaku, ciri khas, atau sifat yang menentukan individualitas seseorang; formatnya bisa berupa checklist, inventori, dan proyektif.

Inventori minat: mengukur kecenderungan seseorang untuk memilih aktivitas sehingga bisa memengaruhi dia dalam memilih pekerjaan.

Prosedur Perilaku: penggambaran dan penghitungan secara obyektif frekuensi perilaku, yang ditujukan untuk mengenali sebab sebuah perilaku muncul dan kemungkinan perilaku yang akan muncul selanjutnya.

Tes Neuropsikologi: mengukur kinerja kognitif, sensori, persepsi, dan motorik seseorang untuk menentukan sejauh mana, dimana, dan dampak kerusakan otak.

Fungsi Tes

Sejauh ini fungsi tes secara umum adalah untuk membuat keputusan mengenai seseorang. Misalnya, institusi pendidikan sering menggunakan tes untuk menentukan tingkat siswa, atau universitas menggunakan tes menentukan apakah seseorang diterima sebagai mahasiswa atau tidak. Lembaga-lembaga itu merujuk kepada skor tes. Lembaga pemerintah juga sangat banyak menggunakan tes terutama untuk melakukan penyeleksian atau penempatan kerja.

Secara umum tes digunakan sebagai pertimbangan dalam membuat keputusan. Namun secara khusus tes berfungsi sebagai:

Klasifikasi

Diagnosis dan perencanaan perlakuan

Pengenalan-diri (Self-Knowledge)

Evaluasi program

Penelitian

Klasifikasi meliputi berbagai cara atau prosedur yang bertujuan mengelompokkan atau membedakan seseorang dengan orang lainnya. Tentu saja klasifikasi ini memiliki tujuan, yang dikaitkan dengan perlakuan selanjutnya. Misalnya pengklasifikasian Indeks Prestasi mahasiswa, klasifikasi IP tertinggi digunakan untuk menentukan mahasiswa yang berhak menerima beasiswa, sedangkan IP terendah dimotivasi lagi untuk meningkatkan kemampuannya.

Dalam dunia terapi, psikolog sering melakukan diagnosis terhadap pasiennya untuk mencari informasi sebagai dasar perlakuan yang akan diambil dalam terapi. Psikolog ingin mengetahui seberapa tinggi tingkat kecemasan, motivasi, atau harga diri pasien sehingga psikolog kemudian bisa dengan tepat menggunakan prosedur atau teknik terapi yang bisa membawa pasiennya ke kondisi yang lebih baik dari sebelumnya. Ini adalah fungsi diagnosi dan perencanaan perlakuan.

Pengenalan-diri bisa dilakukan dengan cara-cara standar dan tidak standar. cara-cara tidak standar seseorang dalam mengenali sendiri selalu dilakukan dilakukan setiap hari karena sebagai manusia mereka akan selalu menilai dirinya dalam hal perilaku, baik perilaku terhadap diri sendiri maupun dengan orang lain. Sementara cara standar adalah dengan menggunakan tes. Tes IQ adalah tes yang paling banyak digunakan atau diinginkan oleh seseorang untuk mengetahui seberapa tinggi tingkat kecerdasan dirinya, begitu pula dengan tes-tes lainnya seperti tes kecerdasan emosi, kejujuran, dan minat.

Dalam perusahaan atau sebuah lembaga sering digunakan tes-tes untuk melakukan analisis jabatan atau evaluasi program. Seringkali tes digunakan untuk menganalisis kerja seseorang atau beberapa orang dalam sebuah kelompok yang menjalankan sebuah program kerja. Untuk tujuan ini tes yang sedang banyak diperbincangkan saat ini adalah tes kecerdasan emosi (Emotional Quotient).

Sebagai sebuah karya, tes adalah karya dari orang-orang yang bergelut dalam dunia keilmuan atau ilmuwan. Maka sangat wajar jika tes digunakan oleh mereka sebagai alat untuk meneliti untuk menambah wawasan keilmuan mereka.

D. RELIABILITAS TES

1. Definisi Reliabilitas

Ide pokok dari reliabilitas tes adalah sejauh mana hasil suatu tes itu dapat dipercaya konsistensinya (Azwar, 2000). Sebuah tes dikatakan reliabel atau dipercaya jika memberikan hasil yang sama dalam atribut yang diukur dari peserta dan tes yang sama (Gebotys, 2003).

Reliabilitas berkaitan erat dengan ketidak reliabelan pengukuran. Misalnya sebuah tes memiliki reliabilitas 0,7 maka ada ketidak reliabelan sebesar 0,3. Ketidak reliabelan ini dapat diartikan sebagai faktor-faktor menebak jawaban tanpa mengetahui jawaban yang sebenarnya dalam menjawab item tes. Jika kejadian main tebak terjadi dalam tes itu kecil maka tes itu bisa dikatakan reliabel dan jika sebaliknya dikatakan tidak reliabel. *“To the extent to which such chance factors predominate, a test is said to be unreliable. Conversely when the influence of chance factors is slight, a test is said to be highly reliable. The chance element ..as measurement error (Nunnally, 1972).*

2. Metode Pengestimasi

Ada banyak metode untuk mengestimasi reliabilitas misalnya bentuk tes ulang, tes belah dua, tes alternatif dan konsistensi internal. Sebuah tes prestasi yang berbentuk pilihan ganda akan cocok dengan pengestimasi konsistensi internal karena dikenakan pada peserta tes hanya sekali dan lebih sering digunakan KR 20 atau Alpha Cronbach. Tes yang pendek (10 – 15 item) bisa dikatakan memuaskan apabila mencapai koefisien reliabilitas minimal sebesar 0,5, sedangkan tes yang besar (sekitar 40 item) dikatakan memuaskan jika mencapai koefisien reliabilitas sebesar 0,8. Tes yang memiliki item lebih dari 40 item, koefisien reliabilitasnya menjadi kurang berarti karena sangat dipengaruhi panjang tes (Kehoe, 1995).

Metode konsistensi internal untuk mengestimasi reliabilitas sebuah tes berdasarkan pada jumlah item pada tes (k) dan rata-rata inter korelasi antara item-item tes (r_{ij}) (Murphy & Davidshofer, 1991):

$$r_{xx} = \frac{k(r_{ij})}{1 + (k-1)r_{ij}}$$

Apabila dalam sebuah tes terdiri dari beberapa sub tes yang mengukur kemampuan yang berbeda, maka ada beberapa analisis reliabilitas atau koefisien reliabilitas sebanyak subtes itu.

Ada beberapa sebab yang memunculkan eror dalam pengukuran yang mengakibatkan kepada rendahnya reliabilitas. Sumber-sumber eror itu adalah 1) penyamplingan item, 2) konstruksi tes, 3) pelaksanaan tes, 4) obyektivitas penyekoran, 5) tingkat kesulitan tes, 6) faktor peserta tes.

Pertama, sebuah tes adalah sebuah kumpulan atau penyampelan dari item-item, dan sampel item-item itu sendiri bisa menjadi sumber kesalahan. Tes yang lebih besar bisa lebih reliabel karena penyampelan yang lebih baik dari sumber isi dan kemampuan peserta tes. Reliabilitas sebuah tes sangat berkaitan dengan sumber varian, semakin besar varian individu maka tes semakin reliabel. Misalnya saja mengukur kemampuan siswa dalam salah bab biologi, dengan memberikan satu pertanyaan. Siswa yang tahu satu pertanyaan ini mempunyai skor sempurna sedangkan yang tidak mereka telah gagal. Padahal satu pertanyaan tidak akan memberikan suatu perkiraan yang reliable dari pengetahuan siswa. Jika pertanyaan terus ditambah maka akan mendapatkan sampel yang lebih fit, dan menghasilkan skor yang lebih akurat yang mencerminkan berbagai kemampuan yang sebenarnya.

Penambahan item dengan demikian akan semakin meningkatkan reliabilitas tes, dan juga mengurangi pengaruh faktor-faktor keberuntungan seperti penebakan jawaban benar atas item soal yang tidak diketahui.

Kedua, penyusunan stem item yang kurang jelas atau membingungkan akan semakin memperkecil reliabilitas. Misalnya item berikut: indeks reliabilitas manakah yang terbaik digunakan oleh guru kelas? Pertanyaan ini kurang jelas karena tergantung dari penyusun tes dalam mengartikan “yang terbaik”.

Ketiga, pelaksanaan tes sangat berpengaruh bagi peserta tes dalam menjawab. Misalnya, faktor lingkungan seperti panas, cahaya, ketenangan, pengarahannya yang membingungkan, dan perbedaan waktu yang diberikan kepada peserta tes.

Keempat, obyektivitas tes juga berpengaruh. Semakin obyektif semakin reliabel. Tes obyektif lebih reliable karena skor tes mencerminkan perbedaan yang sebenarnya dari kemampuan siswa, dan tidak tergantung dari penilaian atau pendapat penyekor. Tes esai lebih memungkinkan mempunyai reliabilitas yang lebih rendah daripada tes obyektif karena tes esai lebih besar pengaruh subyektivitasnya, namun tidak berarti tes esai tidak perlu digunakan karena tes esai juga bisa ditingkatkan reliabilitasnya.

Kelima, tingkat kesulitan yang tinggi atau rendah bisa mengakibatkan rendahnya reliabilitas tes. Tes yang terlalu sulit atau mudah tidak mencerminkan kemampuan siswa secara keseluruhan. Tingkat kesulitan yang mendekati sedang atau beragam akan lebih meningkatkan indeks reliabilitas.

Keenam, faktor siswa berkaitan dengan keadaan fisik atau psikis peserta tes ketika menghadapi tes. Siswa yang sakit atau cemas besar kemungkinan akan mempunyai skor yang lebih rendah dari siswa yang mempunyai kemampuan sama ketika dalam keadaan sama-sama sehat.

Dalam sisi item semakin item homogen maka semakin reliabel (Thordike, 1951). Sumber varian bersumber dari 1) kemampuan pengetahuan dalam menjawab tes, 2) keadaan ketika menghadapi tes, misalnya kesehatan, emosi, kondisi lingkungan eksternal, 3) keberuntungan dalam menebak jawaban yang tidak diketahui (Thordike, 1951).

Sumber eror di atas dapat dikurangi dengan langkah-langkah sebagai berikut: 1) menulis tes yang memiliki item yang lebih banyak, 2) lebih hati-hati dalam menulis pertanyaan-pertanyaan dalam tes, 3) menggunakan petunjuk dan pengarahan yang jelas.

Item tes yang berisikan item yang banyak akan memperbesar reliabilitas tes. Besar kecilnya tes agak sulit ditentukan, karena tergantung kepada kualitas item, tingkat kesulitan, atau faktor-faktor lain. Empat puluh item dalam tes pilihan ganda cukup layak bagi tes kelas.

Lebih berhati-hati dalam menyusun tes dianjurkan dalam usaha meningkatkan reliabilitas. Item tes yang ditulis dengan jelas akan mempermudah peserta tes memahamai apa yang diinginkan oleh penyusun tes, semakin tes itu mendekati apa yang diinginkan atau tujuan penulis maka tes itu diperkirakan semakin reliabel. Waktu

penyusunan tes yang jauh-jauh hari dilakukan juga mempengaruhi kehati-hatian dalam menyusunnya. Waktu penyusunan tes yang semakin dekat dengan pelaksanaan tes membuat tes semakin tidak reliabel.

Menggunakan petunjuk pengerjaan yang jelas dan terarah dan menggunakan prosedur pelaksanaan yang standar. Petunjuk dan pengarahan yang jelas bagi peserta tes akan lebih memunculkan kemampuan yang sebenarnya dari peserta tes. Petunjuk yang membingungkan akan membuat peluang lebih besar peserta tes membuat kesalahan yang akhirnya tidak maksimal dalam mengeluarkan kemampuan yang sebenarnya.

E. VALIDITAS TES

1. Definisi

Tahun 1949, Cronbach menyatakan bahwa definisi validitas adalah sejauh mana sebuah tes mengukur apa yang seharusnya diukur, "*A test is valid to the degree that we know what it measures or predicts* (Cronbach, 1955). Pendapat yang terus dipakai juga diberikan oleh Anastasi & Urbani (1997), validitas adalah apa yang tes ukur dan seberapa baik dia melakukannya dan diakui secara luas. Pada dasarnya bukan apakah tes itu valid tetapi apakah skor tes itu valid? (Stapleton, 1997)

Pendapat-pendapat di atas juga diikuti oleh para ahli psikometri Indonesia, misalnya Azwar (2000) mendefinisikan validitas adalah sejauh mana ketepatan dan kecermatan suatu alat ukur dalam melakukan fungsi ukurnya. Begitu juga dengan Sutrisno Hadi (1997), validitas memiliki pengertian a. seberapa jauh alat ukur dapat mengungkapkan dengan jitu, tepat mengenai sasaran, b. seberapa jauh alat ukur dapat memberikan reading yang teliti, seksama, cermat, dapat menunjukkan dengan sebenarnya status atau besar kecilnya, dan ada tidaknya gejala yang akan diukur.

Pada 1950-an berbagai tipe validitas telah dimunculkan misalnya faktorial, intrinsik, empiris, logis, dan banyak yang lain (lihat Anastasi, 1954). Tipe-tipe itu kemudian diringkas menjadi dua bentuk dasar, yaitu validitas empiris dan logis (Cronbach, 1949).

Validitas logis, penyusun tes mengatur bahan tes mentah, menentukan pendekatan, termasuk analisis isi, pengujian masalah operasional, dan proses pembuatan tes. Pembuat tes diharapkan melakukan studi secara hati-hati terhadap tes itu sendiri, menentukan apa arti skor tes. Kategori diatas kemudian secara luas dikenal sebagai validitas isi.

Validitas empiris meletakkan perhatiannya pada penggunaan analisis faktor (misalnya validitas faktorial Guilford, 1954), dan khususnya tentang korelasi-korelasi antara skor tes (Anastasi, 1954). Sebuah komite dalam psikologi Amerika telah menetapkan empat kategori validitas: isi, konstruk, prediktif, dan konkuren (APA, 1954), kemudian ada revisi bahwa validitas prediktif dan konkuren digabung jadi satu yang disebut validitas kriteria (APA, 1966), kemudian diikuti dengan publikasi Cronbach (1955) mengenai validitas konstruk. Campbel dan Fiske (1959) mengenalkan pendekatan *multitrait-multimethods* untuk menganalisis validitas konstruk.

2. Tipe Validitas

Tipe-tipe validitas ada 3 yaitu; validitas isi, validitas konstruk, dan validitas kriteria.

a. Validitas Isi

Banyak definisi yang dimunculkan, namun dapat diambil sebuah definisi yaitu “*the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose*”. Artinya sejauh mana elemen-elemen instrumen pengukuran itu relevan dan sesuai dengan konstruk yang ditentukan oleh tujuan pengukuran tertentu (Heyness & Richard, 1995).

Validitas isi adalah syarat bagi validitas konstruk. Validitas konstruk menjadi tidak berarti jika validitas isi tidak memenuhi syarat. Validitas isi dapat diterapkan oleh segala metode pengukuran karena menekankan pada data yang diperoleh. Metode validitas isi adalah : (a) definisi dan evaluasi kuantitatif secara hati-hati dari konstruk yang ditargetkan, (b) pendekatan multi elemen, (c) penggunaan populasi dan penyampelan dalam pengembangan item awal, (d) evaluasi kuantitatif dari para ahli dan responden yang berkompeten, (e) evaluasi

item, (f) pelaporan hasil-hasil validitas isi secara detail, dan (g) relevansi untuk validitas isi analisis psikometrik selanjutnya (Heyness & Richard, 1995).

b. Validitas Konstruk

Validitas konstruk adalah tipe validitas yang menunjukkan sejauh mana tes mengungkap suatu trait atau konstruk teoritis yang hendak diukur (Allen & Yen, 1979, Azwar, 2000). Validitas konstruk menggunakan dua analisis validitas, isi dan kriteria. Shepard (1993) dan Anastasi (1986) menyatakan bahwa validitas konstruk mencakup analisis empiris dan logis yang ada dalam validitas isi dan kriteria. Mungkin tipe validitas ini adalah yang paling sulit untuk dipahami. Ini karena berhubungan dengan sejauh mana pemahaman terhadap hubungan antara skor-skor tes dengan model teoritis. Sebuah tes, bisa jadi memiliki derajat validitas prediktif yang tinggi dalam sebuah domain khusus tapi apabila tidak mempunyai atau hanya punya sedikit penghitungan teoritis untuk menjelaskan hubungannya maka tes tidak memiliki derajat validitas konstruk yang tinggi.

Definisi validitas konstruk dipusatkan kepada sebaik apa variabel yang dipilih peneliti mengungkapkan konstruk hipotetik, benar-benar mencakup esensi dari konstruk hipotetik itu dan tentunya harus dibuktikan oleh data (Stapleton, 1997). Sebuah konstruk adalah ide ilmiah dan informatif yang dikembangkan atau dikonstruksi untuk menggambarkan atau menjelaskan perilaku (Cohen, Swedlik, & Phillips, 1996). Gregory (2000) mencatat ada enam pendekatan untuk validitas konstruk, yaitu:

1. analisis untuk menentukan apakah item-item tes atau subtes adalah homogen, maka dari itu mengukur sebuah konstruk tunggal.
2. mencari perubahan-perubahan yang berkembang untuk menentukan apakah mereka konsisten dengan teori konstruk.
3. mencari tahu apakah berbagai perbedaan skor tes adalah konsisten dengan teori.
4. analisis untuk menentukan apakah pengaruh-pengaruh intervensi skor tes adalah konsisten dengan teori.
5. korelasi tes dengan tes-tes dan pengukuran lain yang berkaitan maupun tidak.

6. analisis faktor terhadap skor tes dalam hubungannya dengan sumber-sumber informasi lain.

Selain itu yang perlu diperhatikan dalam menganalisis validitas konstruk yang optimal adalah struktur data yang bisa dilihat dari korelasi antar item. Korelasi antar item yang tinggi akan sangat menguatkan analisis validitas konstruk. Setidaknya korelasi antar item adalah tengah-tengah yaitu 0,20 dan apabila korelasi antar item rendah atau sedang maka bisa diambil korelasi antar item yang tertinggi dan memiliki indeks kesukaran mendekati 50% (Loevinger, 1957).

Brogden (1946) telah membuktikan bahwa jumlah item, indeks kesukaran item, dan interkorelasi antar item sangat mempengaruhi hasil validitas. Misalnya item-item yang berinterkorelasi rata-rata 0,2 dengan indeks kesukaran rata-rata 30% menghasilkan validitas tes sebesar 0,425, sementara dengan rata-rata indeks kesukaran 50% menghasilkan validitas tes sebesar 0,725, dan dengan rata-rata indeks kesukaran 65% menghasilkan validitas sebesar 0,562. Analisis Brogden menunjukkan bahwa semakin rata-rata indeks kesukaran mendekati 50% maka semakin besar validitas tes.

Satu lagi cara untuk menyeleksi item untuk validitas konstruk yaitu lewat analisis faktor pada faktor pertama. Semua item dalam setiap sub tes dianalisis faktor, kemudian dilihat muatan faktor hanya pada faktor pertama dan apabila muatan faktornya besar yaitu minimal 0,40 dengan metode *principal component analysis* dan pada faktor kedua memiliki muatan faktor dibawah 0,200. (Rust & Golombok, 1989; Clark & Watson, 1995).

Analisis item yang paling populer dan paling banyak digunakan adalah korelasi item-total (r_{bis}). Korelasi item-total ini menunjukkan bahwa item-item memiliki kecenderungan atau tidak terhadap tes itu. Item-item yang berkorelasi $\geq 0,30$ dianggap mempunyai kecenderungan terhadap tes secara keseluruhan, dengan demikian item-item itu dianggap homogen. Item-item yang homogen sangat mendukung daya validitas karena pada dasarnya validitas ingin mencari homogenitas perilaku dalam tes (Loevinger, 1957).

c. Validitas Kriteria

Validitas ini ada dua jenis, validitas prediktif dan validitas konkuren. Pola validitas ini seorang peneliti pada dasarnya tertarik dengan beberapa kriteria yang ingin dia prediksi. Dia melakukan tes, dan memperoleh sebuah kriteria independen yang mengukur subyek yang sama, dan menghitung korelasi. Jika kriteria diperoleh beberapa kali setelah tes diberikan, dia meneliti validitas prediktif. Jika skor tes dan skor kriteria ditentukan pada waktu yang sama, dia sedang meneliti validitas konkuren. Validitas konkuren diteliti ketika satu tes diajukan sebagai pengganti bagi yang lain (Cronbach & Meehl, 1955), misalnya, sebuah tes *self concept* yang lama dikorelasikan dengan tes *self concept* baru seperti TSCS (*Tennessee Self-Concept Scale*) (Azwar, 2000).

Ada beberapa faktor yang menyebabkan validitas prediktif rendah yaitu: 1) pembatasan range, yaitu pengurangan range dalam variabel. Hal ini hampir selalu terjadi dalam tes ujian masuk yang dijadikan prediktor prestasi belajar. Tes ujian masuk memilih peserta tes yang mempunyai skor tertinggi dari jumlah peserta yang besar, dengan demikian range skor yang sangat panjang itu dipotong. Pemotongan range ini mengakibatkan penurunan varian dan seterusnya menyebabkan rendahnya koefisien korelasi (Powers, 2001). 2) ketidakreliabelan kriteria. Ketidakadaan jaminan bahwa kriteria yang digunakan adalah reliabel, menjadi sangat melemahkan hasil prediksi keberhasilan prestasi belajar, dalam hal ini indeks prestasi mahasiswa. 3) seleksi kompensatori. Apabila hasil tes menunjukkan sebuah status yang lebih dari satu individu dengan individu yang lain, kemungkinan besar seleksi kompensatori akan terjadi. Status yang lebih rendah (skor yang lebih rendah) itu bisa menyebabkan individu memaksimalkan kemampuan psikologis lain yang dimilikinya misalnya motivasi, atau kedewasaan. Keberhasilan individu dalam belajar di perguruan tinggi yang banyak sekali dipengaruhi oleh faktor-faktor non kognitif menyebabkan rendahnya prediksi tes ujian masuk terhadap prestasi belajar di perguruan tinggi.

Akibat dari rendahnya rata-rata korelasi ujian masuk dengan hasil belajar maka diajukan sebuah pedoman bahwa korelasi validitas prediktif tinggi : $> 0,39$, sedang : $0,25 - 0,39$, dan rendah: $< 0,25$ (Walker, dkk. 2002)

F. Hubungan Reliabilitas dan Validitas

Reliabilitas menjadi penting karena mempunyai keterkaitan dengan validitas. Reliabilitas adalah mengenai konsistensi pengukuran, sedangkan validitas adalah seputar relevansi dan ketepatan apa yang diukur. Mungkin sebuah tes itu reliabel, mengukur hal yang sama secara konsisten tapi bisa jadi pengukuran itu tidak berguna atau invalid (Huitt,1999). Sebuah tes sudah pasti reliabel jika tes itu valid, tapi belum tentu valid jika tes itu reliabel. Artinya sebuah tes yang valid harus mempunyai reliabilitas yang tinggi, namun reliabilitas yang tinggi belumlah cukup untuk menunjang validitas tes.

Suatu hasil koefisien korelasi dalam validitas prediktif menjadi rendah jika prediktor atau kriterianya tidak reliabel. Semakin tinggi koefisien reliabilitas prediktor dan kriteria semakin tinggi pula hasil korelasi antara keduanya. Misalnya tes prediktor yang mempunyai koefisien reliabilitas 0,70 berkorelasi dengan kriteria sebesar 0,30. Jika reliabilitas prediktor ini ditingkatkan menjadi 0,90 maka validitas prediktif akan meningkat dari 0,30 sampai 0,34 (Nunnally, 1972).

Pemilihan item juga mempengaruhi reliabilitas dan validitas. Pemilihan item dengan metode korelasi item-total akan meningkatkan reliabilitas namun akan menurunkan validitas (Azwar,2000).