A. Supratiknya

PENGUKURAN PSIKOLOGIS

Pengukuran Psikologis

A. Supratiknya

Penerbit
Universitas Sanata Dharma

Pengukuran Psikologis

PROF. DR. AUGUSTINUS SUPRATIKNYA
Fakultas Psikologi, Kampus III Universitas Sanata Dharma
Paingan Maguwoharjo Depok Sleman, Yogyakarta.

Diterbitkan oleh:

Penerbit Universitas Sanata Dharma
Jl. Affandi Gejayan (Mrican)
Yogyakarta 55281
Telp. (0274) 513301, 515253;
Ext.1527/1513; Fax (0274) 562383
e-mail: *publisher@usd.ac.id*

Penulis:
A. Supratiknya

Desain Sampul:
Pius Sigit K

Tata Letak:
Thoms

Cetakan Pertama
xvi, 334 hlm.; 148 x 210 mm.
ISBN: 978-602-9187-75-5
EAN: 9-786029-187755

Penerbit USD

Universitas Sanata Dharma berlambangkan daun teratai coklat bersudut lima dengan sebuah obor hitam yang menyala merah, sebuah buku terbuka dengan tulisan "*Ad Maiorem Dei Gloriam*" dan tulisan "Universitas Sanata Dharma Yogyakarta" berwarna hitam di dalamnya. Adapun artinya sebagai berikut.
Teratai: kemuliaan dan sudut lima: Pancasila; Obor: hidup dengan semangat yang menyala-nyala; Buku yang terbuka: iImu pengetahuan yang selalu berkembang; Teratai warna coklat: sikap dewasa yang matang; "Ad Maiorem Dei Gloriam": demi kemuliaan Allah yang lebih besar.

Daftar Isi

Sekapur Sirih

Buku ini sudah penulis impikan sejak tahun 1998. Sebagai salah seorang peserta Proyek Ancangan Aplikasi (AA) Putaran VI yang diselenggarakan oleh Universitas Sanata Dharma bekerjasama dengan Asosiasi Perguruan Tinggi Katolik (APTIK) dan dalam kedudukan sebagai pengampu rumpun mata kuliah pengukuran psikologis di Program S-1 Psikologi Universitas Sanata Dharma, penulis memilih mengembangkan Rencana Kegiatan Belajar Mengajar (RKBM) berikut *Reader* untuk mata kuliah *Konstruksi Tes.*

Saat itu penulis sudah menyadari bahwa dalam kurikulum program studi S-1 Psikologi rumpun mata kuliah pengukuran psikologis yang terdiri dari *Psikometri, Konstruksi Tes,* dan *Penyusunan Skala Psikologis* merupakan semacam "trilogi". *Psikometri* menyajikan dasar-dasar konseptual-teoretis tentang pengukuran psikologis. Penerapannya dalam rangka penyusunan *maximal performance tests* meliputi *aptitude tests* atau tes bakat dan *achievement tests* atau tes prestasi diolah dalam mata kuliah *Konstruksi Tes.* Penerapannya dalam rangka penyusunan *typical performance tests* meliputi berbagai jenis inventori kepribadian diolah dalam mata kuliah *Penyusunan Skala Psikologis.*

Pada tahun 1998 penulis menyusun diktat ringkas untuk mata kuliah *Psikometri.* Dengan tersusunnya RKBM dan *reader* mata kuliah *Konstruksi Tes* pada tahun yang sama, berarti tersedia kurikulum yang kurang lebih konkret untuk dua dari "trilogi" pengukuran psikologis sejak tahun unik tersebut. Selanjutnya kedua sumber bacaan itu praktis masih penulis gunakan sampai dengan semester gasal tahun akademik 2013/2014, dengan aneka tambahan dan pemutakhiran yang bersifat tambal sulam.

Didorong oleh rasa malu karena berulang kali menjanjikan membuat revisi komprehensif terhadap teks-teks yang ada kepada sekian angkatan mahasiswa peserta mata kuliah *Psikometri, Konstruksi Tes* (kemudian menjadi *Konstruksi Alat Ukur*) dan *Penyusunan Skala*

Bab 1
Pendahuluan

Seorang pakar di bidang pengukuran psikologis yang sering dipandang sebagai salah seorang perintis psikologi kuantitatif, Edward Lee Thorndike (1918, dalam Gulliksen, 1974), pernah menyatakan keyakinannya sebagai berikut, *"Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality"* (h. 251). Maksudnya, apa pun yang ada di dunia ini pastilah ada dalam jumlah tertentu, maka untuk mengenalnya secara utuh-menyeluruh kita perlu mengetahui baik kuantitas atau jumlah maupun kualitas atau mutunya. Dalam perkembangan psikologi hingga kini, keyakinan atau pandangan ini diakui menjadi salah satu landasan bagi berkembangnya psikologi kuantitatif pada umumnya maupun bidang kajian psikometri atau pengukuran psikologis pada khususnya. Namun sebelum masuk lebih jauh ke dalam lanskap psikologi kuantitatif maupun lebih mengerucut lagi ke dalam psikometri atau pengukuran psikologis, baiklah kita lihat dulu secara sekilas panorama pemikiran yang berkembang di lingkungan psikologi sejak kelahirannya di Jerman pada paruh kedua abad ke-19 hingga kini.

A. Paradigma

Dalam menjalankan aktivitasnya melakukan penelitian ilmiah, seorang ilmuwan termasuk antara lain yang masih berstatus mahasiswa di perguruan tinggi akan menerapkan metode tertentu. Istilah metode sendiri berasal dari kata Yunani *methodos,* yaitu gabungan antara kata depan *meta* yang antara lain berarti menuju atau melalui, dan *hodos* yang antara lain berarti jalan atau cara (Bakker & Zubair, 1990). Maka, secara harfiah *methodos* atau metode berarti jalan menuju sesuatu, dalam hal ini menuju pengetahuan baru. Maka,

dalam konteks penelitian ilmiah dan dalam arti luas metode berarti "cara bertindak menurut sistem aturan tertentu...agar mencapai hasil optimal"; sedangkan dalam arti khusus berarti "sistem aturan yang menentukan jalan untuk mencapai pengertian baru pada bidang ilmu pengetahuan tertentu" (Bakker & Zubair, 1990; h. 10). Ada banyak metode ilmiah, seperti introspeksi, eksperimentasi, pengukuran psikologis, studi kasus, metode fenomenologis, hermeneutika, dan sebagainya. Pertanyaannya, apa yang menentukan seorang ilmuwan memilih menggunakan metode tertentu dan bukan yang lain dalam melaksanakan aktivitas ilmiahnya?

Salah satu jawabnya adalah bahwa pilihan metode ilmiah ditentukan oleh paradigma yang dianut atau diyakini oleh ilmuwan yang bersangkutan. Paradigma adalah "*a set of basic beliefs (or metaphysics) that deals with ultimates or first principles. It represents a worldview that defines, for its holders, the nature of the 'world', the individual's place in it, and the range of possible relationships to that world and its parts*" (Guba & Lincoln, 1994, h. 105). Artinya, paradigma adalah serangkaian keyakinan dasar (atau metafisika) tentang perkara-perkara yang bersifat ultim atau tentang prinsip-prinsip dasar. Paradigma merepresentasikan sebuah pandangan hidup yang bagi para pemeluk atau penganutnya memberi makna tentang hakikat "dunia," tentang tempat individu di dalam dunia itu, serta tentang kemungkinan relasi dengan dunia beserta bagian-bagiannya yang terbuka bagi individu.

Keyakinan dalam paradigma bersifat mendasar dalam arti bahwa keyakinan itu pada dasarnya diterima berdasarkan kepercayaan semata, betapa pun canggih argumentasi yang dicoba diajukan untuk memberikan pembenaran atau menyokongnya. Dengan kata lain, sesungguhnya tidak ada cara untuk menetapkan kebenaran ultim atau kebenaran terakhir dalam arti kebenaran absolut dari sebuah paradigma (Guba & Lincoln, 1994).

Namun demikian, sekali sebuah paradigma diterima dan diyakini oleh seorang ilmuwan atau peneliti, maka sistem keyakinan dasar atau pandangan hidup itu akan diikutinya sebagai pedoman,

bukan hanya dalam memilih metode namun juga dalam membentuk pandangan ontologis maupun epistemologisnya. Sebagaimana dinyatakan oleh Guba dan Lincoln (1994), sebuah paradigma memang akan mewarnai keyakinan ilmuwan atau peneliti menyangkut tiga pertanyaan fundamental dalam penelitian, yaitu pertanyaan ontologis, pertanyaan epistemologis, dan pertanyaan metodologis.

1. Pertanyaan Ontologis

Pertanyaan ontologis berkisar tentang bentuk (*form*) dan hakikat (*nature*) realitas yang menjadi objek penelitian, serta apa yang bisa diketahui tentang realitas tersebut. Secara garis besar ada dua arus besar paradigma yang memberikan jawaban berbeda terhadap pertanyaan ontologis tersebut (Guba & Lincoln, 1994). Arus paradigma pertama khususnya sebagaimana diwakili oleh ***positivisme*** menganut paham *realisme* yang menyatakan bahwa realitas bersifat *real* atau benar-benar ada dan tunggal. Realitas yang nyata dan tunggal semacam itu dikendalikan oleh aneka hukum dan mekanisme alamiah yang bersifat tetap. Maka, yang bisa diungkap dan diketahui tentang realitas semacam itu adalah "*how things really are*" dan "*how things really work*" atau seperti apakah hakikat dan cara kerja aneka benda-gejala yang terdapat di dalam realitas faktual tersebut (Guba & Lincoln, 1994). Pengetahuan tersebut dapat dirumuskan dalam bentuk aneka generalisasi yang bersifat tidak terikat oleh waktu dan tempat, sebagian di antaranya bahkan dapat dirumuskan sebagai hukum sebab-akibat. Melalui pendekatan yang bersifat reduksionistik dan deterministik sebab-akibat, penelitian ilmiah diyakini mampu mengungkap "kebenaran" yang bersifat tunggal tentang aneka benda dan gejala dalam realitas nyata dunia semesta termasuk kehidupan manusia.

Sebaliknya, arus paradigma kedua dan khususnya sebagaimana diwakili oleh ***konstruktivisme*** menganut paham *relativisme*. Menurut paham ini, realitas bersifat relatif dalam dua pengertian. Pertama, realitas diyakini tidak bersifat tunggal melainkan jamak bahkan

kadang-kadang saling bertentangan, sebab merupakan hasil konstruksi mental yang bersifat lokal dan spesifik dalam arti bahwa bentuk dan isinya ditentukan oleh pengalaman masing-masing orang atau kelompok orang yang meyakini atau mendukungnya. Berbagai konstruksi tentang realitas tersebut tidak bisa dinilai dan tidak bisa saling diperbandingkan dari segi benar-tidaknya, melainkan hanya bisa ditimbang dari segi kecanggihannya. Kedua, berbagai konstruksi tentang realitas tersebut bisa berubah sejalan dengan perkembangan pengalaman dan kemampuan berpikir orang atau kelompok pendukungnya. Maka, tugas ilmu pengetahuan adalah memahami aneka konstruksi tentang realitas tersebut serta mendialogkannya secara dialektis satu sama lain agar aneka konstruksi tentang realitas tersebut mengalami revisi dan rekonstruksi secara terus-menerus.

2. Pertanyaan Epistemologis

Pertanyaan epistemologis berkisar tentang hakikat relasi antara orang yang tahu atau yang akan tahu dan objek atau gejala yang bisa diketahui. Dengan kata lain, pertanyaan epistemologis berkisar pada persoalan tentang bagaimana seharusnya posisi orang yang akan tahu agar mampu mencapai pengetahuan tentang realitas. Tentu saja, jawaban terhadap pertanyaan ini ditentukan oleh jawaban terhadap pertanyaan ontologis yang sudah harus terjawab terlebih dulu (Guba & Lincoln, 1994). Melanjutkan contoh kita sebelumnya, di mata kaum positivis realis seorang ilmuwan harus mengambil posisi atau sikap objektif, mengambil jarak, dan membebaskan diri dari nilai-nilai agar mampu mengungkap kebenaran tentang realitas, yaitu seperti apakah hakikat dan cara kerja aneka objek dan gejala yang membentuk realitas tersebut. Sikap ilmiah semacam ini disebut *dualis* karena memandang peneliti dan objek yang diteliti sebagai dua entitas atau "benda" yang terpisah. Selain itu juga disebut *objektivis* karena berpandangan bahwa seorang ilmuwan harus mampu bersikap objektif dan berjarak dalam melakukan penelitian dalam arti tanpa mempengaruhi atau sebaliknya dipengaruhi oleh objek

yang ditelitinya. Nilai-nilai dan aneka bias subjektif peneliti harus disingkirkan sebab akan mengancam *validitas* hasil penelitian (Guba & Lincoln, 1994).

Sebaliknya bagi kaum konstruktivis, karena tugas ilmu pengetahuan adalah memahami dan membantu terjadinya revisi atas aneka konstruksi tentang realitas secara terus-menerus, maka seorang ilmuwan harus bersikap *transaksional* dan *subjektivis* (Guba & Lincoln, 1994). Peneliti dan subjek yang diteliti dengan seluruh subjektivitas masing-masing dipandang saling terkait secara interaktif sedemikian rupa sehingga temuan-temuan penelitian pada dasarnya diciptakan bersama antara peneliti dan subjek yang diteliti seiring berjalannya proses penelitian. Dengan kata lain, bagi kaum konstruktivis pengetahuan merupakan hasil penciptaan atau konstruksi bersama melalui interaksi antara para subjek yang terlibat dalam proses penelitian, yaitu peneliti dan responden penelitian (Guba & Lincoln, 1994).

3. Pertanyaan Metodologis

Pertanyaan metodologis berkisar tentang cara peneliti atau subjek yang akan tahu menemukan atau mengungkap apa saja yang diyakininya bisa diketahui tentang objek atau gejala yang diteliti (Guba & Lincoln, 1994). Jelas juga bahwa jawaban terhadap pertanyaan ini ditentukan oleh jawaban terhadap pertanyaan ontologis dan pertanyaan epistemologis yang sudah harus terlebih dulu dijawab. Maka bagi kaum positivis-realis cara yang paling jitu untuk mengungkap kebenaran realitas objektif adalah penerapan metode-metode penelitian eksperimental-manipulatif yang memungkinkan peneliti mengontrol atau mengendalikan aneka faktor yang bisa mencemari proses maupun hasil verifikasi atau pembuktian kebenaran atau falsifikasi atau pembuktian kekeliruan aneka hipotesis, khususnya melalui aneka metode yang bersifat objektif-kuantitatif.

Sebaliknya, kaum konstruktivis akan lebih mengandalkan metode-metode hermeneutik-interpretatif yang bersifat dialektik dalam arti metode-metode penafsiran yang memungkinkan pengungkapan tesis-antitesis-sintesis agar mampu melakukan rekonstruksi atau revisi dalam arti pemahaman baru atau berbeda terhadap aneka konstruksi tentang realitas yang selama ini diyakini oleh kelompok subjek yang diteliti sehingga memberikan pencerahan bagi baik peneliti maupun kelompok subjek yang diteliti.

B. Paradigma Positivis-Realis dalam Pengukuran Psikologis

Lantas di manakah tempat psikologi kuantitatif pada umumnya maupun pengukuran psikologis pada khususnya dalam bentangan atau tegangan antara arus paradigma positivis-realis dan arus paradigma konstruktivis-relativis sebagaimana diuraikan di atas? Posisi yang dimaksud dapat kita coba simak dari aneka pengandaian dan keyakinan fundamental yang mendasari pengembangan pengukuran psikologis terkait baik segi ontologis, epistemologis, maupun metodologisnya.

Terkait segi ontologisnya, *pertama* pengukuran psikologis sedikit banyak mengikuti pengandaian kaum realis ketika meyakini bahwa kendati bersifat abstrak namun berbagai atribut psikologis seperti inteligensi, bakat, dan sifat kepribadian sungguh-sungguh *real* nyata ada dalam diri setiap individu atau pribadi manusia. *Kedua,* keberadaan berbagai atribut psikologis tersebut bervariasi khususnya dalam hal kuantitasnya pada diri setiap orang. Artinya, atribut psikologis merupakan *variabel*, entitas yang bisa ada dalam kuantitas yang berlainan, bukan sesuatu yang konstan, khususnya dari orang ke orang. Variasi kuantitas atribut psikologis antar orang ini melahirkan fenomena yang lazim disebut sebagai *individual differences* atau kekhususan individual dan yang memberikan ciri *unik* pada setiap orang. *Ketiga,* keberadaan aneka atribut psikologis dalam

kuantitas yang berlainan dalam diri setiap orang tersebut bersifat relatif permanen atau tetap.

Terkait segi epistemologisnya, pengukuran psikologis menganut keyakinan bahwa agar mampu mengungkap hakikat dan cara kerja aneka atribut psikologis khususnya lewat kegiatan pengukuran, peneliti harus mampu bersikap dualis dalam arti mengambil jarak terhadap objek penelitiannya dan bersikap objektif dalam arti tidak boleh membiarkan subjektivitas mempengaruhi proses dan hasil penelitian atau pengukurannya. Aneka faktor subjektif, baik yang berasal dari peneliti maupun yang berasal dari subjek yang diteliti, dipandang sebagai sumber *error* atau kesalahan yang akan merusak *validitas* atau ketepatan dan *reliabilitas* atau keajekan atau keterandalan hasil-hasil pengamatan atau pengukuran. Maka, faktor-faktor subjektif semacam ini sejauh mungkin harus dikontrol atau dikendalikan atau bahkan dihilangkan.

Maka, terkait segi metodologis dan sesuai namanya pengukuran psikologis mengandalkan metode pengukuran, khususnya dalam rangka pengukuran kekhususan individual. Sebagaimana diketahui selain kekhususan individual bidang penerapan pengukuran psikologis yang lain adalah eksperimentasi. Secara ringkas dan mengutip pendapat Guilford (1954), pengukuran adalah "*the description of data in terms of numbers*" atau pendeskripsian data dalam rupa bilangan (h.1), atau "*the assignment of numerals to objects or events according to rules*" atau peneraan bilangan pada objek atau peristiwa mengikuti seperangkat aturan tertentu (h. 5). Seorang pakar lain menyatakan bahwa "*measurement consists of rules for assigning numbers to objects in such a* way *as to represent quantities of attributes*" (Nunnally, Jr., 1970, h. 7). Dengan kata lain pengukuran adalah kuantifikasi, yaitu peneraan bilangan pada suatu atribut psikologis tertentu untuk menyatakan kuantitas atau jumlahnya. Sebagaimana diketahui, kuantifikasi merupakan primadona dalam *sains* atau ilmu-ilmu alam dan secara umum dipandang sebagai tanda atau tonggak superioritas dalam hirarki ilmu. Maka, bidang-bidang atau jenis-jenis sains yang mengandalkan kuantifikasi seperti fisika dan kimia lazim dijuluki

sebagai *hard sciences* atau ilmu-ilmu keras, sedangkan bidang-bidang ilmu yang kurang mengandalkan kuantifikasi khususnya ilmu-ilmu sosial dijuluki sebagai *soft sciences* atau ilmu-ilmu lunak atau lembek yang berkonotasi *"imprecision and lack of dependability"* atau serba kurang akurat dan kurang bisa diandalkan (Guba & Lincoln, 1994).

Dari sejumlah karakteristik dasar seperti diuraikan di atas, kiranya tampak bahwa sebagai sebuah subdisiplin dalam psikologi pengukuran psikologis cukup jelas cenderung berada dalam gugus paradigma positivis-realis. Bahkan berbagai teknik pengukuran yang terkesan subjektif pun dalam pengukuran psikologis, seperti gugus teknik proyektif, secara ontologis berpegang pada asumsi bahwa ada aneka kategori kehidupan mental seperti emosi, kognisi, identitas, kesadaran, motivasi, dan sebagainya yang bersifat alamiah dalam arti *real* atau nyata dan yang dapat diukur atau dikuantifikasikan. Maka, kendati dalam psikologi yang merupakan disiplin induknya sendiri terjadi perkembangan-perkembangan ke arah pengadopsian paradigma yang lebih konstruktivis-relativis, namun pengukuran psikologis kiranya akan menjadi subdisiplin dalam psikologi yang terus setia pada paradigma positivis-realis.

C. Pengukuran Psikologis di Tengah Menguatnya Orientasi Konstruktivis-Relativis dalam Psikologi

Orientasi konstruktivis-relativis dalam psikologi tampak dalam gerakan yang disebut *psikologi posmodernis* dengan berbagai variannya, sebagai lawan dari psikologi tradisional atau psikologi *mainstream* atau psikologi arus utama yang masih cenderung lekat pada paradigma positivis-pospositivis dan lazim dijuluki *psikologi modernis* (Gergen, 2001; Teo, 2005). Gerakan yang subur mekar di lingkungan psikologi sosial ini secara garis besar memiliki sejumlah karakteristik sebagai berikut. *Pertama,* berbeda dengan psikologi modernis yang memandang pengetahuan sebagai refleksi atau

pantulan realitas yang bersifat objektif, psikologi posmodernis memandang pengetahuan sebagai produk interaksi sosial. Objek atau pengetahuan tentang objek bukanlah sesuatu yang real-nyata seperti diyakini psikologi modernis, melainkan buah atau hasil konstruksi bersama yang tunduk pada aneka konvensi dan aturan bahasa.

Kedua, psikologi posmodernis juga memandang bahwa aneka konsep psikologis seperti identitas, kepribadian, motivasi, dan sebagainya, tidaklah memiliki dasar ontologis atau tidak memiliki rujukan pada entitas psikologis nyata tertentu yang diasumsikan terdapat dalam diri setiap individu seperti diyakini psikologi modernis, melainkan semua itu terbentuk atau tercipta melalui proses historis dan menemukan atau mendapatkan makna masing-masing dalam konteks sosial tertentu.

Ketiga, berbeda dengan psikologi modernis yang memandang bahwa pengalaman individu bersifat unik, bahwa subjek merupakan pusat makna, intensi, serta tindakan, dan bahwa *self* atau diri merupakan suatu entitas yang terisolasi dalam arti berdiri sendiri, psikologi posmodernis memandang bahwa pengalaman individu terletak di tengah jejaring konstruksi yang dibentuk oleh kebudayaan dan sejarah, mencampakkan konsep subjek dan subjektivitas sebagai entitas otonom, sebaliknya memandang bahwa setiap individu berada di tengah jejaring relasi dengan orang lain.

Keempat, berbeda dengan psikologi modernis yang menekankan gagasan tentang identitas sebagai *sense of self* atau rasa diri yang bersifat stabil, koheren, individual, dan berkesinambungan dan karenanya juga menjadi dasar bagi pengukuran psikologis, psikologi posmodernis menekankan konsep identitas yang bersifat ganda, *transient* atau sementara serba mengalir-berubah-bergerak, *nonsubjectified* atau bebas dari perangkap subjek, bersifat *disorganized* atau liar-tak terduga, bersifat *decomposing* atau cair, serta plural.

Maka, *kelima,* tidak seperti psikologi modernis yang meyakini bahwa untuk memahami realitas dan mencapai kebenaran termasuk tentang manusia seorang ilmuwan harus mengandalkan sarana ilmiah *keras* berupa antara lain matematika dan aneka pendekatan kuantitatif,

psikologi posmodernis meyakini bahwa untuk memahami aneka fenomena dan peristiwa apalagi terkait manusia seorang ilmuwan justru harus mengandalkan aneka sarana ilmiah *lunak* meliputi antara lain sejarah, sosiologi, dan aneka pendekatan kualitatif.

Kendati menganut pandangan yang praktis bertolak belakang dengan keyakinan psikologi posmodernis yang lebih mutakhir, namun bukan berarti bahwa pengukuran psikologis tidak lagi memiliki hak hidup baik secara teoretis maupun lebih-lebih jika mempertimbangkan manfaat praktisnya. Dalam kedudukannya sebagai bidang kajian yang cenderung lekat pada paradigma modernis secara epistemologis pengukuran psikologis merupakan salah satu dari antara dua *disiplin* dalam psikologi sebagaimana pernah diuraikan oleh Cronbach (1957). Dalam sambutannya sebagai Presiden Asosiasi Psikologi Amerika pada Konvensi Tahunan *APA* yang ke-65 di New York, 2 September 1957, Cronbach mengatakan bahwa ada "*two historic streams of method, thought, and affiliation...one stream is experimental psychology; the other, correlational psychology*" (h. 671). Ada dua arus metode dan cara berpikir yang berkembang dalam psikologi, yaitu psikologi eksperimental dan psikologi korelasional. Menurut Cronbach (1957), tugas psikologi eksperimental adalah meneliti aneka korelasi yang diciptakan oleh sang ilmuwan sendiri dengan cara mengubah aneka kondisi agar bisa mengamati konsekuensi atau dampaknya. Keunggulan cara kerja ini, sang ilmuwan bisa mengontrol aneka variabel situasional secara ketat sehingga dia akan mampu menguji hipotesis secara ketat dan selanjutnya membuat kesimpulan tentang hubungan sebab-akibat.

Sebaliknya, tugas psikologi korelasional adalah meneliti aneka korelasi yang muncul secara alamiah berupa variasi atau perbedaan antar individu, kelompok sosial, dan spesies. Keunggulan cara kerja ini, sang ilmuwan bisa meneliti hal-hal di luar kendalinya dan yang kiranya tidak akan pernah mampu dikendalikannya. Meminjam kata-kata Cronbach (1957), "*Nature has been experimenting since the beginning of time...*" (h. 672). Sejak awal mula, alam memang telah dan terus bereksperimentasi. Pengukuran psikologis merupakan salah satu cabang dalam psikologi korelasional (Cronbach, 1957). Sebagai

salah satu cabang psikologi korelasional misi pengukuran psikologis adalah mengamati dan mengorganisasikan data hasil eksperimen alam agar selanjutnya mampu membuat aneka keputusan (*decision making)* dan/atau merumuskan aneka prediksi, khususnya tentang orang (Gregory, 2007).

Salah satu bidang penerapan pembuatan keputusan dan/atau prediksi tentang orang yang dimaksud tentu saja adalah klasifikasi atau penempatan seseorang dalam suatu kategori dan bukan dalam kategori yang lain, untuk berbagai tujuan atau keperluan (Gregory, 2007). Aneka tujuan klasifikasi yang dimaksud mencakup *placement* atau penempatan, *screening* atau penyaringan, sertifikasi, dan seleksi. Dalam *penempatan*, pengukuran psikologis bisa membantu sebuah institusi memilah orang-orang ke dalam sejumlah program yang berlainan sesuai kebutuhan atau taraf kemampuan masing-masing. Misal, sebuah universitas melakukan *tes penempatan Bahasa Indonesia* untuk memilah ribuan mahasiswa barunya ke dalam kelompok mahir dan kelompok yang membutuhkan remediasi. Dalam *penyaringan,* pengukuran psikologis bisa membantu sebuah institusi mengidentifikasikan orang-orang yang mungkin memiliki karakteristik atau kebutuhan khusus tertentu sehingga perlu diberi perhatian atau pendampingan khusus. Misal, sebuah SD melakukan tes penyaringan untuk mengidentifikasikan murid-murid baru yang autis. Dalam *sertifikasi,* pengukuran psikologis bisa membantu sebuah institusi menguji kompetensi atau kemampuan seseorang dalam bidang tertentu untuk selanjutnya memberikan *privilese* atau hak istimewa untuk melakukan peran atau tugas dalam bidang terkait sesudah diperoleh bukti bahwa orang tersebut memiliki kemampuan minimum untuk melakukan peran atau tugas terkait. Misal, sesudah melakukan pengujian baik pengetahuan tentang aturan dan tata tertib lalu lintas maupun praktek mengemudikan mobil dan menilai hasilnya memenuhi syarat, Kepolisian menerbitkan *Surat Ijin Mengemudi* kepada seorang pencari *SIM-A.* Dalam *seleksi,* pengukuran psikologis bisa membantu sebuah institusi memilih sebagian dari sejumlah besar calon yang dipandang memenuhi syarat untuk diterima belajar atau

bekerja di lembaga atau organisasinya. Misal, melalui ujian *Seleksi Masuk Perguruan Tinggi* jaringan perguruan tinggi negeri di Tanah Air memilih sebagian dari ratusan ribu lulusan SMA untuk diterima belajar di institusi masing-masing.

Ringkas kata, dalam kedudukannya sebagai bidang kajian yang masih bernaung di bawah kepak sayap paradigma positivis-realis, pengukuran psikologis sebagai salah satu cabang psikologi korelasional tetap memiliki landasan keberadaan yang sah. Lebih-lebih dari segi aplikasi dan manfaat praktisnya, pengukuran psikologis memberikan peran yang kiranya tak tertandingi dan tak tergantikan dalam menyediakan sarana yang cepat, mudah, dan relatif murah untuk membuat keputusan dan prediksi dalam rangka pengklasifikasian orang untuk berbagai keperluan khususnya dalam situasi yang lazimnya menuntut kita memilih demi mendapatkan manfaat terbesar bagi sebanyak mungkin orang di tengah ketersediaan sumber daya yang umumnya terbatas. Kiranya tak bisa dibayangkan kesulitan yang bakal dialami seandainya saat daya tampung berbagai program di seluruh perguruan tinggi di Tanah Air masih sangat tidak sebanding dengan jumlah lulusan SMA/SMK yang terus meningkat dan beraspirasi melanjutkan studi di jenjang tersier, namun demi setia pada perspektif yang lebih konstruktivis-posmodernis kita laksanakan seleksi masuk perguruan tinggi dengan mengandalkan pendekatan yang lebih kualitatif-interpretatif seperti metode wawancara dan *life history,* misalnya. Ψ

Bab 2
Latar Belakang & Beberapa Konsep Dasar

Mengawali penjelajahan kita pada bidang kajian pengukuran psikologis atau psikometri, pada bagian ini akan kita tinjau secara sekilas latar belakang berkembangnya psikometri, pengertian tentang pengukuran dan beberapa konsep dasar yang melekat pada pengertian pengukuran, khususnya pengukuran psikologis.

A. Latar Belakang Berkembangnya Pengukuran Psikologis

Menurut paradigma realis-positivis, agar mampu melahirkan kemajuan setiap disiplin ilmu perlu mengembangkan metodologi untuk mengukur berbagai konstruk yang menjadi objek perhatiannya agar selanjutnya bisa menyimpulkan makna dari hasil pengukurannya tersebut sehingga diperoleh pengetahuan-pengetahuan baru (Browne, 2000). Sebagaimana sudah disinggung, dalam psikologi persoalan ini tidak mudah sebab sebagian besar konstruk yang menjadi objek perhatian psikologi tidak memiliki batasan yang jelas serta tidak bisa diukur secara langsung. Konstruk yang menjadi objek penelitian psikologi merupakan aspek kepribadian atau fungsi psikis tertentu yang melahirkan apa yang lazim disebut *individual differences* atau perbedaan antar individu. Aspek kepribadian atau fungsi psikis yang merupakan konstruk objek kajian psikologi tersebut merupakan variabel laten atau variabel tersembunyi yang tidak bisa diukur secara langsung, melainkan hanya bisa diinferensikan berdasarkan saling hubungan antar berbagai variabel yang merupakan variabel manifesnya (Browne, 2000).

Dalam arti luas psikometri atau pengukuran psikologis merupakan cabang dalam psikologi yang mendalami seluk-beluk kuantifikasi dan analisis berbagai *individual differences* atau perbedaan antar individu. Aktivitas pokok dalam psikometri atau pengukuran psikologis meliputi konstruksi atau penyusunan aneka prosedur untuk mengukur berbagai konstruk psikologis serta pengembangan aneka prosedur analisis data hasil pengukuran berbagai konstruk psikologis tersebut. Secara sempit, psikometri sering diartikan sebagai pengembangan metodologi matematis atau statistis untuk menganalisis data pengukuran aneka konstruk psikologis (Browne, 2000).

Secara lebih rinci, kegiatan ilmiah psikometri atau pengukuran psikologis lazim digolongkan ke dalam tiga wilayah besar, yaitu (1) *mental test theory* atau pengembangan teori tes mental atau tes psikologis, (2) pengembangan analisis faktor dan berbagai metode terkait, dan (3) pengembangan penskalaan multidimensional (Browne, 2000). Teori tes mental menggeluti pengembangan metodologi untuk menganalisis tes psikologis. Hingga kini dikenal dua pendekatan umum untuk menganalisis aneka tes yang terdiri dari item-item dikotomis (pemberian skor 1 untuk jawaban benar atau sesuai kunci jawaban dan skor 0 untuk jawaban salah atau tidak sesuai dengan kunci jawaban), yaitu *classical test theory* atau teori tes klasik dan *item response theory* atau teori respon item. Analisis faktor dipelopori oleh Charles Spearman yang berkreasi mengembangkan model analisis faktor dengan sebuah faktor umum tunggal yang merepresentasikan *general intelligence* atau inteligensi umum sekitar awal abad ke-20. Kini metodologi faktorial bisa digolongkan ke dalam dua kategori, yaitu analisis faktor *eksploratori* sebagaimana dikembangkan oleh Spearman, Thurstone dan lain-lain, serta faktor analisis *konfirmatori* sebagai metode khusus *structural equation modeling* dan pengujian hipotesis sebagaimana terutama dikembangkan oleh Joreskog. *Multidimensional scaling* atau penskalaan multidimensional adalah metode untuk menyelidiki jarak antar berbagai atribut konstruk atau atribut psikologis (Browne, 2000; Carroll, 1992).

Konon kelahiran psikometri atau pengukuran psikologis sebagai cabang psikologi ditandai oleh pembentukan *the Psychometric Society* atau Masyarakat Psikometri pada 1935 di Amerika Serikat di bawah kepemimpinan L.L. Thurstone – seorang pakar psikometri di Universitas Chicago kala itu – sebagai presidennya yang pertama. Organisasi profesi ini menerbitkan sebuah jurnal, *Psychometrica,* dengan tujuan "development of psychology as a quantitative rational science" atau mengembangkan psikologi sebagai sebuah disiplin ilmu rasional-kuantitatif. Langkah ini diikuti dengan penerbitan jurnal-jurnal lain di bidang pengukuran psikologis baik di dalam maupun di luar Amerika Serikat, termasuk *British Journal of Mathematical and Statistical Psychology* di Inggris maupun *Behaviormetrika,* sebuah jurnal pengukuran psikologis berbahasa Inggris di Jepang (Browne, 2000). Sesudah meninjau sekilas latar belakang pengertian maupun sejarah psikometri atau pengukuran psikologis, marilah sekarang kita lihat secara lebih cermat beberapa konsep dasarnya.

B. Pengukuran Psikologis

Secara umum sebagaimana antara lain tercantum dalam Kamus Besar Bahasa Indonesia (2005), kata "mengukur" memiliki makna "menghitung ukurannya (panjang, besar, luas, tinggi, dsb) dengan alat tertentu." Makna ini masih agak jauh dari yang kita perlukan dalam rangka membahas pokok tentang pengukuran psikologis. Dari kepustakaan psikologi, kita menemukan beberapa definisi kata pengukuran sebagai berikut. Pertama adalah definisi yang dikemukakan oleh salah seorang pioner pengembangan pengukuran psikologis, yaitu S.S. Stevens (1946). Menurutnya, pengukuran adalah "*the assignment of numerals to objects or events according to rules*" (h. 677). Maksudnya, pengukuran adalah peneraan bilangan pada objek atau peristiwa menurut aturan tertentu. Definisi kedua, "*measurement consists of rules for assigning numbers to objects in such a way as to represent quantities of attributes*" (Nunnally, Jr., 1970, h. 7). Maksudnya, pengukuran terdiri atas seperangkat aturan untuk

menerakan bilangan pada aneka objek dengan cara sedemikian rupa untuk mencerminkan atau mengungkapkan kuantitas dari aneka atribut. Definisi ketiga, "*measurement is the assigning of numbers to individuals in a systematic way as a means of representing properties of the individuals*" (Allen & Yen, 1979, h. 2). Maksudnya, pengukuran adalah pengenaan atau penetapan bilangan pada individu-individu secara sistematis sebagai cara mencerminkan atau mengungkapkan aneka ciri dari individu-individu yang bersangkutan.

Ketiga definisi dari kepustakaan psikologi di atas memiliki kesamaan, yaitu bahwa pengukuran adalah peneraan bilangan. Perbedaannya, definisi pertama menekankan adanya aturan yang mendasari peneraan tersebut. Definisi kedua dan ketiga sama-sama menekankan setidaknya *tiga* ciri penting pengukuran, yaitu: (1) sifat sistematis, (2) sasarannya adalah atribut, dan (3) proses kuantifikasi.

1. Pengukuran Psikologis bersifat Sistematis

Dalam pengukuran psikologis, peneraan bilangan pada objek atau individu yang dimaksud tidak dilakukan secara sembarangan melainkan "*in such a way*" atau bahkan "*in a systematic way*" khususnya mengikuti "*rules*" atau seperangkat aturan tertentu. Unsur ini memiliki setidaknya dua makna.

Makna pertama, prosedur yang digunakan untuk menerakan bilangan yang dimaksud harus dirumuskan secara eksplisit (Nunnally, Jr., 1970). Misal, mengukur jarak antara titik sepak pinalti dan posisi penjaga gawang dari kesebelasan yang terkena hukuman pada garis di bawah mistar gawang dalam permainan sepak bola, tepat dua belas langkah orang dewasa. Atau, mengukur berat badan dengan *bathroom scale* dalam satuan kilogram. Kedua contoh tersebut memang pengukuran fisik, bukan pengukuran psikologis, namun sekadar memberikan gambaran bahwa prosedur untuk menerakan atau menetapkan bilangan yang dimaksud perlu dirumuskan secara eksplisit.

Makna kedua, selain dirumuskan secara eksplisit prosedur peneraan atau penetapan bilangan tersebut juga harus bersifat *standardized* atau dibakukan. "*A measure is said to be 'well standardized' if different people employ the same measure obtain very similar results*" (Nunnally, Jr., 1970). Maksudnya, sebuah ukuran disebut terbakukan dengan baik bilamana berbagai orang yang menggunakan ukuran yang sama tersebut akan memperoleh hasil pengukuran yang sangat mirip. Jika lima wasit pertandingan sepak bola diminta menghitung dua belas langkah dari garis di bawah mistar gawang ke titik sepak pinalti dan kelimanya sampai ke titik yang kurang lebih atau bahkan persis sama, berarti ukuran dua belas pas atau langkah tersebut standar atau baku.

Contoh lain, jika dua dokter mata mengukur kadar kerusakan penglihatan jauh seorang pasien dengan seperangkat alat yang sama dan sampai pada kesimpulan yang sama, berarti perangkat alat yang dipakai kedua dokter mata tersebut standar atau baku sehingga ukuran kadar kerusakan penglihatan jauh pasien yang dihasilkannya pun juga standar atau baku. Lantas, apa kaitan antara kedua makna dalam unsur pertama dari definisi pengukuran ini? Menurut Nunnally, Jr. (1970), "*formulating explicit rules for the assignment of numbers is a major aspect of the standardization of measures*" (h. 7). Artinya, perumusan secara eksplisit perangkat aturan untuk peneraan atau penetapan bilangan merupakan aspek utama standarisasi atau pembakuan pengukuran.

2. Sasaran Pengukuran Psikologis adalah Atribut

Dalam pengukuran pada umumnya maupun pengukuran psikologis pada khususnya, peneraan atau penetapan bilangan dilakukan dengan tujuan untuk "*to represent quantities of* ***attributes***" atau "*representing* ***properties*** *of the individual.*" Jadi, sasaran pengukuran bukan objek atau individu atau orang, melainkan *properties* atau *attributes* dari objek atau individu. Maksudnya, dalam pengukuran

yang menjadi sasaran adalah *aspek* atau *segi* tertentu dari objek atau orang yang disebut ***atribut,*** bukan objek atau orangnya sendiri. Atribut pada orang dapat bersifat fisik, seperti tinggi badan, berat badan, kekuatan kepalan tangan, dan sebagainya. Atribut fisik lazim bersifat *konkret* dan bisa diamati atau bisa diukur secara langsung. Namun atribut pada orang dapat pula bersifat psikologis seperti kecerdasan, kepemimpinan, kematangan emosi. Atribut psikologis lazim bersifat *abstrak* dan tidak bisa diamati atau diukur secara langsung.

Terkait sifat abstrak atribut psikologis yang menjadi sasaran pengukuran, maka pengukuran psikologis menuntut *proses abstraksi* (Nunnally, Jr., 1970). Melalui proses abstraksi, seorang ilmuwan psikologi yang bermaksud mengukur suatu atribut psikologis tertentu dituntut terlebih dulu *mendefinisikan* hakikat atribut yang akan diukurnya itu. Pendefinisian yang dimaksud berlangsung dalam dua langkah atau tahap, yaitu tahap pertama berupa pendefinisian secara konseptual dan tahap kedua berupa pendefinisian secara operasional.

Pada tahap pertama, sang ilmuwan psikologi yang bermaksud mengukur sebuah atribut psikologis perlu merumuskan ***definisi konseptual*** atau ***definisi teoretis*** atribut yang bersangkutan. Di sini atribut yang hendak diukur tersebut dipandang sebagai sebuah ***konsep*** atau pengertian baru, dan dicoba dijelaskan dengan menggunakan konsep-konsep lain yang sudah lebih dikenal (Blalock, Jr., 1979). Misal, seorang pakar psikologi menjelaskan konsep *reaction time* atau waktu reaksi sebagai "*the time between the onset of a stimulus that serves as a signal and the beginning of a response that is made as quickly as possible to the signal"* (Yaremko *et al.,* 1982). Dalam contoh ini, *waktu reaksi* adalah sebuah atribut yang dinyatakan sebagai sebuah konsep baru. Konsep baru tersebut dijelaskan dengan menggunakan sejumlah konsep (kunci) lain yang sudah lebih kita kenal. Beberapa konsep kunci yang dimaksud adalah "jeda", "stimulus", "tanda", "respon", dan "secepat mungkin." Sehingga konsep waktu reaksi kita pahami sebagai "**jeda** antara kemunculan sebuah **stimulus** yang berfungsi

sebagai **tanda** dan dimulainya sebuah **respon** yang dilakukan **secepat mungkin** terhadap tanda tersebut."

Dalam pengukuran psikologis kebanyakan atribut atau konsep yang menjadi sasaran pengukuran merupakan hasil ***konstruksi*** atau pemikiran para pakar psikologi bertolak dari pengamatan mereka terhadap fenomena perilaku tertentu. Contoh, saat diberi tugas oleh Menteri Pendidikan untuk mengidentifikasikan murid-murid yang mengalami kesulitan menempuh pengajaran dalam sekolah umum di kota Paris, Prancis, Alfred Binet dan rekannya Victor Henri Simon pada 1904 merancang alat yang kemudian menjadi salah satu cikal bakal tes inteligensi, khususnya untuk anak-anak.

Mereka mendefinisikan inteligensi atau kecerdasan sebagai "*judgment, otherwise known as good sense, practical sense, initiative, or the faculty of adapting oneself. To judge well, to understand well – these are the essential wellsprings of intelligence*" (Binet & Simon, 1905, dalam Gregory, 2007, h. 56). Jadi, menurut Binet dan Simon, sumber yang membuat sebagian murid tersebut tidak mampu mengikuti pengajaran di sekolah seperti kebanyakan murid lain terletak pada inteligensi atau kecerdasan mereka yang kurang. Binet dan Simon mendefinisikan kecerdasan sebagai kemampuan memberikan penilaian, kemampuan berpikir, kemampuan mengerjakan hal-hal praktis, kemampuan berinisiatif, atau kemampuan beradaptasi atau menyesuaikan diri, khususnya pada tugas-tugas yang dihadapi dalam pengajaran di sekolah. Jadi, kendati gejala-gejalanya tampak dalam perilaku, namun inteligensi atau kecerdasan seperti juga banyak konsep psikologis lainnya merupakan ***konstruk*** teoretis**,** yaitu hasil konstruksi atau pemikiran konseptual-teoretis pakar psikologi yang bersifat abstrak dalam arti tidak bisa diamati secara langsung.

Sesudah berhasil didefinisikan secara konseptual menjadi sebuah konstruk, sebuah atribut psikologis yang bersifat abstrak tersebut perlu didefinisikan pada langkah atau tahap berikutnya, yaitu tahap ***definisi operasional.*** Definisi operasional adalah definisi yang menyatakan operasi atau prosedur aktual yang dipakai untuk

mengukur atribut yang bersangkutan (Blalock, Jr., 1979). Langkah ini merupakan perwujudan dari apa yang dikenal sebagai *operasionisme* dalam psikologi (Stevens, 1935). Operasionisme berasal dari kata *operasi*. Menurut Stevens (1935), "operation is the performance which we execute in order to make known a concept" (h. 323). Maksudnya, operasi adalah kinerja atau tindakan yang kita lakukan untuk menjelaskan sebuah konsep. Dalam psikologi, konsep yang diungkapkan dalam kata atau pernyataan menjadi bermakna hanya jika kriteria kebenaran atau keberadaannya terdiri atas sejumlah operasi atau tindakan yang bisa dilaksanakan (Stevens, 1936). Konsep kecerdasan misalnya, hanya menjadi bermakna karena kriteria keberadaannya mencakup rangkaian tindakan seperti mampu menyebutkan secara tepat bagian yang kurang atau hilang dalam gambar, mampu merangkai potongan-potongan gambar menjadi sebuah kisah, dan sebagainya.

Salah satu bentuk operasi yang paling fundamental adalah *denoting* atau *pointing to* atau "menunjuk" pada benda atau peristiwa yang dimaksudkan oleh sebuah istilah atau konsep tertentu (Stevens, 1935). Namun jika dicermati, dalam tindakan menunjuk selalu terkandung perbuatan melakukan *diskriminasi* atau pembedaan. Maka ada yang menyatakan bahwa "discrimination is the *sine qua non* of any and every operation including that of denoting". Maksudnya, diskriminasi merupakan kondisi atau syarat mutlak setiap bentuk operasi termasuk penunjukan. Bahkan, "discrimination is the fundamental operation of all science" atau bahwa diskriminasi atau tindakan melakukan pembedaan adalah operasi fundamental semua ilmu pengetahuan (Stevens, 1935, h. 324). Dengan demikian, operasionisme adalah "referring any concept for its definition to the concrete operations by which knowledge of the thing in question is had." Maksudnya, operasionisme adalah tindakan mendefinisikan setiap konsep dengan cara mengacu atau menunjuk pada aneka operasi konkret yang mengandung penjelasan tentang konsep yang dimaksud (Stevens, 1935, h. 323).

Dalam praktek, operasionalisasi seperti diuraikan di atas dapat dilakukan mengikuti langkah-langkah sebagai berikut. Pertama-tama konsep atau konstruk tersebut perlu di-***eksplikasi***-kan, yaitu diidentifikasikan *behavioral indicators* atau indikator-indikator tingkah lakunya berupa bentuk-bentuk tingkah laku spesifik yang bisa diamati dan diukur, baik yang bersifat mendukung (*favorable*) maupun yang bersifat menyangkal atau mengingkari (*unfavorable*) keberadaan konstruk psikologis yang bersangkutan. Langkah ini disebut **eksplikasi konstruk** (Friedenberg, 1995).

Sebagai contoh, sesudah mendefinisikan inteligensi sebagai kemampuan beradaptasi, Binet dan Simon selanjutnya mengidentifikasikan bentuk-bentuk tingkah laku spesifik sebagai indikator-indikator tingkah lakunya. Contoh sebagian bentuk tingkah laku yang dijadikan indikator tingkah laku inteligensi oleh Binet dan Simon adalah sebagaimana disajikan pada Tabel 2.1. Bentuk-bentuk tingkah laku ini memang bersifat mendukung (*favorable)* keberadaan konstruk inteligensi. Jika seorang anak mampu melakukannya dengan berhasil, hal itu mengindikasikan secara positif atau *favorable* (keberadaan atau kebenaran) inteligensinya. Sebaliknya jika dia gagal melakukannya dengan berhasil, hal itu mengindikasikan secara negatif atau *unfavorable* (ketidak-beradaan atau ketidak-benaran) inteligensinya.

Tabel 2.1.
Indikator Tingkah Laku Inteligensi Anak ala Binet dan Simon (1905)

Nomor	Bentuk tingkah Laku
1	Mengikuti dengan kedua mata sebuah objek yang bergerak.
2	Memegang sebuah objek kecil yang disentuh.
3	Mengupas dan memakan sebuah kubus cokelat yang terbungkus kertas.
4	Melaksanakan beberapa perintah sederhana dan menirukan beberapa gerakan sederhana.
5	Mengulang menyebutkan tiga bilangan.
6	Membandingkan berat dua benda.
7	Mengulang sebuah kalimat terdiri atas 15 kata.
8	Menyebutkan persamaan antara dua benda, e.g. "kupu-kupu dan kutu."
9	Membandingkan dua garis dengan perbedaan panjang yang tipis.
10	Merangkai tiga kata menjadi sebuah kalimat.

Sumber: Gregory (2007), h. 57.

Selanjutnya, berdasarkan indikator-indikator tingkah laku baik yang *favorable* maupun *unfavorable* tersebut dikembangkan aneka operasi yang mengandung penjelasan tentang konsep atau atribut yang dimaksud, baik secara positif maupun secara negatif. Pada kasus pengukuran kecerdasan anak yang dilakukan oleh Binet dan Simon (1905), indikator-indikator tingkah laku tersebut sekaligus dipakai sebagai item-item – jadi, sebagai prosedur aktual -- untuk mengukur taraf inteligensi anak-anak. Hasilnya dipakai sebagai dasar untuk memilah mereka yang dipandang mampu mengikuti pendidikan di sekolah umum dan mereka yang dipandang perlu mengikuti pendidikan di sekolah yang khusus diperuntukkan bagi anak-anak yang saat itu disebut memiliki *mental retardation* atau keterbelakangan mental, atau yang kini lebih lazim disebut anak-anak *difabel* atau anak-anak dengan *different abilities* atau memiliki kemampuan yang berbeda dari anak pada umumnya.

3. Pengukuran Merupakan Proses Kuantifikasi

Pengukuran pada umumnya dan pengukuran psikologis pada khususnya pada dasarnya merupakan proses ***kuantifikasi*** (Nunnally, Jr., 1974). Maksudnya, bilangan digunakan untuk menyatakan kuantitas atau jumlah. Melalui kuantifikasi kita hendak menyatakan jumlah atau banyaknya atribut yang terdapat dalam objek. Dalam pengukuran psikologis, kita hendak menyatakan jumlah atau banyaknya atribut psikologis, misal inteligensi atau kecerdasan, yang terdapat dalam diri individu. Penentuan kuantitas atau jumlah atribut ini lazim dilakukan dengan cara *counting* atau menghitung (Nunnally, Jr., 1974). Dalam contoh pengukuran inteligensi murid-murid sekolah di kota Paris yang dilakukan oleh Binet dan Simon pada tahun 1905, penentuan kuantitas inteligensi masing-masing anak dilakukan dengan cara menghitung jumlah atau banyaknya bentuk tingkah laku yang telah ditetapkan sebagai indikator inteligensi, yang dapat dikerjakan dengan tepat atau berhasil oleh masing-masing murid.

Seperti juga tampak dalam contoh pengukuran inteligensi anak yang dilakukan oleh Binet dan Simon di atas, jumlah atau banyaknya inteligensi sebagai sebuah atribut psikologis tersebut bervariasi atau berlainan antar murid. Dengan kata lain, inteligensi seperti juga semua atribut psikologis yang lain merupakan suatu ***variabel***, yaitu suatu entitas yang ada dalam kuantitas atau jumlah yang bervariasi atau berlainan dari murid ke murid atau dari orang ke orang. Inteligensi dan seperti juga semua atribut psikologis yang lain bukan merupakan suatu *constant* atau *konstanta,* suatu entitas yang ada dalam kuantitas atau jumlah yang sama atau tetap pada setiap orang. Variasi atau keberagaman kuantitas atribut psikologis antar orang inilah yang melahirkan gejala yang disebut *individual differences* atau kekhususan antar individu serta yang memberikan ciri atau sifat unik masing-masing individu atau orang. Sebagaimana kita ketahui, studi tentang kekhususan antar individu merupakan salah satu bidang kajian utama pengukuran psikologis.

C. Psikofisika Klasik

Kini ada dua pertanyaan yang menuntut jawaban. *Pertama*, apa dasar Binet dan Simon maupun para pakar-praktisi pengukuran psikologis lainnya mengaitkan tingkah laku konkret murid-murid yang langsung bisa diamati dan dicatat tersebut dengan atribut psikologis inteligensi yang terdapat dalam diri masing-masing anak dan yang bersifat abstrak? *Kedua*, apa dasar Binet dan Simon serta para pakar psikometri lain melekatkan atau menerakan (istilah Inggris, *assign*) bilangan pada tingkah laku konkret dalam rangka mengukur atribut psikologis yang tidak kasat mata tersebut?

Pertanyaan pertama kiranya bisa dijawab dengan meminjam gagasan pokok yang ditawarkan oleh ***psikofisika,*** yaitu cabang ilmu yang menjadi salah satu pilar tradisi pengukuran psikologis (Guilford, 1954). *Psikofisika* sendiri adalah "*an exact science of the functional relations of dependency between body and mind*" atau cabang ilmu eksakta tentang hubungan ketergantungan fungsional antara tubuh dan jiwa, atau "*the science that investigates the quantitative relationships between physical events and corresponding psychological events*" atau cabang ilmu yang menyelidiki hubungan kuantitatif antara aneka peristiwa fisik dan aneka peristiwa psikologis terkait, atau dalam bahasa teknis hubungan kuantitatif antara *stimulus* dan *respon* (Guilford, 1954, h. 3, 20).

Berdasarkan penelitian-penelitian tentang proses *sensasi* atau penginderaan pada subjek manusia, psikofisika klasik berasumsi bahwa setiap peristiwa pengukuran proses penginderaan – misal, membandingkan berat serangkaian objek -- akan melibatkan dua variabel kuantitatif yang membentuk suatu kontinum, yaitu sebuah variabel fisik yang membentuk *kontinum fisik* dan yang secara paralel dan serentak dibarengi oleh sebuah variabel psikologis yang membentuk *kontinum psikologis* (Guilford, 1954). Kontinum sendiri adalah "*a closely graded series, one step merging imperceptibly into the next, the whole forming a straight line signifying changes in a single direction*" (Guilford, 1954, h. 21). Artinya, kontinum adalah suatu rangkaian

bertingkat yang rapat, tingkat yang satu lebur secara mulus ke tingkat berikutnya, sehingga secara keseluruhan membentuk sebuah garis lurus yang mencerminkan rangkaian perubahan dari suatu titik (rendah) menuju ke suatu titik lain (tinggi). Transisi atau perpindahan dari tingkat yang lebih rendah ke tingkat yang lebih tinggi atau sebaliknya lazim berlangsung secara gradual atau bertahap.

Kontinum fisik mencerminkan perubahan dalam variabel fisik tertentu yang dapat diukur dalam satuan ukuran fisik, sedangkan kontinum psikologisnya mencerminkan bentuk pengalaman indera terkait. Dalam contoh kasus membandingkan berat serangkaian objek, kontinum fisiknya adalah berat rangkaian objek dengan satuan ukuran gram, sedangkan kontinum psikologisnya adalah sensasi tekanan dengan kuantitas yang berlainan sesuai berat masing-masing objek. Dengan kata lain, kontinum fisik merupakan *kontinum stimulus,* sedangkan kontinum psikologisnya merupakan *kontinum respon.*

Perubahan pada kontinum fisik atau stimulus bisa langsung diamati berupa perubahan objek dengan berat yang semakin meningkat. Perubahan dalam kontinum psikologis atau respon yang menyertainya berupa sensasi tekanan dengan kuantitas yang juga semakin meningkat tidak bisa diamati secara langsung sebab bersifat *covert* atau tertutup dalam arti berlangsung dalam diri individu. Perubahan pada kontinum psikologis tersebut baru bisa di-*inferensi*-kan atau disimpulkan manakala subjek diminta mengungkapkan pengalaman sensasinya dalam bentuk *verbal report* atau laporan verbal berupa *judgment* atau penilaian. Dengan kata lain, muncullah kontinum ketiga berupa *kontinum penilaian.* Kontinum penilaian ini diasumsikan bersifat paralel dalam arti berkorelasi secara sempurna dengan kontinum respon dan dengan demikian juga terkait dengan kontinum stimulus (Guilford, 1954).

Penerapan gagasan dasar psikofisika dalam pengukuran psikologis pada umumnya kurang lebih akan menghasilkan gambaran sebagai berikut. Penyajian aneka tugas yang harus dikerjakan oleh subjek dalam rangka pengukuran atribut psikologis tertentu akan membentuk sejenis *kontinum fisik*. *Kontinum respon* yang merupakan

manifestasi atau perwujudan pemilikan subjek atas atribut terkait dalam kuantitas atau jumlah tertentu terbentuk dalam diri subjek saat subjek mengerjakan tugas-tugas tersebut. Kontinum respon ini bersifat *covert* atau tidak bisa diamati secara langsung sebab berupa proses psikologis yang berlangsung dalam diri subjek. Keberadaan kontinum respon yang mencerminkan pemilikan subjek atas atribut psikologis yang sedang diukur dalam kuantitas tertentu tersebut baru bisa disimpulkan hanya sesudah subjek mengungkapkan responnya dalam bentuk tingkah laku *overt* atau yang bisa diamati secara langsung, entah berupa ungkapan penilaian (*judgment)* atau ungkapan perasaan (*sentiment)* sebagai bentuk manifestasi atau perwujudan pengerjaan tugas-tugas yang diberikan dalam rangka pengukuran. Ungkapan penilaian atau ungkapan perasaan ini akan membentuk *kontinum penilaian* atau *kontinum perasaan,* dan dari situ kita bisa membuat *inferensi* atau kesimpulan tentang *kontinum respon* yang mencerminkan kuantitas atribut yang sedang diukur yang terdapat dalam diri subjek. Begitulah kurang lebih prinsip dasar kerja *psikometri* atau pengukuran psikologis.

Jawaban atas pertanyaan kedua terkait keabsahan kita melekatkan bilangan pada tingkah laku dalam rangka mengukur sebuah atribut psikologis dan yang merupakan hakikat pengukuran (psikologis), kiranya bisa kita peroleh dari pandangan dua tokoh, yaitu Thurstone (1959, dalam Andrich, 1990) dan Stevens (1946). Menurut Thurstone, salah satu prinsip yang mendasari pengukuran adalah konsep *unidimensionalitas***.** Maksudnya, semua fenomena yang kompleks termasuk atribut psikologis dapat ditempatkan dalam satu kontinum tunggal yang bersifat linear atau membentuk garis lurus. Selanjutnya, menurut Stevens, pelekatan atau peneraan bilangan pada objek atau pengukuran yang sudah ditempatkan dalam satu kontinum tersebut menjadi mungkin dilakukan karena "there is a certain isomorphism between what we can do with the aspects of objects and the properties ot the numeral series" (h. 677). Maksudnya, pengukuran termasuk pengukuran psikologis menjadi mungkin dilakukan karena terdapat isomorfisme atau kesamaan bentuk atau struktur antara apa

yang dapat kita lakukan terhadap aspek-aspek objek dengan ciri-ciri rangkaian bilangan. Menurut Stevens (1946), jenis-jenis operasi atau tindakan empiris yang dapat kita lakukan terkait aspek-aspek objek dan yang memiliki kesamaan struktur dengan ciri-ciri rangkaian bilangan meliputi: (1) penentuan kesetaraan atau kesamaan atau pengklasifikasian, (2) *rank ordering* atau penjenjangan, (3) penentuan kapan perbedaan antar aspek-aspek objek adalah setara atau sama, dan (4) penentuan kapan rasio antar aspek-aspek objek adalah setara atau sama. Sebagaimana akan kita lihat di bagian berikut, jenis operasi atau tindakan empiris yang kita lakukan terhadap aspek-aspek objek yang menjadi sasaran pengukuran kita akan menentukan jenis skala atau taraf pengukurannya (Stevens, 1946).

D. Taraf-taraf Pengukuran

Sudah kita lihat bahwa dalam arti paling luas, pengukuran adalah peneraan atau pelekatan bilangan pada objek (termasuk orang) atau peristiwa menurut aturan tertentu dengan tujuan untuk melukiskan sifat-sifat objek atau peristiwa yang bersangkutan (Lord, 1954). Aturan yang dipakai sebagai dasar peneraan atau pelekatan bilangan pada objek atau peristiwa ini ada beberapa macam. Perbedaan aturan ini menghasilkan jenis skala pengukuran atau lebih mudah dipahami sebagai taraf pengukuran yang berlainan.

Pakar yang berjasa membuat penggolongan tentang jenis skala pengukuran adalah Stevens (1946). Dalam artikelnya yang kini menjadi klasik, dia menggunakan istilah "scales of measurement" atau skala pengukuran. Terkesan ada perbedaan pengertian antara istilah *skala* atau *pensakalaan* dan istilah *pengukuran*. Namun menurut seorang pakar lain, "scaling appears to be virtually indistinguishable in meaning from measurement" (Lord, 1954, h. 375). Maksudnya, makna istilah penskalaan sesungguhnya tidak bisa dibedakan atau sama dengan makna istilah pengukuran. Untuk menghormati jasa Stevens sekaligus mengikuti kelaziman yang kini berkembang, kita

akan menggunakan istilah "jenis skala" atau "taraf pengukuran" untuk mengacu pada temuan Stevens bahwa ada sejumlah jenis skala pengukuran berdasarkan aturan yang diikuti.

Aturan pokok yang menentukan jenis skala atau taraf pengukuran yang dimaksud kiranya adalah "the character of the basic empirical operations performed" (Stevens, 1946) atau ciri atau lebih tepat jenis operasi empiris yang mendasari pelekatan atau peneraan bilangan pada objek atau kejadian dalam kegiatan pengukuran. Seperti sudah disinggung, menurut Stevens (1946) ada empat macam operasi empiris dasar yang kemudian dipakai sebagai aturan dalam pengukuran, yaitu: (1) *determination of equality* atau penetapan kesetaraan atau kesamaan atau pengklasifikasian objek-objek atau kejadian-kejadian, (2) *rank-ordering* atau *determination of greater or less* atau penetapan urutan jenjang atau penetapan mana yang lebih dan mana yang kurang di antara objek-objek atau kejadian-kejadian, (3) *determination of equality of intervals or differences* atau penetapan kesetaraan atau kesamaan interval atau perbedaan di antara objek-objek atau kejadian-kejadian, dan (4) *determination of equality of ratios* atau penetapan kesetaraan atau kesamaan rasio di antara objek-objek atau kejadian-kejadian.

Selain itu, jenis operasi empiris dasar ini kemudian juga berdampak terhadap jenis transformasi atau perubahan terhadap skala yang *admissible* atau diperkenankan, yang menjadi aturan berikut. Sebagaimana kita tahu, skala atau pengukuran menghasilkan bilangan yang mencerminkan karakteristik objek yang diukur. Skala semacam ini tidak boleh ditransformasikan atau diubah dengan cara-cara yang berakibat mengubah dalam arti mengacaukan atau merusak pencerminan karakteristik objek yang diukur. Untuk masing-masing taraf pengukuran, ada jenis transformasi atau perubahan skala tertentu yang *admissible* atau diperkenankan sebab jenis transformasi tersebut tidak mengacaukan atau tetap mempertahankan pencerminan karakteristik objek yang dihasilkan oleh jenis skala sebelumnya (Stevens, 1946; Allen & Yen, 1979).

Dalam bahasa yang disederhanakan, jenis skala atau taraf pengukuran sebagaimana diidentifikasikan oleh Stevens (1946) berdasarkan jenis operasi yang mendasarinya di atas menentukan jenis informasi yang kita peroleh dari bilangan yang kita dapatkan sebagai hasil pengukuran. Jenis informasi terkait dengan operasi empiris yang mendasari peneraan bilangan tersebut mencakup: (1) informasi tentang **identitas** sebagai hasil dari penetapan kesamaan antar objek, (2) informasi tentang **urutan** atau **jenjang** sebagai hasil penetapan mana yang lebih besar dan mana yang lebih kecil di antara objek-objek, (3) informasi tentang **interval** atau **jarak yang sama** sebagai hasil penetapan kesamaan perbedaan di antara objek-objek, dan (4) informasi tentang keberadaan **nilai nol absolut** atau **mutlak** sebagai hasil penetapan kesamaan rasio di antara objek-objek. Tiga kualitas informasi yang terakhir terkait dengan nilai numerik bilangan, yaitu fungsi bilangan dalam menunjukkan kuantitas atau jumlah dari atribut yang menjadi sasaran pengukuran.

Jenis skala atau taraf pengukuran sebagaimana diidentifikasikan oleh Stevens (1946) berdasarkan jenis operasi yang mendasarinya di atas selanjutnya juga menentukan jenis-jenis operasi statistik yang cocok untuk diterapkan pada hasil pengukuran yang diperoleh pada masing-masing jenis skala atau taraf pengukuran.

Selain itu, menurut Stevens (1946), jenis operasi dasar yang menentukan jenis skala dan jenis operasi statistik yang sesuai untuk masing-masing skala tersebut bersifat **kumulatif**. Artinya, setiap operasi dasar yang menentukan jenis skala tertentu beserta jenis operasi statistiknya yang sesuai mencakup aneka operasi dasar lain yang menentukan jenis-jenis skala beserta jenis-jenis operasi statistik masing-masing yang mendahuluinya. Sebagai contoh, dalam mengukur sejumlah objek atau kejadian kita bisa memastikan diri memperoleh sebuah skala interval hanya apabila kita bisa menentukan kesamaan interval, urutan jenjang, dan kesamaan di antara objek atau kejadian yang kita ukur. Sebagai konsekuensinya, jika kita bisa secara sah menghitung *Mean* dan *Standard Deviation* dari hasil pengukuran yang dipastikan kita peroleh pada taraf interval, maka kita pun bisa

secara sah menghitung jenis-jenis operasi statistik lainnya yang lebih sederhana seperti *Median, Mode,* dan *N* atau jumlah kasus. Jenis skala atau taraf pengukuran beserta aneka konsekuensinya tersebut menurut Stevens (1946) adalah: (1) pengukuran nominal, (2) pengukuran ordinal, (3) pengukuran interval, dan (5) pengukuran rasio.

1. Pengukuran Nominal

Pada taraf pengukuran ini bilangan hanya dipakai sebagai *nomen* (kata Latin berarti *nama*), dan tidak memiliki nilai numerik. Bilangan dikenakan pada satu atau serangkaian objek atau peristiwa sekadar sebagai tanda untuk menunjukkan identitas. Jika satu bilangan dikenakan pada satu objek, maka bilangan tersebut dipakai sebagai *label.* Contohnya, setiap pemain sepak bola ditandai dengan bilangan tertentu yang dituliskan pada bagian punggung kostum sepak bola yang dikenakannya sebagai *nomor punggung*. Label bilangan ini dikenakan secara tetap kepada setiap pemain dan menjadi identitasnya. Sebagai contoh, *David Beckham* pemain klub *Manchester United* dari Inggris menyandang nomor punggung 10. *Cristiano Ronaldo* saat menjadi pemain pada klub yang sama menyandang nomor punggung 7. Karena bilangan-bilangan tersebut tidak memiliki nilai numerik atau tidak mencerminkan kuantitas tertentu, tidak bisa dikatakan bahwa *Beckham* dengan nomor punggung 10 melebihi atau mengungguli *Ronaldo* yang bernomor punggung 7, sekalipun dalam kenyataan dan menyangkut atribut tertentu yang tidak terkait langsung dengan sepak bola, seperti misalnya total kekayaan yang dimiliki, hal itu mungkin saja benar. Artinya, David Beckham yang bernomor punggung 10 tersebut memiliki total kekayaan mengungguli Cristiano Ronaldo yang bernomor punggung 7.

Jika satu bilangan dikenakan pada lebih dari satu objek, maka bilangan tersebut dipakai sebagai *kategori.* Sebagai contoh, ada dua jenis kelamin yang secara resmi diakui dalam pencatatan data kependudukan, yaitu lelaki atau pria dan perempuan atau

wanita. Untuk memudahkan pencatatan dan pengolahan data, jenis kelamin lelaki dikenai bilangan 1 dan jenis kelamin perempuan dikenai bilangan 2, atau sebaliknya. Semua penduduk lelaki tanpa memedulikan latar belakang etnik, agama, pendidikan, tingkat sosial ekonomi maupun semua atribut lain kecuali jenis kelaminnya tersebut dikenai bilangan 1, sedangkan semua penduduk perempuan dikenai bilangan 2. Artinya, semua penduduk berjenis kelamin lelaki dimasukkan ke dalam kategori lelaki, dan semua penduduk berjenis kelamin perempuan dimasukkan ke dalam kategori perempuan.

Skala nominal lazim diterapkan dalam pengukuran variabel *diskret,* yaitu variabel yang hanya memiliki nilai-nilai bulat atau utuh (Yaremko, Harari, Harrison, & Lynn, 1982). Variabel diskret tidak mengenal nilai tengah, sehingga ada diskontinuitas, keterputusan, atau jurang antara nilai yang satu dan yang lain. Contohnya, selain jenis kelamin seperti sudah dipaparkan di atas, adalah sejumlah variabel demografik lain seperti agama (Islam, Protestan, Katolik, Hindu, Budha, dsb.), ijazah terakhir (SD, SMP, SMA, S1, dsb.), status perkawinan (kawin, tidak kawin, janda/duda), status pekerjaan (pegawai negeri, pegawai swasta, wiraswasta, dsb.), dan sebagainya.

Stevens (1946) menyebut skala nominal sebagai bentuk peneraan bilangan yang paling tidak mengenal batasan (*unrestricted*) atau bentuk pengukuran primitif yang bahkan konon tidak layak disebut pengukuran sebab menggunakan bilangan sekadar sebagai nama atau label tanpa memanfaatkan nilai numeriknya. Satu-satunya aturan yang membatasi jenis pengukuran nominal adalah: "Do not assign the same numeral to different classes or different numerals to the same class. Beyond that, anything goes with the nominal scale." (Stevens, 1946, h. 679). Artinya, jangan menerakan bilangan yang sama terhadap individu atau kelas yang berbeda atau jangan menerapkan bilangan yang berbeda terhadap individu atau kelas yang sama. Selain itu, khususnya terkait dengan transformasi atau pengubahan skala kita bebas melakukan apa saja dengan skala nominal, seperti menukar bilangan yang satu dengan bilangan yang lain, mengalikan setiap bilangan dengan bilangan konstan atau tetap tertentu, dan

sebagainya. Semua itu tidak akan berdampak apa pun terhadap pencerminan karakteristik objek pengukuran pada taraf pengukuran nominal, sebab sekali lagi bilangan hanya kita pakai sebagai label.

2. Pengukuran Ordinal

Pada taraf pengukuran ini bilangan yang dikenakan pada suatu objek memiliki nilai numerik. Bilangan menunjukkan kuantitas dari atribut yang diukur. Kuantitas itu bergerak secara berkelanjutan dari suatu nilai rendah tak terhingga ke arah suatu nilai tinggi tak terhingga sehingga membentuk sebuah *kontinum* atau bentangan. Makin besar bilangan maka makin besar pula nilai numeriknya dan sebaliknya. Artinya, selain mengandung informasi tentang identitas seperti pada pengukuran nominal, pada pengukuran ordinal bilangan juga bisa dipakai untuk menunjukkan *rank* atau urutan jenjang berdasarkan kuantitas atribut yang diukur. Seorang penjual batu mulia bisa menunjukkan perbedaan tingkat kekerasan (*hardness*) antara tiga jenis batu dagangannya kepada calon pembeli dengan cara menggoreskan masing-masing batu pada sejenis gerinda. Batu yang paling keras tidak mengalami dampak apa pun dari goresan tersebut. Batu yang kedua terkeras mengalami sedikit goresan seperti tampak pada serpihan amat lembut yang tampak di bawah kaca pembesar. Sedangkan batu yang paling lunak mengalami goresan cukup signifikan sehingga bisa diamati dengan mata telanjang. Dengan cara itu penjual batu mulia tersebut bisa membuktikan urutan jenjang kekerasan tiga batu hias dagangannya itu, namun dia tidak bisa menunjukkan apakah interval atau jarak atau perbedaan kekerasan antara batu paling keras dan batu kedua terkeras adalah sama seperti perbedaan kekerasan antara batu kedua terkeras dan batu paling lunak. Dengan kata lain, skala ordinal menunjukkan urutan jenjang terkait atribut tertentu antar sejumlah objek atau kejadian di mana terkait atribut yang sedang menjadi fokus perhatian objek atau kejadian yang satu memiliki kuantitas **lebih** atau **kurang** dibandingkan objek atau kejadian yang lain, namun interval atau jarak antar nilai yang mencerminkan kepemilikan atribut yang

ditempati oleh masing-masing objek atau kejadian tersebut tidak kita ketahui.

Sebagaimana tersirat dalam uraian di atas, pengukuran ordinal hanya bisa mulai diterapkan pada variabel *kontinyu,* yaitu variabel yang bisa ada atau muncul pada nilai berapa pun tanpa batas (Yaremko, Harari, Harrison, & Lynn, 1982). Kebanyakan atribut psikologis yang menjadi objek pengukuran psikologis merupakan variabel kontinyu, seperti inteligensi, sikap terhadap korupsi, nilai hidup, motivasi untuk berprestasi, *locus of control,* dan sebagainya. Pengukuran ordinal tidak dapat diterapkan pada variabel diskret.

Jenis transformasi atau perubahan skala yang diperkenankan untuk dilakukan terhadap skala ordinal disebut *transformasi monotonik,* yaitu jenis transformasi yang tidak mempengaruhi dalam arti mengubah urutan jenjang antar nilai-nilai skala, seperti menambahkan masing-masing skala dengan bilangan konstan tertentu atau mengalikannya dengan sebuah bilangan positif (Allen & Yen, 1979).

3. Pengukuran Interval

Pada taraf pengukuran ini bilangan yang dikenakan pada suatu objek sudah memiliki nilai numerik dan memiliki satuan interval yang sama atau tetap (*equal interval unit)* antar bilangan. Menurut Stevens (1946), pada taraf pengukuran interval inilah sebenarnya kita baru sungguh-sungguh mencapai taraf "kuantitatif" dalam arti yang sebenarnya. Maka selain mengandung informasi tentang identitas dan urutan jenjang, bilangan pada pengukuran interval juga sudah memiliki informasi tentang kesamaan jarak antar bilangan. Seperti pengukuran ordinal, pengukuran interval juga hanya berlaku pada variabel kontinyu. Jika suhu di kota yang terletak di dataran rendah pada ketinggian 100 meter di atas permukaan laut adalah 30 derajat Celsius, sedangkan suhu di kota lain yang terletak di lereng gunung pada ketinggian 500 meter di atas permukaan laut adalah 20 derajat Celsius, maka selisih suhu kedua kota tersebut adalah 10 derajat

Celsius. Perbedaan suhu ini sama seperti perbedaan antara suhu di kota pertama di atas dengan suhu di kota lain lagi yang terletak di tengah gurun pasir sebesar 40 derajat Celsius.

Kendati memiliki kelebihan menunjukkan interval atau jarak antar bilangan yang sama atau tetap, taraf pengukuran ini masih juga punya keterbatasan berupa tidak dimilikinya bilangan nol absolut atau mutlak. Kembali pada contoh di atas, suhu 0 derajat Celsius bukan berarti tidak ada suhu. Bahkan masih ada suhu di bawah 0 derajat Celsius. Selain itu, suhu 0 derajat Celsius setara dengan suhu 32 derajat pada skala pengukur suhu lain, yaitu Fahrenheit. Semua ini menunjukkan keterbatasan pengukuran taraf interval, yaitu hanya mengenal bilangan nol relatif atau tidak mengenal bilangan nol mutlak.

Setiap jenis transformasi linear *admissible* atau diperkenankan untuk dilakukan terhadap skala interval. Rumus transformasi linear adalah $Y = aX + b$, di mana a dan b merupakan bilangan konstan, Y adalah nilai skala baru hasil transformasi, sedangkan X adalah nilai skala asli atau awalnya. Transformasi linear tidak mengubah rasio atau perbandingan jarak antar nilai skala pada skala interval, dengan syarat nilai bilangan konstan a harus lebih besar dari 0 (Allen & Yen, 1979).

Hampir semua operasi statistik yang lazim cocok untuk diterapkan pada data yang diperoleh pada taraf pengukuran interval, kecuali jenis-jenis operasi statistik yang menuntut syarat diketahuinya nilai nol mutlak atau "*true zero point*" atau nilai nol sebenarnya (Stevens, 1946).

4. Pengukuran Rasio

Pada taraf pengukuran ini bilangan yang dikenakan pada suatu objek memiliki kualitas informasi penuh, yaitu menunjukkan identitas, menunjukkan urutan jenjang, menunjukkan interval atau jarak yang sama antar nilai atau bilangan, dan menunjukkan nol mutlak. Jika kita berdiri tepat di tengah-tengah perempatan depan

Kantor Pos Besar di Kota Yogyakarta yang merupakan titik nol kilometer Kota Yogyakarta berarti tidak ada lagi jarak antara tempat kita berdiri dan pusat Kota Yogyakarta. Selanjutnya, dari titik nol itulah jarak antara Kota Yogyakarta dengan tempat-tempat atau kota-kota lain bisa dihitung secara pasti atau eksak. Skala rasio memang lazim diterapkan dalam bidang fisika. Satu-satunya jenis transformasi yang *admissible* atau diperkenankan untuk diterapkan terhadap skala rasio adalah multiplikasi atau pengalian dengan sebuah bilangan konstan. Rumus transformasinya adalah $Y = aX$, di mana bilangan konstan a harus lebih besar dari 0 (Allen & Yen, 1979). Semua jenis operasi atau teknik statistik cocok untuk diterapkan pada data yang diperoleh pada skala rasio (Stevens, 1946).

Ringkasan tentang jenis skala atau taraf pengukuran beserta kualitas informasi dan jenis teknik statistik yang cocok diterapkan pada data pengukuran yang diperoleh pada masing-masing jenis skala atau taraf pengukuran disajikan dalam Tabel 2.2. Seperti sudah disinggung, taraf-taraf pengukuran atau jenis-jenis skala beserta aneka teknik statistiknya yang sesuai bersifat kumulatif, sebagaimana juga tampak pada Tabel 2.2.

Tabel 2.2.
Taraf Pengukuran, Kualitas Informasi, dan Teknik Statistik yang Sesuai

Taraf Pengukuran	Kualitas Informasi				Teknik Statistik yang Sesuai												
	Identitas	Nilai Numerik															
		Urutan	Interval Sama	Nol Mutlak	$\sum$	*P*	*p*	*Rasio*	*Mode*	*Median*	*Persentil*	*Mean*	*SD*	*R*	*r*	*t*	*F*
Nominal	+	-	-	-	+	+	+	+	+	-	-	-	-	-	-	-	-
Ordinal	+	+	-	-	+	+	+	+	+	+	+	-	-	-	-	-	-
Interval	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+
Rasio	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Menurut Stevens (1946), kebanyakan pengukuran psikologis berusaha mencapai skala interval, hanya sebagian kecil saja berhasil. Sementara, upaya merancang operasi untuk menyamakan satuan-satuan skala tidak mudah. Akibatnya, kebanyakan skala yang dipakai secara luas dan efektif di kalangan psikologi merupakan skala ordinal. Mestinya, bahkan perhitungan *Mean* dan *SD* tidak boleh diterapkan pada jenis pengukuran tersebut. Namun praktik yang oleh Stevens (1946) disebut *illegal statisticizing* atau *outlawing* ini lazim dipraktikkan sebab dalam banyak kesempatan ternyata membuahkan hasil yang efektif atau positif. Kenyataan ini semestinya menjadikan para psikolog berhati-hati dalam mengembangkan dan menerapkan praktik pengukuran psikologis. Ψ

Bab 3
Aneka Respon dalam Pengukuran Psikologis

Berbagai atribut psikologis, yaitu dimensi atau aspek kepribadian tertentu yang lazim menjadi objek atau sasaran pengukuran psikologis, seperti kecerdasan, motivasi berprestasi, dan sebagainya, bersifat abstrak dan tidak teramati secara langsung. Terkait situasi ini filsof terkemuka asal Jerman *Immanuel Kant* (Nunnally, Jr., 1970) sampai pernah secara pesimistik menyatakan, "… it would not be possible to have a science of psychology because the basic data could not be observed and measured" (h. 4). Kant pesimis tidak mungkin mengembangkan psikologi menjadi ilmu sebab data mentahnya tidak bisa diamati dan tidak bisa diukur.

Menanggapi pesimisme di atas, lazimnya kalangan psikologi mengemukakan tangkisan argumentatif sebagai berikut:

> "Your subjective experiences – your feelings, sensations, and desires – cannot be observed by others and thus cannot be subjected to measurement. Once you ***do*** something with respect to your feelings – make a judgment, state a preference,or even talk to others about the experience – your behavior meets the requirementsof scientific inquiry, and measurement becomes possible" (Nunnally, Jr., 1970, h. 5).

Aneka pengalaman subjektif kita seperti aneka perasaan, sensasi atau pencerapan, dan hasrat-hasrat kita memang tidak bisa diamati oleh orang lain, sehingga juga tertutup untuk dijadikan objek pengukuran secara langsung. Namun, begitu kita **berbuat** sesuatu terkait dengan perasaan kita itu, misalnya memberikan penilaian, menyatakan pilihan, atau bahkan sekadar mengungkapkan atau membicarakannya dengan orang lain, maka perbuatan atau tindakan kita itu sudah memenuhi syarat sebagai bahan penelitian ilmiah,

sehingga menjadi terbuka untuk diukur. Begitu jawaban para pakar psikometri dalam rangka menepis pesimisme Immanuel Kant.

Dengan kata lain, **respon** dalam arti *overt response* yaitu tindakan atau perbuatan tertentu entah berupa ungkapan penilaian, ungkapan perasaan, atau tingkah laku nyata lainnya terkait dengan suatu proses psikologis tertentu yang bersifat *covert* atau tertutup karena berlangsung di dalam diri kita lazim dipandang sebagai **indikator** dalam arti *behavioral indicators* atau pencerminan atau pengungkapan dalam bentuk tindakan yang bisa diamati dan bisa diukur dari proses atau peristiwa psikologis yang bersifat tersembunyi dan abstrak itu. Prinsip yang dipinjam dari *psikofisika klasik* ini digunakan dalam pengukuran psikologis. Setiap atribut psikologis yang bersifat abstrak dan tidak bisa diamati maupun diukur secara langsung senantiasa memiliki rangkaian tindakan atau perbuatan yang teramati dan terukur yang dipandang sebagai indikator-indikatornya. Dengan mengamati dan mengukur tindakan atau perbuatan tertentu yang ditetapkan sebagai indikatornya, kita bisa melakukan **inferensi** atau penyimpulan tentang atribut psikologis yang dicerminkannya, baik mengenai kuantitas maupun kualitasnya.

Secara sederhana dan sebagaimana sudah kita singgung, pengukuran adalah proses pengenaan atau pelekatan bilangan pada objek tertentu sedemikian rupa mengikuti azas-azas tertentu untuk menunjukkan kuantitas dari aneka atribut yang terdapat di dalam objek yang bersangkutan (Nunnally, Jr., 1970, h. 7). Dalam pengukuran psikologis, yang menjadi objek dalam arti sasaran pengukuran adalah aneka atribut, dimensi, aspek, atau variabel yang bersama-sama membentuk dan memberi ciri unik kepribadian orang. Sebagaimana sudah disinggung, proses pengukuran semacam itu dimungkinkan berkat tersedianya aneka respon yang bisa ditetapkan sebagai indikator perilaku dari masing-masing atribut psikologis yang lazim dijadikan sasaran pengukuran untuk berbagai keperluan. Dalam pengukuran psikologis, respon sebagai indikator perilaku sebuah atribut psikologis bisa dibeda-bedakan menurut beberapa

dasar pengelompokan, khususnya menurut isi, cara, dan taraf atau skala pengukurannya (Nunnally, Jr., 1974).

A. Aneka Respon menurut Isinya

Yang dimaksud isi respon adalah jenis fungsi psikis yang secara dominan mewarnai respon yang bersangkutan. Dua kategori besar fungsi psikis yang mewarnai perilaku kita adalah **kognisi** atau fungsi pikir dan **afeksi** atau fungsi rasa. Dalam pengalaman hidup sehari-hari, setiap tingkah laku selalu merupakan hasil sinergis dari fungsi kognisi dan fungsi afeksi. Namun dalam pengukuran psikologis, tetap bisa dibedakan jenis respon yang secara dominan diwarnai oleh salah satu dari kedua fungsi psikis tersebut. Atas dasar pemikiran tersebut, lazim dibedakan:

1. *Judgment* atau Penilaian

Judgment atau penilaian mencakup semua jenis respon yang bisa dibedakan ke dalam kategori *benar* atau *salah* (Nunnally, Jr., 1970, h. 162). Artinya, terdapat suatu ukuran atau kriteria objektif untuk menentukan apakah suatu penilaian benar atau salah. Dalam pengukuran psikologis, kadang-kadang responden tidak sekadar diminta memberikan penilaian yang bisa ditetapkan secara dikotomis, benar atau salah, melainkan diminta memberikan penilaian terhadap *ketepatan relatif* atau taraf kebenaran atau ketepatan masing-masing dari suatu rangkaian objek atau pernyataan, dengan harapan responden mampu menemukan jawaban yang paling benar. Jelas kiranya, *judgment* atau penilaian merupakan jenis respon yang didominasi oleh fungsi kognisi atau fungsi pikir dan lazim dipakai sebagai dasar untuk melakukan pengukuran terhadap *maximal performance* atau kemampuan maksimal seseorang dalam memecahkan masalah.

2. *Sentiment* atau Perasaan

Sentiment atau perasaan mencakup semua jenis respon yang mencerminkan rasa suka atau tidak suka, sikap, minat, preferensi atau pilihan pribadi, nilai pribadi, dan sejenisnya (Nunnally, Jr., 1970, h. 162). Berbeda dengan respon penilaian, tidak tersedia ukuran atau kriteria objektif untuk menentukan apakah suatu respon perasaan benar atau salah. Sebagai sesuatu yang bersifat subjektif, apa pun pilihan, minat, atau sikap seseorang terhadap objek, orang lain, atau peristiwa tertentu adalah "benar" dalam arti sah. Tentu saja, konsekuensi atau akibat dari pilihan, minat, atau sikapnya itu bisa dinyatakan benar atau salah baik secara umum maupun secara moral, tetapi pilihannya sendiri tidak bisa dinyatakan benar atau salah, dalam arti sah. Sebagai contoh, kita tidak bisa menyatakan benar atau salah terhadap sikap atau tindakan seseorang yang lebih menyukai buah durian daripada mangga, termasuk ketika kemudian terbukti bahwa kesukaannya itu membuatnya terkena serangan *stroke* ringan. Dari namanya, respon perasaan merupakan jenis respon yang didominasi oleh fungsi afeksi atau rasa dan lazim dipakai sebagai dasar untuk melakukan pengukuran terhadap *typical performance* atau kecenderungan seseorang dalam bertingkah laku secara khas atau kepribadiannya.

B. Aneka Respon menurut Caranya

Istilah cara bisa dipakai untuk menunjuk dua makna yang berbeda, yaitu proses psikis yang berlangsung atau ditempuh, atau modalitas yang dipakai saat seseorang memberikan respon. Dari segi proses psikis yang ditempuh, respon dalam pengukuran psikologis bisa dibedakan menjadi: (1) respon absolut atau mutlak, dan (2) respon komparatif. Dari segi modalitas perilaku atau media yang dipakai dalam merespon, respon dalam pengukuran psikologis bisa dibedakan menjadi: (1) respon lisan, (2) respon tertulis, (3) respon *performance* atau kinerja, dan (4) respon termediasikan komputer.

1. Dari Segi Proses Psikis yang Ditempuh

Respon absolut atau mutlak. Dalam respon absolut atau mutlak, subjek dalam pengukuran psikologis memberikan respon langsung berdasarkan hasil penilaian atau perasaan pribadinya, tanpa bantuan berupa serangkaian penilaian atau perasaan alternatif yang sudah disediakan sebagai pembanding. Contohnya adalah mengerjakan item pengukuran psikologis berformat **isian**, baik untuk mengukur penilaian atau perasaan. Di situ subjek harus memberikan respon yang bersifat mutlak, yaitu respon yang harus dirumuskannya sendiri berdasarkan hasil pemikiran atau perasaan pribadinya.

Respon komparatif. Dalam respon komparatif subjek diminta menyatakan penilaian atau perasaannya dengan cara memilih salah satu dari antara beberapa alternatif respon yang telah disediakan atau dengan mendasarkan pada suatu ukuran atau patokan sebagai pembanding. Contohnya adalah mengerjakan item pengukuran psikologis berformat **pilihan wajib** (*forced choice*), **pilihan ganda** (*multiple choice*), atau **menjodohkan** (*matching*). Dalam item berformat pilihan wajib subjek harus memilih salah satu dari antara dua alternatif respon, misal "Ya" atau "Tidak," atau "Benar" atau "Salah". Dalam item berformat pilihan ganda subjek harus memilih salah satu dari antara lebih dari dua alternatif respon yang disediakan. Dalam pengukuran atribut psikologis yang menuntut respon berupa penilaian, subjek bisa dituntut untuk memilih salah satu jawaban yang benar atau yang paling benar. Dalam item berformat menjodohkan subjek harus menemukan pasangan-pasangan yang tepat dari antara serangkaian alternatif objek yang bisa dipasang-pasangkan.

2. Dari Segi Modalitas Perilaku atau Media yang Dipakai dalam Merespon

Respon lisan. Respon lisan atau *verbal report* atau laporan verbal secara lisan merupakan salah satu jenis respon dasar dalam arti paling sering dipakai dalam pengukuran psikologis khususnya menyangkut pengukuran penilaian atau perasaan. Subjek diberi pertanyaan atau

tugas tertentu, dan harus menjawab atau mengerjakan tugas itu secara lisan. Dalam pengukuran kepribadian, karena lazimnya subjek diminta melaporkan keadaan pribadinya maka respon verbalnya tersebut lazim disebut *self-report.* Laporan lisan tentang keadaan diri tersebut selanjutnya dipakai sebagai data untuk melakukan inferensi tentang atribut psikologis tertentu yang sedang menjadi sasaran pengukuran.

Respon tertulis. Respon tertulis adalah juga *verbal report* atau laporan verbal yang diberikan dengan cara dituliskan pada lembar kerja atau lembar jawab. Jenis respon ini pada dasarnya sama seperti respon lisan dan lazim dipakai dalam pengukuran penilaian atau perasaan, namun menuntut responden mampu menulis dan dengan sendirinya juga membaca. Respon tertulis tersebut juga bisa merupakan *self-report.*

Respon kinerja. Dalam pengukuran atribut psikologis yang kompleks dalam arti melibatkan baik penilaian, perasaan, maupun kemampuan menjalankan peran tertentu, misal kemampuan menjalankan peran atau tugas sebagai sekretaris, maka yang dipakai sebagai data pengukuran lazimnya adalah kinerja dalam rangka menjalankan peran atau tugas yang sedang menjadi sasaran pengukuran itu. Peran atau tugas tersebut pertama-tama harus dianalisis atau diuraikan dulu ke dalam rangkaian tindakan tertentu dengan metode *job analysis* dan selanjutnya dirumuskan secara tertulis sebagai *job description*. Selanjutnya, kinerja subjek baik menyangkut ketepatan maupun kecepatan dalam melaksanakan rangkaian tindakan tersebut secara keseluruhan dapat diamati dan dipakai sebagai data pengukuran tentang kemampuannya menjalankan peran atau tugas sebagai sekretaris yang sedang menjadi sasaran pengukuran.

Respon termediasikan komputer. Jenis respon ini bisa berupa laporan verbal tertulis atau kinerja tertentu yang dilakukan pada sebuah komputer. Penggunaan komputer termasuk yang tersambung dengan jaringan Internet semakin lazim dalam pengukuran psikologis (Bray, 2010).

C. Aneka Respon menurut Taraf Pengukurannya

Dalam bab sebelumnya sudah kita bahas tentang aneka jenis skala atau taraf pengukuran. Jenis skala atau taraf pengukuran tersebut ditentukan oleh jenis operasi empiris yang mendasari peneraan atau pelekatan bilangan pada objek atau peristiwa tertentu, sehingga diperoleh jenis informasi beserta teknik statistik yang cocok diterapkan pada masing-masing jenis skala. Dari empat jenis skala yang kita kenal, hanya tiga yang sungguh-sungguh memanfaatkan nilai numerik bilangan, yaitu skala ordinal, skala interval, dan skala rasio, sedangkan skala nominal hanya memanfaatkan bilangan sebagai nama atau label. Begitu pula, kategorisasi respon ini juga hanya berlaku pada respon yang diberikan dalam jenis-jenis skala ordinal, interval, dan rasio. Dengan kata lain, kategorisasi respon ini hanya berlaku pada respon yang diberikan dalam rangka pengukuran variabel kontinyu, dan tidak berlaku pada jenis pengukuran nominal atau pengukuran variabel diskret. Maka, menurut jenis skala atau taraf pengukurannya dikenal tiga jenis respon masing-masing diungkap dengan satu atau lebih metode penskalaan, yaitu (1) respon pada skala ordinal, (2) respon pada skala interval, dan (3) respon pada skala rasio.

1. Respon pada Skala Ordinal

Pada jenis respon ordinal ini, baik lewat penilaian atau perasaan subjek diminta melakukan *ordinal estimation,* yaitu menentukan posisi nilai dari sejumlah objek pada kontinum dalam rangka pengukuran atribut atau variabel tertentu secara ordinal. Ada empat metode estimasi ordinal (Nunnally, Jr., 1970):

a. Metode *rank order* sederhana. Baik lewat penilaian atau perasaan responden diminta menjenjangkan sejumlah objek terkait atribut atau variabel tertentu, mulai dari yang bernilai tertinggi sampai dengan yang bernilai terendah sehingga membentuk sebuah

kontinum atau bentangan. Misal, kepada seorang subjek disajikan lima jenis buah: apel (A), pisang (B), jeruk (C), salak (D), dan jambu biji (E). Selanjutnya dia diminta menjenjangkan kelima buah itu berdasarkan tingkat *favorable*-nya, mulai dari yang dia rasakan paling *favorable* atau yang paling dia sukai sampai dengan yang paling *unfavorable* atau yang paling tidak dia sukai. Sebagai misal, hasilnya adalah pisang pada jenjang 1, apel pada jenjang 2, jeruk pada jenjang 3, salak pada jenjang 4, dan jambu biji pada jenjang 5, atau B-A-C-D-E. Ketika diminta menjenjangkan kelima buah yang sama berdasarkan penilaiannya terhadap tingkat kemanfaatannya bagi kesehatan tubuh berkat jenis-jenis vitamin yang dikandung oleh masing-masing buah hasilnya bisa saja berbeda, misalnya: C-A-E-B-D. Artinya, jeruk ditempatkan pada jenjang 1, apel pada jenjang 2, jambu biji pada jenjang 3, pisang pada jenjang 4, dan salak pada jenjang 5. Sebagai respon pada skala ordinal, kita tidak tahu apakah jarak antar jenjang adalah sama baik ketika respon itu berupa perasaan maupun penilaian.

b. Metode *pair comparisons* atau pembandingan secara berpasangan. Subjek diminta menetapkan posisi serangkaian objek dalam suatu kontinum atau bentangan terkait atribut tertentu dengan cara membandingkan jenjang masing-masing objek berpasangan dengan semua objek yang ada. Kembali pada contoh lima buah sebelumnya. Alih-alih menjenjangkan kelima buah itu begitu saja secara sederhana, subjek diminta menentukan jenjang masing-masing buah dibandingkan buah lain sepasang demi sepasang sampai seluruh buah dibandingkan. Sebagai contoh, hasil pembandingan jenjang yang dilakukan oleh responden X terhadap lima buah di atas adalah sebagai berikut:

Apel : pisang = pisang
Apel : jeruk = jeruk
Apel : salak = apel
Apel : jambu biji = apel
Pisang : jeruk = pisang
Pisang : salak = pisang
Pisang : jambu biji = pisang

Jeruk : salak = jeruk

Jeruk : jambu biji = jeruk

Salak : jambu biji = salak

Dari sepuluh kali pembandingan secara berpasangan, pisang ditempatkan pada jenjang yang lebih tinggi atau diunggulkan sebanyak 4 kali, jeruk 3 kali, apel 2 kali, salak 1 kali, sedangkan jambu biji diunggulkan 0 kali atau tidak pernah diunggulkan. Maka, penjenjangannya adalah sebagai berikut: pisang menempati jenjang 1, jeruk jenjang 2, apel jenjang 3, salak jenjang 4, dan jambu biji menempati jenjang 5. Dari contoh di atas kita juga bisa merumuskan formula untuk mengetahui jumlah pasangan yang akan terbentuk jika kita memiliki sejumlah *n* objek untuk dipasangkan, yaitu = $\boldsymbol{n(n-1)/2}$. Namun seperti halnya penjenjangan dengan metode *rank order* sederhana, kita tidak tahu apakah jarak antar jenjang adalah sama baik ketika respon itu berupa perasaan maupun penilaian.

c. Metode *constant stimuli* atau pembandingan dengan stimulus tetap. Metode ini mirip metode *pair comparisons.* Bedanya, dalam menentukan posisi serangkaian objek pada kontinum terkait atribut tertentu subjek diminta membandingkan masing-masing objek dengan sebuah stimulus tetap atau baku secara random. Subjek diminta melaporkan penilaian atau perasaannya terhadap masing-masing objek dibandingkan dengan sebuah stimulus baku sebagai patokan, yaitu lebih atau kurang, dan seberapa lebih atau kurang masing-masing objek tersebut dibandingkan dengan stimulus bakunya terkait atribut tertentu yang sedang menjadi sasaran pengukuran. Sebagai contoh, untuk mengukur tingkat *favorable* masing-masing dari lima buah dalam contoh di atas bisa dipakai buah pepaya sebagai stimulus baku. Masing-masing buah dipasangkan dengan pepaya secara random untuk selanjutnya dimintakan penilaian dari responden tentang tingkat *favorable* masing-masing buah dibandingkan dengan pepaya. Sebagai contoh, hasil pembandingan lima jenis buah dengan pepaya sebagai stimulus baku yang dilakukan oleh responden Y adalah sebagai berikut:

Apel : pepaya = apel
Pisang : pepaya = pisang
Jeruk : pepaya = jeruk
Salak : pepaya = pepaya
Jambu biji : pepaya = pepaya

Terlihat, tiga jenis buah – apel, pisang, dan jeruk – dinilai lebih *favorable* dibandingkan stimulus patokan pepaya, sedangkan dua buah lainnya – salak dan jambu biji – dinilai kurang *favorable* dibandingkan pepaya. Maka, berdasarkan penjenjangan, penskalaannya menjadi sebagai berikut: apel, pisang, dan jeruk masing-masing menduduki posisi (jenjang) 2 pada skala, yaitu jumlah jenjang 1, 2, dan 3 dibagi rata untuk tiga jenis buah (6 : 3 = 2); sedangkan salak dan jambu biji masing-masing menempati posisi (jenjang) 4,5 pada skala, yaitu jumlah jenjang 4 dan 5 dibagi rata untuk dua jenis buah (9 : 2 = 4,5). Namun lagi-lagi, kita tidak tahu apakah jarak antar jenjang adalah sama baik ketika respon itu berupa perasaan maupun penilaian

d. Metode *successive categories* atau pengkategorian beruntun. Seperti dijelaskan oleh Nunnally, Jr. (1970), dalam metode ini subjek diminta memilah sejumlah besar stimuli ke dalam sejumlah kategori terkait dengan atribut tertentu yang diurutkan mulai dari rendah sampai tinggi. Sebagai contoh, sekitar tahun 2008-2010 terjadi ketegangan antara bangsa Indonesia dan bangsa Malaysia dipicu oleh beberapa peristiwa konflik. Menanggapi situasi itu, seorang peneliti ingin mengetahui sikap sekelompok subjek mahasiswa Indonesia terhadap Malaysia. Dia membuat 100 pernyataan dengan isi pesan yang mencerminkan sikap *favorable* terhadap Malaysia dengan taraf yang berbeda-beda, mulai dari sangat *unfavorable* sampai *sangat favorable.* Subjek diminta memilah pernyataan-pernyataan tersebut ke dalam 10 kategori, diurutkan mulai dari kategori 1 yang berarti "sangat *favorable*" dan menurun secara berturut-turut sampai pada kategori 10 yang berarti "sangat *unfavorable*".

Menurut Nunnally, Jr., ada beberapa variasi dalam melakukan kategorisasi, tergantung dari jenis informasi yang ingin dijaring oleh peneliti. Dua variasi yang cukup terkenal adalah **teknik ordinal**

biasa dan **teknik *Q sort***. Pada teknik ordinal biasa, subjek bebas memasukkan masing-masing pernyataan ke dalam salah satu kategori yang dipilihnya. Hasilnya adalah gambaran sederhana tentang urutan taraf *favorableness* dari masing-masing pernyataan.

Pada teknik *Q sort,* subjek diminta mengkategorikan pernyataan-pernyataan sedemikian rupa sehingga pernyataan-pernyataan tersebut terdistribusikan secara normal dalam sepuluh kategori. Hasilnya adalah gambaran tentang pernyataan-pernyataan dengan taraf *favorable* yang ekstrim pada kedua kutub serta pernyataan-pernyataan dengan taraf *favorable* sedang atau bahkan netral yang terletak di tengah distribusi. Namun lagi-lagi, kita tidak tahu apakah jarak antar kategori *favorableness* dari pernyataan-pernyataan tersebut adalah sama.

2. Respon pada Skala Interval

Ada dua metode terkenal untuk mendapatkan respon pada skala interval, yaitu metode **interval tampak setara (***equal-appearing intervals*) dan metode **estimasi interval.** Menurut Nunnally, Jr. (1970), metode interval tampak setara sebenarnya mirip dengan metode pengkategorian beruntun. Subjek sama-sama diminta memasukkan sejumlah besar stimuli ke dalam sejumlah kategori. Bedanya, pada metode interval tampak setara ada permintaan tambahan, yaitu pemilahan stimuli itu harus dilakukan sedemikian rupa sehingga interval antara kategori-kategori yang tercipta secara subjektif tampak atau terasa setara.

Menurut Nunnally, Jr. (1970) salah satu teknik estimasi interval yang paling terkenal adalah teknik ***bisection***. Contoh aplikasinya, kepada subjek diberikan dua bola peluru besi dengan berat yang berbeda. Selanjutnya dia diminta memilih bola peluru ketiga dengan berat tepat di antara kedua bola peluru sebelumnya. Sebagaimana sudah disinggung berulangkalai, pada jenis skala atau taraf pengukuran interval ini jarak antar kategori adalah sama.

3. Respon pada Skala Rasio

Seperti dinyatakan oleh Nunnally, Jr. (1970), dalam pemberian respon pada skala rasio pada dasarnya subjek diminta menilai besaran absolut stimuli, bisa berupa beratnya, panjangnya, dan sebagainya. Contoh aplikasinya, kepada subjek yang ditempatkan di sebuah ruang gelap ditunjukkan seberkas cahaya dengan intensitas tertentu yang disorotkan pada sebuah layar. Selanjutnya subjek diminta menyorotkan cahaya kedua dengan sejenis lampu sorot dengan intensitas misalnya dua kali lebih terang dibandingkan cahaya yang pertama.

Seperti akan kita lihat, pengukuran psikologis memanfaatkan berbagai jenis respon seperti diuraikan di atas untuk mendapatkan data pengukuran yang dapat dipercaya sehingga sungguh-sungguh memenuhi kegunaannya. Khusus terkait jenis respon berdasarkan jenis skala atau taraf pengukurannya, kebanyakan alat pengukuran psikologis paling tinggi hanya mampu menghasilkan data pengukuran pada taraf interval. Bahkan, konon kebanyakan pengukuran psikologis sesungguhnya hanya menghasilkan data pada taraf pengukuran ordinal kendati secara umum mampu membuahkan hasil yang bermanfaat dalam memetakan kekhususan individual untuk berbagai keperluan (Stevens, 1946). Ψ

Bab 4
Tes Psikologis

Salah satu bentuk alat dalam pengukuran psikologis adalah tes psikologis atau psikotes atau secara ringkas disebut *tes.* Lebih dari sekadar salah satu bentuk alat, psikotes atau tes bahkan praktis sudah identik dengan psikometri atau pengukuran psikologis. Selain itu, baik akibat pesatnya perkembangan layanan psikotes di satu sisi maupun akibat posisi sekunder-komplementer para psikolog terhadap psikiater dalam intervensi psikologis -- khususnya dengan berperan sebagai sekadar pemasok data hasil asesmen dengan psikotes sebagai dasar untuk menetapkan diagnosis, prognosis, dan intervensi dalam terapi psikiatris -- di sisi lain, ada masanya saat secara sarkastis dan *gebyah uyah* psikolog diidentikkan dengan "tukang tes". Label semacam itu tentu terkesan melebih-lebihkan, namun di sisi lain bisa dipandang sebagai indikasi bahwa kendati pada dirinya pun sudah mengandung kontroversi, ternyata psikometri tetap memiliki tempat yang penting dalam psikologi khususnya maupun dalam layanan psikologis yang melibatkan pihak dan keahlian lain seperti psikiatri pada umumnya. Pada bagian ini secara berturut-turut akan kita bahas serba-serbi tes meliputi pengertian, karakteristik dasar, jenis, aneka penggunaan, dan ciri-ciri tes yang baik.

A. Pengertian Tes Psikologis

Ada sejumlah definisi yang bisa kita temukan tentang tes psikologis, psikotes, atau tes. Allen dan Yen (1979) mendefinisikan tes sebagai "a device for obtaining a sample of an individual's behavior" (h. 1). Artinya, tes merupakan piranti, sarana atau alat untuk memperoleh suatu sampel atau contoh perilaku seseorang. Friedenberg (1995) menempatkan definisi tentang tes dalam konteksnya yang lebih luas, yaitu *asesmen.* Menurutnya, asesmen

adalah "any procedure used to gather information about people" (h. 5). Maksudnya, asesmen merupakan setiap jenis prosedur yang digunakan untuk mengumpulkan informasi tentang orang. Secara garis besar ada dua pendekatan dalam asesmen, yaitu pendekatan subjektif dan pendekatan objektif. Pendekatan subjektif mencakup metode observasi dan wawancara. Dalam metode observasi kontak antara asesor dan subjek yang diamati tidak harus terjadi. Sebaliknya dalam wawancara, proses dan hasil asesmen sangat ditentukan oleh keberhasilan asesor menjalin kontak dengan subjek. Namun kedua metode itu tetap disebut subjektif, sebab sarana atau alat utamanya bukan lain adalah si asesor, yaitu pengamat atau pewawancara sendiri. Dua asesor yang melakukan asesmen terhadap subjek yang sama dengan kedua metode tersebut bisa sampai pada kesimpulan yang sedikit atau banyak berbeda, sekali lagi karena bergantung pada faktor subjektif asesor.

Sebaliknya, metode asesmen objektif berusaha meminimalkan bahkan menghilangkan berbagai bentuk subjektivias baik yang berasal dari pribadi asesor maupun subjek yang dikenai asesmen. Berbagai bentuk subjektivitas semacam ini dipandang sebagai *extraneous variables* yang menjadi sumber *error* atau kesalahan dalam asesmen. Tes psikologis merupakan salah satu atau bahkan satu-satunya metode asesmen objektif yang paling luas dipakai, khususnya kategori tes baku. Sebagaimana sudah kita lihat di sana-sini, kualitas suatu tes ditentukan oleh sejauh mana *extraneous variables* berupa antara lain subjektivitas yang bersumber dari pribadi asesor maupun testi berhasil ditekan atau bahkan dihilangkan, sebab dipandang sebagai sumber *error* atau kesalahan pengukuran.

Lebih lanjut Friedenberg (1995) menyatakan bahwa tes merupakan "a type of assessment that uses specific procedures to obtain information and convert that information to numbers or scores" (h. 5). Jadi, tes merupakan salah satu jenis asesmen yang menggunakan serangkaian prosedur khusus untuk memperoleh informasi serta mengonversikan atau mengubah informasi tersebut ke dalam serangkaian bilangan atau skor. Sudah barang tentu, informasi

tentang orang bisa beraneka ragam, meliputi antara lain keadaan fisik dan keadaan psikologisnya. Tes psikologis atau psikotes secara khusus ditujukan untuk memperoleh informasi tentang seseorang terkait aspek tertentu dari keadaan psikologisnya. Terakhir, Gregory (2007) mendefinisikan tes sebagai "a standardized procedure for sampling behavior and describing it with categories or scores" (h. 2). Maksudnya, tes merupakan sebuah prosedur yang dibakukan untuk mengambil sampel perilaku dan selanjutnya mendeskripsikannya dengan menggunakan serangkaian kategori atau skor. Namun dia masih melanjutkan, "most tests have norms or standards by which the results can be used to predict other, more important behaviors" (h. 2). Menurutnya, kebanyakan tes memiliki serangkaian norma atau standar sehingga hasil-hasil tes dapat dimanfaatkan untuk memprediksikan aneka tingkah laku lain yang lebih penting. Maksudnya, tingkah laku dalam menghadapi berbagai tugas kehidupan sehari-hari di luar tes.

B. Karakteristik Dasar Tes Psikologis

Jika kita cermati, ada lima karakteristik dasar tes yang bisa kita temukan dalam tiga definisi di atas, yaitu bahwa tes merupakan: (1) prosedur spesifik atau sistematis yang dibakukan; (2) sampel tingkah laku; menghasilkan (3) kategori atau skor; untuk menginterpretasikan hasilnya membutuhkan (4) norma atau standar; dan fungsi utamanya adalah memberikan (5) prediksi tentang tingkah laku nontes. Masing-masing dari kelima karakteristik dasar tes tersebut akan kita bahas secara lebih rinci dalam bagian-bagian berikut ini.

1. Tes sebagai Prosedur Spesifik atau Sistematis yang Dibakukan

Ada dua komponen penting dalam karakteristik yang pertama ini, yaitu bahwa tes merupakan prosedur yang spesifik dalam arti sistematis dan yang dibakukan. Terkait komponen pertama, tes melibatkan perumusan secara spesifik atau sistematis menyangkut

tiga hal, yaitu: (a) rangkaian tugas atau pertanyaan yang digunakan sebagai item-item tes, (b) aneka kondisi pengadministrasian tes, dan (c) cara penskoran serta cara penginterpretasian jawaban testi atau subjek yang dites (Friedenberg, 1995).

Terkait komponen kedua atau pembakuannya, sebuah tes disebut *standardized* atau dibakukan atau baku jika tiga hal yang sudah dispesifikasikan atau disistematisasikan di atas dibuat uniform atau seragam atau tetap baik antar testi maupun antar asesor, tester, atau penguji. Spesifikasi dan pembakuan rangkaian tugas atau pertanyaan yang digunakan sebagai item-item tes lazimnya dituangkan dalam bentuk *test booklet* atau lazim diindonesiakan menjadi *buku soal* dan/ atau *aparatus* atau piranti tes. Sedangkan pembakuan petunjuk cara pengadministrasian tes, cara penskoran termasuk kunci jawaban, dan cara penginterpretasian hasil tes termasuk norma penilaian, dan informasi lain tentang tes lazim dituangkan dalam *manual* atau buku pegangan.

Sebagai contoh, *Wechsler Adult Intelligence Scale* yaitu tes inteligensi individual untuk subjek dewasa, edisi ke-3 atau disingkat *WAIS-III,* mencakup dua kategori subtes, yaitu kategori subtes *verbal* dan kategori subtes *performance* (Gregory, 2007). Kategori subtes verbal meliputi enam subtes: (a) *vocabulary* atau perbendaharaan kata, (b) *similarities* atau persamaan, (c) *arithmetic* atau berhitung, (d) *digit span* atau rentang bilangan, (e) *information* atau pengetahuan umum, dan (f) *comprehension* atau pemahaman. Bentuk tugas pada enam subtes ini berupa pertanyaan-pertanyaan yang harus dijawab dengan *verbal report* atau secara lisan. Tugas-tugas atau pertanyaaan-pertanyaan tersebut dicantumkan secara terintegrasi dengan informasi lain tentang tes dalam *manual.*

Kategori subtes *performance* meliputi lima subtes: (g) *picture completion* atau menggenapi atau melengkapi gambar, (h) *digit-symbol/coding* atau pengodean atau simbol bilangan, (i) *block design* atau rancangan balok, (j) *matrix reasoning* atau penalaran matriks, dan (k) *picture arrangement* atau menyusun gambar. Kendati bentuk tugas pada lima subtes ini berlainan, namun memiliki kesamaan

yaitu bahwa untuk menjawab atau menyelesaikannya testi harus melakukan sesuatu (*performance)* sehingga penyajiannya pun melibatkan penggunaan media tertentu sebagai *aparatus*. Pada subtes *picture completion* dan *matrix reasoning* bentuk tugas berupa pertanyaan dengan media gambar, testi diminta memberikan jawaban secara lisan. Pada subtes *digit symbol* bentuk tugas berupa soal tercetak pada lembar tugas, testi diminta memberikan jawaban secara tertulis pada lembar tugas yang sekaligus dilengkapi ruang tempat menjawab. Pada subtes *block design* dan *picture arrangement* bentuk tugas bersifat fisik, testi harus memanipulasikan serangkaian objek berupa kubus-kubus berwarna merah putih dan gambar-gambar penggalan kejadian yang dicetak pada kartu-kartu berwarna hitam-putih.

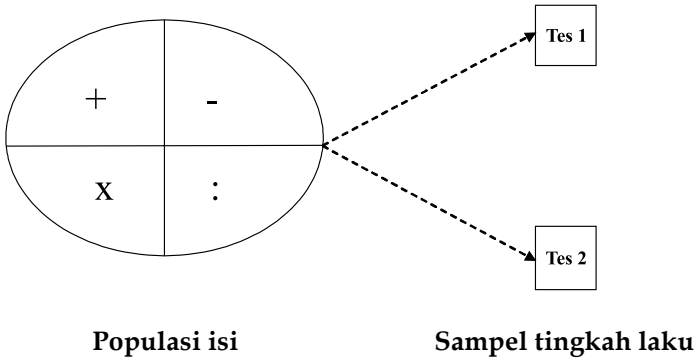
Spesifikasi dan pembakuan aneka kondisi pengadministrasian tes dinyatakan antara lain dalam bentuk penyediaan petunjuk yang jelas terkait waktu pengerjaan dan instruksi pengadministrasian masing-masing subtes baik dalam *buku soal* maupun lebih-lebih dalam *manual*. Spesifikasi dan pembakuan cara penskoran serta cara penginterpretasian jawaban testi atau subjek dinyatakan antara lain dalam bentuk penyediaan kunci jawaban, cara pemberian skor dalam arti *raw score* atau skor kasar, norma penilaian, cara mengonversikan skor kasar menjadi *weighted score* atau skor terbobot, serta cara menginterpretasikan hasil tes secara keseluruhan.

Dengan sistematisasi dan pembakuan seperti diuraikan di atas, di satu pihak seandainya beberapa asesor menguji seorang subjek dengan tes yang sama, maka diharapkan hasilnya pada dasarnya akan sama. Di pihak lain, seandainya seorang asesor menguji beberapa subjek dengan tes yang sama dan ternyata hasilnya berlainan, harapan kita ialah bahwa perbedaan hasil tersebut memang mencerminkan perbedaan kemampuan yang dimiliki oleh masing-masing testi yang bersangkutan. Inti pesan yang bisa kita tangkap dari kedua fakta tersebut ialah bahwa standarisasi prosedur pengadministrasian tes bertujuan memastikan bahwa hasil tes sungguh-sungguh mencerminkan keadaan atribut yang diukur, dan bukan mencerminkan perbedaan prosedur pengadministrasian

antara asesor atau penguji yang satu dan asesor atau penguji yang lain.

2. Tes sebagai Sampel Tingkah Laku

Dari psikofisika klasik kita sudah belajar bahwa respon psikologis yang menandakan kehadiran suatu atribut psikologis dalam jumlah tertentu bersifat *covert* atau berlangsung dalam diri individu sehingga bersifat *intangible* atau tidak bisa diamati secara langsung. Namun respon psikologis yang bersifat tertutup tersebut akan diungkapkan dalam bentuk tingkah laku *overt* atau terbuka berupa *judgment* atau penilaian, *sentiment* atau perasaan, atau *performance* atau kinerja tertentu yang dipandang sebagai *behavioral indicators* atau indikator-indikator tingkah lakunya. Aneka indikator tingkah laku terbuka yang bisa diamati secara langsung ini diasumsikan berkorelasi secara sempurna dengan respon psikologisnya yang bersifat tertutup atau tidak bisa diamati secara langsung. Akibatnya, berdasarkan indikator tingkah laku tersebut seorang asesor bisa melakukan *inferensi* atau kesimpulan tentang atribut psikologis yang diasumsikan mendasari atau melatarinya, baik terkait kualitas maupun kuantitasnya.



Gambar 4.1. Contoh populasi isi berupa indikator tingkah laku kemampuan berhitung dan dua tes berhitung paralel sebagai sampel tingkah laku.

Diasumsikan, sebagian besar atribut psikologis yang menjadi sasaran pengukuran memiliki *universe of content* atau populasi isi berupa indikator tingkah laku yang secara teoretis tidak terbatas jumlahnya. Sebagai contoh, kemampuan berhitung akan mencakup indikator tingkah laku berupa kemampuan menyelesaikan soal-soal pertambahan, perkurangan, perkalian, dan pembagian yang tidak terbatas ragamnya sesuai ragam soal yang bisa dirumuskan dan disajikan oleh asesor atau penguji sebagai *stimulus*. Dari antara populasi indikator tingkah laku yang diasumsikan tidak terbatas jumlahnya tersebut, diambil sebagian kecil – misalnya, 40 buah – sebagai sampel untuk dijadikan tes. Logika atau jalan pikiran tersebut bisa dijelaskan secara visual dengan bagan seperti disajikan di Gambar 4.1.

Sebagaimana tampak dalam Gambar 4.1. secara teoretis mudah membuat tes paralel yang mengukur suatu atribut psikologis. Tinggal mengambil sampel baru dari populasi isi berupa himpunan indikator tingkah laku yang sama, yang untuk sebagian besar atribut psikologis diasumsikan tidak terbatas jumlahnya. Secara teoretis kita juga bisa menyusun tes paralel dalam jumlah yang tidak terbatas tentang suatu atribut psikologis.

3. Tes Menghasilkan Skor

Kita kutip kembali pendapat dua pakar pengukuran psikologis, E.L. Thorndike (1918) dan W.A. McCall (1939). Thorndike menyatakan, "Whatever exists at all exists in some amount." Artinya, apa saja yang ada di dunia ini pasti ada dalam jumlah tertentu. Pernyataan tersebut digenapkan oleh McCall menjadi, "Anything that exists in amount can be measured." Maksudnya, apa saja yang ada dalam jumlah tertentu pasti bisa diukur. Kita pun sudah melihat bahwa mengukur sesuatu, dalam hal ini sebuah atribut psikologis, berarti menerakan bilangan untuk mengungkapkan kuantitas atribut psikologis yang bersangkutan. Dengan kata lain, tes merupakan sarana kuantifikasi

yaitu pengungkapan kuantitas suatu atribut psikologis dalam bentuk bilangan tertentu yang disebut skor.

Pada tes yang bertujuan mengukur aneka jenis kemampuan yang didominasi fungsi kogntif atau fungsi pikir dan yang harus dijawab dengan *judgment* atau penilaian, skor yang dimaksud lazim diperoleh dengan cara menghitung jumlah jawaban benar. Pada tes yang bertujuan mengukur aneka jenis kecenderungan bertingkah laku yang didominasi fungsi afektif atau fungsi rasa dan yang harus dijawab dengan ungkapan *sentiment* atau perasaan, skor yang dimaksud lazim diperoleh dengan cara menghitung jumlah nilai sesuai pola jawaban yang dipilih oleh subjek. Yang perlu dicatat, skor yang diperoleh pada fase ini merupakan *raw score* atau skor mentah atau skor kasar. Skor semacam ini belum bermakna dan tidak bisa diperbandingkan. Sebagai contoh, seorang murid mencapai skor Kemampuan Berhitung 45 dan skor Kemampuan Berbahasa 60. Dapatkah disimpulkan bahwa kemampuan berbahasa murid tersebut lebih baik dibandingkan kemampuan berhitungnya? Ternyata tidak bisa, sebab skor total maksimal yang bisa dicapai untuk tes Kemampuan Berhitung adalah 50, sedangkan untuk tes Kemampuan Berbahasa adalah 100. Bisa jadi, kemampuan berhitung murid tersebut justru lebih baik dari kemampuan berbahasanya. Begitulah ciri skor kasar atau skor mentah.

4. Tes Dilengkapi Norma untuk Menetapkan Kategori

Skor mentah atau skor kasar baru menjadi bermakna sesudah dibandingkan dengan suatu norma atau standar atau kriteria dan dikonversikan menjadi nilai. Dalam asesmen atau pengukuran, fase atau langkah ini memang disebut *evaluasi* atau penilaian, yaitu pembandingan skor mentah dengan sebuah norma atau standar atau kriteria agar bisa diinterpretasikan atau dimaknai dalam arti diputuskan untuk dimasukkan ke dalam kategori kualitas tertentu.

Ada dua macam standar yang lazim digunakan dalam menginterpretasikan skor mentah atau skor kasar menjadi nilai atau kategori, yaitu (a) norma relatif atau disingkat *norma,* dan (b) norma absolut atau lazim disebut *kriteria.* Pada penerapan norma relatif, makna skor mentah yang dicapai oleh seorang subjek ditentukan dengan membandingkannya dengan *average performance* atau kinerja rata-rata dari kelompok sebayanya. Statistik yang dipakai sebagai ukuran rata-rata lazimnya adalah kombinasi antara *Mean* dan *standard deviation.* Proses pengembangan *norma* semacam ini merupakan bagian dari proses standarisasi tes dan akan mencakup dua langkah utama sebagai berikut. *Pertama,* mengadministrasikan tes pada sekelompok besar subjek yang memiliki kesamaan dalam minimal satu karakteristik penting. Karakteristik yang dimaksud bisa bersifat longgar, misal sama-sama berusia dewasa dalam arti berada dalam rentang umur 17 sampai dengan 65 tahun, atau sebaliknya spesifik misal sama-sama berusia 7-12 tahun dan penyandang autisme. Kelompok yang dipakai sebagai acuan untuk menyusun norma ini lazim disebut *sampel standarisasi. Kedua,* sesudah data terkumpul dihitunglah *Mean* dan juga *Standard Deviation* skor kelompok standarisasi tersebut. Sebuah norma penilaian segera bisa disusun dengan menggunakan dua statistik deskriptif tersebut. Dengan menggunakan norma yang merupakan *average performance* dari kelompok sebayanya sebagai standar serta berdasarkan skor yang berhasil dicapai, masing-masing subjek bisa dinilai kepemilikannya atas atribut psikologis tertentu dengan cara memasukkannya ke dalam salah satu kategori, misal *Rendah, Sedang, Tinggi,* atau kategori lain.

Sekadar catatan, norma relatif atau yang secara singkat disebut *norma* ini pada dasarnya bersifat *aposteriori* dalam arti norma tersebut baru bisa disusun sesudah tes diadministrasikan khususnya pada sampel standarisasi. Penggunaan tes dengan penerapan norma relatif semacam ini disebut *norm-referenced testing* atau pengetesan beracuan norma. Kelemahan pendekatan ini ialah bahwa nilai atau kategori yang diperoleh setiap subjek tidak benar-benar mencerminkan kuantitas atribut psikologis yang dimilikinya, melainkan sekadar

menunjukkan seberapa baik atau buruk, seberapa tinggi atau rendah kuantitas atribut psikologis yang dimiliknya itu dibandingkan teman-teman sebayanya.

Pada penerapan norma absolut atau kriteria, makna skor mentah seorang subjek ditentukan dengan membandingkannya dengan sebuah standar yang sudah ditentukan secara *apriori*. Penentuan standar ini lazimnya didasarkan pada pertimbangan teoretis tertentu. Penggunaan tes dengan penerapan norma relatif semacam ini disebut *criterion-referenced testing* atau pengetesan beracuan patokan. Menurut Friedenberg (1995), berdasarkan apa yang dipatok atau dipakai sebagai patokan, penilaian beracuan patokan lazim dibedakan ke dalam tiga jenis, yaitu (a) *content-referenced scoring* atau penilaian beracuan patokan materi, (b) *objective-referenced scoring* atau penilaian beracuan patokan tujuan, dan (c) *pass/fail* atau *mastery scoring* atau penilaian beracuan lulus/gagal atau beracuan penguasaan.

a. Penilaian Beracuan Patokan Materi

Dalam pendekatan yang pertama kali dikemukakan oleh Brown (1983, dalam Friedenberg, 1995) ini, objek yang dipatok atau dijadikan patokan adalah *isi* atau *materi* tes. Isi atau materi tes itu sendiri merupakan turunan langsung dari isi atau materi pengetahuan yang sedang dinilai. Patokannya adalah penguasaan atas isi atau materi tes tersebut, yang tercermin dari *jumlah jawaban benar*. Salah satu jenis nilai beracuan patokan materi yang paling sederhana adalah *percent correct* atau persentase jawaban benar, yang bisa dihitung dengan rumus sebagai berikut (Friedenberg, 1995):

$$\textit{Persentase jawaban benar} = \frac{\textit{Jumlah jawaban benar}}{\textit{Jumlah skor yang mungkin dicapai}} \textit{ X 100} \qquad \textbf{Rumus 4.1.}$$

Angka 85% jawaban benar lazim dipakai sebagai standar dalam penilaian beracuan patokan. Hanya mereka yang mencapai

minimal 85% jawaban benar dinyatakan "Lulus," kurang dari angka itu dinyatakan "Tidak Lulus."

Friedenberg (1995) memberikan dua catatan kritis. Pertama, persentase jawaban benar sekadar menunjukkan kinerja testi menghadapi item-item tes. Jika item-item tes tersebut tidak merepresentasikan secara baik ranah materi bidang studi yang dinilai, maka persentase jawaban benar tidak mencerminkan penguasaan murid atas materi bidang studi yang diujikan. Kedua, makna persentase jawaban benar juga mudah dikacaukan oleh perbedaan taraf kesukaran item-item tes. Dua murid yang sama-sama mencapai nilai 90 tidak bisa disebut memiliki prestasi yang setara jika yang satu memperolehnya lewat tes dengan item-item yang relatif mudah sedangkan yang lain memperolehnya lewat tes dengan item-item yang relatif sukar.

b. Penilaian Beracuan Patokan Tujuan

Untuk mengatasi problem di atas, Thorndike et al. (1991, dalam Friedenberg, 1995) mengganti objek yang dijadikan kerangka acuan, bukan isi atau materi tes melainkan *tujuan* tes atau lebih tepat tujuan pengajaran (*instructional objectives*). Dalam penilaian, tujuan kita adalah menentukan tingkat penguasaan murid atas ranah pengetahuan-ketrampilan tertentu yang bersifat laten atau tidak langsung teramati, dan bukan sekadar mengukur kemampuan murid menjawab seperangkat pertanyaan tes. Item-item tes disusun untuk mengungkap pencapaian seperangkat tujuan pengajaran tertentu. Masing-masing tujuan pengajaran dituangkan ke dalam seperangkat item. Seorang testi atau murid disebut berhasil mencapai atau menguasai tujuan pengajaran tertentu jika dia berhasil menjawab dengan benar sejumlah item yang dimaksudkan untuk mengungkap hal itu. Dalam pendekatan ini kuncinya terletak pada merumuskan sistem penskoran yang mampu menunjukkan persentase tujuan pengajaran yang berhasil dikuasai murid. Toh menurut Friedenberg (1995), pendekatan ini belum terbebas dari jerat persoalan taraf

kesukaran item dan penafsiran skor seperti pada pendekatan pertama.

c. *Pass/Fail* atau *Mastery Scoring*

Penilaian beracuan lulus/gagal atau beracuan penguasaan ini bukan pendekatan yang sama sekali baru, melainkan sejenis aplikasi dari dua pendekatan yang pertama. Maksudnya, nilai yang diperoleh baik melalui penilaian beracuan patokan materi maupun beracuan patokan tujuan sama-sama bisa ditransformasikan ke dalam jenis lain, khususnya untuk memilah testi atau murid menjadi dua kategori: (a) mereka yang kinerjanya dalam tes memenuhi patokan yang ditentukan (*lulus*), dan (b) mereka yang kinerjanya dalam tes tidak memenuhi patokan yang ditentukan (*gagal*). Lazimnya, guru atau pimpinan sekolah memilih salah satu skor tes yang dipandang mencerminkan taraf kinerja (minimal) yang diharapkan. Skor itulah yang dijadikan patokan, semua siswa yang mencapai skor sama atau lebih besar dari patokan dinyatakan *lulus,* dan semua siswa yang mencapai skor kurang dari patokan dinyatakan *gagal.* Perbedaan skor dalam masing-masing kategori *lulus* atau *gagal* tidak lagi memiliki makna.

Kendati tetap memiliki sejumlah kekurangan, pendekatan penilaian beracuan patokan ini secara umum dipandang lebih mencerminkan apa yang benar-benar bisa dilakukan oleh murid terkait pengetahuan, ketrampilan, atau kemampuan yang diajarkan dalam setiap mata pelajaran, dibandingkan dengan pendekatan beracuan norma (Shermis & Di Viesta, 2011).`

5. Prediksi tentang Tingkah Laku Nontes

Tujuan akhir penerapan tes psikologis pada seorang subjek adalah memprediksikan tingkah lakunya di luar konteks tes psikologis itu sendiri, yaitu dalam situasi kehidupan nyata sehari-hari entah dalam konteks pekerjaan, pendidikan sekolah, hubungan perkawinan, dan sebagainya. Asumsinya, pelaksanaan tugas dalam

konteks kehidupan tertentu menuntut penguasaan atau pemilikan atribut psikologis tertentu yang sampel atau contoh indikator tingkah lakunya dipilih sebagai item atau butir tes. Maka dengan mengetahui hasil tes seseorang yang mengukur atribut psikologis tertentu kita bisa memprediksikan keberhasilannya melaksanakan tugas-tugas dalam konteks kehidupan tertentu yang mempersyaratkan pemilikan atribut psikologis yang bersangkutan. Contohnya, keberhasilan menempuh program pendidikan sekolah mempersyaratkan inteligensi atau kecerdasan. Dengan menggunakan hasil tes inteligensi sebagai dasar seleksi calon siswa, kita bisa hanya memilih calon-calon yang memiliki inteligensi cukup sehingga kita prediksikan akan berhasil menempuh program pendidikan yang kita selenggarakan.

Prediksi lazimnya dilakukan dengan menyelidiki korelasi antara hasil tes atribut psikologis tertentu dengan satu atau lebih ukuran tingkah laku yang kita prediksikan. Hasil tes atribut psikologis tertentu yang dipakai sebagai dasar prediksi lazim disebut *prediktor* dan dilambangkan dengan huruf besar Latin *X,* sedangkah ukuran tingkah laku yang diprediksikan disebut *kriteria* dan dilambangkan dengan huruf besar Latin *Y.* Kembali pada contoh kita memprediksikan keberhasilan menempuh program pendidikan berdasarkan hasil tes inteligensi. Prediktor atau *X*-nya adalah hasil tes inteligensi, sedangkan kriteria atau *Y*-nya terlebih dulu harus kita tentukan. Ukuran tingkah laku atau kinerja seseorang dalam menempuh program pendidikan di perguruan tinggi bisa berupa Indeks Prestasi baik yang dicapainya pada semester tertentu (Indeks Prestasi Semester, disingkat IPS) atau yang dicapainya secara akumulatif sesudah beberapa semester atau bahkan sesudah menyelesaikan seluruh program pendidikan (Indeks Prestasi Kumulatif, disingkat IPK), atau berupa keseluruhan masa studi yang ditempuhnya untuk menyelesaikan program pendidikan tersebut entah dihitung dalam bulan atau tahun, atau bahkan lainnya. Kita bisa menggunakan satu atau lebih ukuran tersebut sebagai kriteria, namun lazimnya kita akan memilih salah satu yang ketika dikorelasikan dengan prediktor memberikan prediksi yang terbaik.

Maka, daya prediksi sebuah prediktor terhadap kriteria lazim dinyatakan dalam sebuah koefisien korelasi antara prediktor dan kriteria, dilambangkan r_{xy}. Sebagaimana lazim, pembacaan koefisien korelasi didasarkan pada arah dan besarnya. Arah korelasi ditunjukkan oleh tanda positif atau negatif di depan koefisien korelasi. Korelasi antara prediktor dan kriteria bisa positif atau searah, yaitu peningkatan (atau penurunan) skor pada prediktor diiringi peningkatan (atau penurunan) skor yang kurang lebih setara pada kriteria. Atau sebaliknya, korelasi antara prediktor dan kriteria bisa negatif atau berlawanan arah, yaitu peningkatan skor pada prediktor justru diiringi penurunan skor yang kurang lebih setara pada krtierion. Arah korelasi antara prediktor dan kriteria ditentukan oleh argumentasi teoretis atau konseptual yang mendasarinya. Sebagai contoh, secara teoretis-konseptual inteligensi berkorelasi secara positif dengan prestasi belajar. Sebaliknya, kecemasan berkorelasi secara negatif dengan hasil ujian.

Besar korelasi ditunjukkan oleh bilangan koefisien korelasinya yang berkisar antara ***0*** yang secara teoretis berarti tidak ada korelasi sama sekali, sampai dengan ***1*** entah bertanda plus atau minus yang secara teoretis berarti berkorelasi secara sempurna entah searah atau berlawanan arah. Daya prediksi lazim bergerak di antara kedua bilangan ekstrim tersebut.

Interpretasi terhadap koefisien korelasi, entah positif atau negatif, lazim didasarkan pada besarnya dan ada dua cara. Pertama, dengan menguji signifikansi atau kebermaknaannya. Yang dimaksud signifikansi atau kebermaknaan adalah apakah korelasi antara prediktor dan kriteria yang diamati berdasarkan data sampel tersebut sungguh-sungguh mencerminkan keadaan pada populasinya, sehingga akan diperoleh hasil yang tetap sama bilamana diamati berdasarkan data dari berbagai sampel lain – dengan kata lain, bersifat *sistematis* – atau sekadar karena faktor kebetulan? Uji signifikansi semacam ini lazim dilakukan dengan membandingkan koefisien korelasi yang teramati tersebut (r_o kependekan dari $r_{observed}$) dengan nilai kritisnya pada tabel koefisien korelasi (r_t kependekan dari r_{tabel}) pada taraf signifikansi

tertentu (lazimnya, ***p=0,05*** atau ***p=0,01***) dan dengan memperhatikan ***df*** atau *degree of freedom* atau ***db*** atau *derajat kebebasan*-nya. Korelasi antara prediktor dan kriteria disebut signifikan atau bermakna jika $r_o \geq r_t$ pada taraf signifikansi atau ***p*** tertentu.

Cara kedua menafsirkan koefisien korelasi adalah dengan menghitung *koefisien determinasi.* Pertama, koefisien korelasi yang diperoleh atau teramati dikuadratkan (r^2). Jika koefisien korelasi kuadrat ini kemudian dikalikan 100%, hasilnya menunjukkan besar persentase variasi (skor) dalam kriteria yang dideterminasikan atau ditentukan dalam arti bisa diprediksikan atau bisa dijelaskan oleh variasi (skor) dalam prediktor, atau sebaliknya. Kembali pada contoh di atas, seandainya ditemukan koefisien korelasi antara hasil tes inteligensi sebagai prediktor (*X*) dan IPK sebagai kriteria (*Y*) r_{xy} = ***0,83***, maka koefisien determinasi antara kedua variabel tersebut adalah adalah r_{xy}^2 = ***0,69.*** Berarti 69% variasi IPK bisa dijelaskan berdasarkan variasi pada skor hasil tes inteligensi. Semakin tinggi koefisien korelasi antara prediktor dan kriteria, semakin tinggi koefisien determinasinya, semakin besar persentase variasi pada kriteria yang bisa dijelaskan berdasarkan variasi dalam prediktor, dan berarti semakin besarlah daya prediksi prediktor.

Karakteristik kelima dalam definisi tes psikologis ini memiliki dua implikasi penting. Pertama, bentuk tingkah laku yang dituntut dalam tes sebagai item-item tes tidak harus menyerupai bentuk tingkah laku nyata di luar tes yang dicoba diprediksikan. Kedua, pada akhirnya yang menjadi fokus perhatian seorang asesor adalah *non-test behaviors* atau tingkah laku nontesnya, yaitu tingkah laku dalam menghadapi berbagai tugas kehidupan sehari-hari yang hendak diprediksikan dengan tes yang bersangkutan, dan bukan tingkah laku atau respon subjek dalam tes *per se* atau pada dirinya.

C. Jenis Tes Psikologis

Tes psikologis dapat digolong-golongkan dengan berbagai cara. Beberapa cara yang lazim diterapkan dalam mengklasifikasikan psikotes adalah berdasarkan tujuan, isi, dan format administrasinya (Friedenberg, 1995). Uraian lebih rinci tentang penggolongan tes dengan tiga dasar yang dimaksud adalah seperti dipaparkan di bawah ini.

1. Penggolongan Tes Berdasarkan Tujuan

Tujuan tes dalam arti apa atau siapa yang dituju atau dijadikan sasaran tes lazim dibedakan ke dalam tiga hal (Friedenberg, 1995): (a) *domain* atau ranah atribut yang diukur, (b) *audience* atau khalayak yang akan dikenai tes, dan (c) *types of scores* atau jenis skor, dalam arti bagaimana hasil tes akan digunakan. Marilah kita tinjau pembagian lebih lanjut berdasarkan masing-masing dasar penggolongan di atas.

a. Penggolongan Tes Berdasarkan *Domain* atau Ranah Atribut yang Diukur

Domain atau ranah adalah dimensi kepribadian atau wilayah perilaku yang menjadi fokus atau sasaran pengetesan. Sebagaimana kita tahu, ranah atau wilayah perilaku lazim dibedakan ke dalam tiga kategori, yaitu *ranah kognitif* terkait dengan kemampuan olah pikir atau olah cipta, *ranah afektif* terkait dengan kemampuan olah rasa dan olah karsa, dan *ranah psikomotor* terkait dengan kemampuan olah gerak. Maka, ranah atau wilayah perilaku yang dijadikan sasaran pengukuran tentu saja menentukan *content* atau *isi* tes, di samping sampai batas tertentu juga menentukan jenis format item yang sesuai (Friedenberg, 1995). Berdasarkan ranah atau wilayah perilaku yang diukur, tes dibedakan ke dalam dua golongan besar, yaitu (1) *maximal performance tests* atau tes kinerja maksimal, dan (2) *typical performance tests* atau tes kinerja tipikal atau khas (Friedenberg, 1995).

1). *Maximal Performance Tests*

Maximal performance tests bertujuan mengukur aneka atribut psikologis yang termasuk ke dalam ranah kognitif dan ranah psikomotor, dengan cara menentukan batas maksimal atau batas atas atribut yang dimaksud dalam diri testi (Friedenberg, 1995). Dalam tes semacam ini testi diharapkan menjawab pertanyaan atau mengerjakan tugas tes semaksimal mungkin dan dengan benar atau dengan tepat pula, sehingga mencapai skor setinggi mungkin sesuai batas atas kemampuannya.

Atribut psikologis dalam ranah kognitif yang menjadi sasaran *maximal performance tests* lazim disebut *abilitas* atau kemampuan. Kemampuan dalam arti abilitas menunjuk pada kemampuan seseorang melakukan tindakan tertentu kini, bisa berupa kemampuan menjawab pertanyaan atau melaksanakan tugas tertentu. Tentu saja, kemampuan ini lebih didominasi oleh fungsi kognitif atau pikir. Selain itu, abilitas dipandang bisa terbentuk akibat faktor bawaan atau hasil belajar, atau kombinasi antara keduanya seperti akan kita lihat nanti (Chaplin, 1985; Anastasi, 1982). Yang perlu dikemukakan di sini, ada dua karakteristik penting terkait pengukuran abilitas ini. Pertama, diasumsikan, dari segi kualitas abilitas adalah sama pada setiap orang, dan bervariasi atau berlainan dari orang ke orang hanya dari segi kuantitasnya. Maka, secara khusus tujuan pengukuran abilitas adalah mengungkap perbedaan batas maksimal abilitas yang dimiliki masing-masing testi sebagaimana tercermin dari batas maksimal kinerja yang mampu ditunjukkannya.

Kedua, untuk menjawab item-item tes abilitas pada dasarnya testi harus memberikan respon berupa *judgment* atau penilaian atau bentuk respon berupa tingkah laku lain terkait dengan olah pikir. Maka, pertanyaan atau tugas yang dipakai sebagai item tes *maximal performance* atau abilitas dirumuskan sedemikian rupa sehingga jawaban atau respon testi secara objektif dapat dinyatakan *benar* atau *salah,* atau *tepat* atau *meleset,* tergantung jenis abilitas yang diukur. Selanjutnya jawaban atau respon tersebut akan diskor sedemikian

rupa, lazimnya berupa pemberian skor **1** untuk setiap jawaban benar atau respon tepat dan skor **0** untuk setiap jawaban salah atau respon meleset, sehingga skor tes masing-masing testi identik dengan jumlah jawabannya yang benar atau jumlah responnya yang tepat. Skor total berupa jumlah jawaban benar atau respon yang tepat ini dipandang mencerminkan batas maksimal abilitasnya.

Maximal performance tests lazim dibedakan ke dalam dua kategori, yaitu *achievement tests* atau *tes prestasi* dan *aptitude tests* atau *tes bakat* termasuk di dalamnya tes inteligensi. Masing-masing kategori sendiri masih bisa dibedakan ke dalam dua subkategori berdasarkan aspek yang diukur, yaitu *speed tests* yang bertujuan mengukur aspek kecepatan dalam memberikan jawaban atau respon dan *power tests* yang bertujuan mengukur aspek kekuatan dalam arti kuantitas kemampuan yang dimiliki oleh testi (Friedenberg, 1995). Masing-masing kategori tes beserta subkategorinya akan dibahas seperlunya di bawah ini.

a). *Achievement tests.* *Achievement test* atau *tes prestasi* adalah *maximal performance test* yang bertujuan mengukur kemampuan baru sebagai hasil kegiatan belajar yang baru dijalani. Dengan kata lain, jenis kemampuan yang menjadi sasaran tes prestasi memiliki asal-usul spesifik yang jelas, yaitu *nurture* atau proses pembelajaran yang baru dijalani tentu saja dalam batas-batas bakat yang dimiliki oleh murid atau pelajar yang bersangkutan.

Dalam konteks pendidikan formal, tes prestasi lazim dikaitkan dengan mata pelajaran tertentu untuk tingkat kelas tertentu pada jenjang pendidikan tertentu. Secara lebih eksplisit, Messick (1984) mendefinisikan "*educational achievement essentially refers to what an individual knows and can do in a specified subject area as a consequence of instruction.*" Maksudnya, prestasi pendidikan adalah apa yang diketahui dan bisa dikerjakan oleh seseorang dalam bidang pengajaran tertentu sebagai hasil pengajaran. Dalam konteks pendidikan nonformal, tes prestasi lazim dikaitkan dengan kursus tentang materi pengetahuan tertentu.

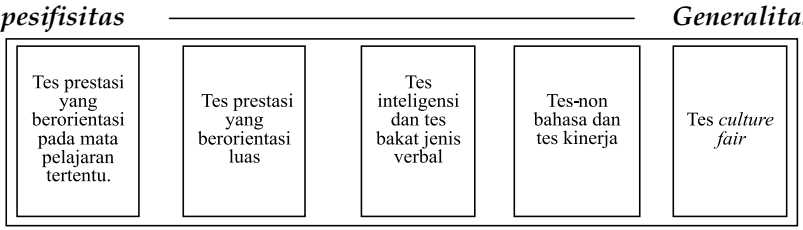
Baik mata pelajaran di sekolah maupun mata kursus di lembaga pendidikan nonformal lazim memiliki cakupan *content* atau isi yang jelas berupa pengetahuan. Isi mata pelajaran atau mata kursus ini lazim dituangkan dalam kurikulum atau silabus yang seringkali bahkan dilengkapi dengan buku bahan ajar atau buku pegangan. Maka, kualitas tes prestasi sebagai sampel tingkah laku yang mencerminkan penguasaan pengetahuan tentang suatu mata pelajaran atau mata kursus ditentukan oleh sejauh mana tes tersebut mewakili secara representatif pengetahuan yang tercakup dalam kurikulum atau yang dirumuskan sebagai tujuan pengajaran dalam mata pelajaran atau mata kursus terkait (Messick, 1984).

Dari segi tujuan, cakupan isi, dan waktu penyelenggaraannya, tes prestasi dibedakan menjadi *tes formatif* dan *tes sumatif* (Friedenberg, 1995). Tes formatif diselenggarakan pada titik-titik waktu tertentu selama pelajaran atau kursus masih berlangsung dengan cakupan materi terbatas, yaitu sebagian dari keseluruhan pelajaran atau kursus yang sudah diajarkan. Tujuannya adalah untuk mendapatkan umpan balik tentang jalannya proses pengajaran atau kursus baik bagi guru atau tutor maupun bagi murid atau peserta kursus. Hasil tes formatif lazim digunakan untuk memutuskan perlu tidaknya dilakukan *remedial teaching* atau pengajaran remedial untuk mengulang atau memperdalam penguasaan bagian materi tertentu yang dipandang masih kurang. Sebaliknya, tes prestasi sumatif diselenggarakan pada akhir pelajaran atau kursus, mencakup seluruh materi yang diajarkan, dan ditujukan untuk menentukan nilai akhir sebagai indikator umum tentang tingkat penguasaan atas pengetahuan yang baru dipelajari.

Di negara maju ada tes baku (*standardized tests*) mata pelajaran untuk jenjang sekolah tertentu, misal tes matematika SMA, tes biologi perguruan tinggi, dan lain-lain, namun bisa diduga bahwa di semua negara tes prestasi berbagai mata pelajaran atau kursus lazimnya dibuat sendiri oleh guru (*teacher-made achievement tests*), sebagaimana juga lazim terjadi di Tanah Air. Apa pun bentuknya, kiranya tes prestasi memang berfokus pada jenis-jenis pengetahuan yang merupakan bagian dari kurikulum akademik (Friedenberg, 1995).

Mengingat pentingnya fungsi tes prestasi dalam dunia pendidikan sekolah, baik sebagai sumber informasi tentang kemajuan belajar murid maupun sebagai sumber informasi tentang keefektivan kurikulum dan program pembelajaran secara keseluruhan, ada indikasi bahwa di lingkungan negara maju tes prestasi dikategorikan sebagai *educational tests* atau tes pendidikan (Embretson, 2007).

b). *Aptitude tests.* *Aptitude tests* atau *tes bakat* adalah *maximal performance tests* yang bertujuan mengukur bakat atau potensi, yaitu potensi seseorang untuk mempelajari pengetahuan baru. Di masa lalu, jenis kemampuan yang menjadi sasaran tes ini dipandang memiliki asal-usul dalam *nature,* berupa hereditas atau pembawaan, lawan dari prestasi sebagai jenis kemampuan yang pembentukannya dipandang lebih ditentukan oleh *nurture* atau proses pembelajaran. Namun pandangan yang lebih progresif cenderung menisbikan pembedaan antara *nature* dan *nurture,* dan memandang baik bakat maupun prestasi sebagai hasil saling pengaruh antara faktor pembawaan dan belajar yang berlangsung sepanjang sejarah kehidupan seseorang bahkan sejak dalam kandungan.



Gambar 4.2. Penggolongan *maximal performance tests* sebagai tes *developed abilities* (Sumber: Anastasi, 1982).

Menurut pandangan baru ini, baik bakat maupun prestasi sama-sama merupakan apa yang oleh Anastasi (1982) disebut ***developed abilities*** atau "kemampuan yang tumbuh seiring perkembangan" hasil saling pengaruh antara faktor pembawaan dan belajar. Kemampuan yang tumbuh seiring perkembangan ini bisa dibedakan ke dalam beberapa kategori atau jenis berdasarkan tingkat

spesifisitas atau kespesifikan dari latar belakang pengalaman yang membentuknya, yang terbentang dalam apa yang oleh Anastasi (1982) disebut *continuum of experiential specificity* atau kontinum kespesifikan pengalaman. Maka, mengikuti konsep baru tersebut pembagian jenis *maximal performance tests* sebagai alat ukur *developed abilities* menjadi lebih kaya sebagaimana disajikan dalam Gambar 4.2.

Seperti tampak dalam Gambar 4.2., berdasarkan konsep baru tentang *developed abilities* sebagaimana dikemukakan Anastasi (1982) maka *maximal performance tests* bisa dibedakan menjadi: (1) *tes prestasi yang berorientasi pada mata pelajaran tertentu,* lazimnya berupa jenis-jenis tes prestasi mata pelajaran untuk jenjang sekolah tertentu yang dibuat oleh guru sendiri, misal tes Matematika semester 1 kelas V SD; (2) *tes prestasi yang berorientasi luas,* lazimnya berupa tes prestasi baku tentang bidang studi tertentu yang disusun oleh pakar bidang studi bekerja sama dengan pakar psikometri, misal tes Biologi untuk SMA; (3) *tes inteligensi* dan *tes bakat* jenis verbal, meliputi aneka tes kecerdasan umum dan tes bakat khusus yang mengandalkan media bahasa, misal *Wechsler Adult Intelligence Scale-III* dan *Differential Aptitude Tests*; (4) *tes non bahasa* dan *tes kinerja,* meliputi aneka tes kecerdasan umum dan tes bakat khusus yang mengandalkan media nonverbal dan ketrampilan atau kinerja, misal *Leiter International Performance Scale-Revised*; dan (5) *tes culture fair,* berupa tes kecerdasan umum atau tes bakat khusus yang ramah terhadap aneka perbedaan budaya, lazimnya mengandalkan simbol-simbol atau pola-pola geometris sebagai media, misal *Culture Fair Intelligence Test.*

Konsep tentang *developed abilities* sebagaimana dikemukakan oleh Anastasi (1982) di atas dimaknai berbeda oleh Messick (1984) dan perbedaan tersebut terletak baik pada substansi maupun asal-usulnya. Dari segi substansinya, menurut Messick (1984) *developed abilities* merupakan abilitas yang bersifat lebih umum sebagai lawan dari abilitas spesifik dalam suatu bidang pelajaran tertentu sebagai prestasi. Dari segi asal-usulnya, berbeda dengan prestasi yang jelas-jelas merupakan hasil pengajaran, *developed abilities* sebagaimana dimaksudkan oleh Messick (1984) mencakup abilitas umum sebagai

hasil belajar dan transfer yang diperoleh dari pengalaman di dalam maupun di luar sekolah serta tanpa memperhitungkan sama sekali pengaruh kapasitas yang bersifat tetap atau bawaan yang terdapat dalam diri setiap orang. Dengan kata lain, konsep *developed abilities* ala Messick mencakup jenis-jenis abilitas paling rendah mulai kategori 2 dalam kontinum kespesifikan pengalaman ala Anastasi dan dengan mengabaikan sama sekali faktor *nature* atau potensi yang bersifat bawaan.

Ada dua catatan tambahan penting khususnya terkait tes inteligensi dan tes bakat. Pertama, pembedaan yang kaku antara inteligensi sebagai kecerdasan umum dan bakat sebagai kemampuan khusus kiranya juga sudah tidak lagi bisa dipertahankan, dengan munculnya konsep *multiple intelligences* sebagaimana dikemukakan oleh Howard Gardner (1999). Kendati menggunakan istilah inteligensi, namun sesungguhnya Gardner merujuk pada sejumlah kemampuan khusus setara bakat dalam pengertian konvensional, meliputi kecerdasan linguistik, kecerdasan matematis, kecerdasan spasial, kecerdasan musikal, kecerdasan kinestetik, kecerdasan interpersonal, kecerdasan intrapersonal, dan kecerdasan naturalistik.

Dalam kenyataan, sejak ditemukan pertama kali sekitar awal abad ke-20 pengukuran abilitas atau inteligensi telah mengalami banyak perkembangan seiring dengan kemajuan di bidang teori kognitif, khususnya terkait cara para pakar psikologi memahami inteligensi. Menurut Woodcock (2002) kemajuan di bidang teori kognitif tersebut secara garis besar telah berlangsung dalam lima taraf sebagaimana diuraikan berikut ini.

Taraf 1: Inteligensi dipandang sebagai sebuah abilitas tunggal. Teori ini menghasilkan rangkaian tes yang sesungguhnya merupakan pengembangan dari tes Binet di Amerika Serikat, yang aslinya berasal dari Prancis. Rangkaian pengembangan tes Binet tersebut diawali dengan penerbitan tes Stanford-Binet oleh Terman pada 1916. Pada tahun 1937 muncul edisinya yang kedua berupa dua bentuk alternatif yang dikerjakan oleh Terman dan Merrill. Edisinya yang ketiga terbit tahun 1960, kembali ke bentuk tunggal. Seluruh

versi tes Stanford-Binet tersebut menghasilkan dua jenis skor, yaitu usia mental (*mental age*) dan rasio *IQ* yang diperoleh dengan cara membandingkan atau membagi usia mental subjek dengan usia kronologisnya (*chronological age*). Dalam praktek, skor IQ tunggal inilah yang lazim dipakai sebagai dasar untuk menafsirkan taraf inteligensi subjek atau testi (Woodcock, 2002).

Taraf 2: Inteligensi dibedakan menjadi sepasang abilitas. Teori ini memandang inteligensi sebagai sepasang abilitas yang saling bertolak-belakang, seperti verbal versus nonverbal. Teori ini terimplementasikan antara lain dalam penerbitan rangkaian tes hasil karya Wechsler, diawali dengan terbitnya tes Wechsler-Bellevue untuk kelompok subjek dewasa yang terbit pada 1939. Tes Wechsler-Bellevue ini lazim diperlakukan sebagai instrumen yang menghasilkan informasi tentang taraf inteligensi umum atau *full scale IQ* (FSIQ), yang merupakan gabungan antara dua jenis abilitas yang lebih spesifik yaitu *verbal IQ* (VIQ) atau IQ verbal dan *performance IQ* (PIQ) atau IQ kinerja. Pada 1949 terbit *Wechsler Intelligence Scale for Children* disingkat WISC untuk mengukur inteligensi umum kelompok subjek anak-anak. Enam tahun kemudian terbit *Wechsler Adult Intelligence Scale* disingkat WAIS menggantikan tes Wechsler-Bellevue untuk mengukur inteligensi umum kelompok subjek dewasa. Seperti tes pendahulunya, kedua tes baru tersebut menghasilkan skor VIQ, skor PIQ, dan FSIQ. Pada tahun 1974 terbit edisi revisi WISC diberi nama WISC-R dan pada 1981 terbit edisi revisi WAIS diberi nama WAIS-R yang terus dipakai secara luas setidaknya sampai laporan ini dibuat (Woodcock, 2002).

Taraf 3: Inteligensi dipandang sebagai serangkaian terbatas sejumlah abilitas. Teori ini mendorong munculnya baterai tes inteligensi yang bertujuan mengukur lebih dari dua jenis abilitas kognitif yang lebih spesifik. Baterai tes pertama adalah *Woodcock-Johnson Tests of Cognitive Abilities* atau WJ yang dikerjakan oleh Woodcock dan Johnson serta terbit pada 1977. Baterai tes ini mengukur empat jenis fungsi kognitif, yaitu abilitas verbal, penalaran, kecepatan perseptual, dan ingatan. Yang kedua adalah *Stanford-*

Binet Intelligence Scale, Fourth Edition atau SB IV yang disiapkan oleh Thorndike, Hagen, dan Sattler serta terbit pada 1986. Baterai tes ini juga bertujuan mengukur empat kategori abilitas, yaitu penalaran verbal, penalaran kuantitatif, penalaran abstrak/visual, dan ingatan jangka pendek. Baterai tes ketiga adalah *Differential Abilities Scale* atau DAS karya Elliot yang terbit pada 1986 dan mengukur tiga jenis abilitas, yaitu verbal, penalaran nonverbal, dan spasial. Baterai tes keempat adalah WISC-III karya Wechsler yang terbit pada 1991, serta mengukur empat jenis abilitas, yaitu pemahaman verbal, organisasi perseptual, kekebalan terhadap gangguan perhatian (*freedom from distraction*), dan kecepatan memroses (*processing speed).* Baterai tes kelima adalah *Cognitive Assessment System* atau CAS karya Naglieri dan Das yang terbit pada 1997. Baterai tes ini mengukur empat jenis abilitas, yaitu perencanaan, perhatian, pemrosesan simultan, dan pemrosesan suksesif. Baterai tes keenam adalah WAIS-III karya Wechsler yang terbit pada 1997, serta mengukur empat jenis abilitas, yaitu pemahaman verbal, organisasi perseptual, ingatan kerja (*working memory),* dan kecepatan pemrosesan (Woodcock, 2002).

Taraf 4. Inteligensi dipandang sebagai rangkaian lengkap abilitas kognitif. Banyak pakar kini merasa bahwa teori abilitas kognitif yang dikenal sebagai teori Cattell-Horn-Carroll (CHC) memberikan gambaran yang paling memuaskan tentang abilitas kognitif. Teori CHC sesungguhnya merupakan kombinasi antara teori Gf-Gc yang dikemukakan oleh Cattell dan Horn serta teori tiga-stratum inteligensi yang dikemukakan oleh Carroll. Secara garis besar, teori CHC menyatakan bahwa abilitas kognitif manusia mencakup tiga stratum atau lapis menyerupai segi tiga yang hirarkis (McGrew, 2009). Pada stratum atau lapis pertama dan yang paling bawah terdapat 70 jenis abilitas spesifik. Pada stratum atau lapis kedua yang terletak di tengah terdapat sembilan jenis abilitas umum, meliputi (1) *Fluid reasoning* atau penalaran lentur (Gf), (2) *Comprehension-knowledge* atau pemahaman-pengetahuan (Gc), (3) *Short-term memory* atau ingatan jangka pendek (Gsm), (4) *Visual-spatial processing* atau pemrosesan visual-spasial (Gv), (5) *Auditory processing* atau pemrosesan auditori

(Ga), (6) *Long-term storage and retrieval* atau penyimpanan dan pemanggilan kembali jangka panjang (Glr), (7) *Cognitive processing speed* atau kecepatan pemrosesan kognitif (Gs), (8) *Reading and writing* atau membaca dan menulis (Grw), dan (9) *Quantitative knowledge* atau pengetahuan kuantitatif (Gq). Pada stratum atau lapis pertama yang merupakan puncak segi tiga hirarki terdapat *g* atau inteligensi umum. Menurut McGrew (2009), teori CHC merupakan taksonomi pertama yang komprehesif dan terbukti kesahihannya secara empiris tentang unsur-unsur kognitif manusia dan yang kini praktis menjadi kisi-kisi utama bagi sebagian besar baterai tes inteligensi masa kini. Contoh baterai tes yang disusun berdasarkan teori CHC adalah *Woodcock-Johnson Tests of Cognitive Ability* revisi tahun 1989, kendati baterai tes ini hanya mengukur tujuh dari sembilan abilitas umum dalam teori CHC (Woodcock, 2002).

Taraf 5. Abilitas kognitif dipandang sebagai rangkaian abilitas umum yang dilatari oleh sejumlah abilitas khusus. Menurut teori ini, ada 60 atau lebih abilitas khusus yang melatari 9 abilitas umum sebagaimana sudah disebut. Masing-masing abilitas khusus merupakan jenis-jenis abilitas yang berbeda secara kualitatif. Ada yang mempersamakan konsep abilitas khusus ini dengan konsep abilitas mental primer yang dikemukakan oleh Thurstone atau dengan abilitas primer *well-replicated cognitive factors* (WERCOF) yang dikemukakan oleh Ekstrom, French dan Harmon. Contoh baterai tes dalam kategori ini adalah *WJ-III Tests of Cognitive Ability* yang mengukur 21 jenis abilitas khusus serta *WJ-III Tests of Achievement* yang mengukur 19 jenis abilitas khusus lain. Contoh-contoh abilitas khusus meliputi antara-lain *language development* atau perkembangan bahasa (LD), *oral production and fluency* atau produksi dan kefasilan lisan (OP), *associative memory* atau ingatan asosiatif (MA), *length estimation* atau estimasi panjang (LE), *perceptual speed* atau kecepatan perseptual (P), dan *memory span* atau rentang ingatan (MS, Woodcock, 2002).

Kedua, jenis-jenis *developed abilities* dengan latar belakang pengalaman yang semakin umum atau luas meliputi baik inteligensi

maupun bakat lazimnya berupa *konstruk psikologis* hasil rumusan para pakar psikologi. Berbeda dari prestasi sebagai *developed ability* dengan latar belakang pengalaman yang spesifik, inteligensi dan bakat sebagai konstruk tidak memiliki batas cakupan isi yang konkret (Cronbach & Meehl, 1955; Friedenberg, 1995). Untuk mengukur sebuah konstruk kita perlu terlebih dulu melakukan apa yang disebut *eksplikasi konstruk,* yaitu mengidentifikasikan bentuk-bentuk tingkah laku, keyakinan, dan sikap spesifik yang bisa dipandang sebagai indikator baik yang bersifat *favorable* atau mendukung maupun yang bersifat *unfavorable* atau menyangkal atau menentang keberadaan konstruk yang bersangkutan (Friedenberg, 1995). Berdasarkan hasil eksplikasi konstruk, kita menemukan sejumlah gugus tingkah laku yang lebih konkret baik yang bersifat mendukung maupun menyangkal kehadiran konstruk yang dimaksud dan yang berperan sebagai sejenis komponen konstruk yang bersangkutan. Karena lebih konkret, maka *content* atau isi masing-masing komponen konstruk tersebut kiranya juga lebih mudah diidentifikasikan melalui *judgment.*

Selanjutnya dan seperti sudah disinggung, masing-masing kategori atau jenis tes tersebut masih bisa dibedakan berdasarkan aspek yang diukur, yaitu ***speed*** atau kecepatan atau ***power*** atau kekuatan. Contoh-contohnya adalah sebagai berikut: (1) tes prestasi kemampuan mengetik bisa berfokus pada *kecepatan*, yaitu mengukur seberapa cepat seseorang mampu menggunakan ketrampilannya untuk mengetik secara akurat, misal satu halaman naskah; (2) ulangan umum akhir semester sebagai tes prestasi aneka mata pelajaran di sekolah lazimnya berfokus pada *kekuatan*, mengukur seberapa "dahsyat" murid mampu menggunakan pengetahuan-ketrampilan yang dimilikinya untuk menjawab pertanyaan-pertanyaan tentang materi pelajaran tertentu yang baru selesai dipelajarinya; (3) tes bakat di bidang *visual-motor coordination* atau koordinasi penglihatan dan gerakan bisa berfokus pada *kecepatan* untuk mengukur seberapa cepat seseorang mampu menggunakan ketrampilan diskriminasi visual-motornya untuk menyalin serangkaian lambang yang tidak lazim; dan (4) tes inteligensi lazim berfokus pada *kekuatan* untuk mengukur

seberapa hebat kemampuan seseorang memecahkan serangkaian masalah baru atau masalah yang tidak lazim (Friedenberg, 1995).

Akibatnya, dalam jenis *tes kecepatan* item-item tes lazim berupa tugas-tugas yang relatif mudah namun waktu pengerjaannya dibatasi secara ketat. Sebaliknya, dalam jenis *tes kekuatan* item-item tes lazim berupa tugas-tugas yang sukar namun testi diberi waktu yang leluasa untuk mengerjakannya. Dalam *take-home exams* atau ujian yang boleh dikerjakan di rumah yang berlangsung di lingkungan pendidikan sekolah, murid tidak hanya diberi keleluasaan waktu melainkan juga keleluasaan menggunakan aneka sumber belajar agar sungguh-sungguh bisa menunjukkan atau membuktikan kekuatannya dalam memecahkan masalah sebagaimana disajikan dalam ujian.

Masih ada satu catatan tambahan. Penyusunan tes *developed abilities* khususnya terkait tes prestasi lazim didasarkan pada asumsi bahwa yang dijadikan sasaran pengukuran merupakan suatu variabel yang bersifat stabil atau statis. Menurut Messick (1984), psikologi kognitif yang lebih mutakhir cenderung menjelaskan proses akuisisi atau pemerolehan aneka ketrampilan kognitif sebagai proses perkembangan. Kecenderungan ini didukung serangkaian teori yang intinya menjelaskan bahwa proses belajar berlangsung dalam tiga fase: (1) fase kognitif, yaitu fase pembentukan pengetahuan deklaratif atau pengetahuan faktual, (2) fase asosiatif, yaitu fase pembentukan pengetahuan konseptual dan pengetahuan prosedural, dan (3) fase otonomi, yaitu fase penyempurnaan pengetahuan menjadi pengetahuan yang semakin kompleks tanpa batas. Menurut Messick (1984), tantangan asesmen dan pengukuran di bidang pendidikan adalah menyusun tes prestasi yang mampu mengungkap kompetensi atau kemampuan sebagai hasil belajar bukan sebagai sesuatu produk akhir yang stabil melainkan sebagai bentuk *expertise* atau keahlian yang terus mengalami perkembangan atau penyempurnaan.

c). Tes ketrampilan atau tes *performance*. Jika abilitas sebagai atribut kepribadian dipandang lebih didominasi oleh fungsi kognitif, maka harus dibahas secara khusus *skills* atau ketrampilan sebagai atribut kepribadian yang lebih didominasi oleh fungsi

psikomotorik dan yang pengukurannya juga masuk ke dalam kategori *maximal performance tests.* Ada dua cara memaknai *skills* atau **ketrampilan** sebagai jenis kemampuan yang lebih didominasi oleh fungsi psikomotorik**,** yaitu secara sempit dan secara luas.

Dalam arti sempit, **ketrampilan** mencakup berbagai jenis kemahiran atau kecakapan yang cukup spesifik seperti ketrampilan menjahit, memasak, merias, menata rambut, bermain sepak bola, bermain badminton, berenang, melompat tinggi, melompat jauh, mendayung, melempar lembing, melakukan senam lantai, dan sejenisnya. Tampak di sini, kendati dinyatakan bahwa jenis-jenis ketrampilan tersebut didominasi oleh fungsi psikomotorik namun tidak bisa dinafikan peran fungsi kognitif bahkan juga fungsi afektif. Sekalipun demikian tetap bisa diterima bahwa dalam semua jenis ketrampilan tersebut pada akhirnya yang menentukan adalah fungsi psikomotorik meliputi kekuatan, kelenturan, ketangkasan, kecekatan, kecepatan, dan kemampuan mengkoordinasikan semua unsur motorik tersebut. Berfungsinya seluruh aspek ketrampilan tersebut sebagai kesatuan akan terwujud dalam *performance* atau kinerja seseorang dalam melaksanakan jenis kecakapan spesifik tertentu seperti menjahit, mengukir, atau bermain tenis meja, misalnya.

Dalam arti luas, **ketrampilan** mencakup gugus kemampuan atau pola tingkah laku dalam rangka melaksanakan *job* atau tugas pekerjaan atau *role* atau peran tertentu. Istilah yang kini sering dipakai adalah **kompetensi.** Ketrampilan dalam arti kompetensi menunjuk pada apa yang mampu dilakukan oleh seseorang dalam situasi kehidupan nyata yang bersifat kompleks. Kompetensi mencakup serangkaian himpunan "kemampuan" berupa ketrampilan (dalam arti relatif lebih sempit), pengetahuan, dan sikap tertentu (CEPH, 2011). Sebagai contoh, kompetensi seorang guru akan mencakup pola tingkah laku yang terkait minimal lima komponen kemampuan (Friedenberg, 1995), yaitu: (1) kemampuan *mengorganisasikan pembelajaran*, meliputi antara lain menjelaskan garis besar kegiatan pada setiap awal pertemuan, membagikan *handout,* menyajikan bahan atau materi pelajaran, dan menunjukkan kaitan antara materi yang sedang dibahas dengan

materi sebelumnya; (2) kemampuan *membuat persiapan,* meliputi antara lain kemampuan menjawab pertanyaan-pertanyaan terkait sumber bacaan yang diwajibkan, kemampuan mengelaborasi topik-topik yang dibahas dalam sumber bacaan yang diwajibkan, dan kemampuan memulai pertemuan kelas baru dengan materi yang belum dibahas dalam pertemuan kelas sebelumnya; (3) kemampuan *berkomunikasi,* meliputi antara lain kemampuan menjelaskan aneka tujuan pembelajaran, menyampaikan tugas-tugas membaca bacaan wajib, berbagai batas waktu mengerjakan tugas, tanggal ulangan atau ujian, dan aneka ketentuan lain terkait pembelajaran; (4) *availability* atau ketersediaan dan kesiapan untuk dihubungi, meliputi antara lain kesediaan menentukan jadwal tetap untuk menerima siswa yang ingin bertanya atau berkonsultasi, konsekuen berada di tempat pada jadwal yang sudah ditetapkan, kesediaan memberikan janji bertemu pada waktu lain bagi siswa yang berhalangan untuk bertemu guru pada jadwal yang sudah ditetapkan; (5) *grading* atau kemampuan memberikan penilaian terhadap hasil belajar siswa, meliputi antara lain kesediaan memberikan aturan penilaian yang jelas untuk pekerjaan rumah, ulangan harian, ujian, dan penentuan nilai akhir, konsekuen menggunakan kriteria yang sama dalam menilai hasil pekerjaan semua siswa, kesediaan memberikan penjelasan jika ada siswa yang merasa kurang puas terhadap penilaian yang diberikan oleh guru. Seperti pada jenis ketrampilan spesifik seperti diuraikan di atas, berfungsinya seluruh komponen kemampuan tersebut sebagai kesatuan akan terwujud dalam *performance* atau kinerja seseorang dalam melaksanakan peran atau tugas pekerjaannya, pada contoh ini sebagai guru.

Untuk mengukur kedua jenis ketrampilan di atas dan khususnya ketrampilan dalam arti luas, lazim dilakukan *appraisal of work performance* dengan menerapkan berbagai tehnik asesmen (Gregory, 2007). Salah satu tehnik yang lazim diterapkan adalah *rating scales* atau skala penilaian terhadap berbagai unsur kemampuan yang membentuk kecakapan atau kompetensi dalam menjalankan peran atau tugas pekerjaan tertentu. Penilaian dengan *rating scales* atau skala

penilaian ini bisa dilakukan oleh diri sendiri (*self-assessment*), oleh kolega atau teman (*peer rating*), atau oleh atasan (*supervisor rating*). Dalam pengukuran ketrampilan dalam arti sempit maupun dalam arti luas sebagai pola tingkah laku ini, tujuannya adalah mengungkap *maximal performance* yang mampu ditunjukkan oleh subjek baik dalam melakukan kecakapan tertentu maupun dalam melaksanakan pola tingkah laku kompleks dalam rangka menjalankan peran atau tugas pekerjaan tertentu.

2). *Typical Performance Tests*

Typical performance tests bertujuan mengukur kebiasaan berpikir, merasa, dan bertingkah laku sehari-hari yang memberikan ciri unik atau khas pada masing-masing orang. Dalam bahasa awam, sasaran pengukuran kategori tes ini adalah kepribadian atau dalam bahasa Inggris disebut *trait,* yaitu disposisi atau kecenderungan bertingkah laku dengan cara tertentu. Ini mencakup jenis-jenis atribut psikologis yang lebih didominasi oleh fungsi afeksi atau masuk dalam ranah olah rasa dan karsa.

Ada dua ciri penting yang melekat pada kategori *typical performance tests.* Pertama, diasumsikan bahwa setiap orang berlainan dari segi kualitas mana yang dominan atau menonjol dan memberikan ciri unik pada cara bertingkah lakunya. Sebagai contoh, dalam hal orientasi nilai misalnya, ada orang yang mengutamakan nilai religius atau kesalehan namun ada pula orang yang mengutamakan nilai ekonomis atau keuntungan materi; dalam hal preferensi atau pilihan makanan kesukaan, ada orang yang menyukai sayur-mayur namun ada pula orang yang menyukai aneka daging; dalam hal sifat atau kecenderungan bertingkah laku, ada orang yang ekstraver namun ada pula orang yang introver. *Typical performance tests* bertujuan mengungkap keunikan disposisi atau kecenderungan bertingkah laku pada masing-masing orang. Maka, ciri kedua, dalam *typical performance tests* tidak dikenal jawaban salah (atau benar). Tentu, setiap item pada *typical performance test* juga memiliki kunci jawaban dan jawaban yang sesuai dengan kunci akan mendapatkan skor.

Namun berbeda dengan skor pada *maximal performance test* yang merupakan jawaban benar dan yang mencerminkan kekuatan atau kuantitas atribut yang terdapat dalam diri testi, skor pada *typical performance test* merupakan jawaban dengan arah isi tertentu yang mencerminkan keunikan kecenderungan testi dalam bertingkah laku. *Typical performance tests* atau tes kepribadian dapat digolongkan ke dalam dua kategori besar, yaitu: (1) tes kepribadian terstruktur, dan (2) tes kepribadian tak terstruktur.

a). Tes kepribadian terstruktur. Meminjam definisi yang dikemukakan oleh Hutt (1945, dalam Meehl, 1945/1971, h. 246) tes kepribadian terstruktur adalah "those in which the test material consists of conventional, culturally crystallized questions to which the subject must respond in one of a very few fixed ways." Artinya, tes kepribadian terstruktur adalah tes kepribadian yang materinya terdiri dari serangkaian pertanyaan yang konvensional, terkristalkan atau terbakukan dalam konteks budaya tertentu, di mana subjek harus memberikan respon dengan salah satu dari sejumlah kecil cara yang sudah ditentukan secara pasti. Konsep pembakuan dalam konteks budaya tertentu menunjuk pada kenyataan bahwa tes kepribadian terstruktur lazimnya berupa tes verbal, dalam arti bahwa baik pertanyaan atau tugas yang disajikan kepada subjek dan jawaban yang kemudian harus diberikan oleh subjek dinyatakan dalam rumusan kata-kata. Konsep pembakuannya sendiri juga menunjuk pada ciri lain tes kepribadian terstruktur, yaitu sifatnya yang *objektif*. Selain itu, tes kepribadian berupa tanya-jawab verbal tersebut pada dasarnya merupakan *self-ratings* atau penilaian-diri (Meehl, 1945/1971). Karena dalam rangka penilaian diri tersebut subjek pada dasarnya diminta melakukan penyelidikan atau pemeriksaan atas dirinya sendiri berpedoman pada pertanyaan-pertanyaan yang disajikan dalam tes, maka tes kepribadian terstruktur ini lazimnya juga disebut *self-inventory* atau inventori-diri. Istilah terakhir ini kini lazim dipakai untuk menyebut kategori tes kepribadian terstruktur atau objektif ini sebagai *inventori kepribadian.*

Penerapan penilaian-diri atau pemeriksaan-diri dalam pengukuran kepribadian didasarkan pada asumsi bahwa penilaian-diri dipandang sebagai "a second best source of information when the direct observation of a segment of behavior is inaccessible for practical or other reasons" (Meehl, 1945/1971, h. 247). Artinya, penilaian diri merupakan sumber informasi kedua terbaik manakala observasi langsung (sebagai sumber informasi terbaik) terhadap sebuah segmen tingkah laku tidak bisa kita peroleh karena berbagai alasan. Dengan kata lain, penilaian-diri atau deskripsi-diri dipandang mampu berfungsi sebagai "surrogate for behavior-sample" atau pengganti sampel tingkah laku. Meehl (1945/1971) memberi contoh, salah satu kriteria orang yang memiliki sifat pemalu adalah mudah memerah mukanya. Maka untuk mengetahui apakah seseorang pemalu, kita cukup bertanya kepada yang bersangkutan apakah dirinya mudah memerah mukanya. Asumsi kita, jika kenyataannya memang demikian pastilah orang itu menyadarinya dan jika dia menyadarinya pastilah dia bersedia mengakui atau mengatakan kepada kita bahwa memang demikian dirinya.

b). Tes kepribadian tak-terstruktur. Tes kepribadian tak-terstruktur identik dengan tehnik proyektif. Istilah tehnik proyektif dan bukan tes proyektif dipandang lebih tepat dipakai, sebab kebanyakan tehnik proyektif yang lazim digunakan dalam layanan klinis tidak memenuhi kriteria tradisional sebagai tes psikologis, antara lain tidak disertai dengan stimuli dan instruksi baku serta tidak dilengkapi dengan norma baku untuk membandingkan respon subjek dengan kelompoknya (Lilienfeld, Wood, & Garb, 2000). Konsep tak-terstruktur menunjuk pada jenis stimulus yang lazim dipakai dalam tehnik proyektif, yaitu *ambigu* atau bermakna ganda dalam arti tergantung cara subjek mempersepsikan dan memaknainya atau *subjektif*.

Ciri pokok yang menjadi definisi tehnik proyektif adalah "present respondents with an ambiguous stimulus, such as an inkblot, and ask them to disambiguate this stimulus" (Lilienfeld,

Wood, & Garb, 2000, h. 28). Artinya, tehnik proyektif menyajikan kepada responden sebuah stimulus ambigu, seperti sebuah gambar bercak tinta, dan memintanya untuk menerangkan stimulus tersebut. Adakalanya tehnik proyektif menuntut partisipan untuk memberikan sebuah respon, misal membuat sebuah gambar, mengikuti serangkaian instruksi yang bersifat terbuka, misal "Gambarlah seorang manusia sekehendak hati Anda." Dengan kata lain, ciri pokok pembeda tehnik proyektif dari tes terstruktur terletak baik pada stimulus maupun respon yang dituntut dari subjek. Stimulus dalam tehnik proyektif lebih ambigu dibandingkan tes terstruktur, sedangkan bentuk atau jenis dan jumlah respon subjek dalam tehnik proyektif lebih bervariasi dibandingkan tes terstruktur (Lilienfeld, Wood, & Garb, 2000).

Rasional yang mendasari kebanyakan tehnik proyektif adalah *hipotesis proyektif* (Lilienfeld, Wood, & Garb, 2000). Hipotesis proyektif menyatakan bahwa "respondents project aspects of their personalities in the process of disambiguating unstructured test stimuli" (h. 29). Artinya, responden memproyeksikan aspek-aspek kepribadiannya dalam proses menerangkan stimuli tes yang bersifat tak-terstruktur. Sebagai konsekuensinya, dalam menafsirkan hasil tehnik proyektif seorang psikolog harus bekerja secara terbalik, yaitu memeriksa jawaban responden terhadap stimuli yang disajikan untuk mendapatkan pemahaman tentang aneka disposisi kepribadiannya (Lilienfeld, Wood, & Garb, 2000).

Istilah proyeksi berasal dari Sigmund Freud (1911) dan dimaknai sebagai sebuah bentuk mekanisme pertahanan diri melalui mana seseorang secara tidak sadar mengatribusikan atau melekatkan aneka sifat dan dorongan negatifnya kepada orang lain. Maka, kelebihan pokok tehnik proyektif terletak pada kemampuannya: (1) menerobos berbagai bentuk pertahanan yang dilakukan oleh subjek secara sadar, dan (2) memungkinkan si psikolog memperoleh informasi psikologis penting (misal, aneka konflik, dorongan) yang tidak disadari oleh si subjek sendiri. Tehnik proyektif lazim dikategorikan ke dalam lima tipe mengikuti taksonomi yang pertama kali dikemukakan oleh

Lindzey (1959, dalam Lilienfeld, Wood, & Garb, 2000). Masing-masing tipe beserta contoh dan deskripsi atau penjelasannya disajikan dalam Tabel 4.1.

Tabel 4.1.
Lima Tipe Tehnik Proyektif beserta Contoh dan Penjelasannya

Tipe	Contoh	Deskripsi
Asosiasi	*Rorschach Inkblot Test* (Rorschach, 19210	Kepada responden disajikan 10 bercak tinta simetris, lima hitam putih dan lima berwarna, dan diminta mengungkapkan kesan dalam arti masing-masing bercak tinta tersebut tampak seperti apa bagi mereka.
	Hand Test (e.g., Wagner, 1962).	Kepada subjek ditunjukkan serangkaian gambar tentang tangan yang bergerak dan diminta menebak apa kiranya yang sedang dilakukan oleh masing-masing tangan.
Konstruksi	*Draw-A-Person Test (*Machover, 1949).	Responden diminta menggambar satu sosok orang pada sehelai kertas kosong, kemudian diminta menggambar satu sosok orang lain berjenis kelamin berbeda dari sosok orang pertama.
	Thematic Apperception Test (Murray & Morgan, 1938).	Kepada responden ditunjukkan serangkaian gambar yang melukiskan aneka situasi sosial yang ambigu dan diminta mengisahkan sebuah cerita tentang tokoh-tokoh yang terdapat dalam masing-masing gambar.
Melengkapi	*Washington University Sentence Completion Test* (Loevinger, 1976).	Kepada responden disajikan serangkaian frase yang merupakan bagian dari sebuah kalimat yang tidak lengkap (e.g. "Ibu saya ...) dan diminta melengkapinya sehingga menjadi sebuah kalimat utuh.
	Rosenzweig Picture Frustration Study (Rosenzweig, Fleming, & Clark, 1947).	Kepada responden disajikan serangkaian gambar kartun yang melukiskan aneka situasi yang menimbulkan frustrasi (e.g. secara tidak sengaja terkena cipratan dari genangan air di permukaan jalan yang terlindas roda mobil yang melintas) dan diminta mengungkapkan responnya secara verbal terhadap masing-masing situasi.
Menyusun/ Memilih	*Szondi Test* (Szondi, 1947).	Kepada responden disajikan serangkaian foto penderita aneka gangguan psikiatrik, dan diminta memilih pasien-pasien mana yang paling mereka sukai dan yang paling tidak mereka sukai.
	Lüscher Color Test (Lüscher & Scott, 1969).	Responden diminta mengurutkan secara berjenjang serangkaian kartu berwarna dari yang paling kurang disukai sampai dengan yang paling disukai.
Eskpresi	*Projective puppet play* (e.g. Woltmann, 1960).	Subjek anak-anak diminta memainkan peran orang lain (e.g. ayah, ibu) atau dirinya sendiri menggunakan serangkaian boneka yang disediakan.
	Handwriting analysis (lihat hasil review Beyerstein & Beyerstein, 1992).	Subjek diminta memberikan secara spontan serangkaian contoh tulisan tangannya.

Sumber: Lilienfeld, Wood, dan Garb (2000), The scientific status of projective techniques, *Psychological Science in the Public Interest, 1*(2), November, h. 30.

b. Penggolongan Tes berdasarkan *Audience* atau Khalayak Sasaran

Khalayak adalah kelompok subjek yang dituju sebagai sasaran penerapan tes. Identifikasi khalayak sasaran ini merupakan salah satu pilar validitas atau kesahihan sebuah tes. Maksudnya, setiap tes hanya akan menghasilkan pengukuran yang valid atau sahih jika diterapkan pada kelompok subjek yang menjadi sasaran yang dituju. Sebagai contoh, *Wechsler Adult Intelligence Scale (WAIS*) adalah tes inteligensi yang ditujukan bagi kelompok subjek dewasa. Versi ketiga dari tes ini (*WAIS-III*) ditujukan bagi kelompok subjek dewasa dalam rentang usia mulai 16 tahun sampai dengan 89 tahun (Gregory, 2007). Tes ini kiranya tidak akan menghasilkan pengukuran yang valid jika diterapkan pada kelompok subjek anak berusia kurang dari 16 tahun. Sebaliknya, *Wechsler Intelligence Scale for Children* adalah tes inteligensi yang ditujukan bagi kelompok subjek anak. Versi keempat dari tes ini (*WISC-IV*) ditujukan bagi kelompok subjek anak berusia antara 6,5 sampai dengan 16,5 tahun. Kiranya juga jelas bahwa tes ini tidak akan menghasilkan pengukuran yang valid jika diterapkan pada kelompok subjek berusia di atas 16,5 tahun.

Penetapan khalayak sasaran tes psikologis lazimnya memang didasarkan pada pembagian kelompok umur untuk populasi subjek normal, khususnya tes untuk khalayak sasaran subjek dewasa dan subjek anak. Di luar itu ada kategori khalayak ketiga, yaitu populasi subjek khusus meliputi subjek bayi dan anak usia dini serta subjek dewasa maupun anak namun dengan kebutuhan khusus atau kemampuan yang berbeda dari populasi subjek dewasa dan anak pada umumnya. Contohnya, *Gesell Developmental Schedules* (1925, 1974, 1987) yaitu tes untuk mengukur kemajuan perkembangan bayi dan anak usia dini mulai umur 4 minggu sampai dengan 60 bulan; *Wechsler Preschool and Primary Scale of Intelligence-III* (2002) yaitu tes kemampuan kognitif untuk kelompok anak usia 2,5 – 7 tahun 3 bulan; *Leiter International Performance Scale-Revised* (1997) yaitu tes inteligensi nonverbal bagi kelompok subjek uaia 2 – 20 tahun 11 bulan yang

mengalami gangguan pendengaran atau gangguan bicara; *Peabody Picture Vocabulary Test-III* (1998) yaitu tes kosakata bagi kelompok subjek usia 2,5 – 90 tahun atau lebih yang mengalami gangguan pendengaran, gangguan bicara, atau gangguan motorik lain termasuk penderita *stroke* (Gregory, 2007).

c. Penggolongan Tes berdasarkan Jenis Skor

Dimensi ketiga tujuan tes adalah untuk tujuan apa skor tes akan digunakan, dalam arti bagaimana tes akan digunakan dalam rangka mengevaluasi testi (Friedenberg, 1995). Sebagaimana sudah disinggung, evaluasi atau penilaian bisa dikatakan merupakan langkah terakhir dalam asesmen atau pengumpulan informasi tentang testi. Dalam evaluasi, hasil pengukuran berupa *raw score* atau skor mentah dikonversikan ke dalam *scaled score* atau skor berskala tertentu dengan cara dibandingkan dengan sebuah kriteria agar menjadi bermakna dan bisa diperbandingkan. Penggunaan tes dalam rangka mengevaluasi testi melalui cara pengonversian skor mentah menjadi skor berskala ini lazim terkait dengan dua hal: (a) apakah tes akan digunakan untuk membandingkan kinerja testi dengan kinerja kelompoknya, atau mengevaluasi kinerja masing-masing testi secara mandiri; dan (b) apakah tes dimaksudkan untuk mengukur sebuah atribut psikologis tunggal, atau mengukur serangkaian atribut psikologis? Kalau pun mengukur serangkaian atribut psikologis, apakah tes akan digunakan untuk mengungkap kekuatan absolut masing-masing atribut atau mengungkap kekuatan relatif masing-masing atribut dibandingkan dengan atribut-atribut yang lain?

Pertanyaan atau isu pertama, yaitu apakah tes akan digunakan untuk mengevaluasi kinerja testi dengan cara membandingkan kinerjanya dengan kinerja kelompoknya, atau mengevaluasi kinerja masing-masing testi secara mandiri dengan membandingkannya dengan sebuah patokan, melahirkan penggolongan tes menjadi kategori tes dengan *norm referenced scores* versus kategori tes dengan *criterion referenced scores.*

1). Tes dengan *Norm-Referenced Scores*

Sebagaimana sudah disinggung, dalam tes yang beracuan norma skor tes digunakan untuk membandingkan kinerja testi dengan kinerja kelompok sebayanya sebagai pencerminan pemilikan suatu atribut psikologis tunggal tertentu yang menjadi sasaran pengukuran. Dengan kata lain, *norm-referenced scoring* lazim diterapkan dalam pengukuran sebuah atribut psikologis tunggal, dan hasilnya digunakan untuk melakukan perbandingan inter-individual yaitu membandingkan kinerja masing-masing testi dengan sebuah kriteria yang bersifat relatif berupa rerata kinerja kelompok sebayanya. Artinya, skor testi digunakan untuk melihat *relative position* atau posisi testi dibandingkan kelompok sebayanya terkait pemilikan atribut psikologis yang diukur. Untuk keperluan itu, *raw score* atau skor mentah masing-masing testi perlu dikonversikan menjadi *scaled score* atau skor terskala atau *weighted score* atau skor terbobot dengan menggunakan norma penilaian tertentu yang disusun berdasarkan *average performance* atau rerata kinerja kelompok yang dipakai sebagai acuan. Prosedur ini lazim dikenal sebagai *norm-referenced evaluation* atau penilaian beracuan norma. Kinerja kelompok yang lazim dipakai sebagai standar dalam penilaian beracuan norma adalah kombinasi antara *mean* dan *SD* atau deviasi standar.

Proses penentuan skor beracuan norma ini akan mencakup dua langkah utama. Langkah utama yang pertama, menyusun norma penilaian khususnya untuk keperluan penilaian beracuan norma. Proses penyusunan *norma* semacam ini merupakan bagian dari proses standarisasi tes dan akan mencakup dua langkah penting. *Pertama,* mengadministrasikan tes pada sekelompok besar subjek yang memiliki kesamaan dalam minimal satu karakteristik penting. Karakteristik yang dimaksud bisa bersifat longgar, misal sama-sama berstatus mahasiswa aktif pada program pendidikan sarjana, atau bisa bersifat lebih spesifik misal sama-sama berstatus mahasiswa aktif program pendidikan sarjana Program Studi Psikologi. Kelompok yang dipakai sebagai acuan untuk menyusun norma ini lazim disebut

sampel standarisasi. Kedua, sesudah data terkumpul dihitunglah *Mean* dan juga *Standard Deviation* skor kelompok standarisasi tersebut. Sebuah norma penilaian segera bisa disusun dengan menggunakan dua statistik deskriptif tersebut. Langkah utama yang kedua, skor kasar masing-masing testi dikonversikan menjadi skor terskala atau terbobot dengan menggunakan norma penilaian yang berhasil disusun. Skor terskala atau terbobot ini merupakan *norm referenced score* yang menunjukkan posisi relatif masing-masing testi di antara kelompok sebayanya, misal apakah tergolong *average* atau rata-rata, *above average* atau di atas rata-rata, atau *below average* atau di bawah rata-rata kelompok sebayanya terkait pemilikan atribut psikologis tunggal tertentu yang sedang menjadi fokus perhatian.

Sebagaimana sudah disinggung di muka, kelemahan pokok skor beracuan norma ialah bahwa skor tersebut tidak sungguh-sungguh mencerminkan kuantitas atribut psikologis yang dimiliki oleh masing-masing testi, melainkan sekadar menunjukkan seberapa tinggi atau rendah kuantitas atribut psikologis yang dimiliknya itu dibandingkan teman-teman sebayanya.

2). Tes dengan *Criterion-Referenced Scores*

Dalam tes yang beracuan patokan, skor tes dimaksudkan untuk membandingkan kinerja testi dengan suatu standar atau kriteria absolut yang sudah ditentukan secara *apriori* sebagai pencerminan pemilikan suatu atribut psikologis tertentu atau *mastery* atau penguasaan atas materi atau kompetensi tertentu yang menjadi sasaran pengukuran. Dengan kata lain, *criterion-referenced scoring* lazim diterapkan dalam pengukuran sebuah atribut psikologis tunggal, dan hasilnya digunakan untuk melakukan perbandingan intra-individual yaitu membandingkan kinerja testi dengan sebuah kriteria yang bersifat absolut. Bisa dikatakan, skor testi digunakan untuk menunjukkan *absolute standing* atau posisi absolut seorang testi terkait pemilikan suatu atribut psikologis tertentu atau *mastery* atau penguasaan atas suatu materi atau kompetensi tertentu. Skor semacam ini diperoleh melalui penerapan *criterion-referenced evaluation* atau

evaluasi beracuan kriteria. Caranya, *raw score* atau skor mentah testi dikonversikan menjadi *scaled score* atau skor terskala atau *weighted score* atau skor terbobot dengan cara dibandingkan dengan sebuah kriteria atau standar absolut yang ditetapkan secara *apriori.*

Penentuan standar atau kriteria absolut ini lazimnya didasarkan pada pertimbangan teoretis tertentu. Berdasarkan apa yang dipakai sebagai kriteria atau patokan, penilaian beracuan patokan lazim dibedakan ke dalam tiga jenis (Friedenberg, 1995), yaitu (1) *content-referenced scoring* atau penilaian beracuan patokan materi, (2) *objective-referenced scoring* atau penilaian beracuan patokan tujuan, dan (3) *pass/fail* atau *mastery scoring* atau penilaian beracuan lulus/gagal atau beracuan penguasaan. Pada pendekatan pertama, yang dijadikan standar atau kriteria adalah penguasaan atas isi atau materi tes, yang tercermin antara lain dari *percent correct* atau persentase jawaban benar. Pada pendekatan kedua, yang dijadikan standar atau kriteria bukan isi atau materi tes melainkan *tujuan* tes atau lebih tepat tujuan pengajaran (*instructional objectives*). Kuncinya terletak pada merumuskan sistem penskoran yang mampu menunjukkan persentase tujuan pengajaran yang berhasil dikuasai murid. Pendekatan ketiga merupakan sejenis aplikasi dari dua pendekatan yang pertama. Maksudnya, nilai yang diperoleh baik melalui penilaian beracuan patokan materi maupun beracuan patokan tujuan sama-sama bisa ditransformasikan ke dalam jenis lain, khususnya untuk memilah testi atau murid menjadi dua kategori: (a) mereka yang kinerjanya dalam tes memenuhi patokan yang ditentukan (*lulus*), dan (b) mereka yang kinerjanya dalam tes tidak memenuhi patokan yang ditentukan (*gagal*). Perbedaan skor dalam masing-masing kategori *lulus* atau *gagal* tidak lagi memiliki makna.

Pertanyaan atau isu kedua, apakah tes dimaksudkan untuk mengukur sebuah atribut psikologis tunggal atau mengukur serangkaian atribut psikologis, dan kalau pun mengukur serangkaian atribut psikologis, apakah tes akan digunakan untuk mengungkap kekuatan absolut masing-masing atribut atau mengungkap kekuatan relatif masing-masing atribut dibandingkan dengan atribut-atribut

yang lain. Isu ini melahirkan penggolongan tes menjadi kategori tes dengan *ipsative scores* versus kategori tes dengan *normative scores.*

3). Tes dengan *Ipsative Scores*

Sejumlah tes bertujuan mengukur serangkaian atribut psikologis yang berlainan sekaligus. Selain itu terhadap serangkaian atribut psikologis yang berbeda-beda tersebut tes ini tidak bertujuan mengukur kekuatan absolut masing-masing atribut, melainkan kekuatan relatifnya dibandingkan dengan setiap atribut yang lain. Tes semacam ini menghasilkan jenis skor yang disebut *skor ipsatif*. Skor ipsatif pada dasarnya bertujuan mengurutkan secara berjenjang kekuatan serangkaian atribut psikologis, bisa berupa kebutuhan, rasa suka, atau kecenderungan bertingkah laku dalam perbandingannya satu dengan yang lain dalam diri seseorang (Friedenberg, 1995). Dalam tes yang mengukur serangkaian atribut dengan skor ipsatif, seorang testi tidak mungkin mencapai skor tinggi pada semua atribut, atau sebaliknya mencapai skor rendah pada semua atribut, atau mencapai skor dengan pola tinggi-rendah sembarang, melainkan akan mencapai pola skor yang mencerminkan jenjang urutan kekuatan dari masing-masing atribut (Friedenberg, 1995). Karena bertujuan membandingkan kekuatan relatif serangkaian atribut, item-item tes dengan skor ipsatif lazim menggunakan format yang menuntut testi membandingkan atribut yang satu dengan atribut yang lain, lazimnya dengan item berformat ***forced-choice*** (Friedenberg, 1995).

Salah satu contoh alat ukur yang menghasilkan skor ipsatif semacam ini adalah *Edwards Personal Preference Schedule*. Tes ini merupakan inventori kepribadian yang bertujuan mengukur jenis-jenis kebutuhan yang merupakan inti teori kepribadian yang dikemukakan Henry Murray. Sebagaimana diketahui, Murray menyatakan bahwa tingkah laku manusia antara lain digerakkan oleh sejumlah kebutuhan dasar, 15 di antaranya meliputi: (1) *achievement* atau kebutuhan untuk berprestasi, (2) *deference* atau kebutuhan untuk tunduk dan mengikuti seseorang yang dikagumi, (3) *order* atau kebutuhan akan ketertiban, (4) *exhibition* atau kebutuhan untuk menarik perhatian dari orang

lain, (5) *autonomy* atau kebutuhan untuk mengatur diri-sendiri, (6) *affiliation* atau kebutuhan untuk menjalin persahabatan dengan orang lain, (7) *intraception* atau kebutuhan untuk memandang dunia sekitar secara hangat-optimistik, (8) *succorance* atau kebutuhan untuk mendapatkan bantuan, perlindungan, dukungan, kasih, penghiburan, dan bimbingan dari orang lain, (9) *dominance* atau kebutuhan untuk mengendalikan atau memimpin orang lain, (10) *abasement* atau kebutuhan untuk mengaku salah dan menerima hukuman, (11) *nurturance* atau kebutuhan untuk melindungi, memberikan bantuan, dan menyejahterakan makhluk atau orang lain yang membutuhkan, (12) *change* atau kebutuhan untuk menghindari kerutinan, mengalami perubahan, (13) *endurance* atau kebutuhan untuk bertekun dan bekerja keras, (14) *heterosexuaity* atau kebutuhan untuk merasa tertarik pada lawan jenis, dan (15) *aggression* atau kebutuhan untuk menyerang atau melukai orang lain, merendahkan, meremehkan, mengolok-olok, memfitnah, menghukum secara kejam, atau melakukan tindakan sadis terhadap orang lain.

Murray sendiri bersama sejumlah kolega awalnya mencoba mengukur aneka kebutuhan tersebut dengan menggunakan *Thematic Apperception Test*, sebuah tes proyektif terdiri dari serangkaian gambar melukiskan aneka tema-peristiwa kehidupan sehari-hari. Kendati menerapkan sistem penskoran, namun karena mengandalkan interpretasi yang bersifat kualitatif-subjektif maka kualitas psikometrik tes proyektif ini oleh sementara pihak dipandang kurang memuaskan. Untuk mengatasinya, sorang psikolog lain bernama Allen L. Edwards (1959) menyusun tes serupa namun dengan menggunakan teknik inventori yang lebih terstruktur dan objektif. Tes yang kemudian terkenal dengan akronim *EPPS* ini terdiri dari 210 item berupa pasangan-pasangan pernyataan. Dalam setiap item pernyataan yang mengukur salah satu kebutuhan dipasangkan dengan pernyataan yang mengukur 14 kebutuhan lainnya. Dengan item berformat *forced-choice* testi diminta memilih salah satu pernyataan yang dirasakan sesuai dengan keadaan dirinya pada setiap item. Hasilnya akan berupa profil skor yang menunjukkan

kekuatan relatif dari masing-masing kebutuhan dibandingkan aneka kebutuhan lainnya. Jelas kiranya, mustahil seorang testi mencapai skor tinggi pada seluruh kebutuhan, atau sebaliknya mencapai skor rendah pada seluruh kebutuhan, atau mencapai skor tinggi dan rendah pada aneka kebutuhan secara sembarang. Secara terstruktur, jika seseorang mencapai skor tinggi pada kebutuhan tertentu pastilah akan mencapai skor rendah pada kebutuhan lain, sehingga terbentuk profil kebutuhan yang mencerminkan keunikan atau kekahasan kecenderungannya dalam bertingkah laku.

4). Tes dengan *Normative Scores*

Skor normatif mencerminkan kekuatan absolut karakteristik atau atribut tunggal tertentu yang terdapat dalam diri testi (Friedenberg, 1995). Dalam bahasa kuantitatif, skor normatif menunjukkan sebesar atau sebanyak apa seorang testi memiliki atribut tertentu yang sedang menjadi sasaran pengukuran. Penskoran normatif dapat diterapkan pada baik jenis tes yang mengukur sebuah atribut psikologis tunggal maupun pada jenis tes yang mengukur serangkaian atribut psikologis sekaligus. Dalam hal ini berlaku prinsip sebagai berikut: (1) sebuah tes yang mengukur sebuah atribut psikologis tunggal pastilah menghasilkan skor normatif; (2) kesatuan rangkaian tes yang mengukur sejumlah atribut psikologis sekaligus hanya menghasilkan skor normatif jika memang dirancang menerapkan sistem penskoran normatif. Dalam hal yang disebut terakhir, maka seorang testi dapat mencapai skor tinggi pada semua tes, atau sebaliknya mencapai skor rendah pada semua tes, atau mencapai skor yang berlainan pada berbagai tes yang disajikan sebagai satu rangkaian namun dengan menerapkan sistem penskoran normatif (Friedenberg, 1995).

Kebanyakan tes psikologis maupun tes pendidikan merupakan jenis tes yang mengukur sebuah atribut psikologis tunggal dan menerapkan penskoran normatif. Contoh yang paling jelas adalah berbagai tes hasil belajar sumatif pada berbagai mata pelajaran di sekolah yang disusun sendiri oleh guru atau dosen, mulai jenjang pendidikan dasar sampai dengan jenjang pendidikan tinggi. Tes

sumatif untuk setiap mata pelajaran atau mata kuliah yang lazim disebut Tes Hasil Belajar (THB) atau Ujian Akhir Semester (UAS) semacam itu dapat dipastikan berdiri sendiri, tidak memiliki kaitan dengan tes sejenis pada setiap mata pelajaran atau mata kuliah lainnya. Hasil tes yang lazim dilaksanakan pada akhir kegiatan pembelajaran atau perkuliahan ini dipandang mencerminkan seberapa tinggi atau banyak setiap murid atau mahasiswa menguasai pengetahuan dan ketrampilan yang selesai dipelajarinya dalam masing-masing mata pelajaran atau mata kuliah secara terpisah satu dari yang lain.

Sejumlah tes psikologis merupakan *test battery* atau baterai tes, yaitu berupa rangkaian yang meliputi sejumlah subtes yang mengukur jenis kemampuan yang berlainan. Kendati disajikan secara bersamaan dalam arti sebagai satu rangkaian, namun setiap tes pada dasarnya tetaplah mengukur satu jenis kemampuan secara mandiri. Salah satu contoh baterai tes yang juga cukup dikenal di Tanah Air adalah *Differential Aptitude Tests* (DAT) yang disusun oleh Bennett, Seashore, dan Wesman pada 1952 (Nunnally, Jr., 1970). Awalnya baterai tes ini dimaksudkan untuk layanan bimbingan belajar dan bimbingan karir bagi murid-murid sekolah setara SMP dan SMA di Amerika Serikat. Selanjutnya tes ini juga dipakai untuk melakukan seleksi calon karyawan. Secara berkala baterai tes ini terus disempurnakan dan hingga kini merupakan salah satu baterai tes bakat yang paling luas digunakan di berbagai negara (Gregory, 2007).

Baterai tes DAT terdiri atas 8 tes yang berlainan (*independent tests*), yaitu (Gregory, 2007): (1) *Verbal Reasoning* atau Berpikir Verbal; item-item tes ini berupa berbagai bentuk analogi verbal yang lebih mengutamakan kemampuan berpikir daripada pemahaman kosa kata; (2) *Numerical Reasoning* atau Berpikir Numerik; item-item tes ini berupa soal-soal komputasi atau hitungan; (3) *Abstract Reasoning* atau Berpikir Abstrak; tes ini pada hakikatnya mirip tes Berpikir Verbal, hanya item-itemnya berupa gambar pola-pola abstrak; (4) *Spatial Relations* atau Hubungan Ruang; tes ini mengukur kemampuan mengimajinasikan bagaimana tampak sejumlah objek jika letaknya dirotasikan serta kemampuan memvisualisasikan objek berdimensi

tiga berdasarkan pola berdimensi dua; (5) *Mechanical Reasoning* atau Berpikir Mekanik; item-item tes terdiri dari gambar-gambar yang menyajikan aneka problem mekanis; (6) *Clerical Speed and Accuracy* atau Tes Cepat-Tepat; tes ini mengukur kecepatan persepsi, testi diminta mengidentifikasikan pasangan-pasangan identik yang terdiri dari bilangan atau pola sederhana tertentu; (7) *Spelling* atau Ejaan; tes ini mengukur penguasaan testi tentang ejaan, tentu aslinya terdiri dari soal-soal tentang ejaan bahasa Inggris; dan (8) *Language Usage* atau Kemampuan Berbahasa; tes ini mengukur penguasaan testi tentang penerapan kaidah tata bahasa dalam penyusunan kalimat. Ada yang menyatakan bahwa dua tes terakhir yang dalam versi-versi awal digabungkan di bawah judul *Language Usage,* lebih tepat disebut tes prestasi atau hasil belajar daripada tes bakat (Nunnally, Jr., 1970).

Masing-masing tes di atas bersifat independen atau berdiri sendiri, dalam arti mengukur kemampuan yang berlainan secara mandiri kendati disajikan bersama tes-tes lain sebagai rangkaian. Sifat independen dari masing-masing tes ini melahirkan setidaknya dua konsekuensi sebagai berikut: (1) masing-masing atau beberapa tes dapat diadministrasikan dan hasilnya diinterpretasikan secara mandiri, tidak harus dalam satu rangkaian utuh bersama tes yang lain; (2) seorang testi dapat mencapai hasil yang sama-sama tinggi atau sebaliknya sama-sama rendah pada masing-masing tes. Kedua konsekuensi tersebut menegaskan sifat normatif dari sistem penskoran yang diterapkan dalam baterai tes DAT.

2. Penggolongan Tes Berdasarkan Isi

Seperti sudah disinggung *content* atau isi tes psikologis terkait dengan *domain* atau ranah, yaitu dimensi kepribadian atau wilayah perilaku tempat atribut psikologis yang sedang menjadi sasaran pengukuran terletak. Sebagaimana kita ketahui, ranah atau dimensi kepribadian atau wilayah tingkah laku tersebut lazim dibedakan menjadi tiga, yaitu ranah kognitif terkait dengan fungsi berpikir, ranah afektif terkait dengan fungsi merasa, dan ranah psikomotor

terkait dengan fungsi gerak tubuh atau anggota tubuh. *Content* atau isi tes psikologis menunjuk pada jenis kemampuan atau jenis atribut psikologis yang terletak dalam masing-masing ranah, dimensi kepribadian atau wilayah tingkah laku tersebut dan yang menjadi sasaran atau objek pengukuran. Istilah "terletak" sebenarnya kurang tepat, sebab setiap bentuk tingkah laku sebagai ungkapan atau perwujudan kemampuan atau atribut psikologis tertentu pastilah melibatkan aneka kemampuan yang terdapat di dalam ketiga ranah tersebut secara serentak. Namun harus diakui bahwa bentuk tingkah laku tertentu lebih didominasi oleh kemampuan yang termasuk ke dalam ranah kognitif, sedangkan bentuk tingkah laku lain lebih didominasi oleh jenis kemampuan yang termasuk ke dalam ranah afektif atau psikomotorik, dan seterusnya. Dalam arti kemampuan dalam ranah mana yang lebih mendominasi itulah kita bisa membedakan jenis-jenis kemampuan yang lazim menjadi *content* atau isi tes psikologis. Berdasarkan isinya, tes psikologis lazim dibedakan ke dalam tiga kategori, yaitu: (a) tes yang mengukur pengetahuan dan proses berpikir pada umumnya; (b) tes yang mengukur disposisi kepribadian atau kecenderungan bertingkah laku, dan (c) tes yang mengukur ketrampilan dan *pola tingkah laku* (Friedenberg, 1995).

a. Tes yang Mengukur Pengetahuan dan Proses Berpikir

Dalam konteks pendidikan sekolah, pembedaan tentang jenis-jenis pengetahuan dan proses berpikir dapat dilakukan dengan beberapa cara. Pertama, jenis-jenis pengetahuan bisa dibedakan berdasarkan *content* atau isi *subject matter* atau mata pelajarannya. Kedua, jenis-jenis pengetahuan dan jenis-jenis proses berpikir juga dapat dibedakan dengan mengikuti taksonomi tujuan pengajaran khususnya untuk ranah kognitif.

Sebelum membahas lebih lanjut jenis-jenis tes berdasarkan jenis pengetahuan dan proses berpikir yang diukur, kiranya perlu kita tinjau secara sekilas makna dan perkembangan taksonomi tujuan

pengajaran. *Taksonomi* adalah sejenis cara mengklasifikasikan objek, dalam hal ini jenis-jenis kemampuan dalam masing-masing ranah. Pada mulanya taksonomi ini dikembangkan dalam rangka membantu para guru merumuskan tujuan pengajaran di sekolah, maka disebut *taxonomy of educational objectives*. Tujuan pengajaran pada dasarnya merupakan rumusan tentang *learning outcomes* atau hasil belajar yang hendak dicapai, lazimnya mencakup dua unsur, yaitu: (1) rumusan tentang isi mata pelajaran tertentu yang dinyatakan dalam kata benda, dan (2) rumusan tentang apa yang harus dikerjakan terhadap isi mata pelajaran tersebut yang dinyatakan dalam kata kerja. Dengan kata lain, rumusan tujuan pengajaran lazim terdiri atas sebuah atau sekumpulan kata benda yang memuat rumusan tentang isi mata pelajaran dan sebuah atau sekumpulan kata kerja yang memuat rumusan tentang proses kognitifnya (Krathwohl, 2002). Sejarah singkat perkembangan taksonomi tujuan pengajaran adalah seperti diuraikan di bawah ini.

Taksonomi Bloom

Salah satu perintis pengembangan taksonomi tujuan pengajaran adalah Bloom (Ed.), Engelhart, Furst, Hill, dan Krathwohl (1956). Mereka pulalah kiranya yang berjasa pertama kali mencetuskan penggolongan kemampuan manusia ke dalam tiga ranah, yaitu kognitif, afektif, dan psikomotorik. Karya rintisan mereka yang termasyhur adalah taksonomi tujuan pengajaran ranah kognitif yang hingga kini dikenal sebagai *Taksonomi Bloom.*

Secara ringkas, Taksonomi Bloom menguraikan kemampuan kognitif manusia ke dalam enam tingkat proses kognitif, masing-masing mencakup sejumlah subkategori kecuali salah satu di antaranya. Keenam tingkat proses kognitif beserta subkategorinya adalah sebagai berikut: (1) *knowledge* atau pengetahuan, meliputi: (a) *knowledge of specifics* atau pengetahuan tentang hal-hal spesifik, (b) *knowledge of ways and means of dealing with specifics* atau pengetahuan tentang cara-cara menangani hal-hal spesifik, dan (c) *knowledge of*

universals and abstractions in a field atau pengetahuan tentang hal-hal yang universal dan berbagai abstraksi yang berlaku dalam suatu bidang tertentu; (2) *comprehension* atau pemahaman, meliputi (a) *translation* atau kemampuan menerjemahkan, (b) *interpretation* atau kemampuan menafsirkan, dan (c) *extrapolation* atau kemampuan membuat ekstrapolasi; (3) *application* atau penerapan, (4) *analysis* atau analisis, meliputi (a) *analysis of elements* atau kemampuan menganalisis unsur-unsur, (b) *analysis of relationships* atau kemampuan menganalisis hubungan-hubungan, dan (c) *analysis of organizational principles* atau kemampuan menganalisis prinsip-prinsip organisasi atau pengaturan; (5) *synthesis* atau sintesis, meliputi (a) *production of a unique communication* atau kemampuan menyusun sebuah komunikasi yang unik, (b) *production of a plan, or proposed set of operations* atau kemampuan menyusun sebuah rencana atau rangkaian tindakan, dan (c) *production of a set of abstract relations* atau kemampuan menyusun serangkaian hubungan yang bersifat abstrak; dan (6) *evaluation* atau evaluasi, meliputi (a) *evaluation in terms of internal evidence* atau kemampuan memberikan evaluasi berdasarkan bukti-bukti internal atau dari dalam, dan (b) *judgments in terms of external criteria* atau kemampuan memberikan penilaian berdasarkan kriteria eksternal atau dari luar (Krathwohl, 2002).

Taksonomi Bloom memiliki tiga ciri penting. *Pertama*, taksonomi ini ber**-dimensi tunggal**, dalam arti masing-masing tingkat proses kognitif sekaligus sudah mencakup baik unsur kata benda maupun unsur kata kerja. Unsur kata kerjanya dinyatakan dalam enam kategori tingkat proses kognitif. Unsur kata bendanya dicantumkan sebagai sub-subkategori dalam masing-masing tingkat proses kognitif, kecuali tingkat proses kognitif ketiga yaitu aplikasi atau penerapan yang tidak memiliki subkategori. *Kedua*, perbedaan antar tingkat proses kognitif ditentukan oleh **perbedaan taraf kesulitan**-nya, mulai dari yang paling mudah atau paling sederhana yaitu *knowledge* atau pengetahuan sampai ke yang paling sulit atau paling kompleks yaitu *evaluation* atau evaluasi. *Ketiga*, keenam tingkat proses kognitif tersebut dipandang membentuk sejenis **hirarki kumulatif.**

Artinya, penguasaan atas setiap tingkat proses kognitif yang lebih sederhana merupakan prasyarat bagi penguasaan tingkat proses kognitif berikutnya yang lebih kompleks (Krathwohl, 2002).

Kendati dipakai secara luas khususnya di bidang pengembangan evaluasi hasil pengajaran dalam beberapa dasawarsa sejak diluncurkan pada tahun 1950-an bahkan hingga kini, Taksonomi Bloom dipandang memiliki sejumlah kelemahan mendasar antara lain terlalu menyederhanakan hakikat proses berpikir dalam kaitannya dengan proses belajar (Marzano & Kendall, 2007).

Taksonomi Revisi Anderson et al.

Pada tahun 2001 Anderson (Ed.), Krathwohl (Ed.), Airasian, Cruikshank, Mayer, Pintrich, Raths, dan Wittrock, menerbitkan revisi terhadap Taksonomi Bloom. Taksonomi revisi yang disusun oleh Anderson dan kawan-kawan ini masih memiliki kesamaan dengan Taksonomi Bloom terkait dua ciri, yaitu pembedaan antar tingkat proses kognitif berdasarkan **taraf kesulitan** dan **hirarki kumulatif** keenam tingkat proses kognitif tersebut. Terkait dimensinya, Anderson dan kawan-kawan memisahkan aspek kata benda dan aspek kata kerja menjadi dua dimensi yang berbeda, yaitu aspek kata benda sebagai **dimensi pengetahuan** sedangkan aspek kata kerja sebagai **dimensi proses kognitif**. Dengan kata lain, taksonomi revisi Anderson dan kawan-kawan memiliki **dua dimensi**, bukan hanya satu dimensi seperti taksonomi Bloom.

Dimensi pengetahuan sebagai aspek kata benda terdiri atas empat jenis: (1) *pengetahuan faktual,* yaitu "basic elements students must know to be acquainted with a discipline or solve a problem in it"; artinya, pengetahuan tentang unsur-unsur dasar yang harus dimiliki oleh murid agar mampu memahami sebuah disiplin ilmu atau memecahkan suatu persoalan dalam disiplin ilmu yang bersangkutan; (2) *pengetahuan konseptual,* yaitu "the interrelationships among the basic elements within a larger structure that enable them to function together"; artinya, pengetahuan tentang saling hubungan antar

unsur-unsur dasar menjadi sebuah struktur yang lebih besar atau luas sehingga bisa berfungsi bersama-sama; (3) *pengetahuan prosedural,* yaitu "how to do something, methods of inquiry, and criteria for using skills, algorithms, techniques, and methods"; artinya, pengetahuan tentang cara melakukan sesuatu, metode-metode penyelidikan, dan kriteria untuk menerapkan aneka ketrampilan, algoritme atau urutan logis pengambilan keputusan untuk pemecahan masalah tertentu, teknik, dan metode; dan (4) *pengetahuan metakognitif,* yaitu "knowledge of cognition in general as well as awareness and knowledge of one's own cognition"; artinya, pengetahuan tentang proses berpikir secara umum serta kesadaran dan pemahaman tentang proses berpikirnya sendiri (Krathwohl, 2002; Marzano & Kendall, 2007).

Dimensi proses kognitif sebagai aspek kata kerja terdiri atas enam jenis proses berpikir, yaitu: (1) *remembering* atau mengingat, yaitu "retrieving relevant knowledge from long-term memory"; artinya, mendapatkan kembali pengetahuan yang diperlukan dari ingatan jangka panjang; (2) *understanding* atau memahami, yaitu "determining the meaning of instructional messages, including oral, written, and graphic communication"; artinya, menentukan makna pesan-pesan pembelajaran yang disampaikan melalui komunikasi baik secara lisan, tertulis, maupun menggunakan gambar-gambar; (3) *applying* atau menerapkan, yaitu "carrying out or using a procedure in a given situation"; artinya, melaksanakan atau menerapkan suatu prosedur dalam situasi tertentu; (4) *analyzing* atau menganalisis, yaitu "breaking material into its constituent parts and detecting how the parts relate to one another and to an overall structure or purpose"; artinya, menguraikan materi atau bahan ke dalam bagian-bagiannya serta menemukan saling hubungan antar bagian-bagian tersebut dan hubungannya dengan sebuah struktur atau tujuan tertentu secara keseluruhan; (5) *evaluating* atau mengevaluasi, yaitu "making judgments based on criteria and standards"; artinya, membuat serangkaian penilaian berdasarkan kriteria atau standar tertentu; dan (6) *creating* atau mencipta, yaitu "putting elements together to form a novel, coherent whole or make an original product"; artinya,

menggabungkan unsur-unsur sehingga terbentuk sebuah kesatuan baru yang bersifat koheren atau membuat sebuah produk yang bersifat orisinal (Krathwohl, 2002; Marzano & Kendall, 2007).

Karena memiliki dua dimensi, taksonomi revisi Anderson et al. ini disebut **Tabel Taksonomi** dengan dimensi *pengetahuan* menempati poros vertikal dan dimensi *proses kognitif* menempati poros horisontalnya. Sebuah tujuan pengajaran dapat ditempatkan pada satu atau lebih sel yang merupakan interseksi atau perpotongan antara kolom kata kerja (dimensi Proses Kognitif) dan baris kata benda (dimensi Pengetahuan), seperti disajikan dalam Tabel 4.2.

Tabel 4.2.
Taksonomi Revisi Anderson et al.

Dimensi Pengetahuan	**Dimensi Proses Kognitif**					
	Mengingat	**Memahami**	**Meng-aplikasikan**	**Meng-analisis**	**Meng-evaluasi**	**Men-ciptakan**
Pengetahuan Faktual	Membuat daftar (*List*)	Membuat ringkasan (*Summarize*)	Mengklasifi-kasikan (*Classify*)	Mengurut-kan (*Order*)	Menentukan *ranking* (*Rank*)	Mengombi-nasikan *(Combine)*
Pengetahuan Konseptual	Mendeskrip-sikan (*Describe*)	Menafsirkan (*Interpret*)	Bereksperi-mentasi *(Experiment)*	Menjelaskan (*Explain*)	Melakukan penilaian (*Assess*)	Merencana-kan *(Plan)*
Pengetahuan Prosedural	Menabulasi-kan (*Tabulate*)	Mempredik-sikan (*Predict*)	Menghitung (*Calculate*)	Membeda-kan (*Differentiate*)	Membuat kesimpulan (*Conclude*)	Membuat komposisi *(Compose)*
Pengetahuan Meta-kognitif	Menggguna-kan secara tepat/ semestinya (*Appropriate use*)	Mengeksе-kusi (*Execute*)	Mengon-struksi (*Construct*)	Mencapai atau meraih pencapaian *(Achieve)*	Melakukan tindakan *(Action)*	Mengaktua-lisasikan *(Actualize)*

Dibandingkan dengan Taksonomi Bloom, taksonomi revisi Anderson et al. ini dipandang lebih menjamin pengungkapan hasil belajar berupa jenis-jenis kemampuan dengan taraf kompleksitas yang tinggi. Namun seperti pendahulunya, taksonomi revisi ini masih disusun berdasarkan pandangan tentang perbedaan taraf kesukaran atau kompleksitas antar masing-masing proses kognitif. Kelemahan ini mendorong sejumlah pakar lain mengemukakan usulan revisi baru.

Taksonomi Revisi Marzano & Kendall

Marzano dan Kendall (2007) berpendapat bahwa proses berpikir tidak bisa diurutkan secara hirarkis berdasarkan tingkat kesulitannya, namun bisa diurutkan berdasarkan *control* atau taraf kendali dalam arti sejauh mana suatu proses berpikir mengendalikan dalam arti memicu dan mengarahkan proses berpikir lainnya. Atas dasar keyakinan tersebut dan dengan menggunakan apa yang mereka sebut *model of behavior* atau model tingkah laku, mereka mengembangkan *Taksonomi Baru* sebagai revisi terhadap dua taksonomi ranah kognitif sebelumnya.

Taksonomi Baru dikembangkan berdasarkan asumsi bahwa belajar berarti melibatkan diri untuk menggeluti sesuatu yang baru. Model tingkah laku yang dipakai dalam taksonomi ini mampu menjelaskan dua hal terkait tingkah laku belajar kita. Pertama, model ini menjelaskan cara seseorang memutuskan untuk mengerjakan atau menghindari suatu tugas baru. Yang dimaksud dengan *tugas baru* adalah "an opportunity to change whatever one is doing or attending to at a particular time" (h. 12). Artinya, tugas baru adalah setiap kesempatan atau peluang untuk mengubah kegiatan atau perhatian pada suatu ketika tertentu dalam kehidupan seseorang. Kedua, model ini juga menjelaskan cara seseorang mengolah informasi baru yang diperolehnya sesudah dia memutuskan untuk mengubah kegiatan atau perhatian, yaitu sesudah dia memutuskan untuk menghadapi tugas baru tersebut dan bukan mengabaikan atau menghindarinya.

Model tingkah laku Marzano dan Kendall (2007) yang dipakai sebagai dasar penyusunan Taksonomi Baru mereka mencakup empat komponen yang terdiri atas tiga sistem mental atau sistem berpikir, yaitu *self-system* atau **sistem diri**, **sistem metakognitif**, dan **sistem kognitif**, serta komponen keempat berupa **pengetahuan**. Makna masing-masing komponen serta prinsip kerja model tingkah laku tersebut adalah seperti diuraikan berikut ini.

Self-system atau **sistem diri** merupakan jaringan atau gugusan keyakinan dan tujuan hidup yang saling terkait satu sama lain dan

yang dipakai seseorang dalam *menilai* perlu-tidaknya melakukan suatu tugas baru. Jika suatu tugas baru dinilai penting, peluang untuk mengerjakannya secara berhasil tinggi, serta bersifat menyenangkan, maka seseorang akan termotivasi untuk menggeluti dalam arti melakukan tugas baru tersebut. Sebaliknya jika tugas baru tersebut dinilai kurang penting, peluangnya untuk sukses kecil, apalagi juga tidak menyenangkan, maka lazimnya seseorang tidak akan termotivasi untuk melakukan tugas baru tersebut. Dengan kata lain, sistem diri berperan sebagai pemicu dalam arti sumber motivasi dalam tingkah laku belajar kita.

Sistem metakognitif. Sesudah keputusan untuk melakukan tugas baru diambil, maka sistem diri akan memicu bekerjanya komponen kedua yaitu **sistem metakognitif.** Sistem metakognitif bertugas melakukan *monitoring* atau pemantauan, evaluasi atau penilaian, serta regulasi atau pengaturan fungsi semua jenis proses berpikir lainnya (Marzano & Kendall, 2007). Secara konkret ada dua hal penting yang harus dilakukan oleh sistem metakognitif terkait pelaksanaan tugas baru tersebut. Pertama, sistem metakognitif bertugas menetapkan **tujuan** atau **tujuan-tujuan** terkait pelaksanaan tugas baru tersebut. Misal, seorang anak tertarik untuk mengetahui isi cerita sebuah novel yang baru dibelikan oleh ayahnya (sistem diri). Selanjutnya dia mungkin memutuskan untuk mengetahui fakta-fakta pokoknya dulu, meliputi siapa tokoh-tokohnya, apa peristiwa-peristiwa pentingnya, serta kapan dan di mana tempat kejadiannya (tujuan-tujuan, sistem metakognitif). Kedua, sistem metakognitif bertugas merumuskan **strategi** atau **strategi-strategi** untuk mencapai tujuan atau tujuan-tujuan yang sudah ditetapkan tersebut. Misalnya, anak itu memutuskan untuk terlebih dulu mempelajari daftar isinya, atau membaca secara cepat bagian awal dan bagian akhir novel tersebut sebagai strategi yang dia tempuh untuk mencapai tujuannya yaitu mengetahui fakta-fakta pokok tentang novel tersebut (strategi, sistem metakognitif).

Sistem kognitif. Sesudah berhasil melaksanakan tugasnya, sistem metakognitif tersebut selanjutnya akan memicu bekerjanya

komponen ketiga dalam model tingkah laku belajar kita, yaitu **sistem kognitif.** Sistem kognitif bertugas memroses aneka jenis pengetahuan yang diperlukan untuk menyelesaikan tugas baru tersebut. Menurut Marzano dan Kendall (2007), ada empat subkomponen dalam sistem kognitif yang terbagi ke dalam empat tingkatan, meliputi: **Tingkat 1.** ***Retrieval*** atau **Pencarian** atau **penelusuran (kembali)**; **Tingkat 2.** ***Comprehension*** atau **pemahaman**; **Tingkat 3.** ***Analysis*** atau **analisis**; dan **Tingkat 4.** ***Knowledge Utilization*** atau **penggunaan** atau **penerapan pengetahuan.**

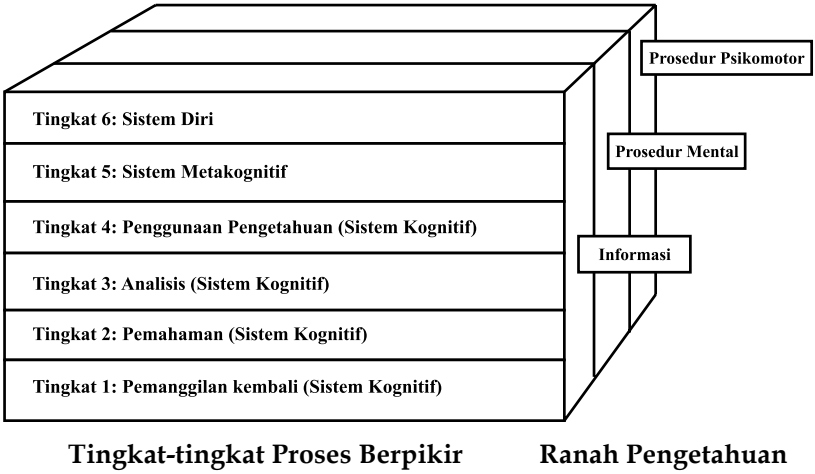
Retrieval atau **pencarian** atau **penelusuran (kembali)** pengetahuan adalah aktivasi atau pengaktivan dan *transfer* atau pengiriman pengetahuan dari ingatan permanen ke *working memory* atau ingatan kerja untuk diproses secara sadar dalam rangka proses pengerjaan suatu tugas. Pencarian pengetahuan dari khazanah pengetahuan dalam ingatan permanen ini bisa berlangsung melalui dua cara, yaitu: (1) *recognition* atau rekognisi atau mengenali kembali, yaitu mencocokkan atau menjodohkan sebuah stimulus (baru) dengan informasi yang sudah tersimpan di ingatan permanen; dan (2) *recall* atau mengingat kembali yaitu memproduksi atau memanggil kembali pengetahuan tertentu yang pernah dipelajari.

Comprehension atau **pemahaman** adalah menerjemahkan pengetahuan ke dalam bentuk yang cocok untuk disimpan di dalam ingatan permanen. Menurut Marzano dan Kendall (2007), aktivitas berpikir ini meliputi dua proses yang saling terkait, yaitu **mengintegrasikan** dan **menyimbolisasikan.** Mengintegrasikan adalah menyaring atau memeras pengetahuan sampai ke ciri-ciri atau unsur-unsur kuncinya serta mengemas atau merumuskannya ke dalam bentuk umum yang ringkas. Sedangkan menyimbolikasikan adalah menerjemahkan pengetahuan yang sudah berhasil diintegrasikan tersebut ke dalam simbol nonbahasa.

Analisis merupakan aktivitas berpikir yang mencakup lima macam proses, yaitu: (1) ***matching*** atau mencocokkan atau menjodohkan, (2) **mengklasifikasikan,** (3) **menganalisis kesalahan**, (4) **menggeneralisasikan**, dan (5) **menspesifikasikan**

(Marzano & Kendall, 2007). Menjodohkan adalah menemukan kesamaan dan perbedaan antara sejumlah komponen pengetahuan. Mengklasifikasikan adalah menata pengetahuan ke dalam kategori-kategori yang bermakna. Menganalisis kesalahan meliputi menguji kelogisan, kemasuk-akalan, atau ketepatan pengetahuan. Sedangkan menggeneralisasikan adalah menciptakan generalisasi baru berdasarkan informasi yang sudah dimiliki.

Penggunaan atau **penerapan pengetahuan** adalah proses berpikir yang digunakan seseorang ketika dia ingin menyelesaikan sebuah tugas spesifik tertentu. Menurut Marzano dan Kendall (2007) aktivitas berpikir ini mencakup empat kategori, yaitu: (1) pengambilan keputusan, yaitu proses memilih salah satu di antara dua atau lebih alternatif; (2) pemecahan masalah, yaitu proses mengatasi hambatan dalam rangka mencapai suatu tujuan; (3) eksperimentasi, yaitu proses merumuskan dan menguji hipotesis dalam rangka memahami fenomena fisik atau psikologis tertentu; dan (4) investigasi, yaitu proses merumuskan dan menguji hipotesis tentang peristiwa-peristiwa di masa lalu, masa kini, atau masa mendatang.



Gambar 4.3. Taksonomi Baru ala Marzano dan Kendall (2007).

Komponen keempat adalah **pengetahuan** atau sering juga disebut **pengetahuan deklaratif.** Menurut Marzano dan Kendall (2007), ranah pengetahuan deklaratif ini mencakup empat jenis pengetahuan yang tersusun secara hirarkis: (1) ***vocabulary terms*** atau kosa kata istilah, yaitu kata atau frase yang dipahami maknanya secara tepat; (2) **fakta,** yaitu informasi tentang orang, tempat, benda, kejadian spesifik tertentu; (3) **generalisasi,** yaitu pernyataan tentang sesuatu yang bisa disertai dengan contoh-contoh; dan (4) **prinsip,** yaitu sejenis generalisasi atau pernyataan tentang hubungan antara dua hal atau lebih. Secara skematis struktur Taksonomi Baru yang dikemukakan oleh Marzano dan Kendall (2007) tersebut dapat disimak pada Gambar 4.3.

Kembali pada tujuan pembahasan kita sebelumnya, kini kita bisa menggolong-golongkan tes berdasarkan isinya, baik berdasarkan isi materi, jenis pengetahuan, maupun jenis proses berpikirnya.

1). Penggolongan tes berdasarkan content atau isi mata pelajarannya. Di sini bisa ditemukan aneka tes mata pelajaran, seperti Tes Matematika, Tes Fisika, Tes Kimia, Tes Geografi, dan berbagai mata pelajaran lain yang masuk dalam kurikulum sekolah mulai jenjang pendidikan dasar sampai jenjang pendidikan tinggi. Semua tes mata pelajaran tersebut lazimnya bisa dibedakan menjadi dua, yaitu: (1) *subject-matter standardized tests* atau tes mata pelajaran yang dibakukan, dan (2) *subject matter teacher made tests* atau tes mata pelajaran yang dibuat sendiri oleh guru. Tes mata pelajaran yang dibakukan lazim disusun oleh lembaga tes atau lembaga pendidikan tertentu dan dibuat dengan memenuhi semua persyaratan perancangan dan ciri psikometrik yang dituntut dalam pengukuran psikologis. Jenis tes semacam ini lazim dimaksudkan untuk mengukur *aptitude* dalam mata pelajaran tertentu serta digunakan untuk berbagai keperluan klasifikasi, seperti seleksi calon siswa/mahasiswa, penempatan siswa/mahasiswa pada jenjang tertentu dalam pembelajaran suatu bidang studi sesuai taraf kemampuan yang dimiliki, dan sertifikasi atau dasar untuk memberikan kesaksian tentang kemampuan seseorang dalam bidang studi tertentu. Sebaliknya, jenis tes bidang

studi yang dibuat sendiri oleh guru lazimnya dimaksudkan untuk mengukur *achievement* sesudah menempuh mata pelajaran tertentu.

2). Penggolongan tes berdasarkan jenis proses berpikirnya. Berpedoman pada salah satu atau gabungan dari berbagai taksonomi tujuan pengajaran ranah kognitif di atas, maka berdasarkan jenis proses berpikir yang diukur, tes psikologis dapat dibedakan menjadi: tes pengetahuan, tes pemahaman, tes kreativitas, dan sebagainya.

3). Penggolongan tes berdasarkan jenis pengetahuannya. Berdasarkan jenis pengetahuan yang diukur dan berpedoman pada salah satu atau gabungan dari aneka taksonomi tujuan pengajaran ranah kognitif di atas, maka tes psikologis dapat digolongkan mjenjadi: tes pengetahuan faktual, tes pengetahuan konseptual, tes pengetahuan prosedural, dan sebagainya.

b. Tes yang Mengukur Disposisi Kepribadian atau Kecenderungan Bertingkah Laku

Dimensi kepribadian yang dipandang menjadi salah satu sumber penggerak dan memberi warna khas cara bertingkah laku seseorang lazim dibedakan menjadi *state, trait,* dan *tipe.* Dari antara ketiganya, yang terpenting dan relevan untuk dibahas di sini adalah *trait.* Guilford (1959, dalam Gregory, 2007) mendefinisikan *trait* sebagai "any relatively enduring way in which one individual differs from another" (h. 351). Artinya, *trait* atau sifat adalah setiap ciri yang secara relatif tetap membedakan seseorang dari setiap orang lain. Dalam pengukuran kepribadian, ciri pembeda seseorang dari setiap orang lain yang menjadi objek perhatiannya adalah disposisi kepribadian atau kecenderungan bertingkah laku secara khas dan yang didominasi oleh fungsi afektif. Nunnally, Jr. (1970) menggolongkan disposisi kepribadian atau kecenderungan bertingkah laku ini ke dalam empat kategori, yaitu: (1) *social traits* atau sifat sosial, (2) *motives* atau motif, atau *needs* atau kebutuhan, atau *drives* atau dorongan, (3) *personal conceptions* atau konsepsi tentang diri, dan (4) *adjustment versus*

maladjustment atau penyesuaian-diri yang baik versus penyesuaian-diri yang salah.

1). Tes yang mengukur *social traits*. *Social traits* atau sifat sosial adalah cara khas seseorang bertingkah laku dalam situasi sosial atau yang melibatkan kehadiran orang lain. Sifat yang dimaksud di sini mencakup apa yang oleh Gordon W. Allport (1961, dalam Hall & Lindzey, 1993) disebut *disposisi sentral,* yaitu aneka kecenderungan khas yang sering muncul atau tampak dalam tingkah laku seseorang. Ada yang menyatakan, dalam setiap komunitas budaya jumlah ragam sifat ini sama banyaknya dengan jumlah kata sifat yang terdapat dalam kamus bahasa yang dimiliki oleh komunitas budaya yang bersangkutan yang lazimnya juga merupakan komunitas penutur bahasa ibu tertentu. Beberapa contoh sifat sosial yang sudah diteliti dan dikembangkan alat ukurnya meliputi antara lain ekstraversi-introversi, *internal vs external locus of control,* individualisme-kolektivisme, sifat asertif, sifat Machiavelianis, kepribadian tipe A dan tipe B.

2). Tes yang mengukur *motives* atau *needs* atau *drives*. *Motives, needs,* atau d*rives* yang dalam bahasa Indonesia disebut dorongan atau kebutuhan adalah kekuatan dari dalam yang menggerakkan sekaligus mengarahkan tingkah laku kita. Dorongan penggerak tingkah laku ini lazimnya dibedakan menjadi dua kategori besar, yaitu yang bersumber dari disposisi biologis dan yang bersumber dari sumber non-biologis khususnya sebagai hasil dari interaksi dengan orang lain. Dalam pengukuran *typical performance* yang lazim dijadikan fokus adalah jenis-jenis dorongan atau kebutuhan non-biologis atau sosial, yaitu jenis-jenis dorongan atau kebutuhan yang terbentuk akibat proses belajar dalam lingkungan sosial. Beberapa contoh dorongan atau kebutuhan sosial yang sudah cukup banyak diteliti dan dikembangkan alat ukurnya meliputi antara lain kebutuhan untuk berprestasi, kebutuhan untuk berafiliasi atau menjalin relasi dengan orang lain, kebutuhan untuk berkuasa, dorongan untuk melakukan agresi atau menyerang.

3). Tes yang mengukur *personal conceptions*. *Personal conceptions* atau konsepsi pribadi adalah cara orang berpiki atau memandang dirinya dan cara orang memandang dunia atau pandangan hidupnya. Beberapa contoh konsepsi pribadi yang sudah diteliti dan dikembangkan alat ukurnya mencakup antara lain konsep-diri, harga-diri, nilai, dan sikap terhadap aneka objek.

4). Tes yang mengukur *adjustment versus maladjustment*. *Adjustment versus maladjustment* atau penyesuaian-diri yang baik versus penyesuaian-diri yang bermasalah adalah taraf sejauh mana seseorang terbebas dari berbagai jenis tekanan emosi dan/atau terbelit oleh kebiasaan bertingkah laku yang menimbulkan gangguan terhadap kesejahteraan pribadinya maupun terhadap relasinya dengan orang lain sehingga pada akhirnya juga mengganggu kesejahteraan pribadinya. Penyesuaian diri yang bermasalah berkisar antara kecenderungan disruptif atau merusak atau mengganggu mulai dari yang relatif ringan seperti neurosis sampai yang relatif berat seperti psikosis. Lawannya adalah penyesuaian diri yang baik yang ditandai antara lain oleh perasaan sejahtera (*subjective well-being*) dan kefektivan dalam menjalankan aneka tugas kehidupan sehari-hari (pribadi yang efektif). Contoh bentuk-bentuk penyesuaian diri yang baik dan yang bermasalah yang sudah diteliti dan dikembangkan alat ukurnya meliputi antara lain *subjective well-being*, depresi, kecemasan.

c. Tes yang Mengukur Ketrampilan dan Pola Tingkah Laku

Dalam pengukuran psikologis, secara garis besar ketrampilan dapat digolongkan ke dalam dua kategori, yaitu ketrampilan yang bersifat spesifik dan ketrampilan berupa pola tingkah laku yang kompleks. Untuk mengidentifikasi jenis-jenis ketrampilan yang spesifik kita bisa menggunakan taksonomi tujuan pengajaran ranah psikomotor yang pernah disusun oleh Simpson (1972, dalam Anderson et al., eds., 2001). Jenis-jenis ketrampilan ini memiliki cakupan yang spesifik, seperti misalnya ketrampilan mengetik

pada komputer dengan dua belas jari, ketrampilan mengoperasikan aplikasi komputer tertentu, atau bahkan lebih spesifik lagi. Contoh tes psikologis yang mengukur jenis ketrampilan spesifik meliputi antara lain *Finger Dexterity Test* yang mengukur ketangkasan jari-jari tangan.

Tabel 4.3.
Taksonomi tujuan pengajaran ranah psikomotor (Simpson, 1972).

Taraf Kemampuan	Uraian
Mampu menghasilkan gerakan baru *(Origination)*	*Mampu menciptakan pola gerakan baru.* Mampu menciptakan aneka pola gerakan baru sesuai tuntutan suatu situasi atau problem khusus tertentu. Hasil belajar yang ditekankan berupa kreativitas yang dilandasi aneka ketrampilan taraf tinggi.
Mampu beradaptasi *(Adaptation)*	*Mampu memodifikasi aneka ketrampilan motor disesuaikan dengan tuntutan situasi baru.* Aneka ketrampilan sudah dikuasai dengan baik, sehingga siswa mampu memodifikasikan pola gerakan agar sesuai dengan tuntutan situasi khusus tertentu.
Mampu melakukan respon kompleks secara lancar *(Complex Overt Response)*	*Menunjukkan tahap agak lanjut menguasai suatu ketrampilan kompleks.* Mampu melakukan tindakan motor secara trampil yang melibatkan pola gerakan yang kompleks. Ketrampilan atau ketangkasan ditunjukkan oleh gerakan yang cepat, akurat, dan sangat terkoordinasi, yang dilakukan dengan energy atau upaya minimum. Kategori ini mencakup mengerjakan tugas tanpa ragu-ragu dan melakukan gerakan secara otomatis.
Mampu melakukan respon secara mekanik *(Mechanism)*	*Mampu melakukan suatu ketrampilan motor yang kompleks.* Merupakan tahap piawai dalam mempelajari suatu ketrampilan kompleks. Hasil belajar sudah menyatu dengan kebiasaan, sehingga gerakan-gerakan bisa dilakukan dengan percaya diri dan lancar.
Mampu melakukan respon tertentu dengan bimbingan guru *(Guided Response)*	*Menunjukkan tahap awal menguasai suatu ketrampilan kompleks, meliputi kemampuan mengikuti contoh atau mencontoh.* Merupakan tahap awal dalam mempelajari suatu ketrampilan kompleks, mencakup kemampuan mencontoh atau coba-salah. Ketrampilan yang memadai akan dicapai lewat latihan.
Memiliki kesiapan untuk bertindak *(Set)*	*Menunjukkan kesiapan untuk bertindak.* Kesiapan untuk bertindak, meliputi kesiapan mental, fisik, dan emosi. Ketiganya merupakan disposisi yang mendasari respon seseorang terhadap berbagai situasi yang dihadapi (kadang-kadang juga disebut *mindset*).

Taraf Kemampuan	Uraian
Mampu mempersepsikan *(Perception)*	*Mampu menggunakan tanda-tanda sensoris untuk membimbing aktivitas fisik tertentu.* Mampu menggunakan petunjuk-isyarat sensoris untuk membimbing aktivitas motor, meliputi kepekaan menangkap stimulasi sensoris kemampuan memilih petunjuk-isyarat sensoris, dan kemampuan menerjemahkannya ke dalam tindakan.

Lain halnya dengan ketrampilan yang mencakup pola perilaku yang kompleks. Jenis ketrampilan ini lazimnya berupa gugusan atau rangkaian kompetensi kompleks yang tercakup dalam sebuah *job* atau pekerjaan atau *role* atau peran tertentu. Sebagai contoh, pekerjaan sebagai guru kiranya akan mencakup rangkaian kompetensi seperti menyusun perencanaan pembelajaran, menyusun evaluasi pembelajaran, melaksanakan pembelajaran, kemampuan bertanya, kemampuan memberikan *feedback*, kemampuan menampilkan diri sebagai model karakter yang baik, dan sebagainya. Contoh lain, peran sebagai ketua kelas akan mencakup rangkaian kompetensi seperti memiliki disiplin diri, memiliki sifat pemberani, mampu mengungkapkan pikiran secara jelas, memiliki rasa tanggung jawab, mampu menampilkan diri sebagai teladan, mampu mendengarkan ungkapan pendapat atau perasaan teman, dan sebagainya. Maka, untuk mengidentifikasi jenis-jenis hasil belajar dengan ranah isi pola perilaku yang kompleks terkait pekerjaan atau peran tertentu, kita perlu menemukan sumber entah berupa sumber tertulis atau pendapat narasumber yang berwenang untuk kita jadikan pedoman agar bisa melakukan analisis tugas secara cermat.

Analisis pekerjaan (*job analysis*) atau analisis tugas (*task analysis*) merupakan proses untuk mendapatkan informasi tentang seluk-beluk suatu pekerjaan atau tugas (McCormick & Ilgen, 1980). Seperti dinyatakan oleh McCormick dan Ilgen (1980), ada beberapa jenis informasi yang lazim dikumpulkan tentang suatu pekerjaan, yaitu: (1) aneka aktivitas yang merujuk pada pekerjaan atau tugas yang harus dilaksanakan, misal mengembangkan kurikulum, mengembangkan materi pembelajaran; (2) aneka aktivitas yang merujuk pada pekerja

yang melaksanakan tugas, misal menyelenggarakan pembelajaran yang mendidik, menyelenggarakan penilaian dan evaluasi proses dan hasil belajar; (3) mesin, perkakas, alat bantu, dan sebagainya yang khas untuk melaksanakan pekerjaan atau tugas, misal *laptop* dan *LCD,* kapur atau spidol, penggaris, peta, dan sebagainya; (4) aneka jenis material yang diproses; (5) jenis pengetahuan yang diperlukan; (6) aneka kondisi kerja; dan (7) aneka persyaratan yang dituntut dari orang yang akan melaksanakan pekerjaan atau tugas yang bersangkutan. Contoh tes yang mengukur pola tingkah laku kompleks dalam rangka melaksanakan pekerjaan sebagai guru adalah Tes Kompetensi Guru.

d. Tes yang Mengukur Aneka Fungsi Psikologis Lain

Ke dalam kategori ini bisa disebutkan tes yang mengukur aneka fungsi psikologis lain yang didominasi oleh salah satu atau gabungan lebih dari satu fungsi psikis (kognitif, afektif, psikomotor) maupun fungsi neuropsikologis, seperti: (1) tes perhatian; (2) tes fungsi sensorimotor, seperti kemampuan visual, auditori; (3) tes persepsi; (4) tes ingatan; (5) tes berpikir abstrak; (6) tes bahasa; dan (7) tes perkembangan dan kematangan karir (AERA, APA, & NCME, 1999).

D. Penggunaan Tes Psikologis

Secara umum, hasil tes psikologis lazim digunakan sebagai dasar pembuatan aneka keputusan terkait orang dalam berbagai sektor kehidupan. Hasil ujian sekolah digunakan untuk memutuskan kelulusan seorang murid pada akhir program pendidikan. Hasil tes seleksi digunakan untuk memutuskan apakah seorang calon pegawai diterima bekerja di sebuah instansi atau perusahaan. Hasil tes kepribadian tertentu digunakan untuk memutuskan dalam arti menentukan jenis gangguan kepribadian yang dialami oleh seorang

pasien, sekaligus menentukan jenis intervensi yang perlu diberikan. Dan masih banyak contoh lain bisa ditemukan. Mengikuti klasifikasi dalam *Standards 1999* (AERA, APA, & NCME, 1999), secara lebih spesifik penggolongan penggunaan tes adalah sebagai berikut: (1) penggunaan tes di lingkungan klinis (*psychological testing*); (2) penggunaan tes di lingkungan pendidikan sekolah (*educational testing*); (3) penggunaan tes untuk pembinaan pegawai dan *credentialing* atau pemberian pengakuan; dan (4) penggunaan tes dalam evaluasi program dan kebijakan publik.

1. Penggunaan Tes di Lingkungan Klinis (*Psychological Testing*)

Penggunaan tes di lingkungan klinis atau *psychological testing* meliputi beberapa bidang kegiatan, yaitu:

a. Diagnosis. Diagnosis adalah proses meliputi pengumpulan dan pengintegrasian hasil-hasil tes dengan informasi lain yang diperoleh sebelumnya maupun kini tentang seseorang berikut aneka keadaan kontekstualnya yang relevan dalam rangka menemukan tanda-tanda apakah yang bersangkutan dalam keadaan sehat secara psikologis atau sebaliknya mengalami gangguan (AERA, APA, & NCME, 1999; h. 126). Tes sangat membantu dalam membuat diagnosis psikologis terhadap seseorang.

b. Perencanaan Intervensi dan Evaluasi Hasilnya. Hasil tes juga sangat membantu dalam merencanakan, melaksanakan, dan mengevaluasi intervensi psikologis. Intervensi semacam ini bisa ditujukan untuk mencegah terjadinya satu atau lebih simptom gangguan psikologis, menstabilkan atau bahkan mengatasi simptom, mengurangi dampak merusaknya, serta memenuhi berbagai kebutuhan fisik, psikologis, dan sosial seseorang (AERA, APA, & NCME, 1999; h. 128).

c. Pengambilan Keputusan Hukum dan Kebijakan Pemerintah. Hasil tes juga bisa sangat membantu dalam pengambilan keputusan di bidang hukum-pengadilan maupun urusan

pemerintahan lainnya. Hasil tes bisa digunakan untuk memberikan informasi penting kepada pihak ketiga, penasehat hukum klien di pengadilan, penasehat hukum pihak lain yang berperkara dengan klien, hakim, atau panitera tentang kondisi psikologis klien yang terkait dengan perkara hukum yang sedang dihadapinya. Hasil tes juga bisa digunakan sebagai dasar untuk memutuskan penerimaan dan penempatan pegawai yang menyandang kebutuhan khusus tertentu atau membuat berbagai jenis keputusan administratif, seperti pembatalan lisensi, pemberian kompensasi, dan sebagainya (AERA, APA, & NCME, 1999; h. 128).

d. Pemahaman Diri, Pertumbuhan, dan Pengambilan Keputusan Pribadi. Tes juga lazim digunakan untuk memberikan informasi dalam rangka membantu seseorang lebih memahami dirinya, menemukan aneka kekuatan dan kekurangannya, serta memperjelas dalam arti menjadi paham tentang berbagai hal terkait diri mereka yang penting agar mampu membuat aneka keputusan yang tepat dan mengalami perkembangan pribadi.

2. Penggunaan Tes di Lingkungan Pendidikan Sekolah (*Educational Testing*)

Sudah barang tentu, tes banyak digunakan di lingkungan pendidikan sekolah mulai dari jenjang taman kanak-kanak atau prasekolah sampai jenjang pendidikan tinggi. Secara garis besar, penggunaan tes di lingkungan pendidikan sekolah dapat digolongkan ke dalam dua kategori, seperti diuraikan berikut ini.

a. Penggunaan Tes untuk Menilai Kinerja Individual. Informasi tentang kinerja individual berdasarkan hasil tes lazim digunakan untuk berbagai tujuan berikut ini: (1) mengevaluasi prestasi dan kemajuan belajar masing-masing murid terkait ranah isi atau mata pelajaran tertentu; (2) mendiagnosis kekuatan dan kelemahan masing-masing murid dalam setiap maupun antar mata pelajaran; (3) merancang aneka bentuk intervensi dan menyusun

rencana pengajaran yang disesuaikan dengan kebutuhan masing-masing murid; (4) menempatkan murid ke dalam program pendidikan yang sesuai; (5) menyeleksi calon untuk diterima mengikuti program dengan daya tampung yang terbatas; dan (6) memberikan sertifikasi terhadap pencapaian prestasi atau perolehan aneka kualifikasi tertentu.

b. Penggunaan Tes untuk Menilai Kinerja Kelompok. Informasi tentang status atau keadaan, kemajuan, dan pencapaian kelompok satuan kerja pendidikan seperti sekolah, dinas pendidikan kota/kabupaten, atau dinas pendidikan propinsi dapat digunakan untuk tujuan seperti (1) menilai dan memonitor kualitas aneka program pendidikan yang ditujukan bagi seluruh murid atau bagi kelompok-kelompok murid tertentu, dan (2) menentukan keberhasilan aneka kebijakan dan intervensi yang telah dipilih untuk dievaluasi (AERA, APA, & NCME, 1999; h. 137).

3. Penggunaan Tes untuk Pembinaan Pegawai dan *Credentialing* atau Pemberian Pengakuan

Di lingkungan dunia pekerjaan, hasil tes psikologis lazim digunakan untuk dua wilayah urusan, yaitu pembinaan pegawai dan pemberian pengakuan atas terpenuhinya kualifikasi seseorang untuk terjun dalam profesi tertentu.

a. Pembinaan Pegawai lazimnya meliputi tiga jenis kegiatan: (1) **seleksi,** yaitu pengambilan keputusan tentang calon-calon pegawai mana yang akan diterima bekerja di lingkungan sebuah organisasi atau perusahaan; (2) ***placement*** atau **penempatan,** yaitu pengambilan keputusan tentang penugasan para pegawai yang sudah diterima ke berbagai pos dalam sebuah organisasi atau perusahaan; dan (3) **promosi,** yaitu pengambilan keputusan tentang para pegawai yang akan ditempatkan dalam pos atau jabatan yang lebih tinggi. Semua kegiatan itu didasarkan pada prediksi dalam arti penilaian tentang kemampuan kerja para pegawai di masa mendatang demi

mengoptimalkan kinerja organisasi atau perusahaan dalam bentuk peningkatan efisiensi, pertumbuhan, produktivitas, serta motivasi dan kepuasan pegawai (AERA, APA, & NCME, 1999).

b. *Credentialing* atau Pemberian Pengakuan. Pemberian pengakuan baik dalam bentuk sertifikasi maupun lisensi adalah bentuk pengesahan bahwa seseorang telah memiliki ketrampilan atau pengetahuan spesifik tertentu yang dipersyaratkan untuk melaksanakan tugas-pekerjaan dalam bidang profesi tertentu. Hasil tes digunakan untuk menentukan apakah pengetahuan dan ketrampilan esensial dalam ranah spesifik tertentu benar-benar sudah dikuasai oleh seorang peserta uji sertifikasi atau lisensi (AERA, APA, & NCME, 1999; h. 156).

4. Penggunaan Tes dalam Evaluasi Program dan Kebijakan Publik

Tes juga banyak digunakan dalam kegiatan evaluasi program dan dalam pengambilan keputusan terkait kebijakan publik. **Evaluasi program** merupakan serangkaian prosedur yang digunakan dalam rangka menilai atau menaksir kebutuhan klien akan program tertentu, cara mengimplementasikan program itu, efektivitas program itu, dan nilai atau manfaatnya bagi klien. **Studi kebijakan** memiliki cakupan yang lebih luas dibandingkan evaluasi program, berupa kajian-kajian yang hasilnya digunakan untuk mengevaluasi aneka rencana, prinsip, atau prosedur yang dilaksanakan untuk mencapai tujuan-tujuan publik yang bersifat luas. Dalam praktek, sering terjadi tumpang tindih antara evaluasi program dan studi kebijakan. Hasil tes digunakan untuk sebagai dasar pertimbangan untuk memulai atau melaksanakan, melanjutkan, memodifikasi, mengakhiri, atau mengembangkan aneka program dan kebijakan (AERA, APA, & NCME, 1999; h. 163). Ψ

Bab 5
Syarat Tes yang Baik

Tes yang baik harus memenuhi sejumlah syarat baik dari segi desain atau rancangan maupun dari segi psikometriknya (Friedenberg, 1995). Pemenuhan syarat pada kedua segi tersebut harus diusahakan pada dua tahap dalam proses konstruksi atau penyusunan tes, yaitu *tahap konseptual-teoretis* dan *tahap empiris-statistis*.

Tahap pertama proses konstruksi atau penyusunan tes adalah **tahap konseptual-teoretis** terkait segi desain atau rancangan. Pada tahap ini penyusun tes harus merumuskan dengan jelas (1) tujuan tes, meliputi apa yang hendak diukur, siapa yang akan dikenai tes, dan bagaimana skor tes akan digunakan; (2) ranah isi yang hendak diukur; (3) prosedur administrasinya; dan (4) prosedur penskorannya. Memadai tidaknya rumusan konseptual-teoretis keempat syarat terkait desain atau rancangan tes ini harus dievaluasi melalui *rational judgement* atau penilaian rasional baik oleh *expert* atau ahli (*expert rational judgment)* maupun oleh awam lazimnya berupa kelompok subjek yang akan dikenai tes (*lay rational judgment*). Sesudah syarat-syarat terkait desain atau rancangan tersebut dipandang terpenuhi secara memadai, maka proses konstruksi atau penyusunan tes dilanjutkan ke tahap kedua, yaitu tahap empiris-statistis untuk memeriksa pemenuhan syarat-syarat psikometriknya.

Tahap kedua proses konstruksi atau penyusunan tes, yaitu **tahap empiris-statistis,** terdiri dari dua gugus aktivitas utama yaitu proses ***try out*** atau **uji coba** tes dan pemeriksaan pemenuhan syarat-syarat psikometrik tes meliputi ***item analysis*** atau **analisis butir** untuk memeriksa ciri-ciri psikometrik item-item tes secara individual atau sendiri-sendiri dan dilanjutkan dengan pemeriksaan pemenuhan syarat-syarat psikometrik item-item secara keseluruhan sebagai kesatuan tes.

Uji coba tes lazim dilaksanakan pada ***standardization sample*** atau **sampel standarisasi,** yaitu kelompok subjek yang memiliki ciri-ciri esensial sama seperti populasi subjek yang akan dikenai tes. Hal-hal penting yang dilakukan dalam uji coba meliputi (1) pemeriksaan efektivitas prosedur administrasi tes; (2) pemeriksaan efektivitas prosedur pernskoran tes; dan (3) pemeriksaan aspek ***face validity*** atau **validitas tampang** tes dengan cara meminta pendapat para subjek apakah tes tersebut memberi kesan mengukur atribut psikologis seperti yang dimaksud. Berbagai kekurangan terkait efektivitas prosedur administrasi, prosedur penskoran, maupun pendapat subjek tentang validitas tampang tes dicatat, agar bisa dilakukan perbaikan untuk diperiksa kembali pada uji coba berikutnya. Data hasil tes yang diperoleh dalam uji coba dipakai sebagai bahan untuk melakukan analisis butir yaitu pemeriksaan pemenuhan syarat-syarat psikometrik item-item tes secara individual, serta pemeriksaan pemenuhan syarat-syarat psikometrik keseluruhan item sebagai kesatuan tes khususnya terkait reliabilitas, validitas, dan daya diskriminasinya (Kline, 1986). Uraian lebih rinci tentang masing-masing pokok di atas akan disajikan pada bagian berikut ini.

A. Segi Desain atau Rancangan Tes

Sebagaimana sudah disinggung, pada segi desain atau rancangannya ada empat syarat yang harus dipenuhi oleh sebuah tes yang baik, yaitu: (1) memiliki tujuan yang dirumuskan secara jelas; (2) memiliki ranah isi yang dirumuskan secara spesifik dan baku; (3) memiliki prosedur administrasi yang baku; dan (4) memiliki prosedur penskoran yang baku. Marilah kita tinjau masing-masing ciri tersebut satu demi satu.

1. Tujuan yang Jelas

Tujuan tes mencakup tiga hal: (a) *atribut psikologis* yang hendak diukur, (b) *populasi subjek* yang akan dikenai tes, dan (c) *jenis skor* dalam arti cara skor hasil tes akan digunakan (Friedenberg, 1995).

a. Atribut Psikologis yang Hendak Diukur. Hal ini terkait dengan *domain* atau ranah yaitu dimensi kepribadian atau wilayah perilaku yang menjadi fokus atau sasaran tes. Sebagaimana sudah diuraikan dalam Bab 4, ranah atau wilayah perilaku bisa dibedakan ke dalam tiga kategori, yaitu *ranah kognitif* terkait dengan kemampuan olah pikir atau olah cipta, *ranah afektif* terkait dengan kemampuan olah rasa dan olah karsa, serta *ranah psikomotor* terkait dengan kemampuan olah gerak. Sebagaimana juga sudah disinggung, berdasarkan ranah atau wilayah perilaku yang diukur tes dibedakan ke dalam dua golongan besar, yaitu (1) *maximal performance tests* atau tes kinerja maksimal, dan (2) *typical performance tests* atau tes kinerja tipikal atau khas. Golongan tes yang pertama cocok untuk mengukur jenis-jenis kemampuan dalam ranah kognitif dan psikomotor, sedangkan golongan tes yang kedua cocok untuk mengukur jenis-jenis kemampuan dalam ranah afektif. Maka sebagai langkah pertama dalam merancang sebuah tes penting sekali merumuskan secara jelas ranah dalam arti dimensi kepribadian atau wilayah perilaku tempat atribut atau kemampuan psikologis yang hendak diukur berada.

b. Populasi Subjek yang Akan Dikenai Tes. Hal ini terkait dengan *audience* atau khalayak, yaitu populasi subjek yang akan menjadi sasaran pengetesan dalam arti dikenai tes. Sebagaimana kita tahu, sebuah tes hanya akan memberikan hasil yang valid bagi populasi subjek tertentu sebagaimana direncanakan sejak awal konstruksi atau penyusunannya. Tes inteligensi untuk populasi subjek dewasa seperti tes *WAIS* misalnya, tidak akan memberikan hasil yang valid jika diadministrasikan atau dikenakan pada sampel subjek anak-anak. Sebaliknya, tes inteligensi yang khusus diperuntukkan bagi populasi subjek kanak-kanak seperti tes *WISC* misalnya, tidak akan memberikan hasil yang valid jika dikenakan pada sampel

subjek dewasa. Maka, khalayak untuk siapa sebuah tes ditujukan harus dirumuskan secara jelas sejak awal. Alasan lain, perbedaan khalayak juga bisa berdampak langsung terhadap jenis tugas yang harus digunakan sebagai item-item tes karena terkait perbedaan kemampuan kelompok subjek. Tes inteligensi untuk populasi kanak-kanak misalnya, harus lebih mengandalkan jenis-jenis tugas yang bersifat nonverbal sebab kemampuan berbahasa populasi kanak-kanak secara umum masih terbatas.

c. Jenis Skor. Sebagaimana sudah kita lihat dalam Bab 4, skor hasil tes dapat digunakan dengan berbagai cara. Dalam *norm-referenced scoring* atau *penskoran beracuan norma*, skor tes digunakan untuk membandingkan kinerja relatif seorang testi dengan kelompok sebayanya. Dalam *criterion-referenced scoring* atau *penskoran beracuan kriteria*, skor tes digunakan untuk membandingkan kinerja seorang testi dengan kriteria absolut untuk menunjukkan taraf penguasaan atau taraf pemilikan atas atribut psikologis tertentu secara absolut. Dalam *penskoran normatif*, skor tes digunakan untuk menunjukkan taraf penguasaan atau taraf pemilikan seorang testi atas suatu atribut psikologis dalam tes yang dimaksudkan untuk mengukur sebuah atribut psikologis tunggal tertentu. Dalam *penskoran ipsatif*, skor tes digunakan untuk menunjukkan taraf penguasaan atau taraf pemilikan relatif seorang testi atas suatu atribut psikologis dibandingkan dengan taraf penguasaan atau taraf pemilikannya atas satu atau lebih atribut psikologis lain dalam tes yang dimaksudkan untuk mengukur lebih dari satu atribut psikologis secara serentak atau bersamaan. Jenis penggunaan skor yang direncanakan seringkali berdampak menentukan format item, taraf kesukaran item, maupun struktur tes secara keseluruhan. Maka, jenis skor yang hendak dipakai harus dirumuskan secara jelas sejak awal dalam konstruksi atau pengembangan tes.

2. Ranah Isi yang Jelas dan Baku

Seperti sudah disinggung, *content* atau isi tes psikologis menunjuk pada gugusan tingkah laku, pengetahuan, ketrampilan, abilitas, sikap, atau karakteristik lain yang hendak diukur oleh tes. Isi yang dimaksud lazimnya dijabarkan dalam sebuah tabel spesifikasi yang rinci serta dipilah ke dalam kategori-kategori antara lain untuk memudahkan pengelompokan item-itemnya (AERA, APA, & NCME, 1999). Seperti sudah disinggung, karena sebagian besar objek atau sasaran tes psikologis berupa konstruk teoretis yang tidak memiliki batas ranah isi yang jelas, maka perumusan isi dalam konstruksi atau pengembangan tes sering harus didahului dengan *eksplikasi konstruk,* yaitu identifikasi bentuk-bentuk tingkah laku, keyakinan, dan sikap spesifik yang menunjukkan maupun sebaliknya yang menyangkal keberadaan konstruk yang dimaksud. Hasil eksplikasi konstruk ini selanjutnya dipakai untuk merumuskan isi tes (Friedenberg, 1995). Selain jelas, isi tes juga harus baku dalam arti semua subjek atau testi harus dikenai item-item yang mewakili cakupan isi yang sama (Friedenberg, 1995).

3. Prosedur Administrasi Baku

Prosedur administrasi tes lazim dituangkan dalam *manual* atau buku instruksi tes, berupa aneka petunjuk bagi testi dan aneka pedoman tentang kondisi pelaksanaan tes. Petunjuk bagi testi antara lain mencakup cara testi memberikan respon, jenis bantuan yang boleh diberikan kepada testi jika mereka tidak memahami pertanyaan atau tugas, cara testi mengubah atau membetulkan respon yang keliru, dan waktu pengerjaan tes. Kondisi pelaksanaan tes mencakup antara lain pedoman tentang pengaturan penerangan dan keheningan ruang tempat tes berlangsung. Prosedur administrasi tes disebut *standardized* atau baku jika petunjuk dan kondisi pelaksanaan tes bagi semua testi sungguh-sungguh mengikuti prosedur rinci yang sudah diuraikan oleh pengembang tes dalam buku instruksi tes (AERA, APA, & NCME, 1999).

4. Prosedur Penskoran Baku

Manual atau buku instruksi tes juga memuat uraian yang rinci dan jelas tentang cara menskor dan melaporkan hasil tes. Pembakuan prosedur penskoran ini diperlukan untuk menjamin ketepatan penskoran dan pelaporan hasil tes. Penskoran tes bisa dilaksanakan dengan mesin, atau dilakukan secara manual oleh petugas. Jika dilakukan oleh mesin, maka ketepatan kerja mesin harus terjamin. Jika dilakukan oleh petugas, maka petugas harus benar-benar terlatih (AERA, APA, & NCME, 1999).

Skor tes dalam bentuk *raw score* atau skor kasar belum bisa ditafsirkan. Untuk menafsirkannya perlu informasi lain berupa norma atau standar, indikasi tentang kemungkinan kesalahan pengukuran, serta deskripsi tentang isi tes. (AERA, APA, & NCME, 1999). Untuk menjamin ketepatan penafsiran semua jenis informasi tersebut juga perlu dibakukan dan diuraikan dengan jelas oleh pengembang tes dalam *manual* atau buku petunjuk tes. Cara pelaporan hasil tes, meliputi skor dan penafsirannya, kepada testi maupun pihak lain juga perlu dibakukan. Pelaporan hasil tes semacam ini sering perlu dilengkapi dengan keterangan tentang keterbatasannya serta kemungkinan kaitannya dengan informasi lain. Pada sejumlah tes tertentu, laporan hasil tes tidak perlu mencakup skor yang dicapai testi, melainkan cukup berupa rumusan penafsiran secara garis besar atau berupa klasifikasi dikotomis seperti "lulus" atau "gagal" (AERA, APA, & NCME, 1999).

B. Segi Psikometrik Tes

Uraian tentang uji coba tes akan dibahas dalam bagian tersendiri di bab lain. Di bagian ini akan langsung dibahas beberapa segi psikometrik tes yang esensial. Yang dimaksud segi psikometrik tes adalah kualitas kinerja tes sebagai alat yang dimaksudkan untuk mengukur atribut psikologis tertentu, khususnya yang harus diperiksa melalui analisis terhadap respon testi terhadap item-item

tes baik secara individual maupun sebagai kesatuan tes dengan menggunakan teknik statistik. Ada empat aspek esensial kualitas yang menentukan kinerja tes sebagai alat ukur, yaitu (1) validitas, (2) reliabilitas, (3) statistik item tes, dan (4) daya diskriminasi tes.

1. Validitas

Validitas adalah kualitas esensial yang menunjukkan sejauh mana suatu tes sungguh-sungguh mengukur atribut psikologis yang hendak diukurnya. Hingga kini pengertian yang paling lazim tentang validitas adalah sebagai berikut. *Pertama*, validitas dipandang sebagai kualitas atau ciri yang melekat pada tes. *Kedua*, validitas bisa dibedakan ke dalam tiga tipe atau jenis, yaitu *content validity* atau validitas isi, *criterion-related validity* atau *criterion-oriented validity* atau validitas terkait dengan kriteria atau validitas yang berorientasi pada kriteria, serta *construct validity* atau validitas konstruk. Pemahaman tentang validitas seperti ini dikenal sebagai ***"trinity" view of validity*** atau ***tripartite view of validity*** atau "pandangan trinitas tentang validitas" atau "pandangan tripartit tentang validitas" yang awalnya dilontarkan oleh dua pakar pengukuran – Lee Cronbach dan Paul Meehl – pada 1955 (Goodwin & Leech, 2003).

Pengertian yang lebih mutakhir tentang validitas diuraikan dalam dokumen *Standards for educational and psychological tests and manuals* yang diterbitkan pada 1999 oleh tiga organisasi profesi yang terkait dengan tes yaitu *American Psychological Association* (APA), *American Educational Research Association* (AERA), dan *National Council on Measurement in Education* (NCME). Dokumen yang selanjutnya dikenal sebagai *1999 Standards* (AERA, APA, & NCME, 1999) ini memberikan rumusan sebagai berikut tentang validitas.

Pertama, validitas bukan ciri atau kualitas yang melekat pada tes melainkan kualitas *konsekuensi sosial* yang ditimbulkan oleh penafsiran hasil tes sesuai tujuan penggunaan tes. Maksudnya, problem utama validitas tes berkisar pada seberapa baik sebuah tes mampu menjalankan tugasnya yang langsung terkait dengan nasib

seseorang, maka disebut sebagai konsekuensi sosial. Ada dua macam konsekuensi (sosial) penggunaan tes, yaitu (a) konsekuensi *deskriptif,* berupa inferensi atau penyimpulan yang menghasilkan pernyataan tertentu tentang testi berdasarkan skor-skor tes, dan (b) konsekuensi *preskriptif,* berupa perumusan tentang keputusan tertentu terkait masa depan testi berdasarkan pernyataan deskriptif yang telah diperoleh (Goodwin & Leech, 2003). Sebagai contoh, berdasarkan skor-skornya pada tes *WISC* seorang anak disimpulkan memiliki taraf kecerdasan *feeble-minded* (konsekuensi deskriptif). Inferensi atau penyimpulan ini sesuai dengan tujuan penggunaan *WISC,* yaitu mengukur taraf inteligensi kelompok subjek anak. Berdasarkan konsekuensi deskriptif tersebut maka anak itu direkomendasikan untuk mendapatkan perhatian dan pendampingan sebagai anak berkebutuhan khusus (konsekuensi preskriptif). Validasi atau pemeriksaan validitas bertugas mengumpulkan evidensi atau bukti-bukti sejauh mana kedua jenis konsekuensi tersebut sungguh-sungguh bisa dibenarkan.

Dengan kata lain, validitas menurut pengertian mutakhir adalah **taraf sejauh mana evidensi atau bukti-bukti empiris maupun teoretis mendukung dalam arti membenarkan cara menafsirkan skor tes sesuai tujuan penggunaan tes.** Dalam validasi atau pemeriksaan validitas, yang dievaluasi adalah **kualitas penafsiran skor tes sesuai tujuan penggunaan tes,** bukan tesnya sendiri. Mengutip kata-kata dalam *1999 Standards,*

> "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests...The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself." (AERA, APA, & NCME, 1999, h. 9).

Kedua, pengertian mutakhir tentang validitas memandang validitas sebagai konsep tunggal. Aneka evidensi yang digunakan untuk mengevaluasi kualitas penafsiran skor tes sesuai tujuan penggunaan tes memang bisa menunjukkan **aspek-aspek validitas**,

namun tidak mewakili **jenis-jenis validitas** yang berbeda. Sekali lagi, validitas merupakan sebuah konsep tunggal, yaitu taraf sejauh mana seluruh evidensi yang berhasil dikumpulkan mendukung interpretasi skor tes sesuai yang dimaksudkan. Meminjam kata-kata dalam *1999 Standards*:

> "...the various sources of evidence that might be used in evaluating a proposed interpretation of test scores for particular purposes...may illuminate different aspects of validity, but they do not represent distinct types of validity. Validity is unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose" (AERA, APA, & NCME, 1999, h. 11).

Ketiga, menurut pengertian mutakhir tentang validitas ada lima jenis evidensi yang perlu dikumpulkan dalam rangka memeriksa validitas penafsiran skor sesuai tujuan penggunaan tes. Kelima jenis evidensi yang dimaksud adalah sebagai berikut (AERA, APA, & NCME, 1999; Goodwin & Leech, 2003):

a. Evidensi Terkait Isi Tes. Salah satu evidensi validitas adalah kesesuaian antara isi tes dan konstruk yang diukurnya. Evidensi ini bisa diperoleh melalui analisis logis atau empiris terhadap seberapa memadai isi tes mewakili ranah isi serta seberapa relevan ranah isi tersebut sesuai dengan interpretasi skor tes yang dimaksudkan. Isi tes mengacu pada tema-tema, pilihan kata, serta format atau bentuk item, tugas, atau pertanyaan yang digunakan dalam tes. Evidensi terkait isi ini juga bisa berupa penilaian pakar atau ahli terhadap kesesuaian antara bagian-bagian tes dan konstruk yang diukur. Secara lebih rinci, aspek-aspek isi tes yang perlu dievaluasi meliputi (1) *sufficiency,* yaitu apakah isi tes tersebut memadai dalam arti mewakili ranah isi spesifik yang hendak diukur; (2) *clarity* atau kejelasan, yaitu apakah isi tes tersebut mencerminkan secara jelas ranah isi spesifik yang hendak diukur dalam arti misalnya tidak mencampuradukkan dengan ranah isi spesifik yang lain; (3) *relevance* atau relevansi, yaitu apakah isi tes tersebut memiliki kesesuaian dengan ranah isi spesifik yang hendak diukur; (4) kesesuaian antara item-item dan tugas-tugas

yang dipakai sebagai stimulus dalam tes tersebut dengan definisi konstruk yang terwakili oleh ranah isi spesifik yang hendak diukur; (5) ada-tidaknya *bias* berupa keberpihakan isi tes pada gender, budaya, umur atau faktor pengelompokan sosial lainnya; dan (6) kemungkinan terjadinya "*construct irrelevant variance*" (varians yang tidak relevan dengan konstruk yang hendak diukur) dan "*construct underrepresentation*" (kurang memadainya keterwakilan konstruk yang hendak diukur), yang menunjukkan sejauh mana kemungkinan tes tersebut mengukur melebihi (*construct irrelevance variance)* atau kurang (*construct underrepresentation)* dari yang semestinya dia ukur (Goodwin & Leech, 2003). Dalam pengertian lama, jenis evidensi ini dikaitkan dengan *validitas isi*.

b. Evidensi terkait Proses Respon yang diberikan oleh Subjek. Evidensi ini didasarkan pada penilaian terhadap kesesuaian antara konstruk yang diukur dengan kinerja atau respon yang diberikan oleh subjek. Sebagai contoh, dalam mengerjakan *typical performance test* subjek sungguh-sungguh menyatakan keadaan dirinya sebagai cerminan konstruk yang diukur dan bukan sekadar memberikan jawaban sesuai norma yang berlaku di tengah masyarakat. Beberapa strategi untuk mengumpulkan jenis evidensi ini meliputi (1) mengobservasi testi saat sedang mengerjakan tugas dalam rangka tes, (2) mewawancarai testi untuk mengetahui alasan mereka memberikan jawaban tertentu terhadap pertanyaan-pertanyaan dalam tes, dan (3) memeriksa atau menyelidiki cara para pengamat dan penilai menerapkan kriteria dalam mencatat dan mengevaluasi tingkah laku, kinerja, atau hasil pekerjaan tertulis testi dalam rangka tes; tujuannya adalah memastikan bahwa kriteria penilaian yang disediakan diterapkan sebagaimana mestinya dan bukan malah menggunakan acuan lain yang tidak sesuai (Goodwin & Leech, 2003). Dalam konsep lama, jenis evidensi ini dikaitkan dengan *validitas konstruk*.

c. Evidensi terkait Struktur Internal Tes. Evidensi ini didasarkan pada penilaian tentang sejauh mana hubungan antar item dan hubungan antar komponen tes sesuai dengan konstruk yang

diukur. Salah satu metode yang sangat lazim ditempuh adalah faktor analisis konfirmatori. Namun diingatkan, terlalu mengandalkan faktor analisis dalam validasi berisiko memperoleh bukti validitas yang kurang kokoh. Metode lain yang disarankan untuk juga ditempuh adalah *differential item function (DIF) techniques* untuk memeriksa kemungkinan terjadinya bias item sebagai evidensi lain invaliditas. *DIF* terjadi jika testi dengan kemampuan yang sama namun termasuk ke dalam kelompok yang berbeda memiliki peluang yang juga berbeda untuk berhasil mengerjakan sebuah item (Goodwin & Leech, 2003). Dalam pengertian lama, jenis evidensi ini dikaitkan dengan *validitas konstruk*.

d. Evidensi terkait Hubungannya dengan Variabel Lain. Evidensi validitas juga bisa diperoleh dengan menganalisis hubungan antara skor tes dan variabel-variabel lain di luar tes itu sendiri. Variabel eksternal bisa mencakup paling tidak tiga hal, yaitu (1) hasil pengukuran terhadap kriteria tertentu yang diprediksikan oleh tes yang bersangkutan; di sini yang perlu dievaluasi adalah seberapa akurat tes tersebut mampu memprediksikan kinerja yang merupakan kriterianya; (2) tes-tes lain yang dimaksudkan untuk mengukur konstruk yang sama seperti yang diukur oleh tes yang bersangkutan; hubungan positif antara skor tes dengan alat ukur lain yang dimaksudkan untuk mengukur konstruk yang sama atau sejenis menghasilkan apa yang disebut *evidensi konvergen;* dan (3) tes-tes lain yang dimaksudkan untuk mengukur berbagai konstruk yang berbeda; hubungan seperti yang diharapkan khususnya berupa hubungan positif namun tidak signifikan atau hubungan negatif dan sifnifikan antara skor tes dengan alat ukur yang dimaksudkan untuk mengukur konstruk yang berbeda menghasilkan apa yang disebut *evidensi diskriminan.* Selain itu, pemeriksaan perbedaan kinerja dalam suatu tes antar dua atau lebih kelompok yang diprediksikan memang berbeda baik melalui *group-comparison studies* atau penelitian tentang perbedaan kelompok dan *experimental research studies* atau penelitian eksperimental yang melibatkan pembandingan kinerja antar kelompok, juga bisa memberikan evidensi validitas terkait hubungan

antara tes dengan variabel lain (Goodwin & Leech, 2003). Dalam pengertian lama, jenis evidensi ini dikaitkan dengan *citerion-related validity* atau *validitas terkait dengan kriteria*.

e. Evidensi terkait Konsekuensi Pengetesan. Konsekuensi, dampak, atau akibat baik yang direncanakan maupun yang tidak direncanakan dari penerapan tes juga bisa digunakan sebagai evidensi validitas. Terkait yang pertama, tes lazim diadministrasikan dengan harapan memperoleh manfaat tertentu dari hasil interpretasi skor yang sudah direncanakan. Contoh manfaat yang direncanakan semacam itu meliputi antara lain bisa memilih bentuk perlakuan yang efektif dalam terapi, bisa menempatkan karyawan pada jenis tugas yang cocok, bisa menghindari risiko meloloskan orang yang tidak memenuhi kualifikasi untuk memasuki sebuah profesi, atau bisa memperbaiki praktik pengajaran di kelas. Tujuan validasi adalah memperoleh evidensi atau bukti bahwa manfaat tersebut sungguh-sungguh terjadi.Terkait yang kedua, penerapan tes seringkali juga diakui memberikan manfaat di luar hasil interpretasi skor yang sudah direncanakan itu sendiri. Contoh manfaat yang tidak direncanakan dari penerapan tes adalah meningkatnya motivasi belajar siswa atau meningkatnya praktik pengajaran di kelas yang dilaksanakan oleh guru sesudah mengetahui hasil tes. Evidensi validitas diperoleh jika manfaat tidak langsung semacam itu sungguh-sungguh tercapai (AERA, APA, & NCME, 1999). Dalam pengertian lama, jenis evidensi ini dikaitkan dengan *validitas konstruk*.

Sebagaimana ditegaskan dalam *1999 Standards,* berbagai evidensi di atas ditambah dengan evidensi lain terkait kualitas teknis sebuah tes meliputi penyusunannya yang dilakukan secara cermat, penciptaan kondisi administrasinya yang baku termasuk cara administrasi dan penskorannya yang tepat, reliabilitas skornya yang memadai, penskalaan dan konversi skornya yang akurat, serta upaya yang cermat menjaga *fairness* atau kesetaraan bagi semua testi, semua itu akan memberikan landasan yang kuat bagi keabsahan cara menafsirkan skor tes sebagaimana direncanakan dan sesuai dengan tujuan pengukurannya.

2. Reliabilitas

Reliabilitas adalah konsistensi hasil pengukuran jika prosedur pengetesannya dilakukan secara berulangkali terhadap suatu populasi individu atau kelompok (AERA, APA, & NCME, 1999). Manfaat hasil pengukuran ditentukan oleh stabilitas kinerja individu atau kelompok yang dikenai tes. Kenyataannya, individu atau kelompok yang sama yang dikenai dengan tes yang sama pada kesempatan yang berlainan bisa dipastikan selalu menunjukkan kinerja atau hasil pengukuran yang berbeda. Penyebabnya, *obtained score* atau skor tercapai yang diperoleh seseorang atau skor rerata yang dicapai oleh sekelompok orang senantiasa mengandung sejumlah kecil *measurement error* atau kesalahan pengukuran.

Artinya, secara hipotetis terdapat sebuah nilai yang bebas dari kesalahan pengukuran dalam skor tercapai yang diperoleh oleh seseorang (AERA, APA, & NCME, 1999). Dalam *classical test theory* atau teori tes klasik nilai yang bebas dari kesalahan dalam skor tercapai ini disebut ***true score*** atau **skor murni**. Skor murni ini dipandang sebagai skor rerata hipotetis yang akan diperoleh jika sebuah tes dikenakan secara berulang-ulang terhadap seseorang atau sekelompok orang yang sama. Dalam *generalizability theory* atau teori generalisabilitas nilai yang bebas dari kesalahan yang terdapat dalam skor tercapai ini disebut ***universe score*** atau **skor populasi.** Dalam *item response theory* (IRT) atau teori respon item, nilai yang bebas dari kesalahan dalam skor tercapai ini disebut ***ability parameter*** atau ***trait parameter.*** Perbedaan atau selisih hipotetis antara skor tercapai seorang testi pada pengukuran atribut psikologis tertentu dengan skor murni atau skor populasinya disebut ***measurement error*** atau **kesalahan pengukuran** (AERA, APA, & NCME, 1999).

Standarisasi tes, berupa penggunaan materi tes yang sama dari testi ke testi, mengikuti secara ketat dan konsisten prosedur administrasi tes, dan penerapan secara ketat dan konsisten aturan penskoran hasil tes, merupakan salah satu strategi untuk meminimalkan kesalahan pengukuran. Kecenderungan untuk menerapkan prosedur tes secara

fleksibel atau longgar demi menyesuaikan diri dengan kebutuhan pengukuran, keadaan testi, maupun situasi pengetesan yang muncul akhir-akhir ini di satu sisi mungkin bisa dipertanggungjawabkan dalam arti mengurangi *construct irrelevance* atau *construct underrepresentation*, namun tetap akan berdampak meningkatkan cakupan maupun besaran kesalahan pengukuran. Artinya, mengorbankan reliabilitas (AERA, APA, & NCME, 1999).

Kesalahan pengukuran yang menjadi sumber irreliabilitas lazim dipandang bersifat ***random*** dalam arti inkonsisten serta *unpredictable* atau tidak bisa diduga sebelumnya. Karena sifatnya itu, kesalahan pengukuran tidak mungkin dihilangkan dari skor tercapai. Namun besaran agregat kesalahan pengukuran semacam itu dapat diringkas dalam arti dirumuskan dengan beberapa cara, yaitu (1) dalam bentuk varians atau deviasi standar kesalahan pengukuran, (2) dalam bentuk koefisien reliabilitas, dan (3) dalam bentuk fungsi-fungsi informasi tes berbasis *IRT.*

a. Varians atau Deviasi Standar Kesalahan Pengukuran. Besaran agregat kesalahan pengukuran bisa dirumuskan dalam bentuk varians atau deviasi standar kesalahan pengukuran. *Standard measurement error* atau kesalahan pengukuran baku adalah deviasi standar distribusi kesalahan pengukuran hipotetis yang diperoleh jika suatu populasi tertentu dikenai sebuah tes atau prosedur tertentu. Keseluruhan varians kesalahan pengukurannya sesungguhnya adalah rerata (*a weighted average*) nilai-nilai yang diperoleh pada berbagai taraf skor murni. Varians pada taraf skor murni tertentu disebut *conditional error variance* sedangkan akarnya disebut *conditional standard error* (AERA, APA, & NCME, 1999). Makin kecil deviasi standar kesalahan pengukuran maka makin kecil pula kesalahan pengukuran, berarti hasil pengukurannya pun makin reliabel.

b. Koefisien Reliabilitas. Cara lain merumuskan besaran agregat kesalahan pengukuran adalah dalam bentuk koefisien reliabilitas. Secara tradisional ada tiga jenis koefisien reliabilitas: (1) koefisien yang diperoleh dari pengadministrasian bentuk-bentuk paralel tes pada kesempatan yang berlainan; hasilnya disebut

alternate-form coefficients atau **koefisien bentuk alternatif**; (2) koefisien yang diperoleh dari pengadministrasian tes yang sama terhadap kelompok subjek yang sama pada kesempatan yang berlainan; hasilnya disebut ***test-retest coefficients*** atau **koefisien tes-retes** atau ***stability coefficients*** atau **koefisien stabilitas**; dan (3) koefisien yang didasarkan pada hubungan antar skor pada masing-masing item atau antar skor pada kelompok-kelompok item dalam tes, yang datanya diperoleh dari satu kali pengadministrasian tes pada sekelompok subjek; hasilnya disebut ***internal conistency coefficients*** atau **koefisien konsistensi internal**. Makin besar tiga jenis koefisien tersebut, makin kecil kesalahan pengukuran sehingga makin tinggi reliabilitas hasil pengukurannya.

Dengan dikembangkannya *generalizability theory* atau teori generalisabilitas seperti yang sudah disinggung, ketiga jenis koefisien di atas bisa dipandang sebagai bentuk khusus dari jenis koefisien yang lebih umum, yaitu *generalizability coefficients* atau koefisien generalisabilitas. Koefisien generalisabilitas sendiri adalah "the ratio of true or universe score variance to observed score variance" atau rasio varians skor murni atau skor universal terhadap varians skor tercapai (AERA, APA, & NCME, 1999; h. 28). Makin besar koefisien generalisabilitas makin kecil kesalahan pengukuran sehingga makin tinggi reliabilitas hasil pengukurannya. Dibandingkan tiga pendekatan reliabilitas yang tradisional, teori generalisabilitas memiliki kelebihan yaitu memungkinkan penyusun tes menspesifikasikan dan mengestimasikan aneka komponen varians skor murni, varians kesalahan, dan varians skor teramati (AERA, APA, & NCME, 1999).

c. Fungsi Informasi Tes. Besaran agregat kesalahan pengukuran juga bisa dirumuskan dalam bentuk fungsi informasi tes mengikuti teori respon item (*IRT*). Menurut teori respon item, fungsi informasi tes secara efisien "summarizes how well the test discriminates among individuals at various levels of the ability or trait being assessed" (AERA, APA, & NCME, 1999; h. 28). Maksudnya, fungsi informasi tes secara efisien menunjukkan seberapa baik tes mampu mendiskriminasikan dalam arti memilah testi pada berbagai

taraf abilitas atau atribut psikologis yang sedang diukur. Dalam teori respon item sebuah fungsi matematis yang disebut *item characteristic curve* atau kurva karakteristik item atau *item response function* atau fungsi respon item digunakan sebagai model untuk menunjukkan proporsi jawaban benar terhadap sebuah item yang meningkat pada berbagai taraf yang secara progresif makin meningkat dari abilitas atau atribut psikologis yang sedang diukur. Fungsi ini bisa dipandang sebagai pernyataan matematis tentang ketepatan pengukuran pada masing-masing taraf atribut psikologis yang sedang diukur. Ketepatan dalam konteks *IRT* adalah setara dengan lawan dari *conditional error variance* dalam konteks *classical test theory* (AERA, APA, & NCME, 1999). Makin tinggi ketepatan makin kecil kesalahan pengukuran, berarti makin tinggi reliabilitas hasil pengukurannya.

Masing-masing pendekatan untuk memeriksa reliabilitas hasil pengukuran dengan sebuah tes memiliki taraf ketepatan dan jenis informasi yang berlainan antara lain terkait sumber kesalahan pengukuran yang menjadi sumber irreliabilitas. Tidak ada pendekatan tunggal untuk menunjukkan besaran reliabilitas. Tidak ada indikator tunggal yang secara memadai mampu mengungkapkan semua fakta yang relevan dengan reliabilitas. Padahal kita tahu, taraf reliabilitas skor-skor hasil tes memiliki dampak langsung terhadap validitas interpretasinya. Selama skor-skor mengandung kesalahan pengukuran yang bersifat random, tidak konsisten, dan tidak dapat diduga sebelumnya, menjadikan kemampuannya untuk memprediksi kriteria, menyusun diagnosis tentang testi, dan memberi landasan bagi pembuatan keputusan menjadi terbatas (AERA, APA, & NCME, 1999). Hal ini perlu disadari betul baik oleh para penyusun maupun pengguna tes psikologis.

3. Statistik Item

Statistik item-item yang membentuk sebuah tes psikologis diperiksa melalui analisis item sesudah item-item tersebut diuji-cobakan pada sekelompok sampel standarisasi. Tujuan analisis item

adalah "to select items that form a homogenous, discriminating scale. The most commonly used method is to correlate each item with the total score and to calculate the proportion of the complete sample who put the keyed response. By selecting items with high correlations with the total score which furthermore have endorsement rates of between 80 percent and 20 per cent, a homogenous and discriminating test can be produced" (Kline, 1986; h. 133). Artinya, tujuan analisis item adalah memilih item-item yang akan membentuk sebuah skala yang bersifat homogen dan memiliki daya diskriminasi yang baik. Cara yang paling lazim ditempuh adalah memeriksa korelasi antara masing-masing item dengan skor total serta menghitung proporsi subjek yang memilih kunci jawaban. Dengan memilih item-item yang memiliki korelasi positif tinggi dengan skor total serta kunci jawabannya pun dipilih oleh antara 20-80 persen subjek penjawab, maka akan diperoleh sebuah tes yang homogen dan memiliki daya diskriminasi yang baik. Berarti, ada dua statistik yang lazim dijadikan indeks bagi item yang baik, yaitu: (a) korelasi antara masing-masing item dengan skor total; dan (b) proporsi subjek yang memilih kunci jawaban.

a. Korelasi Item-Total. Statistik ini menjamin homogenitas tes sebagai kesatuan dengan cara menunjukkan item-item yang paling baik mengukur konstruk atau isi yang sedang diukur. Logikanya, jika sebuah item mengukur atribut yang sama sebagaimana diukur oleh keseluruhan item, maka subjek atau testi yang menunjukkan kinerja yang "baik" pada tes sudah barang tentu juga menunjukkan kinerja yang "baik" pada item yang bersangkutan (Friedenberg, 1995). Dengan cara yang sama statistik ini sekaligus menunjukkan daya diskriminasi item, yaitu kemampuan sebuah item memicu cara menjawab yang berbeda pada diri subjek atau testi dengan tipe yang memang berlainan. Ada beberapa cara untuk memeriksa korelasi item-total (Kline, 1986), yaitu:

(1) *Pearson product-moment correlation* atau korelasi *product-moment* Pearson (*r*). Cara ini cocok untuk diterapkan pada *multi-point items*

atau item-item yang memiliki alternatif jawaban ganda dalam arti lebih dari dua.

(2) *Point-biserial correlation* atau korelasi bi-serial (r_{pbis}). Cara ini cocok diterapkan pada item-item dikotomis, yaitu item-item yang hanya memliki dua alternatif jawaban termasuk item-item "Ya-Tidak" dan "Benar-Salah".

(3) *Phi-coefficient* atau koefisien-phi (φ) . Cara ini cocok diterapkan jika skor total skala dibuat dikotomis, misal menjadi "Lulus-Tidak lulus" atau "Di atas Rerata-Di bawah Rerata" dengan asumsi bahwa kedua kategori tersebut sungguh-sungguh bersifat non-kontinyu.

(4) *Tetrachoric correlation* atau korelasi tetrakorik (r_{tet}). Cara ini bisa diterapkan sebagai ganti atau setara koefisien-phi. Bedanya, kategori "Lulus-Tidak lulus" atau "Benar-Salah" atau sejenisnya diasumsikan bersifat kontinyu. Masalahnya, korelasi tetrakorik memiliki kesalahan baku (*standard error*) yang besar, yaitu dua kali lebih besar dari kesalahan baku korelasi *product moment.*

Terhadap keempat cara di atas Kline (1986) memberikan catatan sebagai berikut. Pertama, mengubah skor total menjadi dikotomis tidak memberikan banyak manfaat bahkan berakibat hilangnya banyak informasi yang berharga. Maka, cara 3 dan 4 cenderung tidak direkomendasikan. Pada cara 1 dan 2 skor total sebagai kriteria dibiarkan bersifat kontinyu. Dari kedua cara tersebut cara 2 (r_{pbis}) dipandang mampu memberikan ukuran atau gambaran korelasi item-total yang paling baik. Karena korelasi item-total ini merupakan prasyarat esensial untuk mendapatkan sebuah tes psikologis yang homogen, maka cara 2 ini sangat direkomendasikan. Namun karena r_{pbis} dan korelasi *product moment Pearson* secara numerik bersifat ekuivalen atau setara, maka cara 1 dan cara 2 praktis sama-sama baik (Kline, 1986).

b. Proporsi Subjek yang Memilih Kunci Jawaban. Statistik ini dihitung dengan cara membagi jumlah subjek pemilih kunci jawaban pada masing-masing item dengan jumlah total subjek

penjawab ($p = \sum n/N$). Dalam *maximal performance tests,* statistik ini disebut ***item difficulty*** atau **taraf kesukaran item.** Istilah ini agak menyesatkan sebab kenyataannya, makin besar p makin rendah taraf kesukaran item atau makin **mudah.** Dalam *typical performance tests* termasuk *projective techniques,* statistik ini menunjukkan ***percent endorsement*** atau besar persentase subjek yang memilih kunci jawaban dan mencerminkan **popularitas** jawaban. Salah satu kendala dalam memeriksa statistik ini adalah kemungkinan adanya subjek yang melewatkan menjawab satu atau lebih item tertentu. Untuk mengatasinya, sebelum memulai analisis butir harus dipastikan bahwa hanya hasil tes yang dijawab secara lengkap oleh masing-masing testi dipakai sebagai data.

4. Daya Diskriminasi Tes

Salah satu statistik yang direkomendasikan untuk memeriksa daya diskriminasi tes adalah koefisien diskriminasi yang disebut ***Ferguson's delta*** atau δ (Kline, 1986). Semua indeks diskriminasi termasuk **delta Ferguson** pada hakikatnya didasarkan pada penjenjangan subjek. Intinya, koefisien diskriminasi menunjukkan seberapa cermat dan konsisten sebuah tes menjenjangkan testi sepasang demi sepasang dalam hal atribut psikologis yang diukur. Hubungan antara skor-skor setiap pasang subjek senantiasa berupa atau perbedaan atau kesamaan. Tentu saja, penjenjangan subjek didasarkan pada perbedaan daripada kesamaan skor di antara mereka. Dengan prinsip semacam itu, rumus koefisien diskriminasi **delta Ferguson** adalah sebagai berikut (Kline, 1986):

$$\delta = (n+1)(N^2 - \sum f_i^2) / nN^2 \qquad \textbf{Rumus 5.1.}$$

δ = delta Ferguson
N = jumlah subjek
n = jumlah item
f_i = frekuensi masing-masing skor

Bisa dirumuskan secara umum bahwa $\delta = 0$ jika seluruh subjek mencapai skor yang sama (tidak terjadi diskriminasi), dan $\delta = 1$ jika terjadi distribusi skor yang bersifat rektangular (terjadi diskriminasi yang sempurna, yaitu pembedaan antara subjek yang mencapai skor di bawah *cutting score* atau skor batas tertentu dan mereka yang mencapai skor di atas skor batas). Dengan kata lain, makin δ mendekati 1 maka makin baik daya diskriminasi tes.

Langkah-langkah perhitungan **delta Ferguson** adalah sebagai berikut (Kline, 1986):

a. Buatlah distribusi frekuensi skor tes yang menjadi objek pemeriksaan.
b. Kuadratkan masing-masing frekuensi dan jumlahkan: $\sum f_i^2$
c. Tambahkan 1 pada jumlah item: $n+1$
d. Kuadratkan jumlah subjek: N^2
e. Kalikan jumlah item dengan jumlah subjek kuadrat: nN^2
f. Masukkan unsur-unsur di atas ke dalam rumus δ dan hitung hasilnya. Ψ

Bab 6
Teori Tes Klasik

Teori tentang tes menyajikan sejenis kerangka umum untuk menjelaskan kaitan antara variabel-variabel yang teramati dalam praktik pengetesan, seperti skor tes dan skor item, dengan variabel-variabel yang tak teramati seperti *true scores* atau skor murni dan *ability scores* atau skor abilitas atau dalam cakupan yang lebih luas, skor atribut. Secara lebih spesifik, salah satu masalah utama dalam praktik pengukuran psikologis yang harus ditangani dengan menggunakan teori tentang tes adalah *measurement errors* atau kesalahan pengukuran. Sebuah teori tentang tes yang baik akan mampu menjelaskan peran atau sumbangan kesalahan pengukuran dalam (1) mengestimasikan abilitas testi dan cara meminimalkan sumbangan kesalahan pengukuran itu sendiri, (2) mempengaruhi korelasi antar variabel, dan (3) mempengaruhi skor murni dan skor abilitas (Hambleton & Jones, 1993).

Teori tentang tes lazim dijabarkan ke dalam sebuah model tentang tes. Model tentang tes semacam ini menjelaskan secara lebih rinci hubungan atau kaitan antar konsep-konsep teoretis yang disajikan dalam teori tentang tes dilengkapi dengan asumsi-asumsi terkait baik aneka konsepnya sendiri maupun saling hubungan antar konsep yang dimaksud. Ketepatan model tentang tes semacam ini bisa diperiksa secara empiris berdasarkan data. Setiap model tentang tes tidak akan pernah mencerminkan keadaan data pengukuran secara sempurna. Maka, persoalan pokok yang melekat pada setiap model tentang tes bukanlah benar atau salahnya, melainkan sejauh mana model tersebut *fits* atau cocok atau sesuai dengan data sehingga memberikan pedoman yang benar dalam keseluruhan proses pengukuran (Hambleton & Jones, 1993).

Kendati ada tiga teori tentang tes sebagaimana sudah disinggung di Bab 5, yaitu teori tes klasik, teori respon item, dan teori

generalisabilitas, secara khusus dalam bab ini hanya akan dipaparkan teori tes klasik beserta model tes yang ditawarkannya. Pilihan ini didasarkan pada beberapa alasan sebagai berikut. Pertama, teori yang berintikan teori kesalahan dalam pengukuran ini dikembangkan bertolak dari serangkaian asumsi yang paling sederhana yang pernah dikemukakan oleh para penyusun tes psikologis, karenanya juga disebut teori tes klasik. Kendati sederhana, namun prinsip-prinsip pokok teori klasik ini diakui masih berlaku secara kokoh. Kedua, prinsip-prinsip pokok teori klasik juga diakui mudah diterapkan dalam penyusunan tes dan dalam kenyataan masih banyak digunakan oleh para penyusun tes psikologis hingga kini (Kline, 1986). Bab ini akan terdiri dari tiga bagian, yaitu: (a) pemaparan ringkas tentang teori tes klasik, (b) reliabilitas menurut model tes klasik, dan (c) pemeriksaan reliabilitas berdasarkan model tes klasik.

A. Model Tes Klasik

Teori tes klasik merupakan bentuk paling sederhana dari *weak true score theory* atau teori skor murni yang lemah (Novick, 1966). Teori ini menawarkan model-model tentang tes yang lazim disebut "weak models" atau model-model tentang tes yang lemah dalam arti longgar. Model-model tentang tes yang ditawarkan oleh teori tes klasik disebut longgar sebab asumsi-asumsinya mudah dipenuhi oleh data tes, seperti asumsi bahwa struktur atribut yang menjadi sasaran pengukuran sama untuk semua bentuk paralel tesnya (Hambleton & Jones, 1993), atau bahkan sama sekali tidak mengajukan asumsi tentang bentuk distribusi dari konsep-konsep pokoknya yaitu *observed score, true score,* dan *error score* (Novick, 1966).

Teori tes klasik menjelaskan skor tes dengan mengajukan tiga macam konsep, yaitu (1) *test score* atau sering disebut *observed score* yang diterjemahkan sebagai **skor tampak** (Azwar, 1997) dan diberi lambang *X*; (2) *true score* yang diterjemahkan sebagai **skor murni** (Azwar, 1997) dan diberi lambang *T;* serta (3) *error score* atau **skor**

kesalahan dan diberi lambang *E***.** Bertolak dari tiga konsep tersebut teori skor klasik menawarkan sejumlah model tentang tes. Salah satu model yang sangat populer dan yang akan kita ikuti dalam pembicaraan kita selanjutnya adalah *classical test model* atau **model tes klasik**.

Model tes klasik adalah sebuah model linear sederhana yang mempostulasikan bahwa *observable test score* atau **skor tampak** (*X*) yang dicapai oleh seorang testi dalam sebuah tes dapat diuraikan ke dalam dua *latent variables* atau variabel laten atau variabel yang tak teramati, yaitu **skor murni** (*T*) dan **skor kesalahan** (*E*), sehingga diperoleh rumus sederhana (Novick, 1966; Hambleton & Jones, 1993; De Champlain, 2010):

$$X = T + E \qquad \textbf{Rumus 6.1.}$$

Dalam praktik administrasi tes, kita hanya memiliki informasi tentang skor tampak *X,* yaitu skor total masing-masing testi dalam suatu tes, sedangkan dua unsur lainnya tidak kita ketahui karena bersifat laten. Kepentingan utama kita adalah mengetahui dua unsur tersebut, yaitu besar skor murni dan skor kesalahan. Semakin besar skor murni atau sebaliknya semakin kecil skor kesalahan, berarti semakin skor tampak *X* mencerminkan abilitas atau atribut psikologis lain yang menjadi sasaran pengukuran. Namun karena dua unsur laten tersebut tidak diketahui, maka Rumus 6.1. di atas tidak mungkin diselesaikan dan kita pun tidak bisa menentukan kualitas dalam arti ketepatan hasil pengukuran yang kita peroleh.

B. Beberapa Asumsi

Untuk mengatasi masalah tersebut dan khususnya agar Rumus 6.1. di atas bisa diselesaikan, model tes klasik mengajukan serangkaian asumsi sebagaimana disajikan di bawah ini (Hambleton & Jones, 1993; Novick, 1966; Allen & Yen, 1979). Ada total tujuh asumsi. Asumsi

pertama sampai dengan asumsi kelima pada hakikatnya adalah definisi model tes klasik tentang *error of measurement* atau kesalahan pengukuran. Menurut model tes klasik, kesalahan pengukuran bersifat *unsystematic* atau tidak sistematis atau random, dan merupakan deviasi atau jarak atau selisih antara skor tampak seorang testi dengan suatu skor tampak yang secara teoretis diharapkan (*a theoretically expected observed score*). Dalam model tes klasik, berbagai kesalahan sistematis tidak masuk dalam sebutan kesalahan pengukuran (Allen & Yen, 1979). Asumsi keenam dan ketujuh adalah tentang bentuk paralel tes.

1. $X = T + E$. Skor tampak berupa skor total yang dicapai seorang testi dalam suatu tes adalah hasil penjumlahan dari dua bagian atau komponen, yaitu skor murni dan skor kesalahan atau kesalahan pengukuran. *T* atau skor murni diasumsikan berupa sebuah nilai tetap, sedangkan *E* atau skor kesalahan bersifat random sehingga bervariasi dalam berbagai kesempatan administrasi tes yang berbeda, sehingga akibatnya begitu pula *X* atau skor tampak. Intinya, menurut model tes klasik, skor murni *T* dan skor kesalahan *E* bersifat saling menjumlahkan sehingga menghasilkan skor tampak *X* (Allen & Yen, 1979).
2. $\varepsilon(X) = T$. *Expected value* atau nilai yang diharapkan, yaitu *mean* populasi, dari *X* adalah *T*. Dengan kata lain, asumsi ini merupakan definisi skor murni *T,* yaitu *mean* dari distribusi teoretis skor-skor *X* yang akan diperoleh jika seseorang dites secara berulang-ulang dalam jumlah kali tak terbatas dengan tes yang sama (Allen & Yen, 1979; De Champlain, 2010). Definisi ini berlaku dengan asumsi bahwa administrasi tes yang berulang-ulang tersebut bersifat independen dalam arti setiap administrasi tes tidak mempengaruhi administrasi-administrasi tes berikutnya. Karena asumsi ini sulit dipenuhi, artinya mustahil menjaga independensi masing-masing administrasi dari tes yang sama terhadap subjek atau testi yang sama, maka skor murni tetap merupakan sebuah konstruk teoretis yang mustahil dihitung secara nyata.

3. $\rho_{ET} = 0$. Skor-skor kesalahan *E* dan skor-skor murni *T* yang dicapai oleh suatu populasi testi pada sebuah tes tidak saling berkorelasi. Artinya, testi yang mencapai skor murni *T* tinggi dalam suatu tes tidak secara sistematis mencapai skor kesalahan *E* lebih positif atau lebih negatif dibandingkan testi yang mencapai skor murni *T* lebih rendah dalam tes yang sama. Skor kesalahan positif berakibat meningkatkan skor tampak *X*, sebaliknya skor kesalahan negatif berakibat menurunkannya.
4. $\rho_{E1E2} = 0$. E_1 adalah skor kesalahan pada Tes 1, sedangkan E_2 adalah skor kesalahan pada Tes 2. Asumsi ini menyatakan bahwa skor-skor kesalahan pada dua tes yang berbeda tidak saling berkorelasi. Artinya, jika seorang testi mencapai skor kesalahan positif dalam Tes 1, maka dia tidak memiliki kemungkinan lebih besar untuk mencapai skor kesalahan positif atau negatif dalam Tes 2.
5. $\rho_{E1T2} = 0$. E_1 adalah skor kesalahan pada Tes 1, sedangkan T_2 adalah skor murni pada Tes 2. Asumsi ini menyatakan bahwa skor-skor kesalahan yang dicapai testi dalam Tes 1 tidak saling berkorelasi dengan skor-skor murni yang mereka capai dalam Tes 2.
6. **Tes-tes yang paralel.** Jika dua tes menghasilkan skor-skor tampak *X* dan *X′* dan skor-skor tampak tersebut memenuhi asumsi 1 sampai dengan asumsi 5, dan jika, untuk semua populasi testi, skor-skor murni $T = T'$ sedangkan varians kesalahan $\sigma_E^2 = \sigma_{E'}^2$, maka kedua tes tersebut disebut ***tes-tes yang paralel***. Varians kesalahan adalah varians dari skor-skor kesalahan dalam suatu tes yang dibuat oleh populasi testi tertentu. Asumsi 6 merupakan definisi model tes klasik tentang tes-tes yang paralel, yaitu bahwa dua atau lebih tes bersifat paralel jika skor murni dan varians kesalahan masing-masing tes adalah sama untuk semua populasi testi yang menempuh tes-tes tersebut. Tes-tes paralel sering disebut *parallel test forms* atau bentuk-bentuk tes paralel atau cukup *parallel forms* atau bentuk-bentuk paralel. Definisi tersebut berimplikasi bahwa bentuk-bentuk paralel akan memiliki *mean* skor tampak, varians,

serta korelasi antar skor tampak yang sama (Allen & Yen, 1979). Asumsi ini secara implisit menyatakan bahwa bentuk-bentuk paralel sebuah tes bisa atau mungkin disusun. Dalam kenyataan sangat sulit kalau bukan mustahil menyusun bentuk-bentuk paralel sebuah tes dengan memenuhi seluruh kriterianya. Maka dikenallah apa yang disebut *essentially τ-equivalent tests* atau tes-tes dengan τ yang secara esensial ekuivalen atau setara dan itulah yang menjadi isi asumsi ketujuh.

7. **Tes-tes dengan τ yang secara esensial ekuivalen.** Jika dua tes menghasilkan skor-skor tampak X_1 dan X_2 dan skor-skor tampak tersebut memenuhi Asumsi 1 sampai dengan 5, dan jika, untuk semua populasi testi $T_1 = T_{2+C12}$, di mana C_{12} berupa suatu bilangan konstan, maka tes-tes itu disebut *essentially τ-equivalent tests* atau tes-tes dengan τ yang secara esensial ekuivalen atau setara. τ (huruf Yunani dibaca "tau") adalah lambang skor murni T dalam populasi. Tes-tes semacam ini memiliki skor-skor murni yang sama dalam arti berselisih dengan sebuah bilangan konstan C_{12} serta bisa memiliki varians-varians kesalahan yang berbeda. Dengan kata lain, tes-tes dengan τ yang secara esensial ekuivalen tidak harus bersifat paralel. Akibatnya, berbeda dengan tes-tes yang paralel tes-tes dengan τ yang secara esensial ekuivalen bisa memiliki kemampuan yang tidak sama dalam mengestimasikan skor-skor murni T (Allen & Yen, 1979).

Berdasarkan uraian di atas, ada sejumlah karakteristik model tes klasik pada khususnya maupun teori tes klasik pada umumnya yang bisa disimpulkan, termasuk kelebihan dan kekurangannya. *Pertama,* berbagai model tes dalam teori tes klasik pada dasarnya dikembangkan pada aras atau tataran skor (total) tes atau *test based* atau berbasis tes. Artinya, estimasi skor-skor murni dilakukan berdasarkan skor-skor (total) tes (Hambleton & Jones, 1993). Untuk keperluan penyusunan atau pengembangan tes, teori dan model tes klasik memang juga mengembangkan sejumlah konsep statistik pada tataran *item* atau butir, khususnya *item difficulty* atau taraf kesukaran

(dilambangkan ***p***) dan *item discriminating power* atau daya beda item atau butir (dilambangkan ***r*** atau tepatnya r_{it}). Namun, seperti juga tampak pada statistik item yang kedua, teori skor klasik pada dasarnya beroperasi pada tataran skor (total) tes.

Kedua, kekurangan atau kelemahan pertama dan utama dari teori dan model tes klasik ini adalah sifatnya yang tergantung pada sampel (*sample dependent*) maupun pada tesnya (*test dependent*) sehingga mengurangi daya gunanya. Berbagai parameter testi (*person parameters*) seperti skor murni maupun parameter item (*item parameters*) seperti taraf kesukaran item (p) dan daya beda item (r_{it}) sangat ditentukan oleh karakteristik tes maupun sampel testinya. Contoh, dua tes yang dimaksudkan untuk mengukur sebuah abilitas yang sama dan yang diadministrasikan pada sampel testi yang sama bisa dipastikan akan terbukti memiliki item-item dengan taraf kesukaran dan daya beda yang tidak sama. Atau, skor tes yang dicapai seorang testi tergantung pada taraf kesukaran item-itemnya. Sebaliknya pula, item-item sebuah tes abilitas yang sama bisa dipastikan akan terbukti memiliki taraf kesukaran dan daya beda yang tidak sama jika diadministrasikan pada dua sampel dengan karakteristik yang berbeda. Terkait faktor sampel, teori dan model-model tes klasik ini memiliki daya guna tertinggi manakala sampel testi memiliki kesamaan karakteristik dengan populasi testi yang menjadi sasaran penyusunan sebuah tes (Hambleton & Jones, 1993).

Ketiga, kelebihan utama teori dan model-model tes klasik adalah bahwa teori ini beserta aneka model turunannya didasarkan pada asumsi-asumsi yang relatif lemah atau longgar sehingga mudah dipenuhi oleh data tes yang lazim kita peroleh. Selain itu, teori dan aneka model tes klasik ini memiliki sejarah yang amat panjang, boleh dikatakan identik dengan sejarah perkembangan metode statistik. Sebagaimana diketahui, statistik sebagai disiplin ilmu mengalami perkembangan pesat sejak ditemukannya konsep *errors in scientific observations* atau kesalahan dalam pengamatan ilmiah sekitar awal abad ke-20. Berikut adalah sejumlah peristiwa yang menjadi tonggak sejarah perkembangan statistik pada umumnya maupun

teori tes klasik pada khususnya (Traub, 1997). *Carl Friedrich Gauss* menemukan distribusi normal saat berusaha membuktikan bahwa *mean* dari banyak pengamatan tentang sebuah kuantitas yang tidak diketahui merupakan kemungkinan nilai yang sesungguhnya dari kuantitas yang bersangkutan. Francis Galton menemukan konsep *correlation* atau korelasi beserta lambang *r*-nya sekalipun mula-mula Galton menggunakan lambang tersebut untuk menjelaskan gejala *regression* (perhatikan, *r* adalah huruf pertama kata "regression" atau "reversion"). Orang yang berjasa pertama kali menyebut lambang *r* sebagai *coefficient of correlation* atau koefisien korelasi adalah *Francis Y. Edgeworth* pada tahun 1892. Pada tahun 1896 *Karl Pearson* berhasil membuktikan bahwa nilai terbaik dari *r* diperoleh dengan cara membagi kovarians dengan jumlah kuadrat deviasi-deviasi standar. Pada 1904 *Charles Spearman* menemukan bahwa koefisien korelasi antar hasil pengukuran sebuah atribut psikologis yang dilakukan secara independen terhadap sekelompok testi tidaklah sempurna dalam arti bervariasi atau berfluktuasi secara random. Maka dia mengusulkan sebuah teknik untuk melakukan *correction for attenuation* atau koreksi terhadap gejala atenuasi atau pelemahan (korelasi). Pada tahun 1910 secara sendiri-sendiri namun dalam waktu bersamaan *Charles Spearman* dan *W. Brown* menemukan rumus untuk menghitung koefisien reliabilitas sebuah tes komposit berdasarkan dua belahannya. Hingga kini rumus ini dikenal sebagai *Spearman-Brown formula* atau rumus Spearman-Brown. Hingga sekarang, banyak tes penting disusun berdasarkan model tes klasik (Hambleton & Jones, 1993).

Kembali pada tema pokok pembahasan kita, bagaimana Rumus 1 yang merupakan proposisi utama model tes klasik tersebut bisa diselesaikan? Di muka sudah disinggung bahwa di antara tiga komponen rumus tersebut, skor tampak *X* diketahui sedangkan sesuai Asumsi 2 skor murni *T* merupakan konstruk teoretis yang mustahil dihitung secara nyata. Berarti satu-satunya tugas yang harus kita lakukan adalah menghitung atau setidaknya mengestimasikan besar skor kesalahan *E* untuk mengetahui ketepatan hasil pengukuran

yang kita peroleh berdasarkan skor tampak *X*. Pada kenyataannya, inilah inti teori dan aneka model tes klasik yaitu menilai sejauh mana sebuah skor tes (skor tampak *X*) sungguh-sungguh mencerminkan atribut yang hendak diukur oleh tes dengan cara mengestimasikan besar skor kesalahan *E*-nya (De Champlain, 2010).

Salah satu cara mengestimasikan besarnya skor kesalahan *E* adalah dengan menghitung apa yang disebut *standard error of measurement* (*SEM*) atau kesalahan pengukuran baku. *SEM* memberikan estimasi skor kesalahan *E* berupa *expected measure of error* atau besar kesalahan yang diharapkan dan yang bisa dihitung dengan rumus seperti disajikan pada Rumus 6.2. (De Champlain, 2010).

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{XX'}}$$ **Rumus 6.2.**

σ_X = *SD* skor total tes
$\rho_{xx'}$ = estimasi koefisien reliabilitas

Sesuai Asumsi 2, rasional dari rumus di atas adalah sebagai berikut: seandainya kita bisa mengetes seorang testi dengan tes yang sama dalam jumlah kali tak terbatas (dengan asumsi setiap pengetesan bersifat independen) maka *mean* dari seluruh skor yang diperoleh adalah sama dengan skor murni *T* sedangkan *SD*-nya adalah sama dengan *SEM.* Namun ada satu komponen dalam Rumus 6.2. yang belum kita ketahui, yaitu estimasi koefisien reliabilitas $\rho_{xx'}$. Pada bab berikut kita akan membahas bagaimana model tes klasik menjelaskan konsep realiabilitas dan cara menghitung atau mengestimasikan koefisien reliabilitas. Ψ

Bab 7
Reliabilitas dalam Model Tes Klasik

Bertolak dari pembahasan kita dalam bab terdahulu, secara umum dapat dikatakan bahwa reliabilitas adalah ketepatan pengukuran tanpa menghiraukan atribut apa yang diukur (Nunnally, 1970). Sebagaimana sudah kita lihat, sejumlah *random measurement error* (*RME*) senantiasa terjadi dalam setiap pengukuran. *RME* semacam ini akan mengurangi ketepatan pengukuran baik secara positif sehingga menghasilkan *over-estimation* maupun secara negatif sehingga menghasilkan *under-estimation* tentang taraf pemilikan atribut yang sesungguhnya.

Secara psikometrik reliabilitas memiliki dua makna (Klein, 1986), yaitu (1) *self-consistency* atau konsistensi internal, dan (2) stabilitas, tes. Konsistensi internal adalah kesesuaian antar bagian-bagian dalam suatu tes. Mengutip pendapat Kline (1986), "if part of a test is measuring a variable, then the other parts, if not consistent with it, cannot be measuring that variable" (h. 2). Maksudnya, jika salah satu bagian dari sebuah tes mengukur suatu variabel tertentu, maka bagian-bagian lainnya jika tidak konsisten dengan bagian yang disebut pertama pastilah tidak mengukur variabel yang sama. Reliabilitas yang didasarkan pada kesesuaian antar bagian-bagian dalam suatu tes semacam ini dikenal sebagai *reliabilitas konsistensi internal*. Menurut Klein (1986), konsep reliabilitas konsistensi internal inilah yang mendasari prinsip umum dalam psikometri yang menyatakan bahwa "high reliability is a prerequisite of validity" (h. 2). Maksudnya, reliabilitas (konsistensi internal yang tinggi) merupakan salah satu prasyarat validitas.

Ada pakar psikometri lain, yaitu Cattell (dalam Klein, 1986), yang tidak sependapat dengan gagasan di atas. Menurut Cattell, "maximum validity is obtained where test items do not all correlate

with each other, but where each correlates positively with the criterion. Such a test would have only low internal consistency reliability" (dalam Klein, 1986, h. 3). Maksudnya, validitas maksimum justru akan diperoleh manakala item-item tes tidak saling berkorelasi satu sama lain melainkan masing-masing item tersebut berkorelasi positif dengan kriterianya; tes semacam itu hanya akan memiliki reliabilitas konsistensi internal yang rendah. Argumen Cattell adalah sebagai berikut: setiap item dalam sebuah tes memiliki cakupan yang lebih sempit dibandingkan dengan kriterianya yang hendak diukur. Jika semua item dalam tes sangat konsisten satu sama lain berarti juga saling sangat berkorelasi satu sama lain. Akibatnya, tes yang reliabel dalam arti reliabilitas konsistensi internal justru hanya mengungkap cakupan sempit dari variabel atau kriteria yang hendak diukur. Hal ini bertolak belakang dengan validitas yang mengandaikan kemampuan sebuah tes mengungkap cakupan yang luas dari variabel atau kriteria yang hendak diukur. Menurut Klein (1986), kendati kritik seperti yang dilontarkan Cattell tersebut sangat beralasan namun dalam praktik dalil umum psikometri bahwa tes yang valid lazimnya memiliki konsistensi (internal) yang tinggi tetap diterima secara luas.

Stabilitas adalah kesamaan skor yang dicapai oleh setiap testi yang sama dalam pengetesan ulang seperti skor yang dicapai dalam pengetesan pertama atau sebelumnya. Reliabilitas antar waktu atau *test-retest reliability,* yaitu kesamaan hasil yang dicapai oleh seorang testi dalam berbagai kesempatan pengetesan dengan tes yang sama ini dipandang sangat esensial dimiliki oleh setiap tes. Reliabilitas ini diperiksa dengan cara mengkorelasikan skor tes yang sama yang diberikan dalam dua kali pengetesan terhadap sekelompok subjek yang sama. Menurut seorang pakar psikometri Guilford, batas minimum koefisien korelasi sebagai bukti *test-retest reliability* yang dipandang cukup memuaskan adalah 0,70. Sebuah tes yang memiliki *test-retest reliability* kurang dari 0,70 dipandang kurang bermanfaat sebab hal itu berarti bahwa *standard error* atau kesalahan baku yang terkandung dalam *skor tampak* adalah sedemikian besar sehingga sulit ditafsirkan (Klein, 1986).

Makna ketiga dari reliabilitas yang merupakan gabungan dari unsur konsistensi internal dan stabilitas sekaligus terdapat dalam konsep *parallel-form reliability* atau reliabilitas bentuk paralel (Klein, 1986). Dua tes yang terdiri dari item-item yang dibuat paralel atau ekuivalen dalam rangka mengukur suatu atribut yang sama diadministrasikan kepada sekelompok subjek yang sama dalam kesempatan yang berbeda. Koefisien korelasi antara skor kedua tes semacam itu menunjukkan konsistensi internal tes yang bersangkutan manakala kedua tes itu dipandang tunggal dari sudut ekuivalensi item-itemnya, sekaligus menunjukkan stabilitas antar waktu dari tes yang bersangkutan manakala dipandang dari sudut perbedaan waktu pengetesan dari tes yang pada hakikatnya tunggal sebab memiliki item-item yang ekuivalen. Namun seperti sudah disinggung di muka, dalam praktik tidak mudah kalau bukan mustahil menyusun dua tes yang benar-benar paralel.

Ketiga makna di atas mendasari aneka strategi untuk mengestimasi taraf reliabilitas tes yang lazim dinyatakan dalam suatu koefisien korelasi dan yang disebut **koefisien reliabilitas.** Aneka strategi mengestimasi taraf reliabilitas tes tersebut akan kita bahas dalam bagian selanjutnya. Namun sebelum sampai ke sana, terlebih dulu akan disajikan beberapa cara menginterpretasikan koefisien reliabilitas berdasarkan model tes klasik.

A. Beberapa Cara Menafsirkan Koefisien Reliabilitas menurut Model Tes Klasik

Konsep reliabilitas atau lebih tepat *unreliability* atau ketiadaan reliabilitas tercakup dalam postulat dasar model tes klasik yang tertuang dalam *Rumus 6.1.* ($X = T + E$). Dibaca dengan cara lain, menurut model tes klasik skor-skor tes mencerminkan pengaruh dari dua faktor (Friedenberg, 1995), yaitu: (a) karakteristik stabil yang terdapat dalam diri testi (karakteristik murni), dan (b) karakteristik

berupa peristiwa *random* atau sembarang yang berasal baik dari dalam diri testi maupun yang berasal dari situasi pengetesannya (*random measurement error,* disingkat *RME*). Peristiwa random atau sembarang adalah peristiwa yang muncul secara tidak menentu pada berbagai kesempatan pengadministrasian sebuah tes yang sama pada sekelompok testi yang sama. Dampak peristiwa random ini berupa *RME* yang mengakibatkan hasil pengukuran tidak reliabel, dalam arti skor tes seorang testi akan berfluktuasi baik secara positif (skor meningkat) maupun secara negatif (skor menurun) kendati dites dengan tes yang sama namun pada kesempatan yang berlainan.

Ada enam cara dalam menafsirkan koefisien reliabilitas menurut model tes klasik (Allen & Yen, 1979):

1. $\rho_{XX'}$. Reliabilitas suatu tes sama dengan korelasi antara skor-skor tampak tes yang bersangkutan dengan skor-skor tampak suatu tes paralelnya.
2. $\rho^2_{XX'}$. Cara kedua dalam menafsirkan koefisien reliabilitas ini merupakan cara baku dalam menafsirkan koefisien korelasi Pearson, yaitu bahwa korelasi antara dua variabel yang dikuadratkan dapat ditafsirkan sebagai proporsi varians dalam salah satu variabel yang dijelaskan oleh hubungan linear dengan variabel lainnya.
3. $\rho_{XX'} = \sigma^2_T/\sigma^2_X$. Koefisien reliabilitas sama dengan rasio varians skor murni terhadap varians skor tampak atau proporsi varians skor tampak yang merupakan varians skor murni. Jika $\rho_{XX'} = 1$, yang berarti reliabilitas tes sempurna, maka $\sigma^2_T/\sigma^2_X = 1$; berarti $\sigma^2_X = \sigma^2_T$; maka $\sigma^2_E = 0$. Sebaliknya jika $\rho_{XX'} = 0$, yang berarti tes tidak memiliki reliabilitas sama sekali, berarti $\sigma^2_X = \sigma^2_E$, maka $\sigma^2_T = 0$. Dengan kata lain, makin tinggi reliabilitas suatu tes, varians skor murni semakin tinggi sedangkan varians skor kesalahan semakin kecil. Jika skor kesalahan makin kecil, maka skor tampak seorang testi makin mendekati skor murninya. Sebaliknya jika skor kesalahan besar, maka skor tampak makin jauh dari skor murninya. Artinya, jika sekor kesalahan besar berarti skor tampak

tidak memberikan estimasi yang baik terhadap skor murni testi (Allen & Yen, 1979).

4. $\rho_{XX'} = \rho^2_{XT}$. Reliabilitas suatu tes sama dengan kuadrat korelasi antara skor-skor tampak dan skor-skor murninya. Artinya, suatu skor tampak akan memiliki korelasi yang lebih tinggi dengan skor murninya sendiri daripada dengan skor tampak suatu tes paralelnya. Maka, korelasi maksimum antara suatu skor tampak dan variabel lain sama dengan $\sqrt{\rho_{XX'}}$ dan berarti sama dengan ρ_{XT}. Jika suatu tes, *X,* dipakai untuk memprediksi suatu kriteria, *Y,* maka ρ_{XY} disebut koefisien validitas. Karena ρ_{XY} tidak mungkin lebih besar daripada ρ_{XT} berarti ρ_{XY} tidak mungkin lebih besar dari $\sqrt{\rho_{XX'}}$. Dengan kata lain, *unreliability* atau reliabilitas (yang rendah) mempengaruhi validitas. Allen dan Yen (1979) memberikan catatan, kendati koefisien validitas tidak mungkin melebihi akar koefisien reliabilitas, namun koefisien validitas bisa melebihi koefisien reliabilitas itu sendiri.
5. $\rho_{XX'} = 1 - \rho^2_{XE}$. Koefisien reliabilitas sama dengan 1 dikurangi kuadrat korelasi antara skor tampak dan skor kesalahan.
6. $\rho_{XX'} = 1 - \sigma^2_E/\sigma^2_X$. Koefisien reliabilitas sama dengan 1 dikurangi proporsi varians kesalahan dalam varians skor tampak. Sebagaimana sudah ditunjukkan, jika $\rho_{XX'} = 1$, maka $\sigma^2_E = 0$, sedangkan jika $\rho_{XX'} = 0$, maka $\sigma^2_E = \sigma^2_X$. Dari sini bisa dipahami dampak taraf homogenitas-heterogenitas sampel yang dipakai dalam pemeriksaan reliabilitas terhadap besarnya koefisien reliabilitas. Dalam sampel yang bersifat homogen varians skor tampaknya σ^2_X adalah kecil. Jika varians skor kesalahan σ^2_E pada sampel yang homogen sama besar seperti pada sampel yang heterogen, maka koefisien reliabilitas pada sampel yang homogen akan lebih kecil dibandingkan pada sampel yang heterogen. Bisa disimpulkan, estimasi realiabilitas dengan menggunakan sampel yang heterogen cenderung lebih tinggi dibandingkan estimasi reliabilitas dengan menggunakan sampel yang homogen (Allen & Yen, 1979).

B. Kesimpulan Umum tentang Reliabilitas

Berdasarkan aneka cara menafsirkan reliabilitas sebagaimana diuraikan di atas, dapat dirumuskan sejumlah kesimpulan umum tentang reliabilitas sebagai berikut (Allen & Yen, 1979):

a. Jika $\rho_{XX'} = 1$, atau reliabilitas tes sempurna, maka:
   1) Pengukuran berlangsung tanpa kesalahan (semua $E = 0$).
   2) Untuk semua testi, $X = T$.
   3) Semua varians skor tampak mencerminkan varians skor murni ($\sigma_X^2 = \sigma_T^2$).
   4) Semua perbedaan atau selisih antar skor tampak mencerminkan perbedaan antar skor murni.
   5) Korelasi antara skor-skor tampak dan skor-skor murni sama dengan 1 ($\rho_{XT} = 1$).
   6) Korelasi antara skor-skor tampak dan skor-skor kesalahan sama dengan 0 ($\rho_{XE} = 0$).

b. Jika $\rho_{XX'} = 0$, atau tes sama sekali tidak memiliki reliabilitas, maka:
   1) Pengukuran hanya berisikan kesalahan random.
   2) Untuk semua testi, $X = E$.
   3) Semua varians skor tampak mencerminkan varians skor kesalahan ($\sigma_X^2 = \sigma_E^2$).
   4) Semua perbedaan atau selisih antar skor tampak mencerminkan kesalahan pengukuran.
   5) Korelasi antara skor-skor tampak dan skor-skor murni sama dengan 0 ($\rho_{XT} = 0$).
   6) Korelasi antara skor-skor tampak dan skor-skor kesalahan sama dengan 1 ($\rho_{XE} = 1$).

c. Jika $0 \geq \rho_{XX'} \geq 1$, atau reliabilitas tes antara 0 dan 1, maka:
   1) Pengukuran mengandung sejumlah kesalahan.
   2) Untuk semua testi, $X = T + E$.
   3) Semua varians skor tampak mengandung baik varians skor murni maupun varians skor kesalahan ($\sigma_X^2 = \sigma_T^2 + \sigma_E^2$).

4) Perbedaan antar skor tampak mencerminkan baik kesalahan pengukuran maupun perbedaan skor murni.
5) Korelasi antara skor-skor tampak dan skor-skor murni sama dengan akar korelasi antara skor-skor tampak suatu tes dan skor-skor tampak tes paralelnya ($\rho_{XT} = \sqrt{\rho_{XX'}}$).
6) Korelasi antara skor-skor tampak dan skor-skor kesalahan sama dengan akar 1 dikurangi korelasi antara skor-skor tampak suatu tes dan skor-skor tampak tes paralelnya ($\rho_{XE} = \sqrt{1 - \rho_{XX'}}$).
7) Reliabilitas merupakan proporsi varians skor murni dalam varians skor tampak ($\rho_{XX'} = \sigma_T^2/\sigma_X^2$).
8) Makin besar $\rho_{XX'}$, makin tinggi kepastian kita dalam mengestimasikan *T* berdasarkan *X*, sebab varians kesalahannya relatif makin kecil.

Kesimpulan a dan b secara konseptual-teoretis terbayangkan, namun dalam praktik mustahil ditemukan. Kesimpulan c secara konseptual-teoretis masuk akal dan secara praktis merupakan kelaziman. Bahkan sebagaimana sudah kita lihat, itulah esensi model tes klasik dan yang telah merangsang perkembangan psikometri hingga kini.

C. Aneka Sumber *Unreliability* menurut Model Tes Klasik

Apa saja yang menjadi biang keladi tidak reliabelnya tes? Secara garis besar, sumber ancaman terhadap reliabilitas suatu tes dapat dikelompokkan dalam tiga kategori, yaitu (Friedenberg, 1995): (a) aneka faktor yang menimbulkan *inconsistent performance* atau kinerja yang tidak konsisten, baik yang berasal dari pihak testi seperti perubahan nyata pada atribut dalam diri testi yang menjadi sasaran pengukuran, perubahan sembarang dalam diri testi akibat berbagai peristiwa (sakit, perubahan suasana hati, perubahan motivasi, dan

sebagainya), dan perbedaan sembarang pada cara testi memberikan respon terhadap item-item tes; maupun yang berasal dari proses administrasi tes dan penskorannya, seperti pengadministrasian tes yang menjadi tidak baku karena gagal mengikuti berbagai petunjuk yang termuat dalam manual tesnya, atau penskoran yang tidak baku bahkan keliru karena gagal mengikuti pedoman akibat kesengajaan maupun ketidaktahuan; (b) faktor *content sampling* yang tidak representatif sebagaimana diwakili oleh item-item tes; seperti yang berlaku umum dalam *sampling* atau pengambilan sampel, makin besar sampel makin baik dalam arti makin representatif sampel tersebut; begitu juga dalam tes, makin besar sampel seperti antara lain tercermin dari jumlah item yang semakin banyak sehingga tes juga menjadi semakin panjang maka kemungkinan sampel item tersebut semakin merepresentasikan populasi itemnya akan semakin besar; tentu selain representatif dalam segi jumlah tes yang baik juga harus representatif dari segi struktur *universe of content* dari atribut yang diukur; namun secara umum bisa dikatakan, makin pendek tes makin besar kemungkinan terjadi apa yang disebut *construct underrepresentation* atau kegagalan tes sebagai sampel merepresentasikan konstruk yang diukur (AERA, APA, & NCME, 1999); dan (c) faktor statistis, khususnya terkait variabilitas skor akibat sifat homogenitas/ heterogenitas sampel testinya; sebagaimana sudah ditunjukkan dalam *Interpretasi 6* tentang reliabilitas, estimasi reliabilitas dengan menggunakan sampel testi yang heterogen cenderung lebih tinggi dibandingkan estimasi reliabilitas dengan menggunakan sampel testi yang homogen; dengan kata lain, makin besar variabilitas skor maka akan semakin tinggi pula reliabilitas tes.

Di bawah ini diuraikan sejumlah sumber khusus *unreliability* tes yang berasal baik dari pihak penyelenggara tes, testi, maupun dari faktor tesnya sendiri (Klein, 1986; Friedenberg, 1995):

a. ***Subjektivitas dalam penilaian.*** Yang dimaksud adalah kecenderungan penilai melakukan penskoran secara subjektif terhadap hasil pengukuran. Akibatnya, terjadi perbedaan penskoran terhadap hasil tes yang sama oleh sejumlah penilai,

bahkan perbedaan penskoran terhadap hasil tes yang sama oleh penilai yang sama pada kesempatan yang berbeda. Bahaya ini terutama menjadi ancaman serius bagi tes dengan format esai subjektif. Namun jika juga terjadi pada tes dengan format objektif misal karena faktor ketidaktelitian, tes semacam itu dipastikan akan memiliki korelasi antar item yang rendah atau memiliki konsistensi internal yang rendah (Klein, 1986).

b. ***Guessing*** atau menebak. Dalam menjawab tes abilitas dengan item-item objektif, kecenderungan testi menebak jawaban yang benar bisa mendistorsi skor yang dia peroleh, sehingga juga bisa berakibat merugikan reliabilitas maupun validitas tes atau pengukurannya. Untuk mengatasi kecenderungan ini, ada yang mengusulkan penerapan *rumus koreksi tebakan* dalam penskoran sebagai berikut (Klein, 1986):

$$X_k = X - \frac{W}{n-1} \qquad \textbf{Rumus 7.1.}$$

X_k = Skor yang dikoreksi dari kemungkinan menebak
X = Jumlah jawaban benar.
W = Jumlah jawaban salah.
n = Jumlah opsi atau alternatif jawaban dalam masing-masing item.

Namun mengikuti pendapat Vernon, seorang pakar psikometri lain, akhirnya Klein (1986) berpendapat bahwa kecenderungan menebak jawaban benar bukan masalah serius dalam penyelenggaraan tes. Penerapan rumus koreksi tebakan pun hanya nyata manfaatnya dalam penskoran tes dengan item-item benar-salah. Namun berdasarkan bukti lapangan tes dengan format item benar-salah semacam itu sesungguhnya tidak direkomendasikan untuk diterapkan.

c. ***Item-item yang kabur.*** Item-item yang dirumuskan secara kabur sehingga sulit dipahami atau bermakna ganda juga akan berdampak menurunkan reliabilitas. Khususnya dalam

tes kepribadian yang berbentuk inventori, item-item yang dirumuskan secara jelas terbukti meningkatkan reliabilitas tes kepribadian (Klein, 1986).

d. ***Panjang tes.*** Makin panjang sebuah tes, makin meningkat reliabilitasnya. Makin panjang tes kiranya juga akan makin terjamin *construct representation* atau keterwakilan (*content* atau isi) *construct*-nya sehingga juga makin terjamin ketepatan pengukurannya sehingga makin reliabel. Jumlah minimum item untuk mendapatkan tes yang reliabel adalah 20 buah (Klein, 1986).

e. ***Rumusan item.*** Rumusan item-item yang jelas dan tidak kabur atau mendua bisa meningkatkan reliabilitas tes. Cara merumuskan item harus sesuai dengan *content* yang hendak diukur. Dalam tes abilitas, *content* yang sederhana bisa menjelma menjadi sebuah item dengan taraf kesulitan tinggi akibat perumusan yang tidak efektif. Dalam semua jenis tes, rumusan item yang kabur akan menghasilkan jawaban yang tidak konsisten dari pihak testi.

f. ***Test-retest unreliability*** atau reliabilitas yang rendah pada dua kesempatan pengetesan dengan tes yang sama. Secara garis besar sumber timbulnya gejala ini bisa dibedakan menjadi dua, yaitu (1) faktor kesalahan pengukuran (*random measurement error)*, misal berupa penskoran yang subjektif atau perubahan sembarang pada aspek tertentu dalam diri testi; atau (2) faktor perubahan nyata status testi terkait atribut yang sedang diukur, misal berupa peningkatan pengetahuan atau kematangan emosi seiring pertambahan usia. Kasus yang kedua tidak boleh dipersamakan dengan gejala *unreliability* akibat *random measurement error* yang merupakan pengertian umum tentang reliabilitas (Klein, 1986).

g. ***Aneka sumber lain.*** Berbagai sumber lain yang menjadi biang keladi rendahnya reliabilitas tes lazimnya dikaitkan dengan aneka faktor dalam diri subjek yang mengakibatkan *inconsistent performance* sebagaimana sudah disinggung di awal bagian ini. Maka, kiranya tidak perlu lagi disinggung secara panjang lebar di sini.

D. Pemeriksaan Reliabilitas Menurut Model Tes Klasik

Secara umum semua analisis reliabilitas melibatkan penggunaan dua rangkaian skor atau hasil pengukuran, lalu menghitung koefisien korelasi antara kedua rangkaian skor tersebut (Friedenberg, 1995). Dua rangkaian skor yang dimaksud bisa diperoleh dengan beberapa cara, yaitu (1) dengan mengadministrasikan ulang tes yang sama pada sekelompok testi yang sama; (2) mengadministrasikan tes dan bentuk paralelnya pada sekelompok testi pada kesempatan yang berbeda; atau (3) membelah menjadi dua rangkaian skor hasil satu kali pengetesan dengan sebuah tes pada sekelompok subjek yang sama. Marilah kita bahas ketiga metode pemeriksaan reliabilitas tersebut satu demi satu.

1. Estimasi Reliabilitas *Test-Retest*

Metode ini pertama kali dan kiranya secara tidak sengaja diperkenalkan oleh William Brown (1910, dalam Traub, 1997), manakala dia mendefinisikan *reliability coefficient* atau koefisien reliabilitas sebagai "the correlation between scores on repeated administrations of the same test" (h. 11). Artinya, koefisien reliabilitas adalah korelasi antara skor-skor yang diperoleh dari pengetesan ulang dengan tes yang sama. Reliabilitas *test-retest* menghasilkan sebuah estimasi reliabilitas $r_{xx'} = \rho_{XX'}$ yang diperoleh dari pengetesan kelompok testi yang sama dua kali dengan tes yang sama kemudian menghitung korelasi hasil pengukurannya. Metode ini memiliki dua kelemahan pokok. Pertama, jika jeda waktu antara pengetesan pertama dan kedua terlalu singkat, sangat mungkin terjadi *carry-over effect* antar pengetesan. Artinya, pengetesan pertama bisa mempengaruhi secara positif hasil pengetesan kedua, antara lain karena pengaruh ingatan dan *practice effects* atau pengaruh latihan. Kedua, jika jeda waktu antara pengetesan pertama dan kedua terlalu panjang, bisa terjadi perubahan dalam bentuk peningkatan nyata dalam diri testi terkait

atribut yang menjadi sasaran pengukuran. Sebagai contoh, pada kelompok subjek anak-anak prasekolah dalam jeda antar pengetesan selama dua tahun bisa terjadi peningkatan penguasaan kosa kata atau kemampuan berhitung secara signifikan. Maka, metode estimasi reliabilitas *test-retest* ini dipandang paling cocok diterapkan untuk memeriksa reliabilitas tes yang mengukur atribut psikologis yang tidak rentan terhadap pengaruh *carry-over* dan bersifat stabil (Allen & Yen, 1979).

2. Estimasi Reliabilitas Bentuk-bentuk Paralel dan Bentuk-bentuk Alternatif

Metode ini pertama kali diperkenalkan oleh T. Kelley yang mendefinisikan reliabilitas sebagai koefisien korelasi antara "comparable tests" atau tes-tes yang sebanding, sepadan, atau setara. Kini istilah yang lebih populer adalah bentuk-bentuk paralel tes, sehingga reliabilitas bentuk paralel tes adalah korelasi, $r_{xx'}$, antara skor-skor tampak dua tes yang paralel. Karena, sebagaimana sudah beberapa kali kita lihat, amat sulit mendapatkan dua tes yang paralel, maka sebagai gantinya dipakai bentuk-bentuk alternatif tes. Bentuk-bentuk alternatif tes adalah dua bentuk tes yang disusun dalam rangka menjadikan keduanya paralel sehingga memiliki *mean-mean* skor tampak, varians, dan korelasi dengan tes lain yang sama atau setidaknya sangat mendekati sama. Korelasi antara skor-skor tampak kedua bentuk alternatif tes ini, r_{xz}, merupakan estimasi reliabilitas masing-masing bentuk alternatif tes tersebut. Korelasi ini menunjukkan sejauh mana kedua bentuk tes itu reliabel sekaligus sejauh mana kedua bentuk tes itu paralel (Allen & Yen, 1979). Seperti metode *test-retest,* metode ini juga rentan terhadap *carry-over effects* jika jeda waktu antar pengetesan terlalu singkat, sebaliknya rentan terhadap efek perubahan nyata pada atribut yang diukur jika jeda waktu antar pengetesan terlalu lama. Selain itu jika kedua bentuk alternatif tes, *X* dan *Z,* benar-benar tidak paralel, maka r_{xz} bukan merupakan estimasi yang baik terhadap baik ρ_{xx} maupun ρ_{zz}.

Maka, korelasi antara skor tampak bentuk-bentuk alternatif tes hanya akan menghasilkan estimasi yang baik terhadap reliabilitas kedua tes yang bersangkutan jika bentuk-bentuk alternatif tersebut sungguh-sungguh paralel dan bila korelasi itu tidak dicemari oleh pengaruh *carry-over effects* maupun oleh perubahan skor akibat perubahan nyata pada atribut yang diukur (Allen & Yen, 1979).

3. Estimasi Reliabilitas Konsistensi-Internal

Berbeda dengan dua metode estimasi reliabilitas sebelumnya yang didasarkan pada dua rangkaian skor hasil dua kali pengadministrasian tes yang sama atau satu tes dan bentuk paralelnya, estimasi reliabilitas konsistensi internal cukup didasarkan pada hasil satu kali pengadministrasian tes. Reliabilitas tesnya sendiri selanjutnya dapat diestimasi dengan dua cara atau metode, yaitu: (a) metode *split-half* atau metode belah dua; dan (b) metode pembelahan sebanyak item yang ada atau lazim disebut metode yang didasarkan pada kovarians item (Crocker & Algina, 2008).

a. Metode Belah-Dua

Sepasang tokoh yang berjasa mempromosikan metode belah dua adalah *Marion Richardson* dan *Frederic Kuder*, yang hingga kini dikenal sebagai penemu dua rumus estimasi reliabilitas konsistensi internal yang dinamai dengan huruf pertama nama masing-masing yaitu Rumus KR 20 dan KR 21. Terhadap metode estimasi reliabilitas *test-retest* (sehingga juga metode bentuk-bentuk paralel atau bentuk-bentuk alternatif), mereka memberikan catatan kritis sebagai berikut:

> "[Using] the same form gives, in general, estimates that are too high because of material remembered on the second application of the test. This memory factor cannot be eliminated by increasing the length of time between the two applications, because of variable growth in the function tested within the

> population of individuals. These difficulties are so serious that the method is rarely used" (dalam Traub, 1997, h. 11).

Intinya, mereka menyatakan bahwa metode *test-retest* menghasilkan estimasi reliabilitas yang terlampau tinggi, maka metode ini jarang digunakan. Sebagai gantinya, mereka merekomendasikan metode *split-half* atau belah dua dalam melakukan estimasi reliabilitas sebab pada kenyataannya metode inilah yang paling banyak digunakan di kalangan para pakar psikometri (Traub, 1979).

Dalam metode ini reliabilitas konsistensi internal diestimasi dengan menggunakan rangkaian skor hasil satu kali pengadministrasian tes untuk menghindari aneka distorsi yang mungkin timbul dalam dua pengadministrasian tes baik dengan tes yang sama maupun dengan bentuk paralelnya. Untuk mendapatkan dua rangkaian skor dari satu kali pengadministrasian satu tes, tes tersebut dibelah menjadi dua bagian. Ada lima metode yang lazim digunakan dalam membelah tes (Allen & Yen, 1979; Crocker & Algina, 2008), yaitu:

(1) *Metode gasal-genap*, yaitu semua item bernomor gasal digabungkan membentuk belahan pertama sedangkan semua item bernomor genap digabungkan membentuk belahan kedua.
(2) *Metode belah tengah,* yaitu tes dibelah persis di tengah: item pertama sampai dengan item yang berada tepat di tengah membentuk belahan pertama, item berikutnya sampai dengan item terakhir membentuk belahan kedua.
(3) *Metode gabungan urutan taraf kesukaran dan gasal genap,* yaitu item-item terlebih dulu diurutkan berdasarkan taraf kesukaran mulai dari yang paling mudah sampai ke yang paling sukar atau sebaliknya, sesudah itu dilakukan pembelahan dengan metode gasal genap sebagaimana sudah diuraikan.
(4) *Metode random,* yaitu secara random membagi item-item menjadi dua belahan.
(5) *Metode matched random subsets.* Metode pembelahan item temuan Gulliksen (1950, dalam Allen & Yen, 1979) ini meliputi sejumlah

langkah. Pertama, dihitung dua macam statistik untuk masing-masing item, yaitu taraf kesukaran (*p*) dan korelasi (*biserial* atau *point-biserial*) antara item dan skor total (r_{it}). Kedua, masing-masing item di-*plot* dalam sebuah grafik berdasarkan kedua statistik tersebut, yaitu *p* sebagai aksis dan r_{it} sebagai ordinat. Ketiga, dua item yang terletak berdekatan dalam grafik dipasangkan. Selanjutnya, keempat, dengan cara tertentu yang diterapkan secara konsisten satu item dalam setiap pasangan digabung membentuk belahan pertama sedangkan satu item lainnya dalam setiap pasangan digabung membentuk belahan kedua.

Dua bagian atau belahan tes yang terbentuk menjadi sejenis bentuk alternatif satu sama lain dan harus diusahakan agar kedua bagian tersebut paralel atau setidaknya secara esensial τ-ekuivalen. Korelasi antara kedua bagian tes tersebut akan menghasilkan estimasi reliabilitas separuh tes ($\rho_{YY'}$). Untuk mendapatkan estimasi reliabilitas keseluruhan tesnya perlu diterapkan **rumus Spearman-Brown** sebagai berikut (Allen & Yen, 1979):

$$\rho_{XX'} = \frac{2\rho_{YY'}}{1+\rho_{YY'}} \qquad \textbf{Rumus 7.2.}$$

Salah satu kelemahan utama metode belah tengah dengan rumus Spearman-Brown di atas ialah bahwa metode ini tidak menghasilkan estimasi unik atau tunggal terhadap koefisien reliabilitas tes secara keseluruhan. Kelima cara membelah tes di atas akan menghasilkan koefisien reliabilitas separuh tes yang berlainan tinggi-rendahnya sehingga ketika dikenai rumus Spearman-Brown akan menghasilkan estimasi koefisien reliabilitas keseluruhan tes yang berlainan pula. Untuk mengatasi problem ini, sejumlah pakar psikometri berusaha merumuskan metode lain yang didasarkan pada pembelahan tes tidak hanya menjadi dua bagian melainkan menjadi sekian bagian sebanyak jumlah itemnya. Tepatnya, metode baru estimasi reliabilitas ini didasarkan pada kovarians item-itemnya (Crocker & Algina, 2008).

b. Metode Berbasis Kovarians Item

Ada tiga metode berbasis kovarians item yang dipakai secara luas untuk mengatasi kelemahan metode belah dua Spearman-Brown, yaitu (1) *Kuder-Richardson 20,* (2) *alpha Cronbach,* dan (3) analisis varians Hoyt. Kendati berlainan bentuknya, ketiga metode tersebut diakui menghasilkan estimasi reliabilitas yang sama atau setara dan koefisien reliabilitas yang dihasilkan ketiganya secara bersama-sama disebut ***koefisien alpha*** (Crocker & Algina, 2008). Kita akan melihatnya satu demi satu.

1) Alpha Cronbach. Pada tahun 1951 Cronbach menyajikan sintesis dari aneka metode estimasi konsistensi internal ke dalam sebuah rumus umum yang selanjutnya dikenal sebagai *alpha Cronbach* dan yang dihitung dengan rumus sebagai berikut (Crocker & Algina, 2008; h. 138):

$$\alpha = \frac{k}{k-1}\left(1 - \sum \alpha_i^2 / \alpha_x^2\right) \qquad \textbf{Rumus 7.3.}$$

k = jumlah item dalam tes
α_i^2 = varians item i
α_x^2 = varians keseluruhan tes

Alpha Cronbach bisa diterapkan untuk mengestimasikan konsistensi internal item-item yang diskor secara dikotomis maupun item-item yang diskor dengan skala yang lebih luas, termasuk penskoran pada inventori kepribadian bahkan pada ujian esai (Crocker & Algina, 2008)

2) Rumus Kuder-Richardson. Kendati pada mulanya kedua pakar psikometri ini mempromosikan metode belah dua namun sesudah menyadari kelemahannya mereka berusaha menemukan rumus baru untuk mengatasi kelemahan pokok metode belah dua, yaitu kegagalan metode tersebut menghasilkan sebuah koefisien reliabilitas tunggal untuk sebuah tes. Rumus baru untuk

memeriksa konsistensi internal tes yang mereka susun mereka namai *Kuder Richardson 20.* Rumus perhitungannya adalah sebagai berikut (Crocker & Algina, 2008; h. 139):

$$KR_{20} = \frac{k}{k-1}(1 - \sum pq/\alpha_x^2) \qquad \textbf{Rumus 7.4.}$$

k = jumlah item dalam tes
α_x^2 = varians keseluruhan tes
Pq = varians item i

Jika semua item tes memiliki taraf kesukaran yang sama, Kuder dan Richardson menyediakan sebuah rumus alternatif yang lebih sederhana dalam arti tanpa perlu menghitung varians masing-masing item. Rumus tersebut dinamai *Kuder Richardson 21,* dan bentuknya adalah sebagai berikut (Crocker & Algina, 2008; h. 139):

$$KR_{21} = \frac{k}{k-1}[\,1 - \mu(k-\mu)/k\alpha_x^2] \qquad \textbf{Rumus 7.5.}$$

μ = *mean* skor total
α_x^2 = varians keseluruhan tes
k = jumlah item dalam tes

Jika seluruh item memiliki taraf kesukaran yang sama, baik KR 20 maupun KR 21 akan menghasilkan estimasi reliabilitas yang sama atau setara. Sebaliknya jika taraf kesukaran item-item berlainan estimasi reliabilitas yang dihasilkan KR 21 secara sistematis akan lebih rendah dibandingkan yang dihasilkan KR 20. Maka tidak boleh hanya melaporkan estimasi reliabilitas KR 21 tanpa disertai KR 20 (Crocker & Algina, 2008).

3) Metode Hoyt. Metode ini berbasis analisis varians serta menempatkan orang atau testi dan item sebagai sumber variabilitas. Memanfaatkan notasi analisis varians baku yang

lazim disajikan dalam tabel ringkasan analisis varians, Hoyt mendefinisikan estimasi reliabilitas sebagai berikut (Crocker & Algina, 2008; h. 140):

$$\rho_{XX'} = \frac{MK_{orang} - MK_{residu}}{MK_{orang}}$$ **Rumus 7.6.**

Jadi, MK_{orang} dipandang mewakili varians skor tampak sedangkan MK_{residu} dipandang mewakili varians skor kesalahan.

Sebagai penutup bab ini akan dikemukakan beberapa hal. *Pertama*, koefisien reliabilitas yang dihasilkan baik dengan rumus Spearman-Brown maupun koefisien α akan besar atau tinggi manakala bagian-bagian tes baik sebagai belahan maupun sebagai item saling berkorelasi tinggi, sebaliknya akan kecil atau rendah manakala bagian-bagian tes tersebut tidak saling berkorelasi. Padahal bagian-bagian tes tersebut akan saling berkorelasi tinggi hanya jika masing-masing mengukur atribut-atribut yang sama atau atribut-atribut yang berkorelasi tinggi. Itulah sebabnya, baik reliabilitas Spearman-Brown maupun reliabilitas koefisien α dipandang sebagai indeks konsistensi internal atau homogenitas suatu tes, yaitu "a characteristic of a test possessed by virtue of the positive intercorrelations of the items composing it" atau salah satu ciri tes yang dimiliki berkat saling korelasi positif antar item-item yang membentuknya (Crocker & Algina, 2008, h. 142; Allen & Yen, 1979).

Kedua, aneka metode estimasi reliabilitas yang sudah dibahas terbukti menghasilkan estimasi koefisien reliabilitas yang berlainan, maka dalam melaporkan taraf reliabilitas suatu tes harus disebutkan dengan jelas metode estimasi yang digunakan. Selain itu, untuk melakukan estimasi reliabilitas *speed tests* atau jenis-jenis tes yang lebih mengutamakan pengukuran kecepatan kerja harus digunakan metode yang berbasis dua rangkaian data seperti *test-retest, alternate forms* atau *parallel forms*. Sebaliknya, metode-metode yang berbasis satu rangkaian data seperti metode Spearman Brown dan koefisien alpha hanya boleh digunakan untuk melakukan estimasi reliabilitas

tes-tes yang homogen atau mengukur satu atribut psikologis sebab metode-metode ini pada dasarnya memang mengungkap homogenitas item. Secara khusus metode Spearman-Brown tidak memberikan estimasi reliabilitas yang akurat bila bagian-bagian tesnya tidak bersifat paralel. Sebaliknya jika bagian-bagian tesnya paralel, selain memberikan estimasi reliabilitas yang akurat untuk keseluruhan tesnya metode Spearman-Brown juga bermanfaat untuk memeriksa dampak perubahan panjang tes terhadap reliabilitasnya (Allen & Yen, 1979).

Ketiga, semua rumus yang disebut di bab ini disajikan dengan tujuan lebih untuk membantu kita memahami konsep atau konsep-konsep di balik aneka metode atau prosedur yang sedang dibahas. Pengoperasian perhitungannya kini tidak lagi perlu dilakukan secara manual, sebab praktis semua sudah tersedia dalam aneka *software* olah data statistik, seperti *SPSS.* Ψ

Bab 8
Validitas dalam Model Tes Klasik

Analisis validitas berfokus pada usaha mengidentifikasikan dan meminimalkan dampak aneka variabel yang menyebabkan perbedaan dalam skor murni (Friedenberg, 1995). Tujuan analisis validitas adalah menentukan sejauh mana skor murni ditentukan oleh sifat atau kemampuan atau atribut yang relevan dengan tujuan tes. Menggunakan terminologi model tes klasik, secara teoretis skor murni memiliki dua komponen: (a) sifat/kemampuan/atribut stabil dalam diri testi yang *relevant* atau relevan dengan tujuan tes; dan (b) sifat/kemampuan/atribut stabil dalam diri testi yang *irrelevant* atau tidak relevan dengan tujuan tes. Jadi kinerja pada tes atau skor tampak seorang testi selain mencerminkan kesalahan pengukuran yang bersifat *random* (*RME*) sebagaimana sudah kita lihat dalam pembahasan tentang reliabilitas, juga mencerminkan dampak dari aneka sifat/kemampuan/atribut dalam diri testi yang berlainan taraf relevansinya dengan tujuan tes termasuk salah satu di antaranya yang relevan. Situasi tersebut dapat dilukiskan dalam rumus sebagai berikut:

$$T = R + I \qquad \textbf{Rumus 8.1.}$$

T = skor murni.
R = sifat/kemampuan/atribut dalam diri testi yang relevan dengan tujuan tes.
I = sifat/kemampuan/atribut dalam diri testi yang tidak relevan dengan tujuan tes.

Sifat/kemampuan/atribut dalam diri testi yang tidak relevan dengan tujuan tes namun yang terukur oleh tes merupakan sifat/

kemampuan/atribut stabil yang selalu akan terukur setiap kali tes tersebut diadministrasikan pada testi. Dampak sifat/kemampuan/atribut stabil yang tidak relevan dengan tujuan tes ini disebut *systematic measurement error* (*SME*), sebab merupakan kesalahan pengukuran yang kehadirannya tak terelakkan. Jadi, *SME* merupakan komponen *error* atau kesalahan dalam skor murni. Sebaliknya, sebagaimana sudah kita lihat, *random measurement error* (*RME*) yang menjadi fokus dalam pemeriksaan reliabilitas merupakan komponen *error* atau kesalahan dari skor tes secara keseluruhan atau skor tampak. Maka, dalam analisis validitas rumus skor tampak perlu dimodifikasi menjadi seperti berikut:

$$X = (R + I) + E$$ **Rumus 8.2.**

X = skor tampak.
R = sifat/kemampuan/atribut stabil dalam diri testi yang relevan dengan tujuan tes.
I = *SME,* dampak sifat/kemampuan/atribut stabil dalam diri testi yang tidak relevan dengan tujuan tes.
E = *RME,* dampak aneka peristiwa yang bersifat *random.*

Begitu pula persamaan varians-nya pun perlu direvisi, sehingga menjadi sebagai berikut:

$$\sigma_X^2 = (\sigma_R^2 + \sigma_I^2) + \sigma_E^2$$ **Rumus 8.3.**

Maka, menurut model tes klasik dan dalam persamaan varians, tes yang valid akan menghasilkan pengukuran dengan ciri: $\sigma_R^2 > \sigma_I^2$. Artinya, skor-skor yang dihasilkan oleh tes yang valid lebih mencerminkan perbedaan dalam hal sifat/kemampuan/atribut stabil yang relevan dengan tujuan tes dalam diri para testi, daripada perbedaan dalam hal sifat/kemampuan/atribut stabil yang tidak relevan dengan tujuan tes.

Maka, menurut model tes klasik dan dalam persamaan varians, estimasi validitas suatu tes = σ_R^2 / σ_X^2. Artinya, validitas suatu tes ditentukan oleh besarnya proporsi varians sifat/kemampuan/atribut stabil yang relevan dengan tujuan tes dalam skor tampak. Sebagai perbandingan, tes yang reliabel memiliki ciri $(\sigma_R^2 + \sigma_I^2)$ atau $\sigma_T^2 > \sigma_E^2$ sehingga rumus estimasi reliabilitas adalah = σ_T^2 / σ_X^2. Artinya, reliabilitas hasil pengukuran yang dihasilkan oleh suatu tes lebih ditentukan oleh besarnya proporsi varians sifat/kemampuan/atribut yang bersifat stabil (skor murni) tanpa menghiraukan relevansinya dengan tujuan pengukuran dalam skor tampak. Dari sini bisa dipahami kesimpulan yang menyatakan bahwa reliabilitas merupakan syarat yang perlu namun tidak mencukupi bagi validitas, atau bahwa tes yang valid pasti reliabel namun tes yang reliabel belum tentu valid.

A. Metode Estimasi Validitas

Model tes klasik memberikan landasan atau kerangka konseptual-statistis tentang validitas, khususnya bagaimana suatu tes bisa menghasilkan pengukuran yang sesuai atau tidak sesuai dengan tujuan tes tersebut disusun. Pemahaman yang baru tentang validitas beserta metode estimasinya, khususnya pasca terbitnya *1999 Standards* (AERA, APA, & NCME, 1999), lebih menekankan sejauh mana penafsiran hasil suatu tes sebagaimana dimaksudkan oleh tes yang bersangkutan sungguh-sungguh dapat dipertanggungjawabkan. Dengan kata lain, kendati masih diturunkan dari kerangka konseptual-statistis yang sama sebagaimana dijelaskan oleh model tes klasik, namun kini validitas tidak lagi dipandang sebagai suatu karakteristik yang melekat pada tesnya melainkan pada soal sejauh mana interpretasi kita atas hasil pengukuran suatu tes sesuai tujuan tes itu sungguh-sungguh bisa diterima atau dipertanggungjawabkan. Maka, validasi suatu tes kini dipandang lebih merupakan soal memperoleh evidensi atau bukti sebanyak mungkin untuk menyokong interpretasi kita terhadap hasil pengukuran suatu tes sesuai maksud atau tujuan tes tersebut disusun. Sebagai contoh, jika suatu tes dimaksudkan

untuk mengukur inteligensi, maka validasi tes tersebut menuntut bahwa interpretasi kita bahwa seorang testi termasuk ke dalam kategori genius atau sebaliknya imbesil didukung oleh bukti-bukti empiris yang sungguh-sungguh meyakinkan sehingga interpretasi kita tersebut bisa dipertanggungjawabkan. Dari rumusan dan contoh validasi ini sekaligus tersirat pengertian bahwa validitas merupakan konsep tunggal namun memiliki sejumlah aspek yang perlu dibuktikan secara empiris keberadaan atau kebenarannya agar interpretasi kita terhadap hasil pengukuran sesuai tujuan tes dapat diterima. Sebagaimana akan kita lihat, berbeda dengan estimasi reliabilitas yang bisa disebut sebagai "kegiatan sekali jadi", validasi atau estimasi validitas suatu tes menuntut pengumpulan bukti-bukti terkait berbagai aspek tes dan pada berbagai tahap penyusunan dan penggunaan tes dalam sejenis proses yang bersifat jangka panjang, berkesinambungan, dan kumulatif.

Seperti sudah diuraikan di Bab 5, menurut pemahaman yang baru tentang validitas pasca terbitnya *1999 Standards* (AERA, APA, & NCME, 1999) ada lima jenis evidensi yang perlu dikumpulkan dalam rangka memeriksa validitas tes, tepatnya validitas penafsiran skor atau hasil pengukuran dengan suatu tes sesuai tujuan penyusunan tes yang bersangkutan. Kelima jenis evidensi yang dimaksud mengacu pada aspek-aspek penting tes yang perlu diperiksa pada berbagai tahap penyusunan dan pengadministrasian tes yang secara kumulatif akan memberikan evidensi yang utuh dan solid terhadap keabsahan kita dalam menafsirkan hasil tes sesuai tujuan tes itu disusun. Kelima evidensi yang dimaksud adalah seperti diuraikan di bawah ini (AERA, APA, & NCME, 1999; Goodwin & Leech, 2003).

1. Evidensi Terkait Isi Tes

Salah satu evidensi validitas adalah kesesuaian antara isi tes dan *konstruk* yang diukurnya. Tentang makna istilah "konstruk" perlu diberikan penjelasan sebagai berikut. Aslinya istilah ini secara khusus dipakai untuk mengacu sifat/kemampuan/atribut psikologis yang

tidak teramati secara langsung dan yang hanya bisa diinferensikan berdasarkan tingkah laku teramati yang dipandang sebagai indikatornya. Pemaknaan ini dipakai misalnya, dalam artikel klasik Cronbach dan Meehl (1955) tentang validitas konstruk. Sebagaimana diuraikan di Bab 2, sebagian besar atribut psikologis yang menjadi objek psikometri merupakan konstruk dalam pengertian seperti antara lain dimaksudkan oleh Cronbach dan Meehl dalam artikelnya tersebut. Dalam pembahasan tentang evidensi terkait isi tes ini dan sebagaimana dipakai dalam *1999 Standards* (AERA, APA, & NCME, 1999) istilah "konstruk" dimaknai secara lebih luas mencakup semua jenis konsep atau sifat yang hendak diukur oleh suatu tes.

Evidensi tentang kesesuaian isi dan konstruk yang diukur oleh suatu tes bisa diperoleh melalui analisis logis atau empiris terhadap seberapa memadai isi tes mewakili ranah isi serta seberapa relevan ranah isi tersebut sesuai dengan interpretasi skor tes yang dimaksudkan. Isi tes mengacu pada tema-tema, pilihan kata, serta format atau bentuk item, tugas, atau pertanyaan yang digunakan dalam tes.

Evidensi terkait isi ini lazim diperoleh melalui penilaian pakar atau ahli terhadap kesesuaian antara bagian-bagian tes dan konstruk yang diukur. Secara lebih rinci, aspek-aspek isi tes yang perlu dievaluasi meliputi (a) *sufficiency* atau kecukupan, yaitu apakah isi tes tersebut mencukupi atau memadai dalam arti mewakili ranah isi spesifik yang hendak diukur; (b) *clarity* atau kejelasan, yaitu apakah isi tes tersebut mencerminkan secara jelas ranah isi spesifik yang hendak diukur dalam arti misalnya tidak mencampuradukkan dengan ranah isi spesifik yang lain; (c) *relevance* atau relevansi, yaitu apakah isi tes tersebut memiliki kesesuaian dengan ranah isi spesifik yang hendak diukur; (d) kesesuaian antara item-item dan tugas-tugas yang dipakai sebagai stimulus dalam tes tersebut dengan definisi tentang konstruk yang hendak diukur; (e) ada-tidaknya *bias* berupa keberpihakan isi tes pada gender, budaya, umur atau faktor pengelompokan sosial lainnya; dan (f) kemungkinan terjadinya "*construct irrelevant variance*" (varians yang tidak relevan dengan konstruk yang hendak

diukur) dan "*construct underrepresentation"* (kurang memadainya keterwakilan konstruk yang hendak diukur), yang menunjukkan sejauh mana kemungkinan tes tersebut mengukur melebihi (*construct irrelevance variance)* atau kurang (*construct underrepresentation)* dari yang semestinya dia ukur (Goodwin & Leech, 2003). Dalam pengertian lama, jenis evidensi ini dikaitkan dengan *validitas isi***.**

Agar bisa mengevaluasi keenam sifat di atas, tentu saja pertama-tama harus ditentukan dulu ranah isi dari konstruk yang menjadi sasaran pengukuran. Secara garis besar, jenis konstruk dalam arti luas yang menjadi sasaran pengukuran psikologis dapat dibedakan ke dalam tiga kategori: (a) pengetahuan atau ketrampilan sebagai hasil belajar; (b) konstruk teoretis (dalam arti sempit) tentang berbagai atribut kepribadian; dan (c) pola tingkah laku kompleks dalam rangka menjalankan peran tertentu. Sesuai jenisnya, ranah isi konstruk yang menjadi sasaran pengukuran dapat diidentifikasikan dengan bantuan tiga macam instrumen atau sarana, yaitu (Friedenberg, 1995): (1) *test plan* atau tabel spesifikasi atau kisi-kisi jika konstruk yang menjadi sasaran pengukuran berupa pengetahuan atau ketrampilan sebagai hasil belajar; (2) *eksplikasi konstruk* jika konstruk yang menjadi sasaran pengukuran berupa konstruk teoretis tentang dimensi atau atribut kepribadian tertentu yang bersifat hipotetis dan yang dipandang membedakan orang yang satu dari yang lain; dan (3) *analisis tugas*, jika konstruk yang menjadi sasaran pengukuran berupa pola perilaku dalam rangka menjalankan peran atau tugas jabatan tertentu. Uraian ringkas tentang masing-masing instrumen atau sarana validasi isi disajikan secara berturut-turut di bawah ini.

a. Menyusun *Test-Plan*, Tabel Spesifikasi, atau Kisi-kisi.

Test-plan, tabel spesifikasi atau kisi-kisi adalah "a written list of the information to be covered by the test items and the behaviors required to answer the questions correctly" (Friedenberg, 1995). Artinya, kisi-kisi merupakan daftar tertulis tentang informasi, yaitu pengetahuan dan atau ketrampilan dalam bidang studi atau mata pelajaran tertentu, yang harus dicakup oleh item-item tes serta jenis-jenis perilaku yang

dituntut untuk menjawab item-item atau pertanyaan-pertanyaan tes secara tepat. *Blue-print* atau cetak biru ini penting khususnya dalam rangka pengukuran pengetahuan dan ketrampilan sebagai hasil belajar dalam bidang studi atau mata pelajaran tertentu, untuk memastikan bahwa isi tes merepresentasikan dan sebaliknya tidak melenceng dari cakupan isi pengetahuan atau ketrampilan dalam bidang studi atau mata pelajaran yang bersangkutan.

Penyusunan tabel spesifikasi atau kisi-kisi lazim didasarkan pada taksonomi tujuan pengajaran tertentu. Salah satu taksonomi tujuan pengajaran yang paling luas dijadikan pedoman dalam pengajaran di sekolah adalah taksonomi tujuan pengajaran ranah kognitif yang disusun oleh Bloom dan kawan-kawan. Taksonomi Bloom telah beberapa kali direvisi. Salah satu revisi yang secara struktural masih mempertahankan bentuk asli Taksonomi Bloom adalah revisi yang dikerjakan oleh Anderson, Krathwohl, Airasian, Cruikshank, Mayer, Pintrich, Raths dan Wittrock (2001) atau yang lebih dikenal sebagai Taksonomi Anderson dan Krathwohl. Secara garis besar, taksonomi revisi ini membedakan pengetahuan sebagai hasil belajar dalam mata pelajaran tertentu ke dalam dua dimensi: (1) dimensi isi atau materi; dan (2) dimensi proses kognitifnya. Dimensi isi atau materinya lebih lanjut dibedakan ke dalam empat kategori pengetahuan: (1) pengetahuan faktual, (2) pengetahuan konseptual, (3) pengetahuan prosedural, dan (4) pengetahuan metakognitif. Dimensi proses kognitifnya lebih lanjut dibedakan ke dalam enam jenis: (1) mengingat, (2) memahami, (3) menerapkan, (4) menganalisis, (5) mengevaluasi, dan (6) mencipta. Tampak bahwa tabel spesifikasi atau kisi-kisi secara rinci menjabarkan ranah isi suatu konstruk meliputi baik jenis pengetahuan maupun jenis kemampuan berpikir tentang bidang studi atau mata pelajaran tertentu yang harus dikuasai seseorang sebagai hasil belajar. Jika berhasil disusun secara tepat, tabel spesifikasi atau kisi-kisi bisa dipakai sebagai kriteria yang handal dalam mengevaluasi berbagai aspek tes dalam rangka mengumpulkan evidensi terkait keabsahan isinya.

b. Melakukan Eksplikasi Konstruk. Jika konstruk yang menjadi sasaran pengukuran suatu tes benar-benar merupakan konstruk teoretis tentang dimensi atau atribut kepribadian tertentu baik berupa abilitas atau kemampuan maupun berupa sifat atau kecenderungan kepribadian yang bersifat abstrak, maka penentuan ranah isinya harus didahului dengan sebuah langkah yang disebut **eksplikasi konstruk**.

Konstruk psikologis bersifat abstrak dan tidak bisa diukur secara langsung. Ada atau tidak adanya atau lebih tepat, seberapa besar kadar keberadaan atribut kepribadian tertentu yang merupakan konstruk dalam diri seseorang harus diinferensikan atau disimpulkan dari bentuk-bentuk tingkah laku spesifik tertentu yang dipandang sebagai indikator atau pencerminan keberadaan konstruk yang bersangkutan.

Dalam eksplikasi konstruk, "the test developer compiles a list of specific behaviors, beliefs, and attitudes that demonstrate the presence of the construct and specific behaviors, beliefs, and attitudes inconsistent with its presence" (Friedenberg, 1995). Artinya, dalam eksplikasi konstruk penyusun tes mengumpulkan daftar perilaku, keyakinan, dan sikap yang mencerminkan keberadaan atau kehadiran konstruk yang diukur (*favorable),* serta perilaku, keyakinan, dan sikap yang bertentangan atau mencerminkan ketiadaan konstruk yang diukur (*unfavorable*).

Dalam praktik eksplikasi konstruk akan melibatkan dua langkah utama: (1) perumusan definisi konseptual konstruk, yaitu merumuskan konstruk yang menjadi sasaran pengukuran sebagai konsep dengan menggunakan konsep-konsep lain yang lebih mudah dipahami; dan (2) merumuskan definisi operasional konstruk, yaitu menjabarkan konstruk sebagai konsep ke dalam serangkaian subkonsep atau komponennya serta mengidentifikasikan indikator-indikator tingkah laku baik yang *favorable* atau positif maupun yang *unfavorable* atau yang negatif. Melalui proses dua langkah eksplikasi konstruk semacam ini maka cakupan isi konstruk yang semula abstrak tersebut menjadi jelas dalam arti kongkret sekaligus operasional.

Jika berhasil disusun secara tepat, eksplikasi konstruk bisa dipakai sebagai kriteria yang handal dalam mengevaluasi berbagai aspek tes dalam rangka mengumpulkan evidensi terkait keabsahan isinya.

c. Melakukan Analisis Tugas. Jika tes dimaksudkan untuk mengukur pola perilaku kompleks terkait pelaksanaan *role* atau peran atau *job* atau jabatan atau *task* atau tugas tertentu, maka penentuan isi tesnya perlu dilakukan melalui *task analysis* atau analisis tugas atau *job analysis* atau analisis jabatan (Friedenberg, 1995). Dalam analisis tersebut, perilaku kompleks dalam rangka menjalankan peran, jabatan atau tugas pekerjaan yang menjadi sasaran pengukuran, misal jabatan *sales manager* atau manajer penjualan, jabatan kepala sekolah, dan sebagainya, diuraikan ke dalam komponen-komponen pokok tingkah laku. Selanjutnya masing-masing komponen pokok diuraikan lebih lanjut ke dalam bentuk-bentuk tingkah laku yang lebih spesifik. Jika berhasil disusun secara tepat dan menyeluruh, hasil analisis jabatan atau analisis tugas bisa dipakai sebagai kriteria yang handal dalam mengevaluasi berbagai aspek tes dalam rangka mengumpulkan evidensi terkait keabsahan isinya dalam rangka mengukur kemampuan entah sebagai potensi atau sebagai prestasi seseorang menjalankan peran, jabatan, atau tugas pekerjaan tertentu.

2. Evidensi Terkait Proses Respon Subjek

Evidensi ini didasarkan pada penilaian terhadap kesesuaian antara respon yang diberikan oleh testi dalam rangka mengerjakan tes dengan konstruk yang diukur oleh tes. Sebagai contoh, dalam mengerjakan *typical performance test* apakah testi sungguh-sungguh menyatakan keadaan dirinya sebagai cerminan konstruk yang diukur atau sekadar memberikan jawaban sesuai norma yang berlaku di tengah masyarakat. Gejala ini lazim dikenal sebagai *response set of social desirability* atau gejala menjawab pertanyaan tes mengikuti penilaian subjektif tentang apa yang diharapkan oleh masyarakat, bukan mengikuti kata hatinya sendiri. Beberapa kecenderungan respon lain

yang bisa merugikan validitas hasil pengukuran mencakup (Klein, 1986): (a) *response set of acquiescence,* yaitu kecenderungan menyetujui atau mengiakan item tanpa menghiraukan isinya; (b) *response set of using uncertain or middle category,* yaitu kecenderungan memilih jawaban di tengah yang mencerminkan *indecision* atau keengganan menunjukkan sikap atau *uncertainty* atau sikap ragu-ragu atau sikap tidak tegas; (c) *response set of using the extreme response,* yaitu kecenderungtan memilih jawaban ekstrem, positif atau negatif, tanpa menghiraukan isi itemnya; (d) *guessing,* yaitu kecenderungan menebak, khususnya dalam menjawab item tes yang bertujuan mengukur konstruk berupa abilitas atau kemampuan. Beberapa strategi untuk mengumpulkan jenis evidensi ini meliputi (1) mengobservasi testi saat sedang mengerjakan tugas dalam rangka tes, dan (2) mewawancarai testi untuk mengetahui alasan mereka memberikan jawaban tertentu terhadap pertanyaan-pertanyaan dalam tes.

Selain bersumber dari respon testi ancaman terhadap validitas dari sisi ini juga bisa berasal dari ketidak-cermatan asesor dalam menjalankan tugasnya. Untuk mengatasinya, salah satu langkah yang direkomendasikan adalah mengevaluasi cara para pengamat dan penilai menerapkan kriteria dalam mencatat dan mengevaluasi tingkah laku, kinerja, atau hasil pekerjaan tertulis testi dalam rangka tes; tujuannya adalah memastikan bahwa kriteria penilaian yang disediakan diterapkan sebagaimana mestinya dan bukan malah menggunakan acuan lain yang tidak sesuai (Goodwin & Leech, 2003). Dalam konsep lama, jenis evidensi ini dikaitkan dengan salah satu aspek *validitas konstruk*.

3. Evidensi Terkait Struktur Internal Tes

Evidensi ini didasarkan pada penilaian tentang sejauh mana item-item dan komponen-komponen dalam tes saling berhubungan sedemikian rupa sesuai dengan konstruk yang diukur. Aspek ini terkait dengan konsistensi internal atau homogenitas tes. Konsistensi internal atau homogenitas tes yang tinggi dipandang merupakan

evidensi yang kuat bahwa tes tersebut mengukur sebuah konstruk, khususnya seperti yang dimaksudkan oleh penyusun tes.

Memang ada pakar psikometri yang tidak sependapat dengan gagasan konsistensi internal atau homogenitas tes sebagai evidensi validitas konstruk sebuah tes. Menurut pakar tersebut, "maximum validity is obtained where test items do not all correlate with each other, but where each correlates positively with the criterion. Such a test would have only low internal consistency reliability" (Cattell, dalam Klein, 1986, h. 3). Maksudnya, validitas tes tidak ditentukan oleh saling korelasi antar item atau antar komponennya melainkan oleh korelasi antara item-item atau komponen-komponen tes tersebut dengan kriterianya. Argumen kontra terhadap konsistensi internal atau homogenitas adalah sebagai berikut: setiap item dalam sebuah tes memiliki cakupan yang lebih sempit dibandingkan dengan kriterianya. Jika semua item dalam tes sangat konsisten satu sama lain berarti cakupan variabel yang diukur menjadi sangat sempit. Hal ini bertolak belakang dengan validitas yang mengandaikan kemampuan sebuah tes mengungkap cakupan yang luas dari variabel atau kriteria yang diukur. Sebagaimana sudah kita lihat, kendati kritik ini sangat beralasan namun dalam praktik dalil umum psikometri bahwa tes yang valid lazimnya memiliki konsistensi (internal) yang tinggi tetap diterima secara luas.

Ada setidaknya dua metode yang lazim ditempuh untuk memeriksa struktur internal tes. Yang pertama adalah **analisis faktor konfirmatori** yang pertama kali dikembangkan oleh Joreskog. Sebagaimana sudah disinggung, gerakan analisis faktor dipelopori oleh Charles Spearman. Sekitar awal abad ke-20 dia mengembangkan model analisis faktor dengan sebuah faktor umum tunggal yang merepresentasikan *general intelligence* atau inteligensi umum. Dalam perkembangannya hingga kini metode faktorial bisa digolongkan ke dalam dua kategori, yaitu analisis faktor *eksploratori* seperti yang dikembangkan oleh Spearman, Thurstone dan lain-lain, serta analisis faktor *konfirmatori* sebagai metode khusus *structural equation modeling* dan pengujian hipotesis seperti khususnya yang dikembangkan oleh

Joreskog. Namun ada yang mengingatkan, terlalu mengandalkan analisis faktor dalam validasi berisiko memperoleh bukti validitas yang kurang kokoh.

Yang kedua adalah *differential item function (DIF) techniques* atau **tehnik *DIF*** untuk memeriksa kemungkinan terjadinya bias item sebagai evidensi lain invaliditas. Dalam pengukuran abilitas, *DIF* terjadi jika testi dengan kemampuan yang sama namun termasuk ke dalam kelompok yang berbeda memiliki peluang yang juga berbeda untuk berhasil mengerjakan sebuah item (Goodwin & Leech, 2003). Dalam pengertian lama, jenis evidensi terkait struktur internal tes ini dikaitkan dengan salah satu aspek *validitas konstruk***.**

4. Evidensi terkait Hubungan antara Tes dan Tes Lain

Evidensi validitas juga bisa diperoleh dengan menganalisis hubungan antara skor tes dan variabel-variabel lain di luar tes itu sendiri. Ada beberapa metode yang tercakup dalam pendekatan ini.

Pertama, analisis hubungan antara skor tes dan skor kriteria yang diprediksikan oleh tes yang bersangkutan. Analisis ini akan memberikan evidensi tentang seberapa akurat tes mampu memprediksikan kinerja atau tingkah laku yang merupakan kriterianya. Dalam pengertian lama jenis evidensi ini dikaitkan dengan *criterion related validity* atau validitas terkait kriteria atau validitas kriteria. Sebagaimana pernah disinggung, tergantung dari saat data kriteria diperoleh jenis validitas ini dibedakan ke dalam *concurrent validity* alilas validitas konkuren jika data kriterianya diperoleh bersamaan waktu dengan diperolehnya data tes, serta *predictive validity* jika data kriterianya diperoleh dalam jeda waktu yang panjang sesudah diperolehnya data tes sebagai prediktor.

Kedua, analisis hubungan antara skor tes dan skor tes-tes lain yang dimaksudkan untuk mengukur konstruk yang sama seperti yang diukur oleh tes yang bersangkutan; dan analisis hubungan antara skor tes dan skor tes-tes lain yang dimaksudkan untuk mengukur

konstruk yang berbeda dari yang diukur oleh tes yang bersangkutan. Pada kasus pertama, hubungan positif antara skor tes dengan skor tes-tes lain yang dimaksudkan untuk mengukur konstruk yang sama atau sejenis menghasilkan apa yang disebut *evidensi konvergen.* Pada kasus kedua, hubungan positif namun tidak signifikan atau hubungan negatif dan sifnifikan antara skor tes dengan skor tes-tes lain yang dimaksudkan untuk mengukur konstruk yang berbeda menghasilkan apa yang disebut *evidensi diskriminan.*

Dalam pengertian lama, jenis evidensi di atas dikaitkan dengan salah satu aspek validitas konstruk. Aslinya metode ini dikembangkan oleh Campbell dan Fiske (1954) dan dinamai *multitrait-multimethod validity approach.* Secara ringkas, evidensi tentang validitas konstruk suatu tes diperoleh dengan cara membandingkan hasil pengukuran dua atau lebih sifat dengan dua atau lebih metode. Sebagai contoh, dua sifat yaitu introversi dan neurotikisme masing-masing diukur dengan dua metode yaitu tes benar-salah dan tes pilihan ganda. Keempat tes diadministrasikan pada sekelompok testi yang sama, lantas korelasi antar skor keempat tes tersebut dihitung dan hasilnya disajikan dalam sebuah *matriks korelasi* yang disebut *multitrait-multimethod validity matrix.* Evidensi tentang *validitas konvergen* diperoleh jika korelasi antara skor tes-tes yang mengukur sifat yang sama benar-benar tinggi. Sebaliknya, evidensi tentang *validitas diskriminan* diperoleh jika korelasi antara skor tes-tes yang mengukur sifat yang berbeda benar-benar rendah.

Ketiga, analisis perbedaan kinerja dalam tes yang sama antara dua atau lebih kelompok testi yang diprediksikan memang akan berbeda berkat hubungan antara konstruk yang diukur oleh tes dan variabel yang mendasari pembagian testi ke dalam kelompok-kelompok. Analisis ini bisa dilakukan melalui *group-comparison studies* atau penelitian tentang perbedaan kelompok atau *experimental research studies* atau penelitian eksperimental yang melibatkan pembandingan kinerja antar kelompok (Goodwin & Leech, 2003). Dalam pengertian lama, jenis evidensi ini secara umum dikaitkan dengan salah satu aspek validitas konstruk.

5. Evidensi Terkait Konsekuensi Pengetesan

Konsekuensi, dampak, atau akibat dari pengadministrasian tes terhadap kinerja atau tingkah laku testi juga dimasukkan sebagai salah satu evidensi penting dalam mengevaluasi validitas pengukuran. Secara garis besar, konsekuensi, dampak atau akibat pengetesan ini dibedakan ke dalam dua kategori, yaitu (a) konsekuensi, dampak, atau akibat yang direncanakan, serta (b) konsekuensi, dampak, atau akibat yang tidak direncanakan.

Terkait yang pertama, yaitu konsekuensi, dampak, atau akibat yang direncanakan, tes lazim diadministrasikan dengan harapan memperoleh manfaat tertentu dari hasil interpretasi skor yang sudah direncanakan. Contoh manfaat yang direncanakan semacam itu meliputi antara lain bisa memilih bentuk perlakuan yang efektif dalam terapi, bisa menempatkan karyawan pada jenis tugas yang cocok, bisa menghindari risiko meloloskan orang yang tidak memenuhi kualifikasi untuk memasuki sebuah profesi, atau bisa memperbaiki praktik pengajaran di kelas. Tujuan validasi adalah memperoleh evidensi atau bukti bahwa manfaat tersebut sungguh-sungguh terjadi.

Terkait yang kedua, yaitu konsekuensi, dampak, atau akibat yang tidak direncanakan, penerapan tes seringkali juga diakui memberikan manfaat di luar hasil interpretasi skor yang sudah direncanakan. Contoh manfaat yang tidak direncanakan dari penerapan tes adalah meningkatnya motivasi belajar siswa atau meningkatnya praktik pengajaran di kelas yang dilaksanakan oleh guru sesudah mengetahui hasil tes. Evidensi validitas diperoleh jika manfaat tidak langsung semacam itu sungguh-sungguh tercapai (AERA, APA, & NCME, 1999). Dalam pengertian lama, jenis evidensi terkait pengetesan baik yang direncanakan maupun yang tidak direncanakan ini dikaitkan dengan salah satu aspek *validitas konstruk*.

B. Hubungan antara Validitas dan Reliabilitas

Kendati sudah disinggung di bagian terdahulu, mengakhiri bab tentang validitas ini akan dibahas kembali secara singkat hubungan antara validitas dan reliabilitas menurut model tes klasik. Secara umum bisa dikatakan bahwa suatu tes yang valid pasti reliabel, sebaliknya tes yang reliabel belum tentu valid. Hal ini bisa dijelaskan berdasarkan pengertian validitas dan reliabilitas sendiri. Sebagaimana kita tahu, validitas ditentukan oleh besarnya proporsi sifat, kemampuan, atau atribut stabil dalam diri individu yang relevan dengan tujuan tes. Sifat, kemampuan, atau atribut stabil yang relevan ini merupakan komponen dari skor murni. Maka cukup jelas, bila komponen yang berkontribusi bagi besarnya skor murni tersebut besar berarti pengukurannya valid, sehingga juga akan reliabel sebab proporsi terbesar skor tampak ditempati oleh skor murni.

Sebaliknya, reliabilitas ditentukan oleh kecilnya komponen *error* yang bersifat random atau tidak stabil dalam skor tampak. Namun kita tahu, jika *random measurement error* kecil yang berarti pengukuran tersebut reliabel, maka reliabilitas atau kestabilan hasil pengukuran tersebut bisa saja lebih ditentukan oleh komponen sifat, kemampuan, atau atribut stabil dalam diri testi yang tidak relevan dengan tujuan tes, bukan ditentukan oleh komponen sifat, kemampuan, atau atribut yang relevan dengan tujuan tes. Ibaratnya, mengukur ketinggian dengan termometer. Hasilnya mungkin reliabel, dalam arti setiap berpindah dari tempat yang rendah ke tempat yang lebih tinggi skala pada termometer senantiasa bergerak turun, sebab ketinggian berhubungan dengan suhu. Tetapi pengukuran semacam itu tentu tidak valid, sebab tidak diketahui seberapa tingginya, sedangkan yang diukur pun bukan lagi ketinggian melainkan suhu. Ψ

Bab 9
Langkah Umum Konstruksi Tes

Dalam bab ini kita akan membahas bagaimana aneka prinsip dan konsep yang sudah kita bahas dalam bab-bab sebelumnya diterapkan dalam konstruksi atau penyusunan tes. Namun sebelumnya perlu dikemukakan catatan sebagai berikut. Sebagaimana sudah kita lihat, secara garis besar tes dibedakan ke dalam dua kategori, yaitu kategori tes yang mengukur abilitas termasuk jenis-jenis ketrampilan spesifik dan diberi nama kategori *maximal performance tests,* serta kategori tes yang mengukur sifat atau kecenderungan kepribadian dan diberi nama kategori *typical performance tests.*

Secara garis besar, jenis kemampuan yang lazim menjadi objek pengukuran *maximal performance tests* bisa dibedakan menjadi dua kategori: (1) *achievement* atau prestasi atau hasil belajar, yaitu jenis *developed abilities* yang bersifat aktual dan dengan latar belakang eksperiensial yang spesifik berupa pengalaman belajar dalam konteks pembelajaran tertentu khususnya di sekolah sebagai faktor pembentuknya, dan (2) *aptitude* atau bakat, yaitu jenis *developed abilities* yang masih berupa potensi dan dengan latar belakang eksperiensial yang luas sebagai faktor pembentuknya.

Sebagai konstruk dalam arti luas, atribut psikologis berupa pengetahuan-ketrampilan yang disebut *achievement* lazim memiliki *content domain* dengan batas-batas yang jelas, khususnya berupa silabus mata pelajaran dan buku teks yang dipakai dalam kegiatan pembelajaran di sekolah. Sebaliknya *aptitude* dan hampir semua atribut psikologis berupa sifat atau kecenderungan kepribadian yang menjadi sasaran *typical performance tests* lazim berupa konstruk dalam arti sempit, dengan ciri utama memiliki *content domain* atau ranah isi

dengan batas-batas yang tidak atau belum jelas (Cronbach & Meehl, 1955).

Dalam penyusunan tes, *content domain* atribut psikologis berupa konstruk teoretis tersebut terlebih dulu harus diidentifikasikan dalam arti dirumuskan dan ditetapkan batas-batasnya. Identifikasi ranah isi atribut psikologis berupa konstruk teoretis semacam ini harus dilakukan melalui *eksplikasi konstruk,* yaitu langkah menetapkan *behavioral indicators* atau indikator-indikator tingkah laku suatu konstruk teoretis berupa bentuk-bentuk tingkah laku spesifik yang bisa diamati dan diukur, baik yang bersifat mendukung (*favorable*) maupun yang bersifat menyangkal atau mengingkari (*unfavorable*) keberadaan konstruk teoretis yang bersangkutan (Friedenberg, 1995).

Ada sejumlah model yang menjelaskan langkah-langkah umum penyusunan tes (Azwar, 1999; Gregory, 2007; Crocker & Algina, 2008). Memanfaatkan unsur-unsur yang baik dari berbagai model tersebut, kita akan menerapkan langkah-langkah umum penyusunan tes sebagai berikut: (a) *defining the test* atau mendefinisikan tes, (b) *preparing test specifications* atau menyusun tabel spesifikasi tes, (c) *selecting a scaling method* atau memilih metode penskalaan, (d) *constructing the items* atau menyusun item-item, (e) memintakan *review* atas item-item dari sejumlah pakar terkait, dan melakukan revisi seperlunya; (f) merakit item-item menjadi bentuk semifinal tes yang siap diuji-cobakan; (g) melakukan uji coba terhadap sampel yang mewakili populasi khalayak yang menjadi sasaran tes, (h) memeriksa ciri-ciri psikometrik skor-skor item melalui analisis item, melakukan seleksi item: item-item yang bisa ditetapkan menjadi calon bentuk final skala dan item-item yang masih perlu direvisi atau bahkan digugurkan, (i) melakukan pemeriksaan reliabilitas, validitas, dan daya diskriminasi bentuk final tes, (j) menyusun manual atau buku pedoman tes dan menerbitkan tes. Marilah kita bahas langkah-langkah tersebut satu demi satu.

A. Mendefinisikan Tes

Langkah ini mencakup minimal tiga sublangkah, yaitu (1) menetapkan khalayak yang akan menjadi sasaran tes; (2) menetapkan jenis skor, yaitu cara skor akan digunakan untuk menafsirkan hasil tes; dan (3) merumuskan *content domain* atau ranah isi tes.

1. Menetapkan Khalayak Tes

Tes selalu valid hanya untuk khalayak tertentu yang ditetapkan sebagai sasaran tes, dan tidak akan valid jika diadministrasikan pada khalayak lain yang bukan merupakan sasarannya. Maka, penentuan khalayak yang menjadi sasaran tes merupakan bagian integral bahkan langkah perdana dalam upaya mengonstruksi sebuah tes yang valid. Penetapan yang jeli tentang khalayak sasaran akan sangat membantu dalam menentukan antara lain jenis tugas atau stimulus yang dipakai sebagai item-item tes, format item, dan bahasa atau media lain yang akan digunakan dalam menyusun item.

2. Menetapkan Jenis Skor

Sebagaimana sudah kita lihat, penetapan jenis skor akan mencakup sejumlah pertimbangan sebagai berikut. Pertama, apakah tes itu hanya mengukur satu atau lebih atribut psikologis sekaligus. Kedua, jika tes itu hanya mengukur satu atribut psikologis, apakah hasilnya akan ditafsirkan dengan cara membandingkannya dengan norma kelompok yang bersifat relatif dan yang baru dirumuskan sesudah pengetesan dilakukan (*norm-referenced scoring*) atau dengan suatu kriteria yang bersifat mutlak dan yang sudah ditentukan sebelum pengetesan dilaksanakan (*criterion referenced testing*). Berbagai tes baku yang mengukur baik abilitas maupun kepribadian lazim menerapkan penskoran beracuan norma, sedangkan tes buatan guru (*teacher made tests*) untuk mengukur capaian kompetensi berbagai mata pelajaran di sekolah lazim menerapkan penskoran beracuan patokan. Ketiga, jika tes hanya mengukur satu atribut berarti harus diterapkan

penskoran normatif (*normative scoring*) untuk menunjukkan jumlah atau kuantitas absolut terkait atribut psikologis tertentu dalam diri masing-masing testi. Sebaliknya, jika tes mengukur lebih dari satu atribut psikologis sekaligus berarti harus diterapkan penskoran ipsatif (*ipsative scoring*) untuk menunjukkan perbandingan kuantitas masing-masing subatribut psikologis satu dengan yang lain. Penetapan yang jeli tentang jenis skor yang akan diterapkan sangat membantu memilih jenis penskalaan yang akan dipakai.

3. Menetapkan Ranah Isi Tes

Dalam pengukuran konstruk dengan batas-batas ranah isi yang jelas, sublangkah ketiga ini cukup sederhana. Dalam *achievement test* atau tes prestasi bidang studi atau mata pelajaran tertentu di sekolah, misalnya, ranah isi atau *universe of content* dari abilitas, kompetensi, atau kemampuan (*knowledge and skills*) yang diukur tercakup dalam buku teks yang digunakan dalam pembelajaran. Dalam pengukuran konstruk berupa konstruk teoretis dengan batas-batas ranah isi yang tidak jelas, sublangkah ini memerlukan metode *eksplikasi konstruk.* Sebagaimana sudah kita lihat, inti eksplikasi konstruk adalah mengidentifikasikan bentuk-bentuk keyakinan, sikap, atau perilaku baik yang mendukung maupun yang mengingkari kehadiran konstruk berupa atribut psikologis yang akan menjadi objek pengukuran (Friedenberg, 1995). Ada beberapa tehnik yang bisa diterapkan untuk melakukan eksplikasi konstruk (Crocker & Algina, 2008), yaitu: (1) *content analysis* atau analisis isi, (2) *review of research* atau telaah hasil-hasil penelitian terdahulu, (3) *critical incidents* atau pengamatan terhadap peristiwa-peristiwa luar biasa, (4) *direct observations* atau pengamatan langsung, dan (5) *expert judgments* atau penilaian oleh sejumlah pakar. Masing-masing tehnik akan dibahas secara lebih rinci di bagian lain.

B. Menyusun Tabel Spesifikasi Tes

Sublangkah ke-3 dalam pendefinisian tes di atas akan menghasilkan identifikasi *content domain* atau ranah isi dari atribut yang hendak diukur sampai ke komponen-komponennya. Langkah selanjutnya adalah menyusun tabel speksifikasi tes, yaitu "a plan for deciding the relative emphasis that each of these components should receive on the test" (Crocker & Algina, 2008; h. 72). Maksudnya, tabel spesifikasi atau *test blue print* atau kisi-kisi tes merupakan sejenis cetak biru atau rencana untuk menentukan bobot relatif setiap komponen dalam tes sehingga terbentuk sebuah keseluruhan tes dengan struktur yang dipandang sungguh-sungguh mencerminkan *content domain* konstruk yang diukur.

Sesuai namanya, tabel spesfikasi akan berupa tabel yang terdiri dari dua kisi, yaitu *kisi horisontal* dan *kisi vertikal*. Kisi horisontal atau datar lazimnya diisi dengan *content domain* berupa komponen-komponen konstruk atau atribut psikologis yang sedang diukur. Kisi vertikal atau tegak diisi dengan dimensi lain sesuai konstruk atau atribut psikologis yang diukur, lazimnya berupa proses. Dalam pengukuran prestasi belajar dalam mata pelajaran tertentu di sekolah, misalnya Matematika, kisi horisontal akan berisi komponen-komponen utama yang membentuk *content domain* kompetensi di bidang Matematika, misalnya kemampuan menjumlahkan, mengurangkan, mengalikan, dan membagi. Kisi vertikalnya lazim diisi dengan *behavioral domain* mengikuti sistematika atau taksonomi tertentu. Sistematika yang lazim diikuti dalam penyusunan tes prestasi di sekolah adalah salah satu atau gabungan dari beberapa versi taksonomi tujuan pengajaran yang dirintis oleh Bloom dan kawan-kawan.

Sudah kita lihat bahwa tes merupakan sampel tingkah laku yang diambil dari populasi tingkah laku yang merupakan indikator-indikator dari atribut psikologis yang sedang diukur. Sebagai sampel tingkah laku, tes harus sungguh-sungguh merepresentasikan populasinya. Representasi tersebut mencakup baik kualitas maupun kuantitasnya. Representasi tersebut akan tercermin dari **struktur**

tes yang tampak dalam tabel spesifikasi. Tabel spesifikasi yang baik harus sungguh-sungguh mampu merepresentasikan atribut yang diukur baik dari segi kualitasnya, yaitu dalam bentuk pemaparan komponen-komponen atribut psikologis yang diukur secara memadai, maupun dari segi kuantitasnya, yaitu berupa distribusi item pada masing-masing komponen yang mencerminkan pandangan penyusun tes tentang bobot masing-masing komponen maupun subkomponen dalam merepresentasikan atribut psikologis yang sedang diukur. Bobot masing-masing komponen dan subkomponen tersebut dinyatakan dalam persentase, sehingga jumlah seluruh komponen dan subkomponen sebagai kesatuan tes harus = 100%. Sesudah berhasil disusun dengan baik, tabel spesifikasi akan menjadi pedoman bagi penyusun tes dalam menyusun item-item.

C. Memilih Metode Penskalaan

Sebelum mulai menyusun item-item berpedoman tabel spesifikasi, terlebih dulu perlu ditentukan metode penskalaan yang akan diterapkan. Pilihan atas metode penskalaan akan menentukan format item yang akan dipilih dalam penyusunan item. Maka, akan dibahas dulu seluk-beluk pemilihan metode penskalaan.

Hakikat pengukuran psikologis adalah menerakan bilangan pada respon testi dalam suatu tes untuk menentukan seberapa banyak testi memiliki atribut psikologis yang sedang diukur (Gregory, 2007). Persoalan pokok dalam penskalaan adalah cara menetapkan aneka bilangan pada aneka respon testi yang dipandang mencerminkan pemilikan atribut psikologis yang sedang diukur dalam jumlah yang berlainan dalam diri testi. Namun perlu kita ingat, "semua bilangan hasil pengukuran dapat ditempatkan pada salah satu dari empat kategori skala yang bersifat hirarkis, yaitu nominal, ordinal, interval, dan rasio; masing-masing kategori mewakili satu taraf pengukuran" (Stevens, 1946). Selanjutnya, taraf pengukuran berkaitan dengan jenis statistik parametrik yang cocok untuk diterapkan. Prinsipnya,

prosedur atau teknik statistik yang makin *powerful* atau makin kuat dan makin *useful* atau makin jamak digunakan (seperti *r* Pearson, analisis varians, regresi ganda) hanya cocok diterapkan pada data yang memenuhi kriteria **skala interval** atau **rasio**. Terhadap data yang berskala nominal atau ordinal hanya cocok dikenakan prosedur statistik nonparametrik yang kurang kuat seperti khi-kuadrat, korelasi tata jenjang, dan tes median.

Sebagian besar atribut psikologis yang bersifat kontinyu dapat diukur pada taraf penskalaan ordinal atau interval. Berbagai atribut geografis yang bersifat diskret seperti jenis kelamin, status perkawinan, dan sebagainya, hanya bisa diukur pada taraf penskalaan nominal. Secara teoretis, tidak satu pun atribut psikologis dapat diukur pada taraf penskalaan rasio. Dalam praktek, sebagian besar instrumen psikologis diasumsikan menerapkan taraf pengukuran interval kendati sangat sulit untuk membuktikannya (Gregory, 2007). Beberapa metode penskalaan, yaitu penetapan aneka bilangan pada aneka respon testi untuk mencerminkan pemilikan atribut psikologis yang sedang dikur dalam jumlah yang berlainan, yang lazim diterapkan dalam pengukuran psikologis adalah (Gregory, 2007; Allen & Yen, 1979): (1) kategorisasi, (2) *rating scales* atau skala penilaian, (3) *expert rankings* atau penjenjangan oleh pakar, (4) skala Likert, (5) skala Guttman atau analisis skalogram, (6) metode *empirical keying* atau penskalaan empiris, dan (7) metode *equal-appearing intervals* atau metode interval tampak setara. Sebagaimana akan kita lihat, aneka metode penskalaan ini beroperasi atau berfungsi pada taraf penskalaan atau ordinal atau interval. Marilah kita lihat satu demi satu.

1. Kategorisasi

Sejumlah subjek penilai diminta mengkategorikan atau memilah sejumlah objek ke dalam kategori-kategori atau himpunan-himpunan atau kelompok-kelompok yang bersifat *mutually exclusive* atau tidak tumpang tindih dan *exhaustive* atau tuntas. Sejumlah

kategori, himpunan, atau kelompok disebut tidak tumpang tindih jika masing-masing objek bisa dimasukkan ke dalam *hanya* satu kategori, himpunan, atau kelompok. Contoh, "perempuan" dan "lelaki" merupakan dua kategori yang tidak tumpang tindih; sebaliknya, "warga negara Indonesia" dan "warga kota Yogyakarta" adalah dua kategori yang tumpang tindih. Kategorisasi disebut tuntas jika seluruh objek yang ada bisa dimasukkan ke dalam salah satu kategori. Misal, kita sedang mengkategorisasikan aneka jenis mobil yang beredar di Tanah Air. Ternyata, mobil dengan merek *KIAT-ESEMKA* tidak bisa dimasukkan ke dalam salah satu dari sejumlah kategori yang berhasil kita susun. Berarti, kategorisasi kita terhadap jenis mobil yang lalu lalang di jalanan di negara kita tidak tuntas. Sesudah kategorisasi selesai dilakukan, masing-masing kategori bisa kita tera atau kita kenai dengan bilangan yang berbeda sebagai nilai skalanya. Penskalaan lewat kategorisasi semacam ini jelas hanya menghasilkan pengukuran pada **taraf nominal.**

2. *Rating Scales* atau Skala Penilaian

Dalam *rating scales* atau skala penilaian, sejumlah testi diminta menyatakan secara langsung jawaban atau pendapat, keyakinan, perasaan atau sikap mereka terhadap pertanyaan atau konsep tertentu dengan tehnik tertentu. Respon atau jawaban testi tersebut selanjutnya ditransformasikan atau diubah ke dalam bilangan atau nilai skala tertentu yang dipandang mencerminkan jumlah pemilikan atribut psikologis yang sedang menjadi objek pengukuran. Ada paling sedikit dua format tehnik untuk mengungkap secara langsung jawaban, pendapat, keyakinan, perasaan atau sikap testi terhadap pernyataan atau konsep tertentu sebagai cerminan pemilikan atribut psikologis yang sedang menjadi objek pengukuran dalam jumlah tertentu. Kedua format tehnik yang dimaksud adalah: (1) skala pilihan, dan (2) *semantic differential* atau diferensial semantik. Marilah kita bahas satu demi satu.

a. Skala Pilihan

Subjek diminta secara langsung memberikan jawaban terhadap pertanyaan atau soal, atau memberikan penilaian terhadap objek pengukuran atau konsep dengan cara memilih salah satu dari antara dua atau lebih jawaban yang sudah disediakan. Beberapa contohnya adalah sebagai berikut (Allen & Yen, 1979; Crocker & Algina, 2008):

a) Akar pangkat dua dari 100 adalah 10. ( ) **Benar** ( ) **Salah**

b) Saya sering merasa cemas tanpa alasan yang jelas. ( ) **Ya** ( ) **Tidak**

c) Siapakah penulis prosa "Pengakuan Pariyem"?
   a. Ahmad Tohari.
   b. Jakob Sumardjo.
   c. Joko Pinurbo.
   d. Linus Suryadi.
   e. Wiji Tukul.

d) Seberapa memuaskankah pekerjaan yang Anda jalani sekarang?
   a. Sangat memuaskan.
   b. Memuaskan.
   c. Cukup memuaskan.
   d. Tidak memuaskan.
   e. Sangat tidak memuaskan.

Masing-masing pilihan jawaban diberi skor, misal skor 1 untuk jawaban yang sesuai dengan kunci jawaban dan skor 0 untuk jawaban yang tidak sesuai dengan kunci jawaban untuk contoh item nomor a), b) dan c). Atau, skor 4, 3, 2, 1, dan 0 untuk jawaban a sampai dengan e pada item nomor d). Skor seseorang pada skala merupakan jumlah skornya pada masing-masing item. Penskalaan dengan tehnik skala pilihan menghasilkan pengukuran pada **taraf ordinal** (Allen & Yen, 1974).

b. *Semantic Differential Scale* atau Skala Diferensial Semantik

Tehnik ini mengharuskan subjek atau testi memberikan penilaian secara langsung terhadap objek pengukuran atau konsep dalam sejenis kontinum skala yang pada kedua ujungnya dibatasi dengan dua kata sifat yang bersifat bipolar seperti contoh berikut ini (Nunnally, Jr., 1970; Kline, 1986):

Mahkamah Konstitusi

Tidak jujur	1	2	3	4	5	6	Jujur
Lemah	1	2	3	4	5	6	Kuat
Lamban	1	2	3	4	5	6	Cepat

Contoh di atas merupakan penerapan tehnik diferensial semantik untuk mengungkap penilaian masyarakat terhadap Mahkamah Konstitusi. Ada beberapa hal yang perlu digaris-bawahi di sini. Pertama, sebagaimana dinyatakan oleh Nunnally, Jr. (1970), logika penerapan diferensial semantik adalah sebagai berikut. Karakteristik gagasan maupun objek lazim dirumuskan dan dikomunikasikan dengan menggunakan aneka kata sifat. Kebanyakan kalau bukan semua kata sifat memiliki lawan atau kebalikannya, seperti "baik-buruk", "besar-kecil", dan sebagainya. Pasangan kata sifat semacam ini dapat digunakan untuk mengukur atau mengungkap cara orang memberikan makna terhadap objek atau konsep tertentu, dalam bentuk penilaian. Menurut Osgood (1962), pakar yang pertama kali mengembangkan tehnik ini, ada tiga jenis faktor makna yang lazim dipakai orang untuk memberikan penilaian terhadap objek atau konsep tertentu, yaitu (1) faktor *evaluasi* atau penilaian, seperti pada item pertama dalam contoh di atas, yaitu evaluasi atau penilaian tentang kejujuran para hakim Mahkamah Konstitusi, antara "tidak jujur" dan "jujur"; (2) faktor *potensi* atau daya atau kekuatan, seperti pada

item kedua dalam contoh di atas, yaitu penilaian tentang kedudukan dan peran Mahkamah Konstitusi dalam penegakan keadilan, antara "lemah" sampai "kuat"; dan (3) faktor *aktivitas* atau keaktivan, seperti pada item ketiga dalam contoh di atas, yaitu penilaian tentang cara kerja Mahkamah Konstitusi, antara "lamban" dan "cepat". Nunnally, Jr. (1970) menambahkan satu faktor lain, yaitu (4) faktor *familiaritas* atau kelaziman atau kejelasan, meliputi "lazim-tak lazim", "jelas-membingungkan", "sederhana-rumit". Dari antara keempat faktor tersebut, faktor evaluasi atau penilaian merupakan yang paling kuat, namun setidaknya faktor potensi dan aktivitas berkorelasi secara positif dan signifikan dengan faktor evaluasi (Nunnally, Jr., 1970). Suatu penskalaan terhadap objek atau konsep tertentu sebaiknya mencakup keempat faktor tersebut. Skor seseorang pada skala merupakan jumlah skornya pada masing-masing item. Penskalaan dengan tehnik diferensial semantik menghasilkan pengukuran pada **taraf ordinal** (Allen & Yen, 1974).

3. *Expert Rankings* atau Penjenjangan oleh Pakar

Inti metode penskalaan ini adalah sebagai berikut. Misalkan, kita akan mengukur suatu atribut psikologis tertentu, katakanlah taraf kemandirian anak usia bawah lima tahun. Penyusunan skalanya – yaitu penetapan aneka bilangan untuk mencerminkan aneka taraf pemilikan atribut psikologis kemandirian – akan mencakup langkah-langkah sebagai berikut. *Pertama,* sekelompok pakar Psikologi Anak diminta mengidentifikasikan sebanyak mungkin jenis tingkah laku yang mencerminkan aneka taraf kemandirian anak balita. *Kedua,* jenis-jenis tingkah laku yang berhasil diidentifikasikan tersebut kemudian diminta agar dikelompokkan ke dalam sejumlah bidang atau wilayah tertentu, misalnya: (a) kemampuan mengurus diri sendiri, mencakup antara lain kemampuan makan sendiri, mengenakan baju sendiri, dan sebagainya; (b) kemampuan beraktivitas secara mandiri, mencakup antara lain kemampuan berada terpisah dari

orang lain yang diakrabi seperti ayah, ibu, kakak, dan sebagainya; (c) kemampuan menghadapi situasi baru entah berupa lingkungan sosial atau lingkungan fisik, mencakup kemampuan berada di tengah orang-orang asing, kemampuan berada di tempat yang belum pernah dikunjungi, dan sebagainya. *Ketiga,* jenis-jenis respon spesifik pada tiap wilayah tersebut kemudian diminta dijenjangkan atau diurutkan dalam kontinum mulai dari yang mencerminkan ketergantungan yang tinggi atau kemandirian yang rendah (skor rendah) sampai kemandirian yang tinggi (skor tinggi). Jumlah dari skor pada masing-masing wilayah tersebut merupakan skor total taraf kemandirian anak. Makin tinggi skor total, makin baik taraf kemandiriannya, dan sebaliknya. Penskalaan dengan metode *expert rankings* atau penjenjangan oleh pakar seperti ini menghasilkan pengukuran pada **taraf ordinal**.

4. Skala Likert

Inti metode penskalaan yang dikemukakan oleh Rensis Likert (1932) ini cukup sederhana. Terhadap pernyataan-pernyataan yang kita susun dalam rangka mengukur atribut psikologis tertentu, testi diminta menyatakan kesetujuan-ketidaksetujuannya dalam sebuah kontinum yang terdiri atas lima respon: "Sangat Setuju" (*Strongly Agree*), "Setuju" (*Agree*), "Tidak tahu" (*Undecided*), "Tidak Setuju" (*Disagree*), dan "Sangat Tidak Setuju" (*Strongly Disagree*). Isi pernyataan dibedakan menjadi dua kategori: (1) pernyataan *favorable,* yaitu "statements whose endorsement indicates a positive or favorable attitude toward the object of interest"; maksudnya, pernyataan-pernyataan yang jika diiyakan menunjukkan sikap positif atau suka terhadap objek terkait; dan (2) pernyataan *unfavorable,* yaitu "statements whose endorsement indicates a negative or unfavorable attitude toward the object"; maksudnya, pernyataan-pernyataan yang jika diiyakan menunjukkan sikap negative atau tidak suka terhadap objek (Anderson, 1990). Jika isi pernyataan bersifat *favorable,* maka masing-masing respon diberi skor berturut-turut 5, 4, 3, 2, dan

1. Sebaliknya jika isi pernyataan bersifat *unfavorable*, maka masing-masing respon diberi skor 1, 2, 3, 4, dan 5. Skor total subjek adalah jumlah skor setiap pernyataan atau item. Karena jawaban subjek terhadap setiap pernyataan atau item pada dasarnya merupakan *rating* atau penilaian dan penilaian tersebut kemudian dijumlahkan untuk mendapatkan pengukuran tentang sikap subjek terhadap objek psikologis atau tentang taraf kepemilikan subjek atas atribut psikologis tertentu, maka seorang pakar psikometri lain (Bird, 1940, dalam Edwards, 1957) menyebut metode penskalaan Likert ini *method of summated ratings* atau metode penilaian terjumlahkan. Penskalaan dengan metode Likert ini menghasilkan pengukuran pada **taraf ordinal**. Uraian lebih lengkap tentang langkah-langkah penyusunan skala Likert akan disajikan di bagian lain.

5. Skala Guttman atau Analisis Skalogram

Skala Guttman dikembangkan oleh Louis Guttman (1944, 1950, dalam Abdi, 2010). Inti penyusunan skala dengan metode Guttman adalah sebagai berikut. Lagi-lagi, misalkan kita akan mengukur sikap terhadap objek tertentu atau pemilikan atribut psikologis tertentu. Yang harus kita lakukan adalah menyusun serangkaian pernyataan. Masing-masing pernyataan menunjukkan sikap terhadap sebuah objek atau pemilikian atribut psikologis tertentu, dan harus dijawab secara *biner* atau dikotomis ("Ya" atau "Tidak") oleh sekelompok testi. Tujuan penyusunan skala Guttman adalah menemukan sebuah dimensi tunggal yang dapat dipakai untuk menentukan posisi baik pernyataan maupun para subjek penjawabnya. Posisi pernyataan dan subjek pada dimensi yang ditemukan selanjutnya bisa dipakai untuk menentukan nilai numerik atau skor mereka (Abdi, 2010). Maka, skala ini memiliki dua ciri: (1) pernyataan-pernyataan mencerminkan perasaan positif yang semakin meningkat terhadap objek sikap atau terkait pemilikan atribut psikologis tertentu; (2) pemilihan (*endorsement*) suatu pernyataan menyiratkan pemilihan (*endorsement*) terhadap setiap pernyataan lain yang memiliki kadar positif yang

lebih rendah. Karena sifatnya ini, ada yang menyebut skala Guttman ini *cumulative scales* (Anderson, 1981, dalam Anderson 1990). Secara tehnis, garis besar langkah-langkah penyusunan skala Guttman adalah sebagai berikut: (1) testi diminta menyatakan setuju (*endorse*) atau tidak setuju (*do not endorse*) terhadap masing-masing dari serangkaian pernyataan yang disajikan dalam rangka mengukur atribut psikologis tertentu; setiap jawaban setuju diberi skor 1 sedangkan jawaban tidak setuju diberi skor 0; dan (2) skor testi adalah jumlah pernyataan yang disetujui atau dipilihnya (*endorsed*). Penskalaan dengan metode Guttman ini menghasilkan pengukuran pada **taraf ordinal** (Allen & Yen, 1979). Uraian lebih lengkap langkah-langkah penyusunan skala Guttman akan disajikan di bagian lain.

6. Metode *Empirical Keying* atau Penskalaan Empiris

Metode penskalaan ini tidak mengandalkan teori atau penilaian pakar melainkan mendasarkan pada proses empiris. Intinya, cara menetapkan skala pada pernyataan yang dipandang mencerminkan pemilikan atribut psikologis yang sedang diukur didasarkan pada perbedaan *endorsement* atau pengiyaan (dan penolakan) yang signifikan antara *kelompok tipikal* atau *kelompok kriteria* dan *kelompok normatif* (Gregory, 2007).

Misalkan, kita akan mengukur *sifat tahan uji.* Kita pilih suatu kelompok subjek yang secara homogen dipandang memiliki atribut psikologis yang sedang menjadi objek pengukuran dalam jumlah atau kadar yang tinggi melebihi populasi kelompok sebayanya. Misalkan kita pilih kelompok taruna AKMIL yang diasumsikan memiliki sifat tahan uji yang tinggi berkat pendidikan kemiliteran yang ketat. Inilah yang disebut *kelompok tipikal* atau *kelompok kriteria,* yaitu kelompok yang secara tipikal mewakili atribut psikologis yang sedang menjadi objek pengukuran, yaitu sifat tahan uji. Sekanjutnya dipilih kelompok subjek lain yang diambil dari antara populasi kelompok sebaya dan yang memiliki karakteristik sama seperti kelompok

tipikal kecuali terkait atribut psikologis yang sedang diukur, yaitu diasumsikan memiliki sifat tahan uji yang tidak menonjol atau rata-rata. Inilah yang disebut *kelompok normatif.* Misal, sebagai kelompok normatif dipilih kelompok mahasiswa universitas umum yang diasumsikan memiliki kadar yang berlainan terkait sifat tahan uji bahkan mungkin rendah berkat aneka kemudahan yang disediakan oleh universitasnya. Selanjutnya kedua kelompok tersebut diminta menanggapi pernyataan-pernyataan dengan isi yang bersifat *favorable* maupun *unfavorable* terhadap sifat tahan uji dalam format tertentu, misal Benar-Salah atau Setuju-Tidak Setuju. Pernyataan-pernyataan yang diiyakan oleh kelompok tipikal dengan frekuensi yang secara signifikan lebih besar dibandingkan kelompok normatif dipilih untuk dijadikan item-item skala. Penskalaan dengan metode *empirical keying* ini menghasilkan pengukuran pada **taraf ordinal**. Uraian lebih lengkap langkah-langkah penyusunan skala empiris akan disajikan di bagian lain.

7. Metode *Equal-Appearing Intervals* atau Interval Tampak Setara

Metode yang dikembangkan oleh L.L. Thurstone (1929, dalam Gregory, 2007) ini merupakan adaptasi metode *paired comparisons* yang dikembangkan oleh pakar yang sama dan yang sudah kita singgung di bab sebelumnya. Cara kerja kedua metode ini didasarkan pada *law of comparative judgments* yang dikemukakan oleh Thurstone, yaitu "a group of persons compares objects with respect to some physical property...and declares which of the pair has more of the property" (Andrich, 1990, h. 330). Metode *paired comparisons* dipandang kurang praktis sebab menuntut subjek memberikan penilaian terhadap pasangan semua pernyataan sehingga menjadi terlalu menguras waktu dan tenaga jika jumlah pernyataan yang harus dinilai secara berpasangan cukup banyak. Sebaliknya, metode *equal appearing intervals* (EAI) hanya menuntut subjek penilai memberikan satu kali penilaian komparatif terhadap setiap pasangan pernyataan sehingga

tidak masalah jika pernyataan yang harus diskala berjumlah besar (Edwards, 1957). Maka metode ini dipandang sangat cocok untuk menyusun skala sikap khususnya maupun inventori kepribadian pada umumnya. Inti metode ini, subjek diminta menilai taraf favorabilitas-unfavorabilitas serangkaian pernyataan terhadap objek sikap yang menjadi sasaran pengukuran, dengan menempatkan masing-masing pernyataan pada salah satu dari 11 kategori yang terentang antara "*extremely unfavorable*" atau "sangat tidak favorable" (kategori 1) dan "*extremely favorable*" atau "sangat favorable" (kategori), sedangkan kategori-kategori lainnya ditempatkan secara berurutan di antara kedua kutub ekstrim tersebut (Gregory, 2007). *Mean* penilaian subjek pada masing-masing pernyataan dijadikan nilai skala pernyataan yang bersangkutan, dengan catatan bahwa pernyataan yang memiliki variabilitas penilaian subjek yang terlampau besar harus digugurkan sebab hal itu menunjukkan bahwa pernyataan tersebut kabur dalam arti kurang jelas kadar *favorable* atau *unfavorable*-nya. Penskalaan dengan metode *equal appearing intervals* atau EAI ini menghasilkan pengukuran pada **taraf interval**. Uraian lebih lengkap langkah-langkah penyusunan skala dengan metode EAI akan disajikan di bagian lain.

D. Menuliskan Item

Berpedoman pada tabel spesifikasi dan dengan mempertimbangkan metode penskalaan yang dipilih, penyusun tes bisa segera memulai langkah selanjutnya yang oleh Klein (1986) disebut "the fundamental aspect of test construction" (h. 24), yaitu *item writing* atau menuliskan item. Ada dua hal penting yang perlu diperhatikan. *Pertama*, pada tahap ini penyusun tes bertugas menyusun *item pool* untuk tesnya. *Item pool* adalah *stock* atau himpunan item dari mana item-item untuk bentuk final tesnya akan diambil. Saat merencanakan tes yang akan disusunnya, penyusun tes lazim sudah memiliki gambaran tentang jumlah item dari bentuk final

tesnya. Gambaran tentang jumlah item bentuk final tes ini antara lain ditentukan oleh sifat dalam arti *content domain* dari konstruk yang diukur dan khalayak yang akan menjadi kelompok sasaran tes. Jika *content domain* konstruknya memang luas dan tes tersebut ditujukan bagi khalayak subjek dewasa, penyusun tes bisa merencanakan jumlah item yang cukup besar untuk bentuk final tesnya. Sebaliknya jika *content domain* konstruknya relatif sempit apalagi tes tersebut ditujukan bagi khalayak subjek anak-anak, maka jumlah item bentuk final tesnya lazimnya juga tidak boleh terlalu besar. Berapa pun jumlah item bentuk final tes yang direncanakan, dalam penyusunan *item pool* ada konvensi yang didasarkan pengalaman empiris bahwa dalam proses penyusunan dan pemeriksaan kualitas item-item sampai dicapai bentuk final tes lazim terjadi *mortality rate* atau angka kegagalan item sebesar 50% atau separo. Konsekuensinya, jika item-item bentuk final tes direncanakan berjumlah *n,* maka jumlah item yang harus disusun sebagai *item pool* harus sebesar minimal 2*n.* Jumlah inilah yang harus dicantumkan dalam tabel spesifikasi.

Kedua, sebagaimana sudah disinggung, pilihan atas jenis skala tertentu akan menentukan format item yang sesuai. Pembahasan tentang format item ini akan disajikan di bagian berikut saat membahas penyusunan item untuk masing-masing kategori tes.

Ada sejumlah kiat praktis yang bisa dikemukakan untuk mempermudah pelaksanaan tugas menyusun item. Pertama, tiap sel yang merupakan interseksi atau perpotongan antara kisi horisontal dan kisi vertikal dalam tabel sepesifikasi sebaiknya diberi kode tertentu dengan menggunakan sejumlah digit atau bilangan. Misalnya, bilangan pertama mewakili komponen kisi horisontal, bilangan kedua mewakili komponen kisi vertikal, dan bilangan ketiga menunjukkan urutan item dalam sel yang bersangkutan. Kedua, penulisan draft pertama item sebaiknya dilakukan dengan pensil di atas sejenis *index card* atau kartu indeks. Karena kartu indeks ini tidak diproduksi dan dipasarkan di toko-toko buku/alat tulis kita, maka kita bisa membuatnya sendiri dengan cara memotong lembar-lembar kertas manila ukuran kuarto menjadi empat potongan sama besar.

Setiap potong menjadi sejenis kartu indeks. Setiap item yang ditulis di atas satu kartu indeks diberi kode atau tanda sesuai kodenya pada tabel spesifikasi.

Manfaat penggunaan kartu indeks dan penulisan dengan pensil adalah sebagai berikut. Pertama, sebagai draft awal sangat mungkin hasil penulisan masing-masing item belum sempurna sehingga masih perlu diperbaiki mungkin hingga berkali-kali. Karena ditulis dengan pensil, suatu item tinggal dihapus seperlunya dengan karet penghapus setiap kali perlu diubah dan diperbaiki. Kedua, penulisan satu item di atas satu kartu indeks akan sangat memudahkan saat penyusun tes harus merakit item-item dengan urutan *random* menjadi bentuk semi-final tes yang siap diuji-cobakan. Kendati disusun secara *random* atau acak, letak masing-masing item dalam struktur tes sebagaimana tercermin dalam tabel spesifikasi tetap mudah dilacak sebab diberi kode atau tanda yang sesuai dengan kode atau tanda pada tabel spesifikasi.

E. Melakukan *Review* dan Revisi Item

Sesudah item-item berhasil disusun berpedoman pada tabel spesifikasi maupun pedoman penulisan item lain yang relevan, item-item tersebut terlebih dulu perlu dimintakan *review* dari nara sumber terkait untuk kemudian dilakukan revisi seperlunya sesuai saran *reviewer* sebelum dirakit menjadi bentuk semi-final yang siap diuji-cobakan. Menurut Crocker dan Algina (2008), *review* atau telaah item tersebut akan mencakup pemeriksaan terhadap hal-hal sebagai berikut: (a) *accuracy* atau ketepatan, yaitu ketepatan rumusan konseptual konstruk atau atribut psikologis yang diukur beserta rumusan operasional sampai ke indikator-indikator tingkah laku, bahkan sampai ke pilihan format item-itemnya, (b) relevansinya dengan tabel spesifikasi, khususnya terkait kesesuaian antara isi item dengan baik komponen isi maupun komponen proses, serta jumlah item sesuai distribusi yang sudah direncanakan dalam tabel spesifikasi, (c) ada-

tidaknya berbagai kesalahan tehnis penyusunan item, seperti adanya lebih dari satu gagasan atau problem dalam satu item, penggunaan bentuk negatif atau kata-kata yang bisa memberi petunjuk arah jawaban seperti "selalu", "tidak pernah", dan sebagainya, (d) tata bahasa dan ejaan, (e) pilihan kata yang bisa menimbulkan kesan menyinggung perasaan atau mendiskriminasikan kelompok tertentu, seperti penggunaan contoh "kereta api" bagi kelompok subjek di Nusa Tenggara Timur atau Papua, dan (f) taraf kesulitan bahasa yang dipakai dibandingkan dengan kelompok khalayak yang akan dikenai tes.

Ada dua saran praktis yang perlu dikemukakan di sini. Pertama, terkait cara penyajian draft *item pool* untuk dimintakan *review* atau telaah. Untuk memudahkan para narasumber melakukan *review* atau telaah, sebaiknya draft *item pool* disajikan secara sistematis mengikuti pembagian baik kisi *content* maupun kisi prosesnya sesuai tabel spesifikasi. Selain itu, draft lengkap rancangan tes mulai dari definisi sampai dengan tabel spesifikasi sebaiknya juga disertakan, agar para narsumber dapat menjalankan tugasnya melakukan telaah *item pool* secara optimal.

Kedua, nara sumber yang perlu dilibatkan bisa bermacam-macam, meliputi pakar dalam aneka bidang terkait maupun awam. Kelompok pakar bisa mencakup ahli spesialis dalam bidang studi atau disiplin tertentu serta guru atau dosen pengampu mata pelajaran atau mata kuliah tertentu. Mereka ini terutama bisa diminta melakukan telaah dalam aspek *accuracy* atau ketepatan rumusan konstruk serta relevansi draft tes dengan tabel spesifikasi. Kelompok pakar juga perlu mencakup ahli psikometri agar bisa memberikan telaah khususnya terhadap ketepatan pilihan format item serta ada tidaknya berbagai kesalahan tehnis penyusunan item. Yang dimaksud awam sesungguhnya merupakan ahli juga, yaitu narasumber yang tidak perlu berstatus pakar namun yang mengenal secara mendalam karakteristik kelompok khalayak yang akan dikenai tes. Secara khusus narasumber awam ini bisa dimintai telaah terkait pilihan kata yang

tidak akan menimbulkan bias tertentu serta taraf kesulitan bahasa yang sesuai bagi kelompok khalayak sasaran.

Semua masukan yang berhasil dikumpulkan dari berbagai narasumber tersebut perlu diolah secara kritis dan dipakai sebagai dasar untuk melakukan aneka perbaikan atau revisi terhadap draft *item pool*. Salah satu kemungkinan kecenderungan kurang baik dari penyusun apa saja, termasuk penyusun tes, adalah bersikap *defensif* merasa lebih tahu dari orang lain dan tidak mudah menerima masukan. Setiap catatan kritis dari nara sumber harus diterima sebagai petunjuk tentang kemungkinan adanya sesuatu yang memang belum beres dalam draft item kita, maka harus dicermati dan ditindak lanjuti dengan melakukan revisi seperlunya.

F. Merakit Item

Sesudah diperoleh *item pool* yaang secara konseptual dipandang sudah memenuhi standar kualitas yang diharapkan, langkah berikutnya adalah merakit item-item tersebut menjadi bentuk semi-final tes yang siap untuk diuji-cobakan baik dalam skala kecil maupun besar. Secara umum langkah ini akan mencakup dua hal, yaitu (1) pemberian petunjuk pengerjaan tes, dan (2) penyusunan item-item secara random dalam format tes.

1. Petunjuk Pengerjaan Tes

Petunjuk pengerjaan tes terdiri atas tiga jenis, yaitu (1) *petunjuk umum* yang lazim disajikan di bagian awal tes; (2) *petunjuk bagian* tentang cara mengerjakan atau menjawab item terkait format item yang lazim disajikan pada awal bagian tes khususnya jika tes dibagi ke dalam beberapa bagian sesuai format itemnya, misal sebuah tes *maximal performance test* dibagi menjadi bagian-bagian meliputi bagian dengan item-item Benar-Salah, bagian dengan item-item *matching* atau menjodohkan, dan bagian dengan item-item pilihan ganda, serta (3)

petunjuk khusus tentang cara mengerjakan item atau kelompok item khusus, misal item pilihan ganda dengan *stem* berupa gambar, tabel, grafik, atau peta dan sejenisnya atau kelompok item yang didahului dengan teks sebagai sumber soal dan jawabannya.

Hal-hal penting yang harus termuat dalam *petunjuk umum* pengerjaan tes adalah sebagai berikut (Gronlund, 1977): (1) informasi tentang tujuan tes, khususnya tes itu hendak mengukur apa; tujuannya adalah membantu testi memiliki orientasi sehingga termotivasi dalam mengerjakan tes, maka informasi ini harus seperlunya dalam arti tidak boleh terlampau rinci sehingga justru berdampak mempengaruhi cara testi menjawab dan merusak validitas tes, (2) tugas yang harus dikerjakan dalam tes, termasuk jumlah item serta waktu yang disediakan untuk mengerjakan tes, (3) cara mengerjakan tes, meliputi cara mengerjakan atau menjawab item, misal dalam item pilihan ganda testi harus memilih jawaban benar di antara alternatif jawaban lain yang salah atau memilih jawaban paling benar di antara alternatif jawaban lain yang sama-sama benar; tempat menuliskan jawaban apakah langsung di buku tes atau dalam Lembar Jawab yang terpisah dari buku tes; dan jika dipandang perlu cara memperbaiki atau mengubah jawaban yang keliru; (4) informasi tentang sikap yang harus dihayati dalam mengerjakan tes, meliputi misalnya sikap spontan apa adanya dan tidak terlalu lama memikirkan jawaban dalam mengerjakan *typical performance tests* atau boleh-tidaknya menebak dalam mengerjakan *maximal performance tests* beserta akibatnya terhadap penskoran; (5) informasi tentang kewajiban testi mengerjakan seluruh item tes, jangan ada yang terlewati; (6) khusus dalam administrasi tes dalam rangka pengumpulan data penelitian, informasi tentang dampak hasil tes terhadap aspek tertentu dari kehidupan testi seperti penilaian dalam pendidikan di sekolah, karir dalam pekerjaan, dan sebagainya, khususnya berupa jaminan bahwa hasil tes tidak memiliki pengaruh apa pun terhadap semua aspek tersebut.

2. Perakitan Item Menjadi Bentuk Semi-final Tes

Persoalan utama dalam langkah ini adalah menemukan cara terbaik untuk mengurutkan item-item sehingga diperoleh bentuk semi-final tes yang siap diuji-cobakan. Secara umum, persoalan tersebut akan berkisar pada sejumlah pertanyaan sebagai berikut: (a) apakah item-item akan dikelompokkan per komponen, jika konstruk yang diukur mencakup sejumlah komponen; (b) apakah item-item akan dikelompokkan per format item, jika item-item disusun menggunakan lebih dari satu format; (c) dalam masing-masing pengelompokan yang diterapkan, apakah item-item perlu diurutkan mengikuti sistematika tertentu, misal taraf kesukarannya untuk *maximal performance tests,* atau diurutkan secara *random* atau acak; atau (d) apakah item-item secara keseluruhan pada dasarnya langsung diurutkan secara acak, kecuali penempatan secara sengaja sejumlah item dengan taraf kesulitan rendah pada awal untuk membangkitkan motivasi testi khususnya pada *maximal performance tests*?

Khusus untuk jenis *maximal performance test* berupa baik *achievement tests* maupun *aptitude tests,* ada yang menyarankan agar item-item dikelompokkan sesuai komponen konstruk atau *learning outcomes* yang diukur dan sesuai format item, serta diurutkan berdasarkan taraf kesukaran masing-masing item dimulai dari yang paling mudah sampai ke yang paling sukar (Gronlund, 1977). Cara pengurutan item sebagaimana disarankan tersebut memang sesuai dengan konsep *maximal performance tests* sendiri, namun tetap terbuka cara lain berupa pengurutan seluruh item langsung secara acak kecuali penempatan secara sengaja sejumlah item yang mudah di awal tes untuk memotivasi testi. Sebaliknya untuk jenis *typical performance tests* item-itemnya lazim langsung diurutkan secara acak secara keseluruhan tanpa menerapkan pengelompokan.

Untuk keperluan pengurutan item-item secara acak, saran untuk menuliskan masing-masing draft item dalam sebuah kartu indeks sebagaimana sudah disinggung di muka akan terbukti sangat

memudahkan pekerjaan menyusun item-item menjadi bentuk semifinal tes. Penyusun tes tinggal mengocok kartu-kartu item sampai diperoleh urutan paling memuaskan. Masing-masing kartu selanjutnya ditandai dengan nomor yang merupakan nomor urut masing-masing item dalam tes. Pencatatan identitas masing-masing item secara jelas baik menurut tempatnya dalam tabel spesifikasi maupun menurut urutan tempatnya dalam tes akan memudahkan penyusun tes saat harus melaporkan struktur tes melalui tabel spesifikasi.

G. Melakukan Uji Coba

Bentuk semi final tes yang berhasil dirakit selanjutnya perlu diuji cobakan. Idealnya uji coba ini dilaksanakan dalam dua tahap, yaitu (1) uji coba pendahuluan dengan sampel kecil, dan (2) uji coba yang sesungguhnya dengan sampel besar (Crocker & Algina, 2008).

1. Uji Coba Pendahuluan

Uji coba pendahuluan perlu dilakukan untuk mendapatkan sejumlah masukan awal. Uji coba pendahuluan ini disarankan dilaksanakan pada sekelompok sampel testi yang memiliki karakteristik seperti populasi khalayak sasaran tes dalam jumlah yang tidak terlalu besar, yaitu sekitar 15-30 orang, dan dilaksanakan secara informal (Crocker & Algina, 2008). Sifat informal yang dimaksud antara lain berupa penyediaan waktu pengerjaan tes yang longgar, sebab penyusun justru bertujuan antara lain menentukan waktu pengerjaan yang sesuai untuk diterapkan dalam uji coba yang sesungguhnya.

Jenis masukan yang perlu dikumpulkan pada tahap uji coba pendahuluan ini mencakup (1) efektivitas petunjuk pengerjaan tes, yaitu apakah petunjuk tersebut dipahami oleh testi sehingga tidak menimbulkan pertanyaan atau bahkan kebingungan atau kesalahan dalam mengerjakan tes; (2) efektivitas item-item tes, yaitu apakah

item-item dipahami dan dikerjakan secara semestinya oleh testi; (3) rerata waktu yang diperlukan oleh testi untuk menyelesaikan tes; dan (4) statistik deskriptif dari distribusi respon testi pada masing-masing item sebagai evidensi tambahan tentang efektivitas item-item.

Untuk memperoleh masukan yang dimaksud dari testi, Crocker dan Algina (2008) menyarankan dua cara, yaitu (1) pengamatan cermat terhadap perilaku testi selama mengerjakan tes, termasuk jika muncul pertanyaan-pertanyaan terkait baik petunjuk pengerjaan maupun item-item tes; dan (2) penyelenggaraan *debriefing,* yaitu pemberian kesempatan kepada testi untuk memberikan komentar kritis dan saran perbaikan terhadap masing-masing item, sesudah administrasi tes selesai dilaksanakan. Seluruh masukan yang diperoleh perlu ditindaklanjuti seperlunya dalam rangka menyempurnakan bentuk semi final tes sebelum melakukan uji coba tes yang sesungguhnya.

2. Uji Coba Sesungguhnya

Bentuk semi final tes sebagai *item pool* yang sudah dicoba disempurnakan melalui uji coba pendahuluan selanjutnya perlu diuji-cobakan dengan sesungguhnya dengan menggunakan sampel testi yang memiliki karakteristik seperti populasi khalayak sasaran tes dalam jumlah yang cukup besar. Lantas seberapa besar sampel testi yang harus digunakan? Ada yang merekomendasikan minimal 50 orang (Allen & Yen, 1979) namun ada pula yang merekomendasikan minimal 200 orang (Crocker & Algina, 2008). Sebagai pedoman kasar sebaiknya digunakan sampel testi sebesar antara 5 sampai 10 kali jumlah bentuk final item sebagaimana disarankan oleh Nunnally (1967, dalam Crocker & Algina, 2008).

Tujuan akhir langkah ini adalah mendapatkan sebuah tes dengan panjang minimum namun mampu menghasilkan pengukuran dengan taraf reliabilitas dan validitas yang memadai sesuai tujuannya (Crocker & Algina, 2008). Tujuan kegiatan ini adalah memperoleh data statistik untuk melakukan pemeriksaan ciri-ciri psikometrik

masing-masing item maupun tes secara keseluruhan melalui kegiatan *analisis item*.

H. Analisis Item

Tujuan utama analisis item adalah memeriksa ciri-ciri statistik respon testi dalam uji coba yang sesungguhnya terhadap masing-masing item untuk keperluan seleksi item, yaitu memutuskan item-item mana yang dipandang langsung memenuhi syarat untuk dimasukkan ke dalam bentuk final tes, mana yang perlu terlebih dulu direvisi dan diuji-cobakan kembali sebelum dimasukkan ke dalam bantuk final tes, dan mana yang harus langsung digugurkan karena memiliki ciri-ciri statistik yang terlalu jauh dari yang dipersyaratkan. Lantas apa saja ciri-ciri statistik atau parameter item yang perlu diperiksa untuk dijadikan dasar dalam seleksi atau pemilihan item untuk bentuk final tes?

Untuk menjawab pertanyaan di atas, terlebih dulu perlu diuraikan jenis pendekatan atau metode konstruksi atau penyusunan tes yang mendasari pemilihan jenis parameter item yang perlu diperiksa sebagai dasar seleksi item. Salah satu metode konstruksi tes yang paling populer dan yang juga akan kita terapkan dalam pembahasan kita adalah **metode rasional** atau lebih tepat **metode konsistensi internal.** Ciri utama metode konstruksi tes ini adalah bahwa *seluruh item tes berkorelasi secara positif satu sama lain dan juga berkorelasi positif dengan skor total tes* (Gregory, 2007). Sesuai ciri utama netode rasional atau konsistensi internal dalam penyusunan tes, maka jenis parameter item yang perlu diperiksa untuk dijadikan dasar atau kriteria dalam seleksi item adalah sebagai berikut (Crocker & Algina, 2008):

1. Aneka indeks yang melukiskan distribusi respon testi terhadap masing-masing item, khususnya proporsi testi yang menjawab item sesuai dengan kunci jawaban.

2. Aneka indeks yang melukiskan taraf hubungan antara respon terhadap item dan kriteria tertentu, khususnya skor total tes sebagai kriteria internal.
3. Aneka indeks yang merupakan fungsi dari varians dan hubungan antara item dengan suatu kriteria, khususnya skor total tes sebagai kriteria internal.

Uraian lebih rinci tentang masing-masing kategori indeks akan disajikan di bagian lain. Kini cukuplah kita tegaskan kembali bahwa berdasarkan aneka parameter tersebut penyusun memilih item-item dalam jumlah dan dengan struktur sebagaimana sudah direncanakan dalam tabel spesifikasi untuk dijadikan bentuk final tes. Langkah ini seringkali tidak bisa sekali jadi, melainkan menuntut beberapa kali melakukan revisi dan menguji-coba item-item sampai diperoleh item-item yang dipandang memuaskan untuk dijadikan bentuk final tes.

I. Memeriksa Reliabilitas, Validitas, & Daya Diskriminasi

Sesudah diperoleh seperangkat item sebagai bentuk final tes sebagaimana direncanakan, langkah berikut adalah memastikan bahwa bentuk final tes yang diperoleh tersebut sungguh-sungguh menghasilkan pengukuran yang bisa ditafsirkan mencerminkan pemilikan atribut psikologis dalam taraf tertentu sebagaimana dimaksudkan. Evidensi tentang hal ini lazim dikumpulkan dengan memeriksa reliabilitas, validitas, dan daya diskriminasi keseluruhan item sebagai bentuk final tes.

1. Reliabilitas

Sebagaimana sudah kita lihat, reliabilitas adalah ketepatan pengukuran tanpa menghiraukan atribut apa yang diukur (Nunnally, 1974). Juga sudah kita lihat, secara psikometrik reliabilitas menunjuk pada dua ciri dalam tes, yaitu (1) *self-consistency* atau konsistensi

internal, yaitu konsistensi antar bagian-bagian dalam tes, dan (2) stabilitas, yaitu konsistensi antar waktu dari hasil tes (Klein, 1986). Sesuai pendekatan rasional atau pendekatan konsistensi internal dalam analisis item yang diterapkan, konsistensi antar bagian-bagian dalam tes sebagai evidensi ketepatan pengukuran memang perlu diperiksa. Evidensi ini sekaligus sudah diperoleh saat melakukan pemeriksaan korelasi antara masing-masing item dengan skor total sebagai kriteria dalam langkah analisis item jika pemeriksaan tersebut dilakukan dengan menggunakan program *SPSS.* Konsistensi hasil pengukuran antar waktu sebagai evidensi ketepatan pengukuran juga bisa diperiksa dengan mengadministrasikan tes yang sama pada kelompok sampel yang sama dalam dua kesempatan berbeda dan memeriksa korelasi antara dua rangkaian skor yang dihasilkan. Menurut Guilford (1956, dalam Klein, 1986), koefisien minimum yang dipandang memuaskan untuk reliabilitas tes adalah 0,70. Di bawah angka tersebut, "a test becomes unsatisfactory for use with individuals because the standard error of an obtained score becomes so large that interpretation of scores is dubious" (Guilford, dalam Klein, 1986, h. 3). Artinya, di bawah koefisien reliabilitas sebesar 0,70 sebuah tes menjadi kurang memadai untuk digunakan bagi perorangan sebab hal itu menunjukkan bahwa kesalahan baku skor tampak sedemikian besar sehingga interpretasi skor menjadi meragukan. Di luar itu semua, kita sudah melihat bahwa reliabilitas merupakan syarat yang perlu untuk sebuah tes yang baik namun ternyata tidak cukup. Tes yang sungguh-sungguh baik terutama haruslah menghasilkan pengukuran yang valid.

2. Validitas

Sudah kita lihat di bab sebelumnya, pemahaman yang baru tentang validitas beserta metode estimasinya lebih menekankan makna validitas sebagai taraf sejauh mana penafsiran terhadap hasil suatu tes sebagaimana dimaksudkan oleh tes yang bersangkutan sungguh-sungguh dapat dipertanggungjawabkan. Pertanggungjawaban

tersebut menuntut pengumpulan evidensi atau bukti-bukti terkait berbagai aspek tes dan yang dilakukan pada berbagai tahap penyusunan dan penggunaan tes dalam sejenis proses yang bersifat jangka panjang, berkesinambungan, dan kumulatif. Seperti sudah diuraikan di Bab 6 ada lima jenis evidensi yang perlu dikumpulkan dalam rangka memeriksa validitas tes, tepatnya validitas penafsiran skor atau hasil pengukuran dengan suatu tes sesuai tujuan penyusunan tes yang bersangkutan, yaitu: (a) evidensi terkait isi tes, (b) evidensi terkait proses respon testi, (c) evidensi terkait struktur internal tes, (d) evidensi terkait hubungan antara tes yang bersangkutan dengan tes lain, dan (e) evidensi terkait konsekuensi atau dampak dari pengetesan.

Pengumpulan evidensi terkait isi tes sesungguhnya sudah terjadi pada beberapa langkah awal penyusunan tes, khususnya pada langkah-langkah pendefinisan ranah isi tes, penyusunan tabel spesifikasi, dan telaah tes khususnya terkait ketepatan rumusan konstruk dan relevansi isi item-item dengan rumusan konstruknya. Pada tahap sesudah diperoleh bentuk final tes ini, hal-hal tersebut kiranya bisa dan perlu ditelaah sekali lagi khususnya terkait relevansi isi item-itemnya mengingat untuk sampai pada bentuk final kemungkinan telah dilakukan revisi berulangkali terhadap sebagian item-itemnya.

Evidensi terkait proses respon testi dalam tes perlu dikumpulkan dari dua pihak, yaitu dari pihak testi dan dari pihak asesor yang bertanggung jawab melaksanakan dan mengolah hasil tes secara penuh. Dari pihak testi, perlu dicermati bahkan diperiksa apakah respon yang diberikan oleh testi dalam rangka mengerjakan tes sunguh-sungguh sesuai dengan konstruk yang diukur oleh tes dan bukan ditentukan oleh berbagai jenis *response sets*-nya seperti *social desirability, acquiescence,* pemilihan jawaban di tengah atau sebaliknya jawaban ekstrim dalam *typical performance tests,* atau *guessing* atau menebak dalam *maximal performance tests.* Sebagaimana sudah disinggung, beberapa strategi untuk mengumpulkan jenis evidensi ini meliputi (1) mengobservasi testi saat sedang mengerjakan tugas

dalam rangka tes, dan (2) mewawancarai testi untuk mengetahui alasan mereka memberikan jawaban tertentu terhadap pertanyaan-pertanyaan dalam tes. Dari pihak asesor, ancaman terhadap validitas terkait proses respon testi bisa bersumber dari ketidak-cermatan asesor dalam menjalankan tugas merekam, menskor, dan menafsirkan hasil kinerja testi. Seperti sudah disinggung, salah satu langkah untuk mengatasinya adalah mengevaluasi cara asesor menerapkan kriteria yang semestinya dalam mencatat dan mengevaluasi hasil kinerja testi dalam tes.

Seperti sudah disinggung, evidensi terkait struktur internal tes menyangkut konsistensi internal atau homogenitas tes, yaitu sejauh mana item-item dan komponen-komponen dalam tes saling berhubungan sedemikian rupa sesuai dengan konstruk yang diukur. Salah satu evidensi untuk aspek ini sesungguhnya sudah diperoleh pada langkah analisis item, yaitu melalui pemeriksaan aneka indeks yang melukiskan taraf hubungan antara respon testi terhadap masing-masing item dengan skor total tes sebagai kriteria internal serta aneka indeks yang merupakan fungsi dari varians dan hubungan antara masing-masing item dengan skor total tes sebagai kriteria internal. Evidensi lain bisa diperoleh melalui penerapan analisis faktor konfirmatori dan tehnik DIF sebagaimana sudah disinggung di bab sebelumnya.

Evidensi terkait hubungan antara tes yang sedang kita susun dengan tes atau tingkah laku lain sebagai kriteria eksternal dapat kita peroleh melalui pemeriksaan efektivitas tes dalam memprediksikan kriterianya baik berupa tes lain maupun tingkah laku nyata tertentu, penerapan metode *multitrait-multimethod* untuk memperoleh evidensi konvergen maupun evidensi diskriminan, serta melakukan *group-comparison studies* terkait konstruk yang diukur oleh tes yang kita susun.

Sebagaimana sudah kita lihat, evidensi terkait konsekuensi atau dampak dari penerapan tes yang kita susun terhadap testi bisa dibedakan ke dalam dampak yang direncanakan dan yang tidak direncanakan. Yang pertama dapat kita peroleh melalui pemeriksaan

secara cermat apakah dampak langsung pengetesan sebagaimana dimaksudkan oleh tes memang terjadi, seperti diperolehnya diagnosis yang tepat tentang keadaan testi terkait atribut psikologis yang diukur oleh tes yang kita susun. Yang kedua dapat kita peroleh melalui pemeriksaan secara cermat apakah dalam diri testi timbul dampak-dampak lain baik positif maupun negatif yang secara konseptual-teroretis dapat dijelaskan kaitannya dengan atribut psikologis yang diukur oleh tes yang kita susun.

3. Daya Diskriminasi

Sesudah berhasil mengumpulkan evidensi secara memadai bahwa tes kita reliabel dan terbukti menghasilkan pengukuran yang valid dalam arti bahwa penafsiran terhadap hasilnya sebagaimana dimaksudkan oleh tes yang bersangkutan sungguh-sungguh dapat dipertanggungjawabkan, masih perlu satu evidensi tambahan tentang sejauh mana tes secara keseluruhan memiliki daya diskriminasi yang baik. Ciri ini lazim diperiksa dengan menghitung koefisien delta Ferguson. Tes yang berdaya diskriminasi baik lazimnya memiliki koefisien delta Ferguson ≥ 0,90. Rumus perhitungan koefisien delta Ferguson adalah sebagai berikut (Klein, 1986):

$$\delta = (n + 1)(N^2 - \sum fi^2) / nN^2 \qquad \textbf{Rumus 9.1.}$$

Langkah-langkah perhitungan koefisien delta Ferguson:

a. Buatlah distribusi frekuensi skor skala.
b. Kuadratkanlah masing-masing frekuensi dan jumlahkanlah: $\sum fi^2$
c. Tambahkanlah bilangan 1 pada jumlah item: $n + 1$
d. Kuadratkanlah jumlah subjek: N^2
e. Kalikanlah jumlah item dengan hasil perhitungan d: nN^2
f. Masukkanlah unsur-unsur di atas ke dalam rumus, dan hitung atau selesaikanlah.

J. Menyusun Manual & Menerbitkan Tes

Sesudah diperoleh bentuk final tes yang kurang lebih memuaskan, langkah terakhir adalah menyusun manual tes dan menerbitkan tes tersebut beserta manualnya. Manual atau buku pedoman tentang tes sebagai *supporting documentation* atau dokumen pendukung berfungsi penting sebagai media bagi penyusun, penerbit, dan distributor tes untuk mengkomunikasikan perihal tes tersebut kepada para pemakai tes. Manual sebagai dokumen pendukung tes bertujuan memberikan informasi yang diperlukan kepada para pemakai tes agar memiliki penilaian yang tepat tentang seluk-beluk dan kualitas tes, skor-skor yang dihasilkan, serta cara merumuskan penafsiran hasil tes berdasarkan skor-skor. Manual yang baik harus memenuhi kriteria *lengkap, tepat, tidak basi,* dan *jelas* (AERA, APA, & NCME, 1999).

Secara garis besar manual tes yang baik harus mencakup dua komponen, yaitu (1) *user's manual* atau manual bagi pemakai tes, dan (2) *technical manual* atau manual teknis. Jenis informasi yang harus tercakup dalam dua komponen manual tersebut meliputi hal-hal sebagai berikut (AERA, APA, & NCME, 1999; Gregory, 2008):

1. *User's manual,* meliputi informasi tentang *the nature of the test* atau tes tersebut mengukur apa, *its intended use* atau manfaat dan kegunaan tes, peringatan tentang kemungkinan bentuk-bentuk penyalahgunaan tes, beberapa contoh hasil penelitian empiris terkait manfaat umum dan manfaat khusus tes, deskripsi tentang populasi subjek sasaran tes, spesifikasi tes, format item, prosedur penskoran, dan proses penyusunan tes.
2. *Technical manual,* meliputi informasi tentang cara penskoran tes, cara menafsirkan skor tes, aneka indeks psikometrik tentang item-item tes, evidensi tentang validitas dan reliabilitas tes, norma dan penskalaan, kualifikasi yang diperlukan untuk mengadministrasikan dan menginterpretasikan tes, dan pedoman administrasi tes.

Sesudah bentuk final tes beserta dokumen pendukungnya tersedia secara lengkap dan memadai, langkah terakhir adalah membawa keseluruhan naskah tersebut untuk diterbitkan dan didistribusikan kepada khalayak pengguna yang berhak. Sayang, di Indonesia praktik semacam ini belum terjadi apalagi menjadi kebiasaan. Ψ

Bab 10
Penyusunan
Maximal Performance Tests

Sebagaimana sudah dibahas di Bab 4, *maximal performance tests* merupakan kategori tes yang bertujuan mengukur aneka atribut psikologis yang termasuk ke dalam ranah kognitif dan ranah psikomotor, dengan cara menentukan batas maksimal atau batas atas atribut yang dimaksud yang terdapat dalam diri testi (Friedenberg, 1995). Dalam bab ini pembahasan akan difokuskan pada jenis-jenis atribut psikologis yang termasuk ke dalam ranah kognitif, khususnya berupa *developed abilities* baik yang memiliki latar belakang pengalaman luas sebagai faktor pembentuknya yang secara popular disebut *aptitudes* maupun yang memiliki latar belakang pengalaman sempit sebagai faktor pembentuknya yang secara popular disebut *achievements.* Kedua pokok bahasan tersebut akan kita bahas satu demi satu.

A. Penyusunan *Aptitude Tests*

Mengikuti cara pemahaman baru tentang *aptitudes* sebagai jenis *developed abilities* yang memiliki latar belakang pengalaman luas sebagai faktor pembentuknya, secara khusus pada bagian ini kita akan membahas penyusunan tes inteligensi dan tes bakat jenis verbal, meliputi aneka tes kecerdasan umum dan tes bakat khusus yang mengandalkan media bahasa. Mengacu pada langkah-langkah umum penyusunan tes yang sudah kita bahas di Bab 9, secara selektif kita hanya akan membahas tiga langkah yang bersifat khas dalam penyusunan tes inteligensi dan tes bakat, yaitu langkah-langkah (1) mendefinisikan tes khususnya sublangkah menetapkan ranah isi tes, (2) menulis item, dan (3) melakukan analisis item.

1. Mendefinisikan Tes

Salah satu sublangkah penting dalam langkah pertama penyusunan tes adalah menetapkan *behavioral content domain* atau ranah isi perilaku dari atribut psikologis yang menjadi objek atau sasaran tes. Sebagaimana sudah disinggung di Bab 9, sebagian besar bahkan mungkin seluruh atribut psikologis yang termasuk ke dalam kategori *aptitudes* merupakan konsep atau konstruk teoretis yang didominasi oleh dimensi atau fungsi kognitif dan yang merupakan temuan atau hasil konstruksi para ahli psikologi yang juga memiliki minat di bidang psikometri atau pengukuran psikologis. Seperti sudah disinggung di Bab 9, konstruk dalam arti sempit ini dipandang memiliki *behavioral content domain* atau ranah isi perilaku dengan batas-batas yang tidak atau belum jelas (Cronbach & Meehl, 1955). Untuk mendefinisikan atau mengidentifikasikan ranah isi perilaku konstruk teoretis semacam itu perlu ditempuh strategi khusus yang disebut *eksplikasi konstruk* (Friedenberg, 1995).

Secara lengkap, eksplikasi konstruk terhadap suatu atribut psikologis tertentu diawali dengan perumusan **definisi konseptual** atau **definisi teoretis** atribut psikologis yang bersangkutan. Mengutip pendapat Blalock, Jr. (1979), di sini atribut psikologis yang hendak diukur tersebut dipandang sebagai sebuah konsep baru, dan perlu terlebih dulu dijelaskan dengan menggunakan konsep-konsep lain yang sudah lebih dikenal. Langkah ini perlu dilakukan secermat dan serinci mungkin sampai berhasil mengidentifikasikan komponen-komponen atau dimensi-dimensi atribut psikologis yang bersangkutan sebagai sebuah konstruk.

Langkah berikutnya, rumusan konseptual tentang atribut psikologis beserta komponen-komponennya yang masih abstrak tersebut perlu dikongkretkan dengan cara merumuskan **definisi operasional-**nya. Pada tahap ini definisi operasional bisa dimaknai sebagai definisi tentang atribut psikologis dalam bentuk rumusan tentang jenis-jenis operasi berupa perilaku atau tindakan aktual yang dipandang sebagai indikator atau perwujudan kongkret

dari keberadaan atribut psikologis yang bersifat abstrak tersebut. Dengan kata lain, inti perumusan definisi operasional adalah mengidentifikasikan jenis-jenis tingkah laku atau tindakan aktual-kongkret yang dipandang sebagai indikator keberadaan atribut psikologis yang menjadi sasaran pengukuran. Sebagaimana sudah disinggung, dalam rangka eksplikasi konstruk identifikasi bentuk-bentuk tingkah laku spesifik yang bisa diamati dan diukur sebagai indikator tingkah laku atribut psikologis yang menjadi sasaran pengukuran perlu mencakup baik yang bersifat mendukung (*favorable*) maupun yang bersifat menyangkal atau mengingkari (*unfavorable*) keberadaan konstruk psikologis yang bersangkutan. Melalui eksplikasi konstruk batas-batas *behavioral content domain* sebuah atribut psikologis hasil konstruksi konseptual-teoretis seorang ahli psikologi yang bersifat abstrak menjadi jelas.

Sebagaimana sudah disinggung di Bab 9, Crocker dan Algina (2008) mengemukakan lima tehnik yang bisa diterapkan dalam rangka mengidentifikasikan *behavioral content domain* sebuah konstruk, yaitu: (a) *content analysis* atau analisis isi, (b) *review of research* atau telaah hasil-hasil penelitian terdahulu, (c) *critical incidents* atau pengamatan terhadap peristiwa-tindakan luar biasa, (d) *direct observations* atau pengamatan langsung, dan (e) *expert judgments* atau penilaian oleh sejumlah pakar. Tehnik-tehnik tersebut bisa dipandang sebagai variasi metode dalam melakukan eksplikasi konstruk. Modifikasi pemaknaan tehnik-tehnik tersebut sebagai variasi metode dalam eksplikasi konstruk akan diuraikan di bawah ini.

a. *Content Analysis* atau Analisis Isi

Inti analisis isi adalah memilah respon berupa ungkapan verbal khususnya yang diberikan secara tertulis dari sekelompok subjek ke dalam sejumlah kategori sesuai makna atau isinya. Misal, kita akan menyusun tes tentang bakat berbahasa untuk populasi subjek remaja seusia siswa SMA. Sejumlah sampel siswa SMA di beberapa kota besar kita minta secara bebas-terbuka menuliskan sebanyak mungkin bentuk-bentuk perilaku orang seusia mereka yang mencerminkan

kemampuan tinggi dalam berbahasa maupun bentuk-bentuk perilaku yang mencerminkan kemampuan rendah dalam berbahasa. Hasil pekerjaan mereka dapat kita pilah dalam beberapa tahap: (1) tahap pertama, kita pilah dulu berdasarkan relevansinya: jawaban yang tidak relevan kita coret, sedangkan yang relevan kita pertahankan; (2) tahap kedua, jawaban-jawaban tersebut kita pilah ke dalam komponen-komponen kemampuan berbahasa, misal *kekayaan kosa kata, kepekaan ejaan, kepekaan tanda baca,* dan sebagainya; dan (3) tahap ketiga atau terakhir, dalam masing-masing komponen jawaban-jawaban tersebut kita pilah berdasarkan *favorableness*-nya, yaitu jawaban yang mencerminkan kemampuan tinggi dalam berbahasa kita masukkan ke dalam kategori *favorable,* sedangkan jawaban yang mencerminkan kemampuan rendah dalam berbahasa kita masukkan ke dalam kategori *unfavorable*. Hasilnya adalah *behavioral content domain* dari konstruk kemampuan berbahasa kelompok subjek usia SMA.

b. *Review of Research* atau Telaah Hasil Penelitian

Melanjutkan contoh penyusunan tes bakat berbahasa, alih-alih meminta sejumlah sampel subjek untuk menuliskan aneka bentuk perilaku yang mencerminkan kemampuan berbahasa atau sebaliknya lalu menganalisis isinya, dalam metode ini penyusun tes menelaah kepustakaan baik berupa buku-buku maupun khususnya artikel-artikel hasil penelitian dan hasil pemikiran yang membahas tentang bakat atau kemampuan berbahasa. Menurut Crocker dan Algina (2008), telaah pustaka ini bisa dilakukan secara *eklektik* dengan memeriksa semua bahan pustaka yang relevan dari berbagai penulis-peneliti yang bisa ditemukan, atau sebaliknya *terfokus* mengandalkan karya-karya tulis seorang tokoh yang dikenal sebagai ahli bahasa sekaligus ahli dalam bidang pendidikan bahasa. Tujuannya adalah mengidentifikasikan indikator-indikator perilaku yang mencerminkan taraf pemilikan bakat berbahasa baik secara *favorable* maupun secara *unfavorable.* Hasilnya adalah *behavioral content domain* dari konstruk

kemampuan berbahasa yang bisa juga diterapkan untuk kelompok subjek siswa SMA.

c. *Critical Incidence* atau Identifikasi Contoh Perilaku Ekstrem

Inti metode ini adalah meminta sejumlah orang yang dipandang ahli tentang konstruk yang akan diukur untuk menyebutkan aneka contoh perilaku yang bisa ditempatkan pada kedua ujung ekstrem dari kontinum perilaku yang mencerminkan taraf pemilikan atribut psikologis yang menjadi sasaran pengukuran. Hasilnya adalah daftar bentuk-bentuk perilaku ekstrem yang mencerminkan keberadaan atribut psikologis, misal bakat berbahasa, baik secara *favorable* (menunjukkan keberadaan kemampuan berbahasa dalam jumlah tinggi) maupun secara *unfavorable* (menunjukkan keberadaan kemampuan berbahasa dalam jumlah rendah). Konon metode ini pertama kali dirancang dan diterapkan oleh Flanagan (1954, dalam Crocker & Algina, 2008) terhadap sekelompok *supervisors* perusahaan yang diminta mendeskripsikan situasi-situasi saat seorang pegawai menjalankan tugas secara sangat efektif maupun sebaliknya saat seorang pegawai menjalankan tugas secara sangat tidak efektif. Hasilnya adalah daftar *critical behaviors* yang kemudian bisa dimanfaatkan untuk melakukan penilaian *job performance* pegawai.

d. *Direct Observations* atau Observasi Langsung

Melanjutkan contoh penyusunan tes bakat berbahasa di atas, untuk mengidentifikasikan indikator-indikator perilaku baik yang *favorable* maupun yang *unfavorable* terhadap kemampuan berbahasa si penyusun tes melakukan observasi langsung terhadap hasil karya para ahli bahasa meliputi penyair, cerpenis, novelis, esais dan sebagainya baik yang disajikan secara lisan maupun dalam bentuk karya tulis, serta hasil karya tulis murid-murid di berbagai jenjang pendidikan berupa teks pidato, puisi, cerpen, esai atau bentuk karangan lain seperti makalah, skripsi, dan sebagainya baik yang disajikan secara lisan maupun khususnya dalam bentuk tertulis.

Hasilnya adalah sejenis daftar *critical behaviors* yang mencerminkan kemampuan berbahasa baik secara *favorable* berupa pilihan kata yang tepat dan kaya, penulisan ejaan yang tepat, penggunaan tanda baca yang cermat, penggunaan kalimat dan paragraf yang efektif, dan sebagainya maupun secara *unfavorable* berupa berbagai bentuk kesalahan berbahasa, dan yang dapat dipandang sebagai *behavioral content domain* dari konstruk kemampuan berbahasa.

e. *Expert Judgments* atau Pendapat Pakar

Dalam metode ini, penyusun tes meminta pendapat satu atau lebih pakar di bidang bahasa untuk mengidentifikasikan bentuk-bentuk tingkah laku yang dipandang sebagai indikator baik yang *favorable* maupun yang *unfavorable* terhadap kemampuan berbahasa. Pengumpulan pendapat pakar tersebut bisa dilakukan dengan bantuan kuesioner tertulis atau wawancara tatap muka (Crocker & Algina, 2008), atau melalui *focus group discussion*.

2. Menuliskan Item

Sesudah *behavioral content domain* atau ranah isi perilaku atribut psikologis yang menjadi sasaran pengukuran berhasil diidentifikasikan atau dirumuskan, langkah berikutnya adalah menyusun tabel spesifikasi untuk memperoleh evidensi tentang kecukupan isi dari tes yang akan disusun sekaligus sebagai pedoman penulisan item. Penyusunan tabel spesifikasinya sendiri tidak perlu kita bahas di sini sebab sudah cukup diuraikan dalam langkah umum konstruksi tes di Bab 9. Kita akan membahas langkah berikutnya, yaitu penulisan item.

Kita akan fokus pada jenis tes *developed abilities* dengan latar belakang pengalaman luas sebagai faktor pembentuknya khususnya berupa tes inteligensi dan tes bakat jenis verbal, meliputi aneka tes kecerdasan umum dan tes bakat khusus yang mengandalkan media bahasa. Berbagai jenis tes inteligensi umum lazim didasarkan pada teori Spearman (1927, dalam Kline 1986) tentang inteligensi yang

terkenal dengan nama teori faktor *g*. Analisis faktor terhadap faktor umum ini menghasilkan dua faktor, yaitu g_f kependekan dari *fluid ability* atau abilitas cair dan g_c kependekan dari *crystallized ability* atau abilitas padat. Abilitas cair mirip faktor *g* Spearman berupa kemampuan mempersepsikan hubungan pada materi yang bersifat bebas dari pengaruh budaya lokal. Sebaliknya, abilitas padat merupakan kemampuan yang sama seperti abilitas cair namun pada materi yang terikat pada faktor budaya lokal. Dengan kata lain, dalam jenis-jenis tes *developed abilities* khususnya berupa inteligensi yang didasarkan pada teori Spearman *content* yang digunakan sebagai materi tes adalah hubungan antar benda, peristiwa, atau hal yang dipakai sebagai item-item tes (Kline, 1986). Untuk mengungkap kemampuan mempersepsikan hubungan tersebut, ada tiga jenis format item yang dipandang efektif maka juga lazim diterapkan dalam penyusunan tes inteligensi, yaitu: (a) analogi, (b) *odd-man-out* atau pilih yang beda, dan (c) *sequences* atau deret (Kline, 1986).

a. Item Analogi

Analogi atau persamaan adalah format item berupa perbandingan antara dua hal yang didasarkan pada hubungan persamaan tertentu. Agar mampu menjawabnya testi harus mampu mempersepsikan hubungan persamaan antar dua hal yang diperbandingkan. Beberapa contoh item hasil adaptasi contoh-contoh item yang dikemukakan oleh Kline (1986) adalah sebagai berikut:

Contoh 1

Merpati berbanding burung seperti gurami berbanding …

a. binatang
b. semut
c. ikan
d. burung
e. reptil

Jawaban yang benar adalah "*c. ikan*". Hubungan persamaannya adalah keanggotaan dalam suatu golongan binatang. Merpati adalah anggota golongan burung, maka gurami juga harus ditempatkan dalam golongan yang memiliki hubungan analog dengan hubungan antara merpati dan burung, yaitu ikan.

Contoh 2

25 berbanding 10 seperti 53 berbanding …

a. 2
b. 8
c. 31
d. 15
e. 24

Jawaban yang benar adalah "*d. 15*". Hubungan persamaannya adalah bahwa 10 merupakan hasil perkalian antara 2 dan 5 yang membentuk bilangan "25", maka jawaban yang benar adalah hasil perkalian antara 5 dan 3 yang membentuk bilangan "53", yaitu 15.

b. Item *Odd-Man-Out* atau Pilih Satu yang Beda

Item ini berupa daftar serangkaian objek, kata, bentuk, bilangan atau apa pun yang berhasil dirumuskan oleh penyusun tes. Tugas testi adalah "mengeluarkan" dalam arti menemukan salah satu dari antara rangkaian objek, kata, bentuk, bilangan, atau apa pun itu yang berbeda atau tidak sesuai dengan lainnya dalam rangkaian. Agar mampu menemukan yang berbeda atau yang tidak sesuai tersebut, testi harus mampu mempersepsikan hubungan atau benang merah antar objek atau apa pun itu yang menjadi dasar kesamaan dan perbedaannya. Beberapa contoh item hasil adaptasi contoh-contoh item yang dikemukakan oleh Kline (1986) adalah sebagai berikut:

Contoh 1

Sapi, kerbau, kuda, anjing, ular

Jawaban yang benar, yaitu nama binatang yang harus dikeluarkan karena tidak sesuai dengan empat lainnya, adalah "ular". Empat binatang lainnya memiliki kesamaan sebagai jenis binatang menyusui dan berkaki empat pula, sedangkan ular bukan termasuk keduanya.

Contoh 2
24, 63, 10, 48, 35

Jawaban yang benar, yaitu bilangan yang harus dikeluarkan karena tidak sesuai dengan empat lainnya, adalah "10". Empat bilangan lainnya memiliki kesamaan yaitu berupa kuadrat bilangan tertentu dikurangi 1, seperti $24 = 5^2 - 1$, $63 = 8^2 - 1$, $48 = 7^2 - 1$, dan $35 = 6^2 - 1$, sedangkan 10 bukan seperti itu.

c. *Sequences* atau Deret

Item ini berupa sekuensi atau deret atau rangkaian yang belum selesai yang terdiri dari objek, kata, bilangan, atau apa pun yang berhasil dirumuskan oleh penyusun tes. Tugas testi adalah menemukan objek, kata, bilangan atau apa pun yang terakhir untuk melengkapi deret atau rangkaian tersebut. Agar mampu menemukan jawaban yang diminta, testi harus mampu menangkap hubungan atau benang merah yang mendasari rangkaian objek, kata, bilangan atau apa pun itu. Beberapa contoh item hasil adaptasi contoh-contoh item yang dikemukakan oleh Kline (1986) adalah sebagai berikut:

Contoh 1
12, 15, 17, 20, 22 …

Jawaban yang benar adalah "25". Deret bilangan tersebut meningkat dengan selisih 3 dan 2 secara bergantian: dari 12 ke 15 meningkat 3, dari 15 ke 17 meningkat 2, dari 27 ke 20 meningkat 3, dari 20 ke 22 meningkat 2, maka dari 22 harus meningkat 3 menjadi 25.

Contoh 2
Sangat tidak setuju, Tidak setuju, Ragu-ragu, Setuju…

Jawaban yang benar adalah "Sangat setuju". Deret atau rangkaian tersebut merupakan kontinum *agreement* atau kesetujuan mulai dari "Sangat tidak setuju" sampai dengan "Sangat Setuju" sebagaimana lazim dipakai dalam skala Likert.

Ada sejumlah kemungkinan variasi baik untuk format item analogi, *odd-man-out,* maupun sekuensi atau deret, namun kiranya akan terlalu makan waktu dan energi jika kita bahas di sini. Sebaiknya hal itu diuraikan dalam buku lain yang secara khusus membahas seluk-beluk penyusunan tes inteligensi umum. Selain itu, langkah ketiga berupa analisis item kita tunda dulu pembahasannya sampai kita membahas dua langkah pertama yang sama dalam penyusunan tes jenis *developed abilities* yang lain, yaitu *achievement tests* atau tes prestasi.

B. Penyusunan *Achievement Tests*

Achievement tests atau *attainment tests* (Kline, 1986) atau tes prestasi atau tes hasil belajar termasuk ke dalam kategori *developed abilities* dengan latar belakang pengalaman yang spesifik atau sempit sebagai faktor pembentuknya. Seperti sudah disinggung, kategori tes yang oleh Kline (1986) disebut "more simple" atau lebih sederhana khususnya jika dibandingkan dengan tes inteligensi ini, mencakup : (1) tes prestasi yang berorientasi pada mata pelajaran tertentu, lazimnya berupa jenis-jenis tes prestasi mata pelajaran untuk jenjang sekolah tertentu yang dibuat oleh guru sendiri, misal tes Matematika semester 1 kelas V SD; (2) tes prestasi yang berorientasi luas, lazimnya berupa tes prestasi baku tentang bidang studi tertentu yang disusun oleh pakar bidang studi bekerja sama dengan pakar psikometri, misal tes Matematika untuk SMA.

Kline (1986) menyatakan, penentu kualitas tes prestasi terletak pada *content* atau isinya. Terkait *form* atau formatnya, yang penting format item yang dipilih sungguh-sungguh menjamin objektivitas dalam penskorannya. Di sisi lain juga sudah disinggung bahwa dibandingkan dengan bakat atau inteligensi yang merupakan konstruk teoretis, *content domain* atau ranah isi tes prestasi lazim dipandang sudah jelas yaitu tercakup dalam silabus mata pelajaran beserta buku teks atau buku ajar yang dipakai dalam aktivitas pembelajarannya. Yang lebih penting untuk dicermati adalah *behavioral domain*-nya, dalam arti pada bentuk-bentuk tingkah laku atau pada tingkat-tingkat proses berpikir mana penguasaan atas berbagai komponen *content* atau materi pelajaran tersebut akan diukur.

Dalam praksis penyusunan tes prestasi atau hasil belajar berbagai mata pelajaran di sekolah, agar mampu mengidentifikasikan *content domain* dan *behavioral domain* sebagai indikator penguasaan berbagai kompetensi atau kemampuan yang diajarkan dalam berbagai mata pelajaran di sekolah orang lazim menggunakan *taxonomy of instructional objectives* atau taksonomi tujuan pengajaran sebagai dasar dalam menyusun tabel spesifikasinya. Pembahasan lebih lanjut akan berfokus pada penyusunan tes prestasi mata pelajaran dalam konteks pendidikan sekolah dan secara lebih spesifik lagi kita hanya akan berfokus pada tes prestasi mata pelajaran buatan guru.

1. Mendefinisikan Tes

Tes prestasi dalam lingkungan pendidikan sekolah hampir selalu dimaknai sebagai pengukuran hasil kegiatan pembelajaran dalam suatu mata pelajaran tertentu dan yang disusun sendiri oleh guru pengampu mata pelajaran yang bersangkutan. Ini pulalah yang akan menjadi fokus pembahasan kita di bagian ini. Dengan kata lain dan secara khusus dalam konteks pembahasan kita sekarang, tes prestasi di sekolah selalu didefinisikan oleh mata pelajaran tertentu yang diselenggarakan dalam satuan waktu tertentu, pada tingkat kelas tertentu, pada jenjang pendidikan tertentu, seperti misalnya tes

prestasi Matematika semester gasal kelas V Sekolah Dasar. Bahkan masih ada satu pembatas penting lain, yaitu yang disusun sendiri oleh guru (*teacher made achievement test*).

Salah satu kelaziman lain dalam penyusunan tes prestasi mata pelajaran yang disusun sendiri oleh guru adalah pemanfaatan taksonomi tujuan pengajaran tertentu dalam menyusun tabel spesifikasi tes. Sebagaimana sudah kita bahas di Bab 5 ada beberapa versi taksonomi tujuan pengajaran untuk tiga ranah kemampuan, yaitu ranah kognitif atau pengetahuan dan ketrampilan berpikir, ranah afektif atau pembentukan sikap, dan ranah psikomotor atau pembentukan aneka ketrampilan motorik yang bersifat spesifik. Kesamaan yang mempersatukan semua jenis taksonomi tujuan pengajaran adalah bahwa masing-masing taksonomi sekaligus mendefinisikan "both the specific content on which items should focus and the nature of the tasks the examinees should be able to perform" (Crocker & Algina, 2008; h. 68). Artinya, setiap taksonomi tujuan pengajaran atau pembelajaran senantiasa mendefinisikan dengan jelas baik *content* atau materi berupa pengetahuan yang harus dikuasai maupun jenis tugas dalam arti taraf proses berpikir yang dituntut untuk menyatakan penguasaan atas materi yang dimaksud. Dalam bahasa para penyusun taksonomi, taksonomi tujuan pengajaran senantiasa mencakup aspek **kata benda** yang mengacu pada materi dan aspek **kata kerja** yang mengacu pada tingkat proses berpikir yang perlu dilibatkan dalam mengolah materi yang bersangkutan sebagai wujud kompetensi tertentu yang diperoleh sebagai hasil kegiatan pembelajaran.

Sumber penting dalam mendefinisikan tes dalam arti menentukan ranah isi dan perilakunya untuk kategori tes hasil belajar adalah silabus mata pelajaran dan buku teks atau buku ajar yang menyertainya. Silabus dan buku teks atau buku ajar semacam itu akan memberikan informasi minimal tentang hal-hal sebagai berikut: (a) nama mata pelajaran, (b) semester, (c) tingkat kelas pada jenjang pendidikan sekolah tertentu, dan (d) jenis-jenis kompetensi

terkait dengan *content* atau materi tertentu yang dijadikan tujuan pembelajaran.

Rumusan kompetensi lazimnya mencakup *content domain* atau ranah materi mata pelajaran dan *behavioral domain* atau ranah perilaku sebagai bentuk perwujudan kompetensi atas materi tertentu. Ranah materi pelajaran lazim bisa diidentifikasikan dengan mengikuti pembagian komponen-komponen materi dalam silabus dan buku teks atau buku ajar. Ranah perilaku terkait masing-masing materi lazim bisa diidentifikasikan dengan mengikuti salah satu atau gabungan beberapa taksonomi tujuan pengajaran. Kedua kategori ranah tersebut selanjutnya dipakai sebagai dasar dalam menyusun tabel spesifikasi.

Tabel spesifikasi sendiri pada dasarnya merupakan sebuah matriks dua dimensi (Crocker & Algina, 2008). Dimensi pertama lazim memuat ranah materi dan ditempatkan pada sisi horisontal, sedangkan dimensi kedua lazim memuat ranah perilaku dan ditempatkan pada sisi vertikal. Interseksi atau perpotongan antara kedua dimensi tersebut akan membentuk sel-sel untuk mengidentifikasikan atau menentukan jumlah item. Jumlah item dalam masing-masing sel mencerminkan bobot yang diberikan oleh penyusun tes terhadap materi dan bentuk tingkah laku yang diharapkan mencerminkan tingkat penguasaan tertentu atas materi yang bersangkutan. Prinsip umum yang berlaku, pemberian bobot yang makin tinggi terhadap materi tertentu pada tingkat proses berpikir tertentu perlu dinyatakan dalam persentase jumlah item yang makin tinggi pula. Sebaliknya, pemberian bobot yang makin rendah atau kurang terhadap materi tertentu pada tingkat proses berpikir tertentu perlu dinyatakan dalam persentase jumlah item yang makin kurang pula, bahkan nol. Sebagaimana tersirat, bobot-bobot tersebut perlu dinyatakan dalam persentase. Jumlah keseluruhan bobot-bobot ini harus sama dengan 100%. Contoh tabel spesifikasi Tes Hasil Belajar Mata kuliah Psikologi Kepribadian 1 bagi mahasiswa Program S1 Psikologi disajikan pada Tabel 10.1.

Sekadar informasi tentang tabel spesifikasi seperti disajikan di Tabel 10.1., mata kuliah Psikologi Kepribadian 1 membahas konsep kepribadian menurut 12 perspektif teoretis mengikuti uraian dalam sumber utama yang dipakai dalam mata kuliah ini, yaitu seri buku teks Psikologi Kepribadian yang merupakan terjemahan Bahasa Indonesia dari buku karangan Hall dan Lindzey (1978). Kedua belas perspektif teoretis yang dimaksud meliputi: (a) psikologi individu Gordon Allport, psikologi konstitusi William Sheldon, teori faktor Raymond Cattell, dan teori tipologi kebudayaan Eduard Sranger mewakili kelompok teori sifat atau teori tipe; (b) psikoanalisis klasik Sigmund Freud, psikologi analitik Carl Jung, psikologi ego Erik Erikson, psikologi sosial Alfred Adler, Erich Fromm, Karen Horney, dan Harry Sullivan, mewakili teori-teori yang berorientasi psikodinamik; serta (c) personologi Henry Murray mewakili teori-teori yang berorientasi holistik-organismik (Supratiknya, 2006).

Tabel 10.1.
Tabel Spesifikasi Tes Hasil Belajar Mata Kuliah *Psikologi Kepribadian 1*

No.	Komponen Materi	Tingkat Proses Berpikir			Jumlah
		Pengetahuan	**Pemahaman**	**Penerapan**	
1	Allport	1	5	1	7
2	Sheldon	2	4	1	7
	Cattell	1	5	1	7
	Spranger	1	1	1	3
3	Freud	5	8	4	17
4	Jung	2	1	5	8
	Erikson	2	4	2	8
5	Adler	2	4	1	7
	Fromm	2	2	4	8
6	Horney	1	5	1	7
	Sullivan	1	6	1	8
7	Murray	6	5	2	13
	Jumlah	26	50	24	100

(Sumber: A. Supratiknya, 2006).

Ada beberapa catatan yang perlu dikemukakan terkait tabel spesifikasi. *Pertama*, tabel spesifikasi mencerminkan struktur tes. Struktur tes meliputi cakupan baik materi maupun tingkat proses

berpikir yang hendak diungkap oleh tes. Selain itu, struktur tes juga menunjukkan pandangan penyusun tes tentang bobot masing-masing sel yang merupakan interseksi antara baris yang berisi komponen materi dan kolom yang berisi komponen proses berpikir, sebagaimana tampak dari besar persentase jumlah item yang direncanakan oleh penyusun tes untuk masing-masing sel yang bersangkutan.

Kedua, tabel spesifikasi akan berfungsi sebagai pedoman dalam penyusunan atau penulisan item-item tes. Dalam proses penyusunan tes sangat jarang bahkan mungkin mustahil diperoleh bentuk final tes sekali jadi. Yang lazim terjadi, dalam proses memperoleh item-item yang efektif melalui pemeriksaan sejumlah parameter item akan terbukti sejumlah item masih perlu direvisi atau bahkan digugurkan atau disingkirkan sama sekali karena sangat tidak memenuhi syarat kualitas. *Mortality rate* atau tingkat persentase gugurnya item-item ini konon bahkan mencapai kisaran 50%. Maka item-item yang direncanakan dalam tabel spesifikasi lazim dimaksudkan sebagai *item pool* atau sejenis bank item dari mana akan diperoleh item-item terbaik dalam jumlah sebagaimana direncanakan sebagai bentuk final tes. Mengingat bahwa tingkat persentase gugurnya item-item berkisar 50%, maka jumlah item sebagai *item pool* yang direncanakan dalam tabel spesifikasi minimal harus sama dengan dua kali dari jumlah item yang direncanakan sebagai bentuk final tes (Kline, 1986). Dalam contoh penyusunan tes hasil belajar mata kuliah Psikologi Kepribadian 1 di atas, karena bentuk final tesnya direncanakan terdiri atas 60 item maka sebagai *item pool* direncanakan disusun 145 item dengan distribusi mengikuti distribusi persentase sebagaimana direncanakan dalam tabel spesifikasi di atas.

Dengan tersusunnya tabel spesifikasi tes berarti ranah isi tes dalam arti luas mencakup baik komponen materi maupun komponen tingkah laku atau proses berpikirnya berhasil didefinisikan. Langkah selanjutnya adalah menyusun atau lebih tepat menuliskan item-item berpedoman pada tabel spesifikasi.

2. Menuliskan Item

Seorang pakar pengukuran pendidikan bernama Popham (1981, dalam Crocker & Algina, 2008; h. 76) menggolongkan format item yang efektif dan yang lazim dipakai dalam pengukuran jenis-jenis *developed abilities* ke dalam dua kategori: (a) format **isian,** yaitu format item yang menuntut testi merumuskan sendiri respon atau jawabannya, dan (b) format **pilihan,** yaitu format item yang menyediakan dua atau lebih kemungkinan jawaban dan menuntut testi memilih salah satu di antaranya sebagai jawaban yang benar atau yang paling benar. Kategori format item yang pertama menuntut proses penskoran yang sedikit banyak bersifat subjektif sehingga lazim disebut item-item **tes subjektif.** Kategori format item yang kedua memungkinkan proses penskoran dengan meminimalkan unsur subjektivitas penilai sehingga lazim disebut item-item **tes objektif**. Kita akan berfokus pada kategori kedua, yaitu item-item tes objektif.

Ada dua catatan sebelum kita membahas lebih lanjut penyusunan item-item tes objektif. *Pertama,* sebagaimana dinyatakan oleh sekelompok pakar pengukuran, kendati mendapat banyak kritik namun item-item tes objektif atau format pilihan ganda tetap memainkan peran penting dalam penilaian hasil belajar baik yang berskala kecil seperti penilaian hasil belajar oleh guru di kelas maupun yang berskala besar seperti ujian nasional untuk berbagai jenjang pendidikan dasar dan menengah di Tanah Air (Haladyna, Downing, & Rodriguez, 2002). *Kedua* dan seperti sudah disinggung, dalam penyusunan tes hasil belajar yang paling krusial adalah menentukan *content* berupa kompetensi yang merupakan kombinasi antara komponen materi dan tingkat proses berpikir tertentu, sedangkan terkait formatnya yang penting menjamin penilai mampu melakukan penskoran secara objektif (Kline, 1986). Namun sebagaimana akan tampak, jenis format item tertentu lebih sesuai untuk mengungkap jenis kompetensi tertentu berhubung terkait dengan taraf kesulitannya. Maka, tugas penyusun tes adalah memilih format item yang paling

tepat sesuai dengan taraf kesulitan *content* dalam arti kompetensi yang hendak diukur.

Crocker dan Algina (2008) menyebut tiga jenis format objektif yang dipandang paling lazim diterapkan untuk menyusun tes *developed abilities* khususnya hasil belajar, yaitu (a) *alternate choice* atau format dua pilihan (benar-salah atau ya-tidak), (b) *multiple choice* atau pilihan ganda, dan (c) *matching* atau menjodohkan. Haladyna, Downing, dan Rodriguez (2002) menggolongkan ketiga format tersebut ke dalam satu kategori besar *multiple choice* atau pilihan ganda dan menunjukkan jenis-jenisnya. Sekali lagi, kita hanya akan berfokus pada jenis format pilihan dan akan mengikuti penggolongan Haladyna, Downing, dan Rodriguez (2002). Salah satu kelebihan penggolongan mereka adalah bahwa penggolongan tersebut didasarkan pada pengamatan empiris terhadap 27 buku-teks dan 27 laporan penelitian maupun tinjauan kritis (*review)* tentang *educational testing* atau seluk-beluk penilaian dengan tes di lingkungan pendidikan sekolah.

Jenis-jenis item pilihan ganda mengikuti penggolongan Haladyna, Downing, dan Rodriguez (2002) adalah sebagai berikut: (a) *conventional multiple choice* atau pilihan ganda konvensional; (b) *alternate choice* atau dua pilihan; (c) *true-false* atau benar-salah; (d) *multiple true-false* atau benar-salah ganda; (e) *matching* atau menjodohkan; (f) *complex multiple choice* atau pilihan ganda kompleks; dan (g) *context-dependent item set* atau rangkaian item tergantung konteks. Penjelasan tentang masing-masing jenis item pilihan ganda yang dimaksud beserta contoh masing-masing adalah seperti diuraikan berikut ini.

a. *Conventional Multiple Choice* atau Pilihan Ganda Konvensional

Manakah dari antara pernyataan berikut ini paling jelas mendefinisikan proses polinasi?

a. Bertemunya sel telur dan sel sperma.
b. Berpindahnya serbuk sari ke putik.
c. Terurainya makanan dan dilepaskannya energi.

Menurut Haladyna, Downing, dan Rodriguez (2002) jenis item pilihan ganda (IPG) ini banyak dipakai dalam penyusunan tes prestasi di semua jenjang pendidikan dan untuk kebanyakan jenis mata pelajaran atau mata kuliah. Jenis item pilihan ganda ini terdiri dari "a question or problem statement, also referred to as the **stem**, and a series of four or five responses, of which only one is the correct answer. Incorrect responses are referred to as **distracters** and should be written as *equally plausible answers* to the question" (Tarrant, Knierim, Hayes, & Ware, 2006). Jadi, jenis item pilihan ganda ini terdiri dari sebuah **stem** berupa *question* atau pertanyaan, atau *completion* atau pernyataan yang tidak lengkap sebagai masalah dan diikuti tiga, empat, atau lima **opsi** sebagai kemungkinan jawaban. Opsi berisi jawaban yang benar (atau paling benar) disebut **kunci**, sedangkan opsi sisanya yang berisi jawaban salah (atau kurang benar) disebut **distraktor**.

Ada dua variasi item pilihan ganda, yaitu: (1) *question stem* atau *question format* atau stem berformat pertanyaan, dan (2) *completion stem* atau *sentence-completion format* atau stem berformat melengkapi kalimat. Pada varian pertama, stemnya berupa pertanyaan. Pada varian kedua, stemnya berupa bagian sebuah kalimat yang akan menjadi kalimat utuh jika dilengkapi dengan opsi-opsi yang tersedia. Menurut Haladyna, Downing, dan Rodriguez (2002), "we prefer the question format over the sentence-completion format for its directness in getting to the central idea of the test item" (h. 323). Artinya, mereka lebih menyukai format pertanyaan dibandingkan format melengkapi kalimat, sebab format pertanyaan lebih langsung merujuk pada gagasan pokok yang dipersoalkan.

b. *Alternate Choice* atau Dua Pilihan

Manakah dari benda di bawah ini yang akan paling efektif menghambat proses pernafasan pada tanaman?

a. Air dingin.

b. Air beriak.

Menurut Haladyna, Downing, dan Rodriguez (2002), jenis item pilihan alternatif (IPA) pada hakikatnya merupakan *a two-option multiple choice item* atau item pilihan ganda beropsi dua. Selanjutnya menurut mereka, kendati lebih sederhana dibandingkan item pilihan ganda (IPG) ternyata item pilihan alternatif (IPA) mampu berfungsi sama efektifnya seperti item pilihan ganda (IPG) dalam penilaian kelas.

c. *True-False* atau Benar-Salah

Ibu kota Uruguai adalah Montevideo.

Jenis item benar salah (IBS) ini juga sangat populer. Nama lain untuk IBS adalah item pilihan alternatif (IPA), item dua pilihan (IDP), atau item pilihan biner (IPB). Variasi jenis item ini berupa sebuah pernyataan yang harus dijawab dengan cara memilih salah satu antara "ya" atau "tidak", "benar" atau "salah", "fakta" atau "pendapat", atau pasangan istilah bipolar lain yang bisa dinyatakan benar atau salah. Menurut Haladyna, Downing, dan Rodriguez (2002), kendati ada sejumlah masalah yang belum berhasil diatasi dalam penerapannya namun item benar salah (IBS) masih sering dipakai dalam penilaian kelas.

d. *Multiple True-False* atau Benar-Salah Ganda

Anda seorang petani organik ahli. Anda menguasai rahasia mengembangkan tanaman yang kuat dan sehat. Mana dari antara pernyataan berikut ini menjelaskan cara Anda bercocok tanam? (Tandailah dengan huruf A jika pernyataan yang bersangkutan benar, B jika pernyataan yang bersangkutan salah).

a. Saat menanam pohon buncis Anda memastikan agar tanaman tersebut tumbuh di tempat yang teduh agar terkena sedikit atau bahkan tidak terkena sinar matahari sama sekali.

b. Saat menanam benih Anda menyiraminya secara cukup dan menjaga agar tanah tetap lembab.

c. *Anda hanya menanam benih pada suhu udara yang tepat.*
d. *Karena memahami proses penyerbukan, Anda menyemprot tanaman dengan insektisida untuk mencegah lebah dan serangga lain merusak tanaman.*

Jenis item benar-salah ganda (IBSG) yang kurang lazim ini merupakan *hibrida* atau hasil persilangan antara item pilihan ganda (IPG) dan item benar salah (IBS). Stem berupa pernyataan atau skenario dan diikuti sejumlah opsi. Testi diminta menilai masing-masing opsi, benar atau salah. Menurut Haladyna, Downing, dan Rodriguez (2002), menulis item benar salah ganda (IBSG) lebih mudah dibandingkan menulis item pilihan ganda (IPG) konvensional, namun item benar salah ganda (IBSG) lebih rentan terhadap godaan menebak dan cenderung mengungkap kemampuan kognitif yang lebih sederhana dibandingkan item pilihan ganda (IPG) biasa. Maka, kurang disarankan.

e. *Matching* atau Menjodohkan

Jodohkanlah masing-masing istilah di sisi sebelah kanan dengan deskripsinya di sisi sebelah kiri.

1. Menarik lebah	*a. Serbuk sari*
2. Menghasilkan serbuk sari	*b. Putik*
3. Melindungi sel telur	*c. Bunga*
4. Terbentuknya benih	*d. Benang sari*
5. Mengandung indung telur	*e. Indung telur*
	f. Putik

Item menjodohkan (IM) terdiri dari serangkaian opsi, lazim diletakkan di sisi sebelah kanan, disertai serangkaian stem jodoh atau padanannya berupa pernyataan, pertanyaan, atau frase tertentu yang lazim ditempatkan di sisi sebelah kiri. Jumlah opsi lazimnya dibuat satu lebih banyak dibandingkan jumlah stemnya. Dengan kata lain, rangkaian opsi lazimnya mengandung satu distraktor. Tugas testi

adalah menjodohkan masing-masing stem dengan opsi padanannya. Menurut Haladyna, Downing, dan Rodriguez (2002), kendati bukan mustahil menyusun item menjodohkan (IM) yang baik, namun mereka hanya merekomendasikan penggunaan jenis item ini untuk penilaian kelas, bukan untuk *high-stakes testing* atau penilaian untuk keperluan pengambilan keputusan yang lebih penting seperti seleksi pegawai dan sejenisnya.

f. *Complex Multiple Choice* atau Pilihan Ganda Kompleks

Manakah dari antara benda-benda berikut ini merupakan jenis buah?

1. Tomat
2. Mentimun
3. Mete

a. 1 & 2
b. 2 & 3
c. 1 & 3
d. 1, 2, & 3

Secara umum, jenis item pilihan ganda kompleks (IPGK) yang juga disebut item *Tipe K* dipandang tidak efisien karena beberapa alasan. Pertama, IPGK lebih sulit dibandingkan IPG konvensional namun daya diskriminasinya ternyata tidak lebih baik. Kedua, IPGK lebih panjang dibandingkan semua jenis item pilihan-ganda lainnya, sehingga waktu administrasinya juga menjadi lebih lama. Memperhatikan semua kekurangan tersebut, Haladyna, Downing, dan Rodriguez (2002) merekomendasikan agar jenis item ini tidak digunakan.

g. *Context-Dependent Item Set* atau Rangkaian Item Tergantung Konteks

Bayangkanlah Anda anggota delegasi dari Negara bagian Massachusetts ke Constitutional Convention. Anda diberi wewenang untuk bertindak atas nama Negara bagian Anda.

1. Anda pasti akan mendukung
   a. New Jersey Plan
   b. Virginia Plan
2. Anda akan menentang kompromi tiga per-lima sebab
   a. Negara bagian Anda sangat pro-abolisi.
   b. Anda akan dikucilkan di Kongres
      oleh Negara-negara bagian utara.
   c. Anda hanya menghendaki dewan perwakilan tunggal.
3. Anda mendukung usulan agar Kongres mengenakan pajak
   a. Impor.
   b. Ekspor.
4. Akibat pengalaman Negara bagian Anda dengan Pemberontakan Shay, Anda merasa bahwa
   a. Petani tidak semestinya menanggung beban pajak bagi orang kota.
   b. Syarat bagi tercapainya perdamaian adalah terjaminnya rasa aman warga asli Amerika.
   c. Kelompok Tori harus membayar ganti rugi.

Bentuk baku item tergantung-konteks (ITK) terdiri atas sebuah skenario, vinyet, tabel, peta, grafik, penggalan bacaan, atau materi visual lain diikuti sebuah pertanyaan atau tugas. Variasinya berupa rangkaian item (R-ITK) seperti contoh di atas. Rangkaian item semacam ini sering juga disebut *testlets, interpretive exercises* atau soal interpretatif, atau *superitem* (Haladyna, Downing, & Rodriguez, 2002). Kelebihan jenis item ini adalah efektivitasnya untuk mengukur jenis-jenis kemampuan berpikir taraf tinggi. Selain digunakan dalam penilaian kelas, jenis item ini juga makin banyak digunakan dalam aneka jenis penilaian berskala besar, maka juga sangat direkomendasikan (Haladyna, Downing, & Rodriguez, 2002).

Taksonomi Pedoman Penulisan Item Pilihan-Ganda

Sebenarnya ada berbagai pedoman penulisan item pilihan-ganda seperti yang dikemukakan oleh Azwar (1999) atau penulis lain. Kita akan menggunakan pedoman penulisan item pilihan-ganda yang dikemukakan oleh Haladyna, Downing, dan Rodriguez (2002) karena tiga alasan: (1) pedoman ini khusus dibatasi untuk penulisan jenis item pilihan-ganda dengan aneka variasinya seperti diuraikan di atas, bukan jenis item objektif pada umumnya; (2) pedoman atau tepatnya mengikuti istilah yang dipakai oleh para pengarangnya sendiri, taksonomi pedoman penulisan item pilihan-ganda ini sudah diuji efektivitasnya secara empiris berdasarkan pengamatan terhadap 27 laporan penelitian maupun tinjauan kritis (*review)* tentang *educational testing* atau seluk-beluk penilaian dengan tes di lingkungan pendidikan sekolah; dan (3) sebagai taksonomi, pedoman ini dipilah ke dalam sejumlah komponen dalam penulisan item meliputi aneka aspek terkait isi, format, gaya bahasa, stem, dan opsi. Taksonomi pedoman penulisan item pilihan-ganda yang dimaksud, yang sesungguhnya merupakan revisi terhadap taksonomi yang mula-mula disusun oleh Haladyna dan Downing (1989), dan dengan penyesuaian seperlunya, adalah seperti disajikan dalam Tabel 10.2.

Tabel 10.2.
Taksonomi Pedoman Penulisan-Item Pilihan Ganda

No.	Uraian
	Terkait Isi
1	Setiap item harus mengandung materi dan jenis proses berpikir spesifik, sebagaimana ditentukan dalam spesifikasi tes (cetak biru tes).
2	Setiap item harus mengukur materi yang penting untuk dipelajari; hindari mengukur materi yang kurang penting.
3	Untuk mengukur taraf belajar yang tinggi pakailah rumusan yang baru. Rumusan yang dipakai dalam buku teks atau yang dipakai dalam pembelajaran perlu diparafrasekan atau dirumuskan kembali dengan menggunakan kata-kata baru jika digunakan sebagai isi item untuk menghindari mengukur sekadar kemampuan mengingat.
4	Materi setiap item harus bersifat independen atau tidak bersangkut paut dengan materi item-item lain dalam tes.

No.	Uraian
5	Jangan menggunakan materi yang terlampau spesifik atau sebaliknya terlampau umum dalam item-item pilihan ganda.
6	Jangan menggunakan item-item yang berisi pendapat.
7	Jangan menggunakan item yang berisi jebakan.
8	Gunakan kosa kata yang sederhana sesuai kemampuan kelompok testi yang diuji.
	Terkait Format
9	Gunakanlah versi pertanyaan, pernyataan yang tidak lengkap, dan jawaban paling benar untuk menyusun item-item berformat pilihan ganda konvensional, dua pilihan, benar-salah, benar-salah ganda, menjodohkan, item tergantung konteks, dan item rangkaian, dan jangan menggunakan format pilihan ganda kompleks (Tipe K).
10	Susunlah opsi-opsi item secara vertikal, jangan secara horizontal.
	Terkait Gaya Bahasa
11	Editlah dan periksalah kembali setiap item.
12	Gunakan tata bahasa, tanda baca, penggunaan huruf kapital, dan ejaan secara tepat.
13	Minimalkan tuntutan membaca pada setiap item.
	Terkait Stem
14	Pastikan bahwa perintah dalam stem sungguh-sungguh jelas.
15	Tempatkanlah gagasan pokok dalam stem, bukan pada opsi-opsinya.
16	Jangan menggunakan kata-kata yang hanya dimaksudkan untuk hiasan.
17	Rumuskanlah stem secara positif, jangan menggunakan kata yang bermakna negatif seperti TIDAK, BUKAN, atau KECUALI. Jika sangat terpaksa, gunakanlah kata-kata negatif secara sangat hati-hati dan tuliskan kata negatif tersebut dengan HURUF KAPITAL dan dengan CETAK TEBAL.
	Terkait Opsi
18	Tuliskanlah opsi yang efektif sebanyak mungkin dalam setiap item, namun penelitian menunjukkan bahwa TIGA opsi yang efektif sudah memadai.
19	Pastikanlah bahwa hanya salah satu opsi merupakan jawaban yang benar.
20	Variasikanlah penempatan jawaban yang benar sesuai jumlah pilihan.
21	Susunlah opsi-opsi dalam urutan logis atau numeris.
22	Masing-masing opsi dalam setiap item harus bersifat independen satu sama lain; opsi-opsi juga tidak boleh saling tumpang tindih.
23	Opsi-opsi dalam setiap item harus bersifat homogen baik dari segi isi maupun susunan kalimatnya.
24	Panjang opsi-opsi dalam setiap item harus kurang lebih sama.
25	Opsi "tidak satu pun benar" hanya boleh digunakan jika sangat terpaksa.
26	Jangan pernah menggunakan opsi "semua di atas benar".
27	Rumuskanlah opsi-opsi dalam setiap item secara POSITIF; jangan menggunakan rumusan negatif seperti TIDAK atau BUKAN.

No.	Uraian
28	Jangan memberikan petunjuk apa pun ke arah jawaban yang benar, seperti a. Penggunaan kata determinan seperti SELALU, TIDAK PERNAH, SEPENUHNYA, dan SECARA MUTLAK. b. Asosiasi bunyi, pilihan kata yang sama atau mirip dalam opsi-opsi dengan kata-kata dalam stem. c. Ketidak-konsistenan tata bahasa yang bisa memberi petunjuk ke arah opsi yang benar. d. Opsi yang benar secara mencolok. e. Rangkaian dua atau tiga opsi yang bisa memberi petunjuk ke arah jawaban yang benar. f. Opsi-opsi yang absurd atau aneh.
29	Jawaban terhadap suatu item tidak boleh memberi petunjuk kea rah jawaban yang benar pada item yang lain.
30	Buatlah semua distraktor terkesan layak dipilih.
31	Gunakanlah kesalahan-kesalahan yang sering dilakukan oleh testi sebagai distraktor.
32	Gunakanlah humor jika hal itu sesuai dengan karakter guru dan lingkungan belajarnya.

Terkait cara penskoran, secara umum ada dua cara penskoran item (Kline, 1986), yaitu (a) cara penskoran dikotomis, dan (b) cara penskoran nondikotomis. Sebuah item merupakan variabel dikotomis jika kemungkinan nilai untuk skor item tersebut adalah 0 atau 1. Cara penskoran ini bisa diterapkan baik dalam *maximal performance tests* maupun dalam *typical performance tests,* baik untuk item-item pilihan ganda dengan dua opsi maupun dengan lebih dari dua opsi. Sebaliknya, sebuah item merupakan variabel nondikotomis jika kemungkinan nilai untuk skor item tersebut tidak terbatas antara 0 dan 1. Cara penskoran ini lazim diterapkan dalam *typical performance tests* khususnya pada jenis-jenis tes yang menerapkan skala Likert atau sejenisnya, sedangkan dalam *maximal performance tests* cara penskoran nondikotomis ini hanya lazim ditemukan pada jenis item esai atau jawaban pendek. Penskoran item-item pilihan ganda yang kita bahas di atas jelas menggunakan cara penskoran dikotomis, yaitu pemberian skor "1" jika jawaban benar dan "0" jika jawaban salah. Skor testi dalam tes atau skor total merupakan jumlah skor pada item-item yang dijawab dengan benar.

3. Analisis Item

Tes merupakan kumpulan item, maka kualitas tes sebagai kesatuan dengan sendirinya juga ditentukan oleh kualitas masing-masing item yang membentuknya. Dalam analisis item sejumlah parameter item diperiksa untuk selanjutnya dipakai sebagai dasar untuk melakukan seleksi item, yaitu memilih item-item dengan parameter yang memenuhi syarat untuk dimasukkan ke dalam bentuk final tes. Lantas apa saja jenis parameter yang perlu diperiksa dalam analisis item? Seperti sudah disinggung di Bab 9, jenis parameter item yang perlu diperiksa untuk dijadikan dasar atau kriteria dalam seleksi item adalah sejumlah parameter terkait distribusi jawaban testi terhadap item, terkait hubungan antara jawaban testi terhadap item dan skor total tes sebagai kriteria internal, serta terkait fungsi variabilitas dan hubungan antara skor item dan skor total tes sebagai kriteria internal (Crocker & Algina, 2008).

Sebelum melanjutkan dengan pembahasan tentang aneka parameter item, perlu dikemukakan catatan sebagai berikut tentang analisis item. Analisis item hanya merupakan salah satu metode untuk mendapatkan item-item yang memenuhi syarat dalam konstruksi atau penyusunan tes. Melalui analisis item akan diperoleh item-item yang akan membentuk sebuah tes yang homogen dan memiliki daya beda yang baik (Kline, 1986). Untuk itu, analisis item mengandalkan pemeriksaan korelasi antara masing-masing item dengan skor total dan proporsi subjek dalam sampel yang memilih kunci jawaban pada setiap item. Salah satu kelemahan analisis item yang mengandalkan korelasi item-total semacam ini ialah bahwa kendati mampu menghasilkan tes yang homogen namun tidak menjamin kemurnian faktor yang diukur, sebab dua item yang sesungguhnya mengukur dua faktor yang berbeda bisa sama-sama terpilih semata-mata karena memiliki korelasi yang tinggi dengan skor total. Untuk mengatasi kekurangan itu, seleksi item yang didasarkan pada analisis faktor merupakan jawabnya. Namun penerapan analisis faktor bukan tanpa kendala. Di samping adanya sejumlah masalah tehnis serius yang hingga kini

belum berhasil diatasi secara tuntas, kendala lain penerapan analisis faktor dalam seleksi item ialah diperlukannya sampel subjek dalam jumlah yang besar. Maka kendati tidak sempurna, analisis item tetap dipandang sebagai langkah awal yang ampuh dalam konstruksi atau penyusunan tes (Kline, 1986). Itulah sebabnya metode ini juga tetap kita pilih.

a. Distribusi Jawaban terhadap Item: Taraf Kesukaran Item

Salah satu parameter penting yang melukiskan distribusi jawaban testi terhadap sebuah item adalah *Mean* skor item. Pada item-item yang diskor secara dikotomis, *Mean* skor sebuah item adalah sama dengan proporsi testi yang menjawab item tersebut dengan benar atau sesuai kunci jawaban, dan diberi lambang p_i atau proporsi (penjawab benar) terhadap item *i*. Nilai p_i bergerak antara 0.00 sampai dengan 1.00. Dalam kategori *maximal performance tests* yang mengukur abilitas atau hasil belajar, p_i atau *Mean* skor item semacam ini disebut **taraf kesukaran** item yang bersangkutan. Istilah taraf kesukaran sendiri sering dipandang menyesatkan, sebab makin tinggi p_i berarti makin banyak testi yang mampu menjawab item dengan benar berarti item tersebut mudah. Sebaliknya, makin rendah p_i berarti makin sedikit testi yang mampu menjawab item dengan benar berarti item tersebut sukar.

Item-item dengan p_i yang ekstrem tinggi, misal $p_i = 1.0$ berarti semua testi mampu menjawab item yang bersangkutan dengan benar, atau sebaliknya dengan p_i yang ekstrem rendah, misal $p_i = 0$ berarti tidak satu pun testi mampu menjawab item dengan benar, tidak efektif sebab item semacam itu tidak mampu membedakan taraf pemilikan atribut yang sedang diukur di kalangan testi. Sebuah item memberikan informasi maksimum tentang perbedaan di kalangan testi terkait atribut yang sedang diukur jika memiliki $p_i = 0.50$, sebab varians skor-skor item yaitu $\sigma_i^2 = p_i(1 - p_i)$ akan maksimal. Namun jika semua item dalam suatu tes memiliki $p_i = 0.50$ sedangkan saling korelasinya pun sempurna yang berarti bahwa separo testi memperoleh skor total =

0 sedangkan separo testi lainnya memperoleh skor total sempurna, berarti tes semacam ini juga gagal memberikan informasi maksimum tentang perbedaan di kalangan testi terkait atribut yang sedang diukur (Allen & Yen, 1979).

Lantas sebaiknya berapa besar p_i masing-masing item untuk mendapatkan sebuah tes yang baik dalam arti mampu menunjukkan dengan baik perbedaan di antara testi terkait atribut yang sedang diukur? Jawabnya tergantung pada jenis item atau tesnya. Untuk tes bakat maupun prestasi yang skornya dirancang untuk ditafsirkan dengan acuan norma (*norm-referenced score interpretation*) taraf kesukaran item-itemnya yang ideal terletak dalam *range* atau kisaran 0.60 – 0.80 (Crocker & Algina, 2008). Jika rerata korelasi antara skor item dan skor total pada item-item dikotomis berkisar 0.30 – 0.40, maka taraf kesukaran item-itemnya yang ideal terletak dalam kisaran 0.40 – 0.60 (Henryssen, 1971, dalam Allen & Yen, 1979, h. 121). Jika tes dimaksudkan untuk menyeleksi kelompok atas yang mampu melampaui *cutting score* atau nilai lulus tertentu, maka item-itemnya sebaiknya memiliki $p_i = 0.50$ dalam arti mampu dijawab dengan benar oleh sekitar separo testi yang memperoleh skor total setara nilai lulus (Allen & Yen, 1979). Secara umum dapat dikatakan, untuk bisa mendapatkan diskriminasi maksimum pada berbagai taraf pemilikan atribut yang sedang diukur di antara testi sebaiknya item-item tes memiliki taraf kesukaran dalam kisaran 0.30 – 0.70 (Allen & Yen, 1979).

b. Hubungan Antara Jawaban Terhadap Item dan Skor Total Tes sebagai Kriteria Internal: Daya Diskriminasi Item

Mengutip pendapat Crocker dan Algina (2008), tujuan kebanyakan tes adalah menunjukkan perbedaan di antara testi baik dalam hal atribut yang diukur oleh tes maupun dalam hal kriteria eksternal berupa atribut psikologis lain tertentu yang diprediksikan oleh tes tersebut. Untuk kedua keperluan tersebut parameter yang digunakan untuk menyeleksi item adalah sebuah indeks yang

menunjukkan **daya diskriminasi** item, yaitu kefektivan sebuah item dalam membedakan testi yang secara relatif menempati posisi tinggi dan testi yang secara relatif menempati posisi rendah dalam hal kriteria atau atribut psikologis yang sedang menjadi objek pengukuran. Dalam banyak kasus satu-satunya ukuran yang tersedia untuk menunjukkan tinggi-rendahnya posisi seorang testi dalam atribut psikologis yang sedang diukur semacam itu adalah skor total tes yang dicapai atau diperolehnya. Artinya, jumlah skor pada seluruh item yang dicapai oleh seorang testi dipakai sebagai definisi operasional tentang posisi relatifnya dalam hal atribut psikologis yang sedang diukur. Dengan menggunakan skor total semacam itu sebagai **kriteria internal**, tugas penyusun tes dalam analisis item adalah mengidentifikasikan item-item yang memiliki probabilitas tinggi untuk dijawab secara benar oleh testi yang mencapai skor total tinggi serta item-item yang memiliki probabilitas rendah untuk dijawab secara benar oleh testi yang mencapai skor total rendah. Ada enam parameter yang lazim dipakai sebagai indikator daya diskriminasi item, satu parameter didasarkan pada perbedaan antara proporsi penjawab benar dari kelompok peraih skor total tinggi dan kelompok peraih skor total rendah (Allen & Yen, 1979; Crocker & Algina, 2008) sedangkan lima parameter lainnya berupa korelasi antara skor item dan skor total tes (Allen & Yen, 1979; Kline, 1986; dan Crocker & Algina, 2008). Marilah kita bahas kelima parameter tersebut satu demi satu.

1) *Item-discrimination Index* atau Indeks Diskriminasi Item

Perhitungan parameter ini menuntut langkah-langkah sebagai berikut. Pertama, dengan menggunakan skor total sebagai **kriteria internal** kita urutkan distribusi skor total testi mulai dari yang tertinggi sampai dengan yang terendah. Kedua, kita tetapkan salah satu atau dua nilai skor total sebagai *cutting score* atau *cutting scores.* Jika hanya digunakan satu *cutting score* berarti distribusi skor total tersebut dibelah menjadi dua, masing-masing belahan mencakup 50% testi. Separo jumlah testi yang mencapai skor total di atas *cutting*

score disebut **kelompok atas** sedangkan separo jumlah testi sisanya yang mencapai skor total di bawah *cutting score* disebut **kelompok bawah.** Strategi ini cocok ditempuh jika jumlah keseluruhan testi tidak terlampau besar, misal sekitar 40-an. Jika jumlah keseluruhan testi besar, misal 100-an atau lebih, lebih tepat diterapkan strategi kedua yaitu menggunakan dua *cutting scores* masing-masing untuk menentukan kelompok atas dan kelompok bawah. Secara umum kelompok testi yang mencapai skor total dalam kisaran 10% - 33% teratas dan terbawah dalam distribusi skor total yang sudah diurutkan dari yang tertinggi sampai dengan yang terendah lazim digunakan sebagai dasar penetapan kelompok atas dan kelompok bawah (Allen & Yen, 1979). Namun jika skor total testi tersebut terdistribusikan secara normal, indeks diskriminasi item yang optimal akan diperoleh jika penetapan kedua kelompok ekstrem atas dan bawah tersebut didasarkan pada 27% dari jumlah testi yang mencapai skor total tertinggi dan 27% dari jumlah testi yang mencapai skor total terendah sebagaimana disarankan oleh Kelley(1939, dalam Allen & Yen, 1979; Crocker & Algina, 2008). Kendati demikian, jika sampel testi cukup besar penggunaan 30% atau 50% dari jumlah testi dengan skor total tertinggi dan terbawah masing-masing sebagai kelompok atas dan kelompok bawah akan menghasilkan indeks diskriminasi yang sama baiknya tanpa menghiraukan bentuk distribusi skor totalnya (Crocker & Algina, 2008).

Sesudah diperoleh kelompok atas dan kelompok bawah, indeks diskriminasi masing-masing item dapat dihitung dengan rumus sebagai berikut (Allen & Yen, 1979; Crocker & Algina, 2008):

$$d_i = (U_i - L_i)/n_i \qquad \textbf{Rumus 10.1.}$$

d_i = indeks diskriminasi item *i*
U_i = proporsi kelompok atas yang menjawab dengan benar item *i*
L_i = proporsi kelompok bawah yang menjawab dengan benar item *i*

Nilai-nilai d_i berada dalam kisaran -1.00 sampai dengan +1.00. Nilai-nilai d_i yang positif menunjukkan bahwa item-item yang bersangkutan menjalankan fungsinya dengan baik yaitu menunjukkan keunggulan kelompok atas dibandingkan kelompok bawah dalam menjawab item-item yang bersangkutan. Sebaliknya nilai-nilai d_i yang negatif menunjukkan kegagalan item-item yang bersangkutan dalam menjalankan fungsinya, sebab dalam menjawab item-item tersebut kelompok bawah justru lebih unggul jumlah dibandingkan kelompok atas (Crocker & Algina, 2008).

Lantas berapa atau seperti apa nilai d_i yang ideal? Mengutip hasil penelitian Ebel yang dipublikasikan pada 1965, Crocker dan Algina (2008; h. 315) menyajikan pedoman cara menafsirkan nilai-nilai d_i seperti disajikan dalam Tabel 10.3.

Tabel 10.3.
Cara Menafsirkan Indeks Diskriminasi Item (d_i)

d_i	Penafsiran
≥ 0.40	Item yang bersangkutan memiliki daya diskriminasi yang memuaskan.
0.30 – 0.39	Item yang bersangkutan bisa dipertahankan atau perlu dikenai revisi kecil.
0.20 – 0.29	Item yang bersangkutan kurang memiliki daya diskriminasi dan perlu direvisi.
≤ 0.20	Item yang bersangkutan harus digugurkan atau direvisi secara total.

Seperti sudah disinggung, empat indeks diskriminasi item lain yang akan kita bahas selanjutnya secara berturut-turut didasarkan pada korelasi antara skor item dan skor total tes sebagai kriteria internal.

2) Korelasi *Pearson Product Moment*

Rumus korelasi ini cocok diterapkan untuk menghitung koefisien korelasi antara skor item dan skor total tes khususnya pada jenis-jenis *multi-point item*, yaitu jenis item yang memiliki kisaran skor 1-4 atau lebih (Kline, 1986; Crocker & Algina, 2008).

3) Korelasi *Point-Biserial*

Rumus korelasi ini cocok diterapkan untuk menghitung koefisien korelasi antara skor item dan skor total tes khususnya pada jenis-jenis item dikotomis, termasuk jenis item lain yang penskorannya pada dasarnya bisa direduksi menjadi "benar/salah" atau "sesuai kunci jawaban/tidak sesuai kunci jawaban" yaitu diberi skor 1 jika "benar" atau "sesuai kunci jawaban" atau 0 jika "salah" atau "tidak sesuai kunci jawaban" (Kline, 1986).

4) Korelasi *Biserial*

Rumus korelasi ini cocok diterapkan untuk menghitung koefisien korelasi antara skor item dan skor total tes khususnya jika asumsi-asumsi berikut ini terpenuhi: (1) variabel laten yang mendasari kinerja testi pada setiap item, dengan kata lain skor total tesnya, terdistribusikan secara normal; (2) regresi antara skor item dan skor total tes sebagai kriterianya bersifat linear (Crocker & Algina, 2008; Kline, 1986).

5) Koefisien Phi

Rumus korelasi ini cocok diterapkan untuk menghitung koefisien korelasi antara skor item dan skor total tes khususnya jika skor itemnya bersifat dikotomis sedangkan skor total tesnya pun dikonversikan menjadi dua kategori yang juga bersifat dikotomis seperti "lulus/tidak lulus" atau "di atas rerata/di bawah rerata" atau dua kategori lain yang bersifat non-kontinyu (Kline, 1986; Crocker & Algina, 2008). Bahkan menurut Crocker dan Algina (2008), koefisien phi ini paling tepat diterapkan manakala kedua variabel yang dikorelasikan bersifat murni dikotomis.

6) Korelasi Tetrakorik

Rumus korelasi ini cocok diterapkan untuk menghitung koefisien korelasi antara skor item dan skor total tes khususnya jika baik skor item maupun skor total tesnya yang bersifat dikotomis

merupakan hasil dikotomisasi dari distribusi skor yang sesungguhnya bersifat kontinyu. Rumus yang sering disetarakan dengan koefisien phi ini jarang digunakan sebab selain perhitungannya rumit kesalahan bakunya juga besar, yaitu dua kali lipat kesalahan baku korelasi *product moment* (Kline, 1986; Crocker & Algina, 2008). Karena menuntut dikotomisasi skor total tes, baik r_{tet} maupun Φ dipandang membuang sejumlah data penting tertentu, maka cenderung tidak direkomendasikan (Kline, 1986).

Lantas indeks diskriminasi item mana sebaiknya kita gunakan dalam analisis item? Crocker dan Algina (2008) memberikan saran sebagai berikut:

a) Jika item-item memiliki taraf kesukaran sedang, semua indeks diskriminasi item sama baiknya untuk diterapkan. Jika perhitungan dilakukan secara manual kiranya pilihan akan jatuh pada indeks diskriminasi item d_i. Karena sudah tersedia fasilitas olah data dengan menggunakan komputer, pilihan lain pun bahkan bisa menjadi lebih mudah.
b) Jika penyusun tes ingin memperoleh item-item dengan taraf kesukaran ekstrem tinggi (mudah) atau sebaliknya ekstrem rendah (sukar), maka disarankan menerapkan korelasi biserial dengan syarat asumsi normalitas distribusi skor total terpenuhi.
c) Jika penyusun tes memperkirakan bahwa kelompok testi yang akan dikenai tes memiliki abilitas yang berbeda dari sampel testi yang dipakai untuk analisis item, maka disarankan menerapkan rumus korelasi biserial sebab indeks diskriminasi item yang dihasilkan cukup stabil untuk diterapkan pada aneka sampel dengan abilitas yang berlainan.
d) Jika penyusun tes cukup yakin bahwa kelompok testi yang akan dikenai tes memiliki abilitas yang setara dengan sampel testi yang dipakai untuk analisis item, sedangkan penyusun tes juga berharap memperoleh item-item dengan konsistensi internal yang tinggi, maka mengikuti hasil penelitian Lord dan Novick (1968, dalam Crocker & Algina, 2008; h. 320) disarankan menerapkan rumus korelasi *point biserial*.

e) Jika baik item-item maupun skor total tes diskor secara dikotomis dan distribusi keduanya bisa diasumsikan normal, maka koefisien phi atau koefisien tetrakorik sama baiknya untuk diterapkan. Secara khusus koefisien tetrakorik disarankan untuk diterapkan jika data korelasi yang diperoleh akan dianalisis lebih lanjut dengan analisis faktor.

Namun karena menurut model klasik tentang kesalahan pengukuran, korelasi item-total setara dengan rerata korelasi antara sebuah item dengan semua item lain maka rumus yang paling ampuh untuk diterapkan pada semua situasi sesungguhnya adalah rumus korelasi *point biserial* (r_{pbis}). Menurut Kline (1986), r_{pbis} memberikan parameter atau ukuran terbaik tentang korelasi antara skor item dan skor total tes yang merupakan syarat esensial dalam menyusun sebuah tes yang homogen.

c. Variabilitas Skor Item serta Korelasi Skor Item dan Kriteria: Indeks Reliabilitas Item & Indeks Validitas Item

Kategori ketiga parameter item yang sering dipakai sebagai dasar seleksi item adalah yang disebut **indeks reliabilitas item** dan **indeks validitas item.** Kedua parameter ini merupakan fungsi dari gabungan antara variabilitas skor item dan korelasi antara skor item dengan sebuah kriteria (Crocker & Algina, 2008).

Jika kriteria yang digunakan adalah kriteria internal berupa skor total tes, maka parameter yang dihasilkan disebut **indeks reliabilitas item** dan diberi lambang $s_i r_{iX}$; di mana s_i adalah deviasi standar skor item [$s_i = \sqrt{pi(1 - pi)}$]; r_{iX} adalah korelasi *point-biserial* antara skor item dan skor total tes. Sebagai dasar seleksi item, parameter ini dipandang akan menghasilkan tes dengan reliabilitas konsistensi internal yang lebih baik dibandingkan jika pemilihan item didasarkan pada korelasi item-total biasa.

Jika kriteria yang digunakan adalah kriteria eksternal berupa ukuran dari atribut psikologis tertentu yang hendak diprediksikan dengan tes yang bersangkutan, maka parameter yang dihasilkan

disebut **indeks validitas item** dan diberi lambang $s_i r_{iY}$; di mana s_i adalah deviasi standar skor item [$s_i = \sqrt{pi(1 - pi)}$] seperti pada indeks reliabilitas item; r_{iY} adalah korelasi *point-biserial* antara skor item dan skor kriteria eksternal. Sebagai dasar seleksi item, parameter ini dipandang mampu menghasilkan tes dengan validitas terkait kriteria (*criterion-related validity*) yang lebih baik dibandingkan jika pemilihan item didasarkan pada korelasi item-total biasa.

Mungkinkah menerapkan kedua parameter tersebut secara bersamaan demi memperoleh sebuah tes yang memiliki reliabilitas konsistensi internal yang baik sekaligus memiliki validitas terkait kriteria yang juga baik? Menurut Allen & Yen (1979), langkah semacam itu mustahil. Menurut mereka, "if items are chosen to maximize validity, the resulting test may not have good internal-consistency reliability" (Allen & Yen, 1979; h. 125). Maksudnya, jika item-item dipilih dengan cara sedemikian rupa demi memaksimalkan validitas (terkait kriteria), tes yang dihasilkan kiranya tidak akan memiliki reliabilitas konsistensi internal yang baik. Begitu pula kiranya sebaliknya, jika item-item dipilih dengan cara demi memaksimalkan reliabilitas konsistensi internal maka tes yang dihasilkan kiranya tidak akan memiliki validitas terkait kriteria yang bagus. Dengan kata lain, jika mau menggunakan parameter item kategori ketiga ini sebagai dasar seleksi item seorang penyusun tes harus memilih mana yang akan diutamakannya, reliabililtas konsistensi internal atau validitas terkait kriteria dari tes yang hendak dihasilkannya. Tidak bisa memilih keduanya sekaligus.

Menghadapi dilema semacam itu, Crocker dan Algina (2008) memberikan jalan keluar yang lebih radikal sebagai berikut. Menurut mereka, "as long as items with medium difficulties are chosen, there is little practical advantage in using the item reliability index instead of the item total score correlation" (h. 320). Maksudnya, sejauh item-item yang dipilih sebagai bentuk final tes memiliki taraf kesukaran sedang maka penggunaan korelasi item-total biasa sebagai dasar seleksi item tetap mampu menghasilkan tes yang sama baiknya dalam hal reliabilitas (konsistensi internal) dan validitas (terkait kriteria) seperti

tes yang akan dihasilkan jika seleksi itemnya didasarkan pada indeks reliabilitas item (dan indeks validitas item). Dengan kata lain, dengan hanya mengandalkan korelasi item-total biasa pun dalam melakukan seleksi item seorang penyusun tes tetap akan mampu menghasilkan sebuah tes dengan reliabilitas konsistensi internal dan validitas terkait kriteria yang baik tanpa perlu menerapkan indeks reliabilitas item maupun indeks validitas item. Pada bagian selanjutnya secara khusus akan kita bahas langkah melakukan seleksi item untuk mendapatkan bentuk final tes yang kita susun. Namun sebelumnya akan dibahas dulu sebuah parameter item tambahan, yaitu efektivitas distraktor.

d. Analisis Efektivitas Distraktor

Di luar tiga kategori parameter item di atas, masih ada satu parameter tambahan yang sesungguhnya juga penting dipertimbangkan dalam seleksi item khususnya pada item-item dengan format pilihan ganda, yaitu efektivitas distraktor masing-masing item. Parameter tersebut lazim diperiksa lewat analisis efektivitas distraktor atau analisis daya distraktor (Friedenberg, 1995).

Sebagaimana sudah disinggung, dalam format pilihan ganda item atau soal akan berupa pernyataan atau pertanyaan yang disebut *stem* dan diikuti dengan sejumlah *opsi* atau alternatif jawaban. Salah satu dari antara opsi tersebut merupakan jawaban benar atau paling benar yang disebut *kunci jawaban*, sedangkan opsi sisanya merupakan jawaban yang kurang benar atau bahkan salah dan disebut *distraktor.* Distraktor yang efektif dalam arti menjalankan fungsi atau perannya harus mampu menarik perhatian testi untuk memilihnya. Maka, daya atau efektivitas distraktor dievaluasi berdasarkan persentase testi yang memilih masing-masing opsi yang salah dalam menjawab item (Friedenberg, 1995).

Ada beberapa metode analisis efektivitas distraktor, salah satu di antaranya dan yang akan kita uraikan di sini adalah yang bisa kita beri nama **metode Friedenberg** (1995). Metode ini didasarkan pada sejumlah konsep dasar sebagai berikut. *Pertama,* item pilihan ganda

yang baik memiliki dua ciri: (1) testi yang memiliki kemampuan yang diujikan akan memilih jawaban yang benar; dan (2) testi yang kurang atau tidak memiliki kamampuan yang diujikan akan menebak dalam menjawab dengan cara memilih secara acak dari antara opsi yang tersedia. Pada kasus yang kedua dan jika distraktor menjalankan fungsinya dengan baik, secara teoretis jawaban testi akan tersebar secara merata pada masing-masing opsi dan sebagian di antaranya berhasil menebak jawaban secara tepat.

Kedua, berdasarkan konsepsi di atas dalam analisis efektivitas distraktor lazim dibedakan antara *expected distractor power* atau efektivitas distraktor yang diharapkan dan *actual distractor power* atau efektivitas distraktor yang nyata. *Efektivitas distraktor yang diharapkan* adalah jumlah testi yang diharapkan memilih masing-masing distraktor atau opsi yang salah secara acak. Rumus perhitungannya adalah sebagai berikut:

$$\textit{Efektivitas distraktor yang diharapkan} = \frac{\textit{jumlah testi yang menjawab salah pada item}}{\textit{jumlah distraktor}}$$

Rumus 10.2.

Efektivitas distraktor yang nyata adalah jumlah testi yang nyata-nyata memilih masing-masing distraktor atau opsi yang salah. Efektivitas sebuah distraktor dianalisis dengan cara membandingkan efektivitas distraktor nyata dengan efektivitas distraktor yang diharapkan.

Ketiga, berdasarkan konsep-konsep di atas secara umum kriteria untuk menentukan efektivitas distraktor sebuah item adalah sebagai berikut: (1) sebuah item adalah baik jika efektivitas distraktor nyata dari masing-masing distraktornya adalah sama atau makin mendekati efektivitas distraktor yang diharapkan; (2) dalam sebuah item distraktor yang tidak dipilih oleh satu pun testi, atau yang dipilih oleh testi dalam jumlah yang terlampau kecil atau sebaliknya terlampau besar dibandingkan jumlah testi yang diharapkan, kiranya perlu dilihat kembali atau bahkan perlu ditulis ulang (Friedenberg,

1995). Penerapan metode Friedenberg dalam analisis efektivitas distraktor tersebut akan mencakup langkah-langkah sebagaimana tersirat dalam contoh seperti disajikan di Tabel 10.4.

Tabel 10.4. adalah contoh penerapan metode Friedenberg dalam analisis distraktor sebuah tes yang terdiri dari *n* item pilihan ganda dengan 4 (empat) opsi yang diadministrasikan pada sampel testi yang terdiri dari 30 orang. Item nomor 1 cukup mudah; distraktor B dan D cukup baik sebab berselisih tidak terlalu besar (0.67) dari efektivitas distraktor yang diharapkan; namun distraktor C kurang baik sebab jumlah pemilih nyatanya melebihi jumlah pemilih yang diharapkan secara cukup signifikan. Item nomor 5 cukup sukar; ketiga distraktornya kurang efektif sebab distraktor A dan B memiliki pemilih nyata jauh melampaui yang diharapkan sedangkan distraktor D tidak ada yang memilih sama sekali.

Tabel 10.4.
Contoh Analisis Efektivitas Distraktor

Item	Jumlah Testi Menjawab Benar	Jumlah Testi Menjawab Salah	Jumlah Testi Diharapkan Memilih Masing-masing Distraktor	Jumlah Testi yang Memilih			
				A	B	C	D
1	25	5	1.67	*	1	3	1
5	16	14	4.67	7	7	*	0

Diadaptasikan dari Lisa Friedenberg (1995), *Psychological testing. Design, analysis, and use,* Boston: Allyn & Bacon, h. 287.

e. Melakukan Seleksi Item

Langkah krusial terakhir dalam analisis item adalah melakukan seleksi item berdasarkan aneka parameter item yang sudah diperoleh, khususnya taraf kesukaran item (p_i) dan korelasi item-total (r_{it}) sebagai indeks daya diskriminasi item. Menurut Kline (1986), untuk bisa memilih item yang sungguh-sungguh efektif terlebih dulu perlu diketahui kriteria umum sebuah tes yang efisien sekaligus efektif. Kriteria yang dimaksud adalah sebagai berikut (Kline, 1986):

1) Panjang tes. Untuk memenuhi syarat reliabilitas, tes harus terdiri dari antara 20 dan 30 item.
2) *Content* atau isi tes. Untuk menjamin validitas, tes harus terdiri dari item-item dengan cakupan keragaman isi yang luas.
3) Korelasi item-total. Parameter ini merupakan kriteria utama. Makin tinggi korelasi item-total, makin baik. Idealnya, semua item harus memiliki koefisien korelasi item-total di atas 0.20.
4) Taraf kesukaran. Parameter ini juga sangat penting untuk tes abilitas pada umumnya maupun tes prestasi khususnya. Sebagai patokan umum, item-item dengan taraf kesukaran dalam kisaran 0.80 – 0.20 dipandang memuaskan. Yang tidak kalah penting, sebaiknya tes terdiri dari item-item dengan taraf kesukaran yang berlainan.
5) Sebagai tambahan, perlu juga kiranya diperhatikan efektivitas distraktor masing-masing item, dalam artian perlu dihindari memilih item-item dengan satu atau lebih distraktor yang tidak efektif.

Berpedoman para kriteria di atas, berikut ini adalah langkah-langkah yang disarankan dalam melakukan seleksi item untuk menyusun bentuk final sebuah tes abilitas khususnya tes hasil belajar di sekolah (Kline, 1986):

1) Pilihlah semua item yang dalam analisis item terbukti memenuhi dua syarat utama, yaitu memiliki r_{pbis} dan p yang ideal, serta memiliki distraktor-distraktor yang cukup efektif.
2) Cermatilah item-item yang tidak memenuhi salah satu dari antara kedua syarat utama tersebut, khususnya memiliki r_{pbis} tinggi namun dengan p yang terlampau tinggi atau sebaliknya terlampau rendah. Cobalah diperiksa, benarkah gejala tersebut hanya disebabkan oleh karakteristik khusus tertentu dari sampel yang digunakan untuk analisis item. Jika memang begitu, mungkin item semacam itu bisa tetap dipertahankan.
3) Periksalah *content* atau isi dari item-item yang berhasil lolos seleksi untuk melihat benarkah item-item tersebut sudah mencakup

seluruh aspek dari atribut yang ingin kita ukur dengan tes tersebut. Jika belum, cobalah periksa item-item yang tidak lolos seleksi untuk melihat adakah di antaranya yang mengukur aspek yang belum tercakup tersebut dan mendekati syarat statistis yang dituntut. Jika ada beberapa, ambil saja untuk dimasukkan ke dalam bentuk final tes. Jika tidak ada dan kekurangan tersebut kita pandang serius, maka harus ditulis item-item baru dan harus kita uji cobakan kembali.

4) Hitunglah jumlah item-item yang berhasil lolos seleksi. Jika sudah memenuhi syarat jumlah seperti tercantum dalam kriteria (20 – 30 item) dan memiliki cakupan *content* atau isi yang sudah memuaskan, segera hitunglah koefisien reliabilitas konsistensi internalnya (dengan rumus K-R20). Jika koefisien reliabilitas yang diperoleh ≥ 0.70, berarti reliabilitas tes tersebut memuaskan. Berarti proses penyusunan tes bisa kita akhiri sebab sudah kita peroleh sebuah tes yang homogen dan reliabel serta dengan *content* atau isi yang relevan dengan tujuan tes.
5) Jika reliabilitas tes yang kita peroleh kurang tinggi (< 0.70), cobalah kita tambahkan item-item baru, diambilkan dari yang terbaik secara statistis dari antara item-item yang tidak lolos seleksi, lantas kita hitung kembali koefisien reliabilitas konsistensi internalnya (K-R20). Langkah ini bisa kita ulangi secukupnya, sampai penambahan item baru tidak lagi berdampak meningkatkan koefisien reliabilitas konsistensi internalnya.
6) Jika sudah berhasil kita peroleh sebuah bentuk final tes yang reliabel dan memiliki cakupan *content* atau isi yang memuaskan, periksalah distribusi skor totalnya. Distribusi skor total tes ini harus bersifat kurang lebih simetrik atau normal. Selain itu, periksalah variansnya. Tes yang baik harus memiliki varians yang tinggi atau besar sebab berarti tes tersebut menjalankan fungsi diskriminasi yang baik. Jika variansnya terlampau kecil, tambahkanlah item-item baru untuk meningkatkan daya diskriminasi tes.
7) Jika akhirnya sudah kita peroleh sebuah tes yang memuaskan dari segi besar variansnya, hitunglah delta Ferguson-nya. Jika

delta Ferguson-nya > 0.90 berarti bentuk final tes tersebut benar-benar memiliki daya diskriminasi yang baik.

8) Jika semua hal di atas sudak OK, lakukanlah validasi silang terhadap bentuk final tes tersebut dengan menggunakan kelompok sampel yang baru. Jika semua item mencapai r_{pbis} dan p yang memuaskan, berarti bentuk final tes kita memang sudah bagus sehingga proses penyusunan tes benar-benar bisa kita akhiri. Ψ

Bab 11
Penyusunan
Typical Performance Tests

Seperti sudah diuraikan di Bab 4, *typical performance tests* bertujuan mengukur *trait* atau kepribadian, yaitu disposisi atau kecenderungan bertingkah laku dengan cara tertentu mencakup jenis-jenis atribut psikologis yang lebih didominasi oleh fungsi afeksi atau rasa dan karsa. Dari segi isi atribut psikologis yang menjadi objek atau sasaran pengukurannya, *typical performance tests* atau tes kepribadian dapat digolongkan ke dalam empat kategori, yaitu tes kepribadian yang mengukur (1) *social traits* atau sifat sosial, (2) *motives* atau motif, *needs* atau kebutuhan, dan *drives* atau dorongan, (3) *personal conceptions* atau konsepsi tentang diri, serta (4) *adjustment versus maladjustment* atau penyesuaian-diri yang baik versus penyesuaian-diri yang salah. Dari segi struktur atau format tesnya, *typical performance tests* atau tes kepribadian dapat digolongkan ke dalam dua kategori besar, yaitu: (1) tes kepribadian terstruktur lazimnya berupa *inventori kepribadian*, dan (2) tes kepribadian tak terstruktur lazimnya berupa aneka jenis tes proyektif.

Dalam bab ini pembahasan akan difokuskan pada penyusunan kategori tes kepribadian terstruktur khususnya berupa inventori kepribadian. *Self-inventories* atau inventori kepribadian merupakan salah satu metode yang paling luas digunakan dalam asesmen kepribadian (Nunnally, Jr., 1974). Inventori kepribadian lazimnya berupa "printed tests in which the individual is required to describe himself" (Nunnally, Jr., 1974, h. 358), yaitu tes tulis atau tercetak yang menuntut individu mendeskripsikan dirinya. Sesuai namanya, metode ini paling cocok untuk mengukur aneka jenis *personal conceptions* meliputi cara orang berpikir tentang atau memandang dirinya sendiri dan dunia sekelilingnya. Beberapa contoh atribut kepribadian yang mencerminkan cara orang memandang dirinya sendiri meliputi

konsep-diri dan harga-diri. Beberapa contoh atribut kepribadian yang mencerminkan cara orang memandang dunia sekelilingnya meliputi minat, sikap, dan nilai kehidupan. Kenyataannya, metode inventori juga lazim dipakai untuk mengukur atribut kepribadian yang masuk dalam kategori *social traits, motives,* maupun a*djustment-maladjustment* (Nunnally, Jr., 1974). Maka, penyusunan inventori kepribadian ini kita pilih sebagai fokus pembahasan dalam bab ini dengan keyakinan bahwa metode tersebut cocok digunakan untuk mengukur semua jenis disposisi kepribadian.

Seperti dalam pembahasan tentang penyusunan *maximal performance tests* khususnya *achievement tests* di Bab 10, pembahasan tentang penyusunan *typical performance tests* di bab ini juga hanya akan difokuskan pada langkah-langkah yang bersifat khas dalam penyusunan inventori kepribadian. Langkah-langkah khas dalam penyusunan *typical performance tests* meliputi (a) mendefinisikan tes, (b) memilih metode penskalaan, (c) menuliskan item, (d) uji coba dan analisis item, dan (e) merevisi item.

A. Mendefinisikan Tes

Dalam penyusunan inventori kepribadian, pendefinisian tes mencakup tiga hal penting. *Pertama* tentang kelompok sasarannya, inventori kepribadian atau skala psikologis ini hendak ditujukan bagi kelompok subjek seperti apa sebagai kelompok sasaran. Kita ingat, sebuah tes pengukur atribut psikologis tertentu hanya valid digunakan untuk diterapkan pada kelompok subjek tertentu. Contoh, inventori atau skala tentang konsep diri untuk kelompok subjek remaja tidak akan valid diterapkan pada kelompok subjek anak, dan begitu juga sebaliknya.

Kedua, tentang keunikan atau kekhasan inventori kepribadian yang hendak disusun, khususnya jika sudah terdapat inventori kepribadian lain sejenis yang mengukur atribut psikologis yang sama. Keunikan tersebut bisa dalam hal teori atau konsep tentang

atribut psikologis yang dipakai, metode penskalaan yang diterapkan, kelompok sasaran yang dipilih, dan sebagainya.

Ketiga, sudah barang tentu yang utama adalah tentang atribut psikologisnya sendiri yang hendak dijadikan objek atau sasaran pengukuran, inventori kepribadian atau skala psikologis ini hendak mengukur atribut psikologis apa? Hampir seluruh atribut psikologis berupa *trait* atau disposisi kepribadian yang menjadi objek atau sasaran pengukuran dengan inventori kepribadian merupakan konstruk dalam arti sempit sebagaimana didefinisikan oleh Cronbach dan Meehl (1956), yaitu konstruk teoretis hasil konstruksi para pakar psikologi yang belum jelas batas-batas *behavioral content domain* atau ranah isi perilakunya. Maka langkah pendefinisian tes yang sekaligus mengandung sublangkah pengidentifikasian ranah isi perilaku konstruk atau atribut psikologis yang menjadi objek pengukurannya akan melibatkan strategi *eksplikasi konstruk* sebagaimana dikemukakan oleh Friedenberg (1995). Karena langkah ini secara lengkap sudah diuraikan di Bab 9, maka di sini tidak akan diuraikan lagi. Pembaca dipersilakan membaca bagian yang menguraikan hal tersebut di Bab 10.

Terkait eksplikasi konstruk, kiranya sangat relevan dikemukakan tentang pendekatan khas yang lazim diterapkan dalam penyusunan inventori kepribadian. Menurut Burisch (1984), ada tiga pendekatan dalam menyusun inventori kepribadian, yaitu (1) *external approach* atau pendekatan eksternal, yang sering juga disebut *empirical approach* atau *criterion group approach,* (2) *inductive approach* atau pendekatan induktif, yang sering juga disebut *internal approach, internal consistency approach,* atau *itemetric approach,* dan (3) *deductive approach* atau pendekatan deduktif, yang sering juga disebut *rational approach, intuitive approach,* atau *theoretical approach.* Penjelasan ringkas tentang masing-masing pendekatan tersebut adalah seperti diuraikan di bawah ini.

1. Pendekatan Eksternal

Pendekatan ini didasarkan pada sejumlah asumsi sebagai berikut (Burisch, 1984). *Pertama*, orang cenderung mengelompok dalam himpunan-himpunan, yaitu himpunan orang dengan karakteristik tertentu dan himpunan orang dengan karakteristik lawannya sehingga terbentuk sejenis pasangan yang berlawanan sifat. Misal kelompok orang yang bersifat maniak dan lawannya kelompok orang yang bersifat depresif; kelompok orang yang bersifat skizofrenik dan lawannya kelompok orang yang bersifat histeris, dan sebagainya. *Kedua*, pasangan kelompok orang yang berlawanan sifat semacam itu bisa dibedakan dengan menggunakan sebuah inventori kepribadian, namun seperti apa perbedaan mereka dalam menjawab item-item dalam inventori kepribadian tidak bisa diketahui sebelumnya. Maka, secara umum penyusunan inventori kepribadian yang menggunakan pendekatan eksternal akan mencakup langkah-langkah sebagai berikut: (1) menyusun atau mengumpulkan sejumlah besar item yang bersifat heterogen terkait atribut psikologis tertentu yang sedang diukur, (2) mengadministrasikan atau mengenakan item-item tersebut pada dua kelompok subjek yang merupakan pasangan yang berlawanan sifat, yaitu satu kelompok subjek yang diasumsikan memiliki atribut yang sedang diukur dalam jumlah yang signifikan atau mencolok dan kelompok subjek lawannya, (3) membandingkan jawaban kedua kelompok subjek tersebut pada masing-masing item, untuk melihat perbedaan jawaban mereka item demi item, dan (4) item-item yang dijawab secara berbeda dengan mencolok oleh kedua kelompok subjek dipilih untuk dijadikan item-item inventori kepribadian untuk mengukur atribut psikologis yang sedang diteliti.

Pendekatan ini disebut eksternal, sebab penentuan masuk-tidaknya setiap item ke dalam inventori kepribadian didasarkan pada faktor-faktor di luar ranah kuesioner atau inventorinya (Burisch, 1984). Gregory (2007) menyebut pendekatan ini *method of empirical keying* atau metode penetapan kunci (jawaban) secara empiris, dan

membahasnya dalam rangka menguraikan aneka metode penskalaan (*scaling method*) dalam penyusunan tes.

2. Pendekatan Induktif

Pendekatan ini didasarkan pada dua asumsi sebagai berikut (Burisch, 1984): (1) ada struktur dasar tertentu yang bersifat universal sesuai hukum alam dan bisa diungkap yang mendasari perbedaan individual antar orang, (2) struktur tersebut belum diketahui secara rinci, namun bisa dipastikan bahwa sifatnya *simple* atau sederhana, dan (3) perbedaan individual antar orang pada tingkat respon terhadap masing-masing item belum diketahui.

Untuk mengungkap secara induktif struktur dasar yang bersifat alamiah-universal tersebut akan meliputi langkah-langkah: (1) mengumpulkan *item pool* atau himpunan item dalam jumlah besar, dengan cara menyusun (*to invent)* item-item baru maupun meminjam item-item dari tes-tes yang sudah ada; (2) himpunan besar item tersebut selanjutnya diadministrasikan pada sekelompok subjek dengan jumlah sebesar mungkin; (3) data yang terkumpul dianalisis secara kompleks dengan tujuan untuk mengungkap atau menemukan struktur yang dimaksud; (4) dengan menggunakan teknik yang disebut "matrix staring," plot-plot atau pola-pola yang muncul dicoba dimaknai, dan akhirnya, (5) item-item dikelompok-kelompokkan ke dalam sejumlah skala.

Pendekatan ini disebut "induktif" sebab jumlah dan sifat skala-skala yang dihasilkan ditentukan oleh analisis datanya. Dengan kata lain, dalam pendekatan ini peneliti memulai dengan mengumpulkan item-item satu per satu, membiarkan data yang diperoleh "berbicara sendiri," sehingga akhirnya diperoleh sejumlah skala pada taraf abstraksi yang lebih tinggi. Pendekatan ini jelas berbeda dengan pendekatan eksternal, di mana peneliti terlebih dulu menentukan jenis skala seperti apa yang akan dia susun. Pendekatan ini jelas mengandalkan metode analisis yang berbasis analisis faktor. Seperti sudah disinggung, kendati memiliki keunggulan penerapan analisis

faktor dalam penyusunan tes dipandang memiliki sejumlah kendala baik yang bersifat substansial maupun praktis yang menjadikannya kurang direkomendasikan.

3. Pendekatan Deduktif

Berbeda dari dua pendekatan sebelumnya, dalam pendekatan ini peneliti bertolak dari asumsi-asumsi sebagai berikut (Burisch, 1984): (1) orang bisa menyusun skala untuk sifat kepribadian apa pun sebagaimana terwakili oleh semua kata sifat yang bisa ditemukan dalam bahasa percakapan sehari-hari; tentu saja, dalam menemukan sifat kepribadian yang hendak ditelitinya orang juga bisa mendasarkan diri pada teori kepribadian tertentu, misal teori Murray tentang aneka kebutuhan manusia, namun yang lebih ditekankan adalah mencari inspirasi dari kosa kata sifat dalam bahasa sehari-hari; (2) nama-nama sifat (*trait names*) lazimnya bersifat kabur, dalam arti masing-masing kurang didefinisikan secara ketat dan berupa kategori-kategori dari aneka kecenderungan tingkah laku lebih sederhana yang saling tumpang tindih.

Maka, penyusunan sebuah inventori kepribadian dengan pendekatan deduktif akan mencakup langkah-langkah sebagai berikut. *Pertama,* menentukan sifat yang akan diteliti dan merumuskan definisinya. Langkah pertama ini lazimnya akan mencakup strategi sebagai berikut: konstruk yang masih bersifat global tersebut perlu dijabarkan ke dalam sejumlah subkonstruk yang lebih spesifik. Burisch (1984) memberikan contoh, misal seorang peneliti ingin menyusun skala tentang *gregariousness* atau sifat suka berteman atau bergaul. Dia mengamati, ada orang yang sangat ingin memiliki banyak teman namun kurang memiliki ketrampilan bergaul maka orang semacam itu lebih sering sendirian. Sebaliknya, ada orang lain yang sebenarnya pandai bergaul namun lebih suka sendirian. Dengan kata lain, konstruk suka berteman mencakup dua subkonstruk, yaitu "kebutuhan untuk menjalin kontak dengan orang lain" dan "ketrampilan menjalin kontak dengan orang lain," atau sejenis itu.

Langkah ini mirip dengan apa yang oleh Friedenberg (1995) disebut *eksplikasi konstruk.* Dalam eksplikasi konstruk, suatu konstruk yang masih bersifat global atau kabur dalam arti tidak jelas batas-batas cakupannya itu dicoba disederhanakan dalam arti diperjelas atau dipertajam cakupannya dengan cara diidentifikasikan *behavioral indicators* alias indikator-indikator tingkah lakunya berupa bentuk-bentuk tingkah laku spesifik yang bisa diamati dan diukur, baik yang bersifat mendukung (*favorable*) maupun yang bersifat menyangkal atau mengingkari (*unfavorable*) keberadaan konstruk psikologis yang bersangkutan (Friedenberg, 1995).

Kedua, sesudah diperoleh definisi atau definisi-definisi yang memuaskan tentang konstruk yang akan diukur, penyusun tes menuliskan item-item sesuai definisi atau definisi-definisi tersebut. Burisch (1984) menyarankan agar dalam bentuk final tesnya hanya digunakan item-item yang sungguh-sungguh mengena ("hit the core") dengan definisinya. Dengan menggunakan metode eksplikasi konstruk Friedenberg (1995), langkah ini sangat dipermudah, sebab penyusun skala tinggal menjabarkan indikator-indikator baik yang *favorable* atau positif maupun yang *unfavorable* atau negatif ke dalam sub-subindikator yang semakin operasional dalam arti mudah diamati dan diukur.

Pendekatan ini disebut "deduktif" sebab pemilihan dan pendefinisian konstruk atau subkonstruk-subkonstruk mendahului dan membimbing formulasi atau penulisan item-item. Burisch (1984) memberikan catatan bahwa menyusun inventori tentang konsep-konsep sifat kepribadian yang diturunkan dari sebuah teori kepribadian seringkali lebih jitu dibandingkan dengan mengandalkan sumber psikologi sehari-hari ("folk" psychology). Namun, mengutip seorang pakar lain, dia juga menyatakan bahwa "regardless of the theoretical considerations which guide scale construction or the mathematical elegance of item-analytic procedures, the practical utility of a test must be assessed in terms of *the number and magnitude of its correlations with nontest criterion measures"* (Wiggins, 1973; dalam Burisch, 1984, h. 215). Maksudnya, pada akhirnya manfaat praktis

sebuah tes harus diukur dari besarnya korelasi tes itu dengan ukuran-ukuran kriteria yang tidak berupa tes, misalnya berupa kinerja atau perilaku dalam kehidupan nyata. Burisch (1984) mengakui, dalam praktek banyak peneliti menggabungkan unsur-unsur baik dari masing-masing pendekatan.

Sesudah atribut psikologis sebagai konstruk yang hendak diukur berhasil dieksplikasikan dengan baik sedangkan *behavioral content domain*-nya pun berhasil diidentifikasikan dengan jelas, maka langkah selanjutnya adalah menuliskan item-item. Namun perlu digaris-bawahi di sini bahwa dalam penyusunan inventori kepribadian pemilihan format item sesungguhnya sangat terkait dengan pemilihan jenis skala yang hendak diterapkan. Maka, sebelum benar-benar mulai menuliskan item-item terlebih dulu harus ditentukan jenis skala yang hendak diterapkan.

B. Memilih Metode Penskalaan

Perlu kita ingat bahwa "semua bilangan hasil pengukuran dapat ditempatkan pada salah satu dari empat kategori skala yang bersifat hirarkis, yaitu nominal, ordinal, interval, dan rasio; masing-masing kategori mewakili satu taraf pengukuran" (Stevens, 1946). Selanjutnya, taraf pengukuran yang kita pilih berkaitan dengan jenis statistik parametrik yang cocok untuk diterapkan. Prinsipnya, prosedur atau teknik statistik yang makin *powerful* atau makin kuat dan makin *useful* atau makin jamak digunakan (seperti *r* Pearson, analisis varians, regresi ganda) hanya cocok diterapkan pada data yang memenuhi kriteria **skala interval** atau **rasio**. Terhadap data yang berskala nominal atau ordinal hanya cocok dikenakan prosedur statistik nonparametrik yang kurang kuat seperti khi-kuadrat, korelasi tata jenjang, dan tes median. Dalam praktek, sebagian besar instrumen psikologis termasuk inventori kepribadian diasumsikan menerapkan taraf pengukuran interval kendati sangat sulit untuk membuktikannya (Gregory, 2007). Dengan latar belakang pemikiran

semacam itu, ada beberapa metode penskalaan yang dapat kita terapkan dalam menyusun inventori kepribadian sebagaimana diuraikan di bagian berikut ini.

1. *Expert Rankings* atau Penetapan Urutan Jenjang oleh Ahli

Salah satu contoh penerapan metode penskalaan ini adalah *behavioral rankings of experts* atau penetapan urutan jenjang tingkah laku oleh ahli yang diterapkan dalam penyusunan *Glasgow Coma Scale* (Teasdale & Jennet, 1974, dalam Gregory, 2007). Langkah-langkahnya adalah sebagai berikut: sekelompok neurolog atau ahli neurologi diminta mengidentifikasikan sebanyak mungkin jenis tingkah laku yang mencerminkan aneka taraf kesadaran pasien yang mengalami koma; jenis-jenis tingkah laku yang berhasil diidentifikasikan tersebut kemudian diurutkan dalam sejenis kontinum taraf kesadaran, mulai dari taraf koma berat (*deep coma*) sampai taraf memiliki orientasi dasar alias sadar.

Dalam penyusunan *Glasgow Coma Scale* (selanjutnya disingkat GCS) tersebut ditemukan bahwa jenis tingkah laku yang ditetapkan oleh para neurolog sebagai indikator taraf koma atau sadar bisa dikategorikan dalam tiga wilayah, yaitu (1) respon membuka mata, (2) respon verbal, dan (3) respon motor atau gerak; dan masing-masing wilayah meliputi beberapa jenis respon spesifik. Jenis-jenis respon spesifik pada masing-masing wilayah tersebut kemudian diurutkan dalam kontinum mulai dari yang mencerminkan koma berat sampai ke taraf sadar, masing-masing diberi skor mulai dari 1 (koma berat) sampai dengan 4 atau 5 (sadar). Jumlah dari skor pada masing-masing wilayah tersebut merupakan skor total taraf kesadaran pasien. Makin tinggi skor total, makin baik taraf kesadarannya.

Tabel 11.1.
Contoh Identifikasi Jenis Tingkah Laku sebagai Indikator Taraf Kesadaran dalam Penyusunan *Glasgow Coma Scale* (Teasdale & Jennet, 1974)

	Wilayah Respon (Dimensi/ Komponen)	Jenis Respon	Skor
Skala koma Glasgow (Teasdale & Jennet, 1974)	Membuka mata	Secara spontan	4
		Menanggapi ucapan orang lain	3
		Menanggapi rasa sakit	2
		Nihil	1
	Memberikan respon verbal	Nyambung	5
		Kacau	4
		Slenco	3
		Tidak bisa dipahami	2
		Nihil	1
	Memberikan respon motor/ gerak	Mengikuti perintah	5
		Menunjukkan tempat rasa sakit	4
		Flexion to pain	3
		Extension to pain	2
		Nihil	1

Diadaptasikan dari Robert J. Gregory, (2007), *Psychological testing. History, principles, and applications,* Boston: Pearson, h. 146.

Penskalaan dengan metode *expert rankings* seperti di atas menghasilkan pengukuran pada *taraf ordinal*.

2. Metode *Equal-Appearing Intervals* atau Interval Tampak Setara

Sebagaimana sudah disinggung, metode yang dikembangkan oleh L.L. Thurstone (1929, dalam Gregory, 2007) ini merupakan adaptasi metode *paired comparisons* yang juga dikembangkan oleh pakar yang sama. Kedua metode ini didasarkan pada *law of comparative judgments* atau prinsip penilaian komparatif yang dikemukakan oleh

Thurstone. Inti hukum atau prinsip tersebut berbunyi: "a group of persons compares objects with respect to some physical property...and declares which of the pair has more of the property" (Andrich, 1990, h. 330). Maksudnya, sekelompok orang diminta membandingkan sejumlah objek terkait ciri fisik tertentu dan diminta menyatakan mana dari antara pasangan objek tersebut memiliki ciri yang dibandingkan dalam jumlah lebih besar.

Metode *pair comparisons* dipandang memiliki dua kelemahan utama, yaitu menuntut terlampau banyak waktu dan tenaga dari pihak subjek yang diminta menjadi *judges* atau penilai. Sebagaimana kita tahun, jika jumlah pernyataan yang harus diskala adalah *n* buah, maka rumus perhitungan jumlah pasang pernyataan yang harus dinilai oleh subjek adalah $n(n\text{-}1)/2$. Akibatnya, jika dalam rangka mengukur atribut psikologis tertentu kita menyusun 10 pernyataan, maka subjek harus menilai $10(10\text{-}1)/2$ pasang pernyataan atau 45 pasang pernyataan. Bisa dibayangkan beban penilai jika jumlah pernyataan yang harus diskala atau dinilai sebanyak 20 atau 30 buah atau bahkan lebih. Maka metode *paired comparisons* memang hanya cocok diterapkan untuk menskala pernyataan-pernyataan dalam pengukuran sikap atau inventori kepribadian lainnya jika jumlah pernyataannya tidak terlalu besar (Edwards, 1957).

Sebaliknya, metode *equal appearing intervals* (selanjutnya disingkat EAI) dipandang lebih sederhana dan sangat cocok untuk menyusun skala sikap khususnya maupun inventori kepribadian pada umumnya, sebab antara lain dalam metode EAI setiap subjek hanya dituntut memberikan satu penilaian komparatif terhadap setiap pernyataan sehingga tidak masalah jika pernyataan yang harus diskala berjumlah besar (Edwards, 1957).

Langkah-langkah penyusunan skala sikap atau inventori kepribadian dengan metode EAI adalah sebagai berikut:

a. Rumuskan sebanyak mungkin pernyataan benar-salah (*true-false statements*) yang mencerminkan variasi sikap positif (*favorable*) dan negatif (*unfavorable*) (termasuk yang moderat atau di antara positif dan negatif) terhadap objek atau atribut psikologis

tertentu yang menjadi sasaran pengukuran. Mengutip pendapat pakar lain Edwards (Thurstone & Chave, 1929, dalam Edwards, 1957) menyarankan bahwa setiap pernyataan tersebut sebaiknya ditulis atau dicetak pada sebuah kartu indeks. Terkait jumlah pernyataannya, ada yang menyarankan 100 buah (Andrich, 1990), namun juga harus mempertimbangkan *test blue-print* yang dikembangkan berdasarkan hasil eksplikasi konstruk khususnya untuk pengukuran atribut psikologis tertentu.

b. Mintalah sejumlah subjek untuk menilai taraf favorabilitas-unfavorabilitas masing-masing pernyataan terhadap objek atau atribut psikologis yang menjadi objek pengukuran. Terkait jenis dan jumlah subjek yang akan diminta menjadi *judges* atau penilai ada dua kemungkinan. Kemungkinan pertama, penilai itu terdiri dari *experts* atau ahli. Jika demikian, jumlahnya cukup sekitar 10 orang (Gregory, 2007). Kemungkinan kedua, penilai itu terdiri dari subjek dengan ciri-ciri seperti subjek yang akan dijadikan kelompok sasaran alat ukur tersebut. Jika demikian, jumlah subjek penilai tersebut harus merupakan sampel yang representatif sehingga jumlahnya harus cukup besar bahkan ada yang menyarankan sebesar 200 sampai 300 orang (Andrich, 1990).

c. Prosedur penilaiannya adalah sebagai berikut: Setiap subjek penilai diminta menempatkan setiap pernyataan dalam salah satu di antara 11 kategori yang terentang antara "*extremely unfavorable*" atau "sangat tidak favorable" (kategori 1 dan ditempatkan di sisi paling kiri) dan "*extremely favorable*" atau "sangat favorable" (kategori 11 dan ditempatkan di sisi paling kanan), sedangkan kategori-kategori lainnya ditempatkan secara berurutan di antara kedua kutub ekstrem tersebut (Gregory, 2007). Subjek diminta agar penilaian atau pemilahan pernyataan-pernyataan itu dilakukan sedemikian rupa sehingga interval antara kategori-kategori pernyataan yang tercipta secara subjektif tampak atau terasa setara.

d. Mengikuti cara yang pernah ditempuh sendiri oleh Thurstone bersama Chave (dalam Edwards, 1957) pengarang lain menyarankan, kategori-kategori tersebut diberi label huruf mulai dari huruf A bagi kategori "*extremely unfavorable*" dan yang ditempatkan di sisi paling kiri, diikuti kategori berikutnya dengan label huruf B dan seterusnya, sampai kategori "*extremely favorable*" di sisi paling kanan dan diberi label huruf K (Edwards, 1957). Dalam pelaksanaannya bisa digunakan kotak kosong atau sekadar kartu yang diberi label A sampai dengan K untuk menampung hasil penilaian subjek terhadap setiap pernyataan. Kontinum psikologis yang terentang dari "sangat tidak favorable" ke "sangat favorable" dipandang bersifat kontinyu dan memiliki interval yang *equidistant* alias sama atau setara (Edwards, 1957; Gregory, 2007).

e. Hitunglah nilai skala dan ukuran variabilitas penilaian subjek terhadap masing-masing pernyataan. Sebagaimana dilaporkan oleh Edwards (1957), Thurstone dan Chave menggunakan *median* sebagai nilai skala dan *interquartile range* atau Q (di mana $Q = Q_{75} - Q_{25}$) sebagai ukuran variabilitas distribusi penilaian subjek terhadap masing-masing pernyataan. Namun pengarang lain menunjukkan cara yang lebih sederhana, yaitu menggunakan *mean* dan *SD* distribusi penilaian subjek pada setiap pernyataan, masing-masing sebagai nilai skala dan ukuran variabilitas penilaian subjek terhadap pernyataan yang bersangkutan. Tabulasi data dalam rangka perhitungan nilai skala dan ukuran variabilitas setiap pernyataan akan berbentuk seperti disajikan dalam Tabel 11.2.

Tabel 11.2.
Penskalaan Pernyataan dengan Metode EAI
(Isikan frekuensi penilaian pada setiap kategori/interval untuk setiap pernyataan)

Pernyataan	Kategori Penilaian											Nilai Skala (*mean*)	*SD*
	A	B	C	D	E	F	G	H	I	J	K		
	1	2	3	4	5	6	7	8	9	10	11		
1													
2													
Dst.													

f. Pernyataan dengan *SD* besar atau tinggi harus digugurkan sebab hal itu menunjukkan bahwa pernyataan atau item tersebut ambigu alias kabur, terbukti dari besarnya variabilitas penilaian subjek penilai. Dalam pengukuran sebuah objek atau atribut psikologis, idealnya diperoleh antara 20 sampai 30 item atau pernyataan yang terdiri atas kira-kira separuh pernyataan *favorable* dan separuh lainnya berupa pernyataan-pernyataan *unfavorable*.

Dalam pengadministrasian skala baik dalam rangka uji coba maupun penggunaannya sesudah terbukti memiliki ciri psikometrik yang baik, pernyataan disajikan dalam format dikotomis, khususnya "Setuju" atau "Tidak Setuju", sehingga subjek tinggal diminta membububuhkan tanda centang untuk menyatakan kesetujuan atau ketidaksetujuan mereka terhadap setiap pernyataan. Skor subjek pada skala adalah *mean* nilai skala dari item-item atau pernyataan-pernyataan yang disetujui oleh subjek. Penskalaan dengan metode *equal appearing intervals* (EAI) atau interval tampak setara ini menghasilkan pengukuran pada *taraf interval*.

3. Skala Likert

Metode penskalaan yang dikemukakan oleh Rensis Likert (1932, dalam Anderson, 1990b) ini jauh lebih sederhana daripada metode EAI Thurstone. Intinya, terhadap setiap pernyataan atau item dalam rangka mengukur atribut psikologis tertentu subjek diminta

menyatakan kesetujuan-ketidaksetujuannya dalam sebuah kontinum yang terdiri atas lima respon: "Sangat Setuju" (*Strongly Agree*), "Setuju" (*Agree*), "Tidak tahu" (*Undecided*), "Tidak Setuju" (*Disagree*), dan "Sangat Tidak Setuju" (*Strongly Disagree*).

Dalam metode penskalaan Likert, isi pernyataan dibedakan menjadi dua kategori: (1) pernyataan *favorable,* yaitu "statements whose endorsement indicates a positive or favorable attitude toward the object of interest"; maksudnya, pernyataan-pernyataan yang bila disetujui atau diiyakan menunjukkan sikap positif atau menyukai objek yang menjadi sasaran perhatian; dan (2) pernyataan *unfavorable,* yaitu "statements whose endorsement indicates a negative or unfavorable attitude toward the object"; maksudnya, pernyataan-pernyataan yang bila disetujui atau diiyakan mencerminkan sikap negatif atau tidak menyukai objek yang menjadi pusat perhatian (Anderson, 1990b).

Jika isi pernyataan bersifat *favorable,* maka masing-masing respon mulai "Sangat Setuju" sampai dengan "Sangat Tidak Setuju" diberi skor berturut-turut 5, 4, 3, 2, dan 1. Sebaliknya jika isi pernyataan bersifat *unfavorable*, maka masing-masing respon mulai "Sangat Setuju" sampai dengan "Sangat Tidak Setuju" diberi skor 1, 2, 3, 4, dan 5. Skor total subjek adalah jumlah skor setiap pernyataan atau item. Karena jawaban subjek terhadap setiap pernyataan atau item pada dasarnya merupakan *rating* atau penilaian dan penilaian tersebut kemudian dijumlahkan untuk mendapatkan pengukuran tentang sikap subjek terhadap objek psikologis atau tentang taraf kepemilikan subjek atas atribut psikologis tertentu, maka seorang pakar psikometri lain (Bird, 1940, dalam Edwards, 1957) menyebut metode penskalaan Likert ini *method of summated ratings* atau metode penilaian terjumlahkan.

Secara lengkap, ada tujuh langkah yang perlu diikuti dalam menyusun inventori kepribadian dengan skala Likert (Anderson, 1990b):

a. Merumuskan pernyataan secara *favorable* atau *unfavorable.* Setiap pernyataan dituliskan pada sebuah kartu indeks.

b. Meminta kepada sejumlah *judges* atau penilai yang dipilih dari populasi yang akan dikenai skala atau inventori, untuk memeriksa pernyataan-pernyataan dan memilahnya ke dalam tiga kategori: *favorable, unfavorable,* atau *neither* alias netral.
c. Mengeliminasi atau menggugurkan pernyataan-pernyataan yang gagal diklasifikasikan sebagai *favorable* atau *unfavorable* (jadi, diklasifikasikan sebagai netral) oleh sebagian besar penilai.
d. Menuliskan pernyataan-pernyataan yang lolos seleksi atau tidak tereliminasi dengan urutan *random* menjadi bentuk skala atau inventori siap uji coba, serta melengkapinya dengan petunjuk yang berisi:
    1) Permintaan kepada subjek agar menunjukkan perasaannya terhadap setiap pernyataan dengan mencentang salah satu dari lima alternatif jawaban yang tersedia, mulai "Sangat Setuju" sampai dengan "Sangat Tidak Setuju."
    2) Penjelasan tentang tujuan skala atau inventori, khususnya skala atau inventori kepribadian ini mengukur apa.
    3) Penegasan bahwa tidak ada jawaban benar atau salah.
e. Mengadministrasikan versi awal skala (versi uji coba) pada sampel populasi yang menjadi sasaran skala. Agar memperoleh data yang bermakna, sebaiknya besar sampel adalah beberapa kali lebih besar dari jumlah pernyataan.
f. Melakukan seleksi akhir terhadap pernyataan-pernyataan dengan cara:
    1) Menghitung korelasi antara respon subjek pada setiap pernyataan dengan skor total skala (r_{it}).
    2) Pernyataan yang berkorelasi secara positif namun tidak signifikan atau secara negatif dan tidak signifikan, apalagi berkorelasi secara negatif dan signifikan dengan skor total skala dieliminasi atau digugurkan.
    3) Memeriksa reliabilitas himpunan pernyataan yang lolos seleksi sebagai satu skala dengan koefisien alpha Cronbach.

    Syarat bahwa setiap pernyataan harus berkorelasi secara positif dan signifikan dengan skor total tes untuk terpilih dimasukkan

ke dalam bentuk final inventori atau skala disebut "*Kriteria konsistensi internal ala Likert*".

g. Jika bentuk final skala dipandang memuaskan (jumlah pernyataan antara 20 sampai 30 atau lebih dan memiliki koefisien reliabilitas alpha yang tinggi), maka skala siap digunakan.

Dalam perkembangannya secara umum ada dua jenis modifikasi terhadap skala Likert (Anderson, 1990b):

a. Modifikasi pada opsi jawaban, bukan hanya 5 tetapi 2, 3, 4, 6, atau bahkan 7. Alasannya:
   1) Penggunaan jumlah genap opsi jawaban, untuk memaksa subjek memilih antara jawaban *favorable* atau *unfavorable.* Artinya, tidak memberi kesempatan kepada subjek memberikan jawaban netral.
   2) Penggunaan opsi jawaban dalam jumlah banyak (>5), untuk meningkatkan konsistensi internal skala (setara dengan meningkatkan jumlah item pada tes kognitif berskor dikotomis).
   3) Penggunaan opsi jawaban dalam jumlah sedikit (<5), agar lebih sesuai bagi kelompok subjek anak dan/atau kelompok dewasa yang kurang berpendidikan.
b. Modifikasi pada format pernyataan, berupa *incomplete statements* atau melengkapi pernyataan. Contoh:

Saat sekolah diliburkan karena cuaca buruk, saya merasa
(a) sangat senang (b) senang (c) sedih (d) sangat sedih

Beberapa kelebihan skala Likert adalah: (a) mudah penyusunannya, (b) mudah diterapkan pada aneka objek, situasi, atau *setting,* dan (c) mampu mengungkap baik arah (*favorable* versus *unfavorable*) maupun intensitas sikap atau atribut psikologis. Namun salah satu kelemahannya adalah pola jawaban yang berlainan bisa menghasilkan skor total yang sama (Anderson, 1990b).

4. Skala Guttman atau Analisis Skalogram

Skala Guttman dikembangkan oleh Louis Guttman (1944, 1950, dalam Abdi, 2010). Skala ini terdiri atas serangkaian pernyataan, semua menunjukkan sikap seseorang terhadap sebuah objek atau menunjukkan pemilikian seseorang atas atribut psikologis tertentu, dan harus dijawab secara *biner* atau dikotomis ("Ya" atau "Tidak") oleh sekelompok subjek. Tujuan analisis dengan skala Guttman adalah menemukan sebuah dimensi tunggal yang dapat dipakai untuk menentukan posisi baik pernyataan maupun para subjek penjawabnya. Posisi pernyataan dan subjek pada dimensi yang ditemukan selanjutnya bisa dipakai untuk menentukan nilai numerik atau skor mereka (Abdi, 2010). Maka, skala ini memiliki dua ciri: (1) pernyataan-pernyataan mencerminkan perasaan positif yang semakin meningkat terhadap objek sikap atau terkait pemilikan atribut psikologis tertentu; (2) pemilihan (*endorsement*) suatu pernyataan menyiratkan pemilihan (*endorsement*) terhadap setiap pernyataan lain yang memiliki kadar positif yang lebih rendah. Karena sifatnya ini, ada yang menyebut skala Guttman ini *cumulative scales* (Anderson, 1981, dalam Anderson 1990a).

Pengadministrasian dan penskoran skala Guttman secara garis besar adalah sebagai berikut: (a) subjek diminta menyatakan setuju (*endorse*) atau tidak setuju (*do not endorse*) setiap pernyataan; (b) skor subjek adalah jumlah pernyataan yang disetujui atau dipilihnya (*endorsed*).

Dalam praktek, jarang memperoleh data yang cocok dengan model penskalaan Guttman secara sempurna. Salah satu cara paling sederhana untuk mengatasi masalah ini adalah penerapan *metode Goodenough-Edwards* (Abdi, 2010). Metode ini menyatakan bahwa penyimpangan dari skala ideal tersebut merupakan *random errors* atau kesalahan random. Maka, tujuan penerapan metode ini adalah memulihkan atau membersihkan skala Guttman dari data yang cemar. Secara garis besar ada tiga langkah dalam menetapkan apakah

sebuah skala merupakan skala Guttman dengan menerapkan metode Goodenough-Edwards (Anderson, 1990a; Abdi, 2010):

a. **Langkah 1. Memeriksa jumlah "pola jawaban yang tidak sesuai."** Misal, sebuah skala terdiri atas 5 pernyataan, diurutkan dari *least positive* sampai *most positive.* Skala ini diadministrasikan pada sekelompok subjek berjumlah *n.* Pola jawaban yang sesuai adalah **1 1 1 0 0**; sedangkan pola jawaban yang tidak sesuai adalah **1 1 0 1 0**. "Any response pattern in which a **'1'** appears to the right of a '**0**' is an inappropriate response pattern" (Anderson, 1990). Selanjutnya, pada setiap pola jawaban yang tidak sesuai, sebuah kesalahan dihitung setiap kali ada angka '**1**' muncul di sebelah kanan angka '**0**'.
b. **Langkah 2. Menghitung jumlah kesalahan pada semua pola jawaban dari seluruh sampel responden atau subjek.**
    1) Jumlah total jawaban = jumlah pernyataan dalam skala x jumlah responden.
    2) Persentase kesalahan = jumlah total kesalahan / jumlah total jawaban.
    3) *Coefficient of Reproducibility (*CR) = 100 – Persentase kesalahan. Syarat skala Guttman adalah: CR ≥ 90.
c. **Langkah 3. Menghitung *Coefficient of Scalability* (CS).** Caranya:
    1) Mengurangkan *minimum marginal reproducibility* (MMR) dari *coefficient of reproducibility* (CR). Hasilnya disebut *percent improvement* (PI), yaitu selisih antara CR aktual dan CR minimal. MMR sendiri adalah *a chance score,* yaitu persentase *appropriate response patterns* yang terjadi *by chance.*
    2) Mengurangkan MMR dari 100%. Hasilnya disebut *possible percent improvement* (PPI), yaitu selisih antara CR maksimal dan CR minimal.
    3) Membagi PI dengan PPI. Hasilnya adalah *coefficient of scalability* (CS). CS menunjukkan sejauh mana CR secara substansial melampaui angka yang bisa diharapkan *by chance.* Menurut Guttman, besarnya CS harus >60 (Anderson, 1990a).

Berdasarkan uraian di atas dapat ditarik beberapa kesimpulan. *Pertama,* kriteria CR ≥ 90 dan CS >60 merupakan dasar untuk menentukan keabsahan skala Guttman. Jika kriteria tersebut tidak terpenuhi, maka sebuah skala Guttman gagal disusun. Berarti, "a scale either is or is not a Guttman scale"; maksudnya, sebuah skala merupakan skala Guttman atau bukan skala Guttman sama sekali. *Kedua,* skala Guttman sulit disusun, namun jika berhasil ada minimal dua kelebihan, yaitu: (a) "it is possible to determine the entire pattern of responses to the statements from a single total score"; maksudnya, kita bisa menentukan keseluruhan pola jawaban testi terhadap pernyataan-pernyataan dalam tes hanya berdasarkan sebuah skor total tunggal; (b) "the cumulative nature of Guttman scales makes it feasible to assess change in attitude"; maksudnya, sifat kumulatif skala Guttman memungkinkan kita mengukur perubahan sikap atau atribut psikologis lain yang menjadi objek pengukuran (Anderson, 1990a).

5. Metode *Empirical Keying* atau Penskalaan Empiris

Tidak seperti empat metode sebelumnya, metode ini tidak mengandalkan teori atau penilaian ahli melainkan mendasarkan pada proses empiris. Intinya, pemilihan pernyataan-pernyataan untuk menjadi item-item skala didasarkan pada sejauh mana *kelompok kriteria* atau *kelompok tipikal* memberikan jawaban yang berbeda secara signifikan terhadap setiap pernyataan dibandingkan jawaban yang diberikan oleh *sampel normatif* atau kelompok subjek dari populasi umum yang dipakai dalam rangka uji coba dan penyusunan norma. Langkah-langkah penerapan penskalaan empiris dalam pengukuran sebuah atribut psikologis akan mencakup sebagai berikut (Gregory, 2007):

a. Memilih suatu kelompok subjek yang secara homogen dipandang memiliki atribut psikologis yang sedang menjadi objek pengukuran dalam jumlah atau kadar yang tinggi melebihi

kelompok subjek pada umumnya. Kelompok semacam ini bisa disebut *kelompok tipikal* atau *kelompok kriteria,* yaitu kelompok yang secara tipikal mewakili atribut psikologis yang sedang menjadi objek pengukuran.

b. Mengadministrasikan pernyataan-pernyataan dalam format "benar-salah" pada kelompok tipikal dan kelompok subjek umum dengan karakteristik setara seperti kelompok tipikal, kecuali dalam hal atribut yang sedang diukur. Kelompok subjek umum ini disebut *sampel normatif.*
c. Membandingkan frekuensi jawaban yang mengiyakan (*endorsement*) di kalangan kelompok tipikal dan frekuensi jawaban serupa di kalangan kelompok normatif pada setiap pernyataan.
d. Pernyataan-pernyataan yang memiliki perbedaan signifikan atau mencolok dalam hal frekuensi pengiyaan (*endorsement*) antara kelompok tipikal dan kelompok normatif, dalam arti frekuensi *endorsement* di kalangan kelompok tipikal secara signifikan lebih besar atau tinggi dibandingkan frekuensi *endorsement* di kalangan kelompok normatif, dipilih untuk dijadikan item-item skala dengan penafsiran isi searah dengan *endorsement* kelompok tipikal.
e. Skor kasar skala untuk atribut yang sedang menjadi objek pengukuran adalah jumlah item atau pernyataan yang diiyakan atau dibenarkan atau dijawab sebagai benar.

C. Penulisan Item

Sebelum mulai menyusun item, ada beberapa hal yang perlu dikemukakan sebelumnya, yaitu tentang: (1) beberapa masalah yang mengancam validitas tes, (2) aneka format item, dan (3) aneka petunjuk penulisan item.

1. Beberapa Masalah yang Mengancam Validitas Tes

Dalam penyusunan *typical performance tests* pada umumnya dan inventori kepribadian pada khususnya, ada paling sedikit enam masalah yang perlu dicermati bahkan diatasi dalam menyusun item, agar tes atau inventori kepribadian yang kita susun terjamin validitasnya (Kline, 1986). Keenam masalah yang dimaksud bisa dikelompokkan ke dalam dua kategori, yaitu (1) masalah yang bersumber pada *response sets,* dan (2) masalah yang terkait proses validasi tes. Dua kategori masalah beserta upaya yang bisa ditempuh untuk mengurangi atau bahkan menghilangkan dampak negatifnya adalah seperti dipaparkan di bawah ini.

a. Masalah yang Bersumber pada *Response Sets*

Yang dimaksud *response sets* adalah "stylistic consistencies, stimulated by the form of response of personality inventory items" (Cronbach, 1946, dalam Kline, 1986). Jadi, *response sets* adalah keajegan gaya (dalam menjawab) yang dipicu oleh bentuk respon terhadap item-item inventori kepribadian. Secara lebih sederhana, *response sets* adalah kecenderungan menjawab item-item inventori kepribadian dengan cara tertentu secara ajeg tanpa memedulikan isi item atau pertanyaannya. Ada empat jenis *response sets* penting yang perlu diwaspadai dalam menyusun item inventori kepribadian:

1) ***Response set of acquiescence* atau kecenderungan mengiyakan item.** Yang dimaksud adalah kecenderungan subjek untuk menyatakan setuju dengan setiap item tanpa memedulikan isinya. Cronbach (1946, dalam Klein, 1986) menyebut gejala ini *stylistic consistencies* atau konsistensi gaya menjawab yang lazim dipicu oleh format item inventori kepribadian. Menurut Guilford (1959, dalam Kline, 1986), kecenderungan ini akan makin kuat manakala item-item bersifat kabur atau mendua makna.

Ada dua cara yang lazim disarankan untuk mengatasi masalah ini: (1) cara yang disarankan oleh Messick (1962, dalam Kline, 1986), yaitu penerapan *balanced scales* atau skala berimbang, yaitu mengusahakan agar terdapat jumlah item yang sama untuk dijawab "Ya" dan "Tidak", atau "Benar" dan "Salah"; (2) cara yang disarankan oleh Guilford (1959), yaitu menuliskan item secara "clear, unambiguous, and referred to specific behavior" atau menuliskan item secara jelas, tidak kabur atau mendua makna, dan mengacu pada tingkah laku spesifik. Namun menurut Kline (1986), kecenderungan mengiyakan item semacam ini sering tetap muncul kendati sudah menerapkan skala berimbang, sedangkan di pihak lain sering tidak mudah merumuskan item secara benar-benar jelas. Intinya, perlu cermat dan penuh pertimbangan dalam merumuskan item.

2) ***Response set of social desirability* atau kecenderungan memberikan jawaban mengikuti selera masyarakat**. Ini adalah kecenderungan seorang subjek untuk memberikan jawaban terhadap item-item inventori kepribadian mengikuti apa yang dia pikir dipandang baik oleh masyarakat. Semakin suatu jawaban dipandang sesuai dengan apa yang diterima atau dipandang baik oleh masyarakat, semakin jawaban tersebut akan dipilih oleh subjek dalam menjawab sebuah item.

Ada empat cara yang lazim disarankan untuk mengatasi masalah ini: (1) cara yang disarankan oleh Edwards (1959), yaitu penggunaan *forced-choice items of matched social desirability;* di sini dengan sengaja penyusun tes menggunakan format item berupa sepasang pernyataan yang sama-sama mengandung isi yang sejalan dengan norma sosial, kemudian subjek diminta memilih salah satu di antaranya; sebagaimana kita ketahui, Edwards menempuh cara ini dalam menyusun *Edwards' Personal Preference Schedule* atau *EPPS* (1959, dalam Kline, 1986) yang terkenal itu; yang perlu diperhatikan, kesejalanan dengan norma sosial masing-masing pernyataan dalam setiap pasang harus benar-benar setara, sebab bahkan perbedaan kecil akan menjadi besar dan mencolok

manakala kedua pernyataan tersebut disajikan bersama-sama sebagai pasangan (Kline, 1986); (2) cara yang disarankan oleh Kline (1986) sendiri, yaitu menghindari perumusan item dengan isi yang jelas-jelas sejalan (*desirable*) atau sebaliknya bertentangan (*undesirable*) dengan norma sosial; (3) cara yang disarankan oleh Eysenck (1976, dalam Kline, 1986), yaitu penggunaan sebuah *lie scale* atau skala kebohongan; konkretnya dengan menyisipkan sejumlah item khusus yang bertujuan mengungkap ada tidaknya kecenderungan *social desirability* dalam jawaban subjek; Eysenck menggunakan item-item yang berisi *peccadillos* atau dosa kecil yang sekali tempo dilakukan oleh banyak orang; subjek yang memperoleh skor tinggi pada skala kebohongan ini dipandang menerapkan kecenderungan *social desirability* dalam menjawab, maka hasil tesnya digugurkan; (4) menerapkan *analisis butir* dan *validasi tes* secara cermat; ketiga cara sebelumnya harus diterapkan pada saat penyusunan item, cara keempat ini bisa dikatakan penyempurna ketiga cara sebelumnya dan baru bisa ditempuh sesudah item-item diuji-cobakan; seperti ditegaskan oleh Kline (1986), sebuah item yang sejalan dengan norma sosial pastilah item yang bias sehingga jawaban subjek akan terdistribusi secara non-random, maka melalui *analisis butir* item-item semacam ini disarankan untuk digugurkan; selanjutnya, jika melalui proses *validasi* diperoleh evidensi atau bukti bahwa sebuah tes benar-benar valid maka kita tidak perlu lagi khawatir tentang kemungkinan beroperasinya kecenderungan *social desirability;* salah satu alat yang bisa dipakai sebagai kriteria *social desirability* adalah *Crowne-Marlowe Social Desirability Scale* (Crowne dan Marlowe, 1964, dalam Kline, 1986); jika sebuah tes tidak berkorelasi secara signifikan dengan CM-SDS, berarti tes tersebut valid dalam arti bebas dari bias *social desirability* (Kline, 1986).

3) ***Response set of using uncertain or middle category* atau kecenderungan memilih jawaban tidak tentu atau kategori tengah**. Jika tersedia kesempatan untuk memilih jawaban kategori tengah yang mencerminkan ketidakpastian atau netralitas, banyak subjek cenderung memilihnya demi mencari aman. Kecenderungan ini berdampak menurunkan validitas item tes sebab kebanyakan metode analisis item didasarkan pada jawaban ekstrem (Kline, 1986).

Menurut Kline (1986), cara terbaik mengatasi masalah ini adalah menggunakan format item dikotomis ("Ya-Tidak" atau "Benar-Salah"). Kelemahannya, kadang-kadang memang ada pernyataan yang sungguh-sungguh tidak bisa dijawab dengan "Ya atau Tidak", "Benar atau Salah", melainkan di antara keduanya.

4) ***Response set of using the extreme response* atau kecenderungan memilih jawaban ekstrem**. Kecenderungan semacam ini lazim muncul dalam inventori yang menggunakan *multi-point rating scale* atau skala penilaian bernilai jamak. Sejumlah subjek cenderung menyukai jawaban ekstrem tanpa memedulikan isi item atau pernyataannya (Vernan, 1964, dalam Kline, 1986).

Satu-satunya cara yang disarankan oleh Kline (1986) untuk mengatasi masalah ini adalah dengan menghindari atau tidak menerapkannya, menggantinya dengan format item yang bersifat dikotomis. Namun jika seorang penyusun tes tetap memilih untuk menerapkannya, satu-satunya cara lain untuk mengatasinya adalah melakukan analisis butir dan uji validasi secara cermat untuk mengeliminasi item-item yang paling rentan terhadap kecenderungan memberikan jawaban ekstrem semacam itu (Kline, 1986).

b. Beberapa Masalah Terkait Validasi Tes

Menurut Kline (1986) ada empat jenis masalah terkait proses validasi tes yang perlu diwaspadai, yaitu (a) masalah terkait *face*

validity item-item; (b) masalah terkait *sampling from the universe of items* atau pengambilan sampel dari populasi item; (c) masalah terkait *sampling from the universe of subjects* atau pengambilan sampel dari populasi subjek; dan (d) masalah terkait penetapan kriteria yang memadai untuk menguji validitas. Marilah kita bahas secara singkat keempat masalah tersebut satu demi satu.

1) **Masalah terkait *face validity of items* atau validitas muka item-item**. Dalam inventori kepribadian, jawaban subjek terhadap sebuah item harus dipandang atau diterima sebagai jawaban yang sebenarnya. Namun Cattell dan Kline (1977, dalam Kline, 1986) menunjukkan bahwa respon subjek dalam kuesioner termasuk inventori kepribadian sesungguhnya mencakup dua aspek, yaitu aspek yang bisa dipandang mencerminkan respon subjek yang sesungguhnya terhadap item tes (Q), dan aspek yang mungkin memuat atau tidak memuat faktor tertentu, tanpa ada kaitannya dengan respon sesungguhnya subjek terhadap item tes (Q^1). Beroperasinya aspek kedua dalam respon subjek terhadap item-item sebagai data inventori tentu saja mengancam validitas tes;
2) **Masalah terkait *sampling from the universe of items* atau pengambilan sampel dari populasi item.** Sudah berulangkali disinggung, hampir seluruh atribut psikologis yang menjadi objek pengukuran dengan inventori kepribadian berupa konstruk teoretis dalam arti sempit sebagaimana dimaksudkan oleh Cronbach dan Meehl (1955). Dari sudut pandang teori tes klasik, konstruk teoretis semacam ini memiliki *behavioral content domain* atau populasi item dengan batas-batas yang tidak jelas. Akibatnya, pengambilan sampel item-itemnya untuk dijadikan bentuk final tes atau inventorinya pun menjadi problematis dalam artian tidak mudah. Kita tahu, pemilihan sampel item yang baik merupakan kunci untuk menjamin validitas hasil pengukuran yang akan tercermin dari diperolehnya skor murni yang benar-benar bisa dipercaya (Kline, 1986).

3) **Masalah terkait *sampling from the universe of subjects* atau pengambilan sampel dari populasi subjek.** Penyusunan inventori kepribadian yang ditujukan bagi kelompok subjek normal sesungguhnya menuntut penggunaan sampel subjek yang mewakili seluruh kemungkinan skor yang bisa dicapai dalam inventori yang bersangkutan. Hal ini berarti perlu digunakannya sampel subjek dengan jumlah yang jauh lebih besar dengan mempertimbangkan keterwakilan berbagai strata populasi subjek yang ada pula (Kline, 1986).
4) **Masalah terkait *establishing adequate criteria for validity* atau menemukan kriteria yang memadai untuk menguji validitas**. Dalam mengukur aneka atribut kepribadian, kita lazim mengandalkan *ratings* atau penilaian sebab tidak tersedia ukuran eksternal lain seperti dalam pengukuran abilitas. Maka lazimnya kita mengandalkan pemeriksaan validitas konstruk melalui inferensi berdasarkan analisis multivariat antara tes yang kita susun dengan tes yang mengukur aneka variabel atau atribut lain atau berdasarkan pengujian hubungan dengan kelompok tipikal yang diasumsikan memiliki skor tertentu dalam hal variabel atau atribut yang sedang kita susun alat ukurnya (Kline, 1986). Namun kita tahu, seringkali tidak mudah menemukan variabel lain atau kelompok tipikal semacam ini yang sungguh-sungguh dapat kita andalkan sebagai kriteria eksternal.

2. Aneka Format Item Inventori Kepribadian

Dalam penyusunan inventori kepribadian, format item perlu dipilih secara cermat demi mendapatkan item-item yang mampu mengungkap jawaban subjek secara *genuine* atau jujur dan objektif sehingga hasil pengukurannya pun menjadi valid sehingga juga reliabel. Perlu dikemukakan, item inventori kepribadian lazimnya berupa *self-report items* atau item pelaporan-diri. Di situ subjek dituntut melaporkan atau mengungkapkan keadaan dirinya, dan

harus memberikan jawaban sejujur mungkin (Kline, 1986). Pemilihan format item bertujuan memperoleh jawaban jujur apa adanya dari subjek. Menurut Kline (1986), aneka format item yang lazim diterapkan dalam penyusunan inventori kepribadian secara garis besar bisa dikelompokkan ke dalam tiga kategori, yaitu: (1) item dikotomis, (2) item trikotomis, (3) item dengan skala penilaian, dan (4) format lain.

a. Item Dikotomis

Dalam format item ini subjek diminta menjawab pertanyaan atau pernyataan dengan memilih salah satu dari dua jawaban yang disediakan. Ragamnya adalah sebagai berikut (Kline, 1986):

1) **"*Yes-No item*" atau "*Item Ya-Tidak*"**. Item berupa pertanyaan yang harus dijawab oleh subjek dengan jawaban "Ya" atau "Tidak." Format item ini dipandang mudah dan cepat disusun, serta mudah dipahami oleh subjek.
2) **"*True-False item*" atau "*Item Benar-Salah*"**. Item berupa pernyataan (sering menggunakan kata ganti orang pertama), dan subjek diminta menyatakan apakah pernyataan tersebut "Benar" atau "Salah" dalam mencerminkan keadaan dirinya.
3) **"*Like-Dislike item*" atau "*Item Suka-Tidak suka*"**. Item berupa sebuah kata atau frase yang melukiskan benda, keadaan, atau hal lain terkait atribut yang sedang diukur, dan subjek diminta menyatakan perasaannya terhadap benda, keadaan, atau hal lain tersebut: "Suka" atau "Tidak suka."
4) **"*Forced-choice items*" atau "*Item Pilihan Wajib*"**. Item berupa dua pernyataan dan subjek diminta memilih salah satu yang dia pandang sesuai dengan keadaan dirinya.

b. Item Trikotomis

Dalam format item ini subjek diminta menjawab pertanyaan atau pernyataan dengan memilih salah satu dari tiga jawaban yang disediakan. Ragamnya adalah sebagai berikut (Kline, 1986):

1) **"*Yes ? No item*" atau "*Item Ya ? Tidak*"**. Format ini merupakan varian dari item "Ya-Tidak". Bedanya, dalam format

ini dilengkapi dengan opsi jawaban ragu-ragu "?". Ada subjek yang merasa tidak nyaman dipaksa hanya memilih antara "Ya" dan "Tidak" saat sebenarnya mereka tidak merasa sepasti itu. Masalahnya dan seperti sudah disinggung, jawaban tengah seperti ini cenderung menggoda orang untuk memilihnya padahal jarang jawaban semacam itu sungguh-sungguh bermakna. Di pihak lain ada bukti bahwa format item dikotomis dan trikotomis tidak menghasilkan perbedaan hasil yang signifikan, maka lazimnya lebih disarankan menerapkan format item dikotomis karena lebih menjamin subjek yang peragu untuk memilih memberikan jawaban yang jelas atau pasti (Kline, 1986).

2) **Aneka item trikotomis**. Format item ini menyajikan opsi jawaban yang lebih bervariasi daripada sekadar "Ya-?-Tidak", seperti misalnya: "Lazimnya, kadang-kadang, tidak pernah"; "Benar, ragu-ragu, salah"; "Setuju, ragu-ragu, tidak setuju", dan sebagainya.

3) **Item trikotomis dengan pilihan**. Format item ini berupa pernyataan tak lengkap yang harus dilengkapi dengan memilih salah satu dari tiga frase yang disediakan sebagai jawaban. Contoh:

   Jika saya menang undian sebesar satu milyar rupiah maka akan saya pakai untuk:

   (a) mendirikan rumah ibadat.

   (b) mendirikan perpustakaan.

   (c) mendirikan pusat kuliner.

c. *Items with Rating Scales* atau Item dengan Skala Penilaian

Format item ini berupa pernyataan atau kalimat yang dilengkapi dengan skala penilaian. Ada banyak ragam skala penilaiannya, baik terkait *jumlah* maupun *isi* skalanya. Contoh-contohnya adalah sebagai berikut:

1) Skala lima butir berisi kesetujuan: "Sangat Tidak Setuju, Setuju, Ragu-ragu, Tidak Setuju, Sangat Tidak Setuju".

2) Skala tujuh butir berisi frekuensi: "Selalu, Sangat Sering, Sering, Kadang-kadang, Jarang, Sangat Jarang, Tidak Pernah".

Format item ini memiliki kelebihan terasa lebih wajar dibandingkan misalnya format item dikotomis. Namun sekaligus memiliki beberapa kelemahan, yaitu lebih rentan terhadap ancaman *response sets* dan penafsiran subjek terhadap isi skala bisa berlainan. Kendati begitu, format item ini cukup lazim diterapkan (Kline, 1986).

d. Format Lain

Salah satu format di luar format-format berbentuk objektif seperti sudah dipaparkan di atas adalah format berbentuk proyektif, kendati tetap diskor dan diolah secara objektif. Salah satu contoh format item ini berupa pasangan kata dan subjek diminta menyatakan mana yang lebih disukainya. Ada yang menyebutnya sebagai versi lain dari format item *like-dislike* (Kline, 1986).

3. Aneka Petunjuk Penulisan Item

Sesudah mengetahui aneka format item dan berbagai kelebihan, kekurangan, dan kemungkinan ancaman masing-masing, berikut ini disajikan sejumlah petunjuk dalam menuliskan item inventori kepribadian demi mendapatkan item-item yang benar-benar efektif berdasarkan keyakinan bahwa "sebuah tes tidak mungkin lebih baik (namun bisa lebih buruk) dibandingkan item-itemnya" (Kline, 1986):

a. *Usahakanlah sejauh mungkin agar subjek tidak menangkap maksud setiap item***.** Alasannya demikian: jika subjek mempersepsikan atau menangkap bahwa sebuah item bermaksud mengukur sifat X, maka jawabannya akan mencerminkan pandangan pribadinya tentang keadaannya terkait sifat yang sedang diukur itu, dan bukan mencerminkan keadaannya yang nyata. Masalahnya, banyak subjek memiliki pandangan yang keliru tentang kepribadian mereka. Untuk mengatasi bias ini, kiranya baik memperhatikan saran Guilford (1959, dalam Kline, 1986) berikut

ini dalam menyusun item: "the ideal is to score a subject on traits which he does not know, by asking questions about what he does know."

b. *Tuliskan item secara jelas, tidak mendua***.** Langkah ini penting untuk mengurangi kesalahan akibat subjek gagal memahami item.
c. *Item perlu ditulis mengacu pada tingkah laku yang spesifik, bukan tingkah laku yang umum atau terlalu luas***.** Istilah olah raga memiliki makna atau cakupan yang lebih luas dibandingkan, misalnya sepak bola.
d. *Setiap item harus hanya berisi satu pertanyaan atau satu pernyataan***.**
e. *Sejauh mungkin hindari penggunaan istilah-istilah yang bermakna frekuensi.* Makna sering, kadang-kadang dan sejenisnya akan berlainan bagi setiap orang.
f. *Sejauh mungkin item harus mengacu ke tingkah laku, bukan perasaan.*
g. *Melalui petunjuk yang jelas, yakinkan subjek untuk menjawab setiap item secara cepat, tidak perlu menimbang-nimbang***.**

Kini, sebelum benar-benar mulai menuliskan item-item untuk inventori kepribadian kita, buatlah *blue-print*-nya terlebih dulu berdasarkan eksplikasi konstruk yang sudah kita buat sebelumnya. *Blue-print* ini penting untuk memastikan struktur alat ukur yang kita susun. Sesudah puas dengan *blue-print* kita, kita bisa mulai menuliskan item satu demi satu. Setiap item yang kita susun kita tuliskan pada sebuah kartu indeks dilengkapi dengan kode-kode yang jelas untuk mengidentifikasikan isi dan arahnya sesuai *blue-print.*

Sesudah mempertimbangkan dengan cermat hal-hal di atas, penulisan item segera bisa dimulai dengan memperhatikan beberapa saran praktis berikut ini:

a. Cermatilah kembali *test blue print* Anda sampai Anda merasa benar-benar puas, dalam arti bahwa struktur *test blue print* tersebut secara konseptual sungguh-sungguh terderivasikan atau terturunkan dari definisi operasional dan definisi konseptual konstruk yang sedang menjadi objek pengukuran.

b. Tandailah setiap bagian yang merupakan irisan dari berbagai komponen pembagian yang Anda pakai dengan sebuah bilangan.
c. Tulislah setiap item atau pernyataan di atas sebuah kartu indeks dengan menggunakan pensil. Penggunaan kartu indeks disarankan, agar memudahkan proses selanjutnya khususnya dalam penyusunan item-item menjadi skala dengan urutan random. Penggunaan pensil disarankan agar bila terjadi kesalahan atau ingin mengubah rumusan, bisa dihapus dengan bersih, sampai diperoleh rumusan final yang dipandang memuaskan.
d. Tandailah setiap item dengan dua bilangan: bilangan pertama menunjukkan posisi item yang bersangkutan dalam struktur *blue print,* sedangkan bilangan kedua menunjukkan nomor urut item yang bersangkutan dalam kelompok item sesuai struktur *blue print* sebagaimana ditunjukkan oleh bilangan pertama.
e. Untuk setiap kelompok item yang dimaksudkan mengukur bagian tertentu dalam struktur *blue print,* tulislah dalam jumlah sekitar dua kali lipat dari jumlah yang Anda harapkan dalam bentuk final skala Anda. Dengan kata lain, susunlah sebuah *item pool* dengan jumlah item dua kali lipat dari jumlah item bentuk final skala yang Anda rencanakan. Menurut Kline (1986), untuk mendapatkan sebuah skala kepribadian yang homogen dan reliabel dibutuhkan antara 20-30 item.
f. Lengkapilah skala Anda dengan petunjuk yang lengkap dan jelas. Hal-hal yang perlu dicantumkan dalam petunjuk meliputi antara lain:
    1) Tujuan skala: mengukur apa, dijelaskan secara umum untuk memotivasi.
    2) Cara menjawab dan di mana: sebaiknya sediakan Lembar Jawab. Bila perlu berilah satu contoh cara menjawab.
    3) Tekankan: Tidak ada jawaban yang salah dan pentingnya menjawab secara spontan, jangan memikirkan jawaban terlalu mendalam, tuliskan apa yang pertama kali melintas dalam benak subjek.

4) Waktu mengerjakan, jika sudah dapat diperkirakan.
5) Tekankan pentingnya menjawab seluruh item, jangan ada yang dilewati.

D. Uji Coba Skala & Analisis Item

Mengikuti pendekatan *rational scale construction* atau konsistensi internal dalam penyusunan tes pada umumnya maupun penyusunan inventori kepribadian pada khususnya, sesudah item-item berhasil ditulis dan disusun menjadi sebuah skala atau inventori kepribadian langkah selanjutnya adalah mengujicobakan skala untuk keperluan melakukan analisis item (Gregory, 2007; Kline, 1986). Sama seperti dalam penyusunan *maximal performance tests* dan sebagaimana dinyatakan oleh Kline (1986), dalam penyusunan *typical performance tests* dan khususnya inventori kepribadian tujuan analisis item adalah memilih item-item yang akan membentuk sebuah skala yang homogen dan berdaya diskriminasi tinggi, dalam arti mampu membedakan secara signifikan antara subjek yang memiliki atribut yang diukur dalam kadar yang rendah dan subjek yang memiliki atribut yang diukur dalam kadar yang tinggi. Inti metode *rational scaling* atau metode *konsistensi internal* ini adalah mengupayakan agar seluruh item skala berkorelasi positif satu sama lain dan juga berkorelasi positif dengan skor total skala (Gregory, 2007).

Secara umum dua parameter yang lazim diperiksa dalam analisis item untuk dijadikan dasar dalam seleksi item adalah p_i yaitu proporsi subjek atau testi yang memilih kunci jawaban dalam menjawab item, dan r_{it} yaitu korelasi antara skor item dan skor total tes sebagai kriteria internal. Dalam penyusunan *maximal performance tests* khususnya berupa aneka jenis tes abilitas, p_i disebut *taraf kesukaran* item dan menjadi salah satu dasar penting dalam melakukan seleksi item. Dalam *typical performance tests* khususnya inventori kepribadian, p_i sekadar menunjukkan popularitas jawaban yang mengacu pada atribut yang sedang diukur baik pada item yang dirumuskan secara

favorable maupun secara *unfavorable* terhadap atributnya. Dalam artian atau situasi semacam itu, sepanjang $0 < p_i < 1$ untuk suatu item berarti item tersebut berfungsi baik dalam arti memiliki daya diskriminasi. Tentang berapa persis besarnya p_i untuk setiap item, lazim kurang begitu dihiraukan. Dengan kata lain, dalam penyusunan inventori kepribadian p_i lazim tidak terlalu dipertimbangkan sebagai dasar dalam melakukan seleksi item. Dengan kata lain pula, parameter yang dipakai sebagai indeks daya diskriminasi item dan yang dipakai sebagai satu-satunya dasar pertimbangkan dalam melakukan seleksi item dalam penyusunan inventori kepribadian adalah r_{it} atau koefisien korelasi antara skor item dan skor total tes.

Maka, secara umum langkah-langkah analisis item dalam penyusunan inventori kepribadian lebih sederhana atau ringkas dibandingkan dalam penyusunan tes abilitas. *Pertama,* hitunglah korelasi antara skor masing-masing item dengan skor total skala, dengan tehnik sebagai berikut: (a) *Pearson product moment correlation,* jika item-itemnya berformat *multi-point*; hanya, konon penerapan tehnik ini sebenarnya meragukan jika itemnya menggunakan *five-point scale* atau skala lima poin (Kline, 1986); dan (b) *point-biserial correlation,* jika item-itemnya berformat dikotomis. Makin tinggi korelasi antara skor item dan skor total skala, makin baik item yang bersangkutan. Item-item yang berkorelasi negatif atau berkorelasi positif namun rendah dengan skor total disingkirkan. Sebagai patokan, semua item yang berkorelasi ≥ 0,20 dengan skor total layak dipertahankan. Dengan kriteria semacam itu, usahakan diperoleh antara 20-30 item dengan struktur atau sebaran yang baik sesuai *blue-print* untuk bentuk final skala.

Kedua, sesudah diperoleh bentuk final skala yang memuaskan, yaitu terdiri dari sekitar 20-30 item dengan $r_{it} \geq 0.20$, hitunglah koefisien reliabilitas konsistensi internalnya dengan tehnik KR-20 atau alpha Cronbach. Koefisien reliabilitas yang dipandang memuaskan adalah ≥ 0,70 (Kline, 1986). Jika semua ini tercapai, berarti kita telah memperoleh sebuah skala pengukur atribut psikologis tertentu yang homogen, reliabel, dan diharapkan juga valid. Jika koefisien

reliabilitasnya belum memuaskan, cobalah tambahkan item-item baru diambilkan dari item-item yang memiliki korelasi dengan skor total kedua terbaik dari *item pool.* Hitunglah kembali koefisien reliabilitas konsistensi internalnya, sampai penambahan item baru tidak lagi berdampak meningkatkan koefisien reliabilitas (Kline, 1986).

Ketiga, sesudah diperoleh bentuk final skala yang sungguh-sungguh memuaskan, dalam arti memiliki koefisien reliabilitas konsistensi internal ≥ 0,70 dan memiliki struktur isi yang baik sehingga bisa dipandang valid, cobalah memeriksa distribusi skor skala tersebut. Skala yang baik dalam arti reliabel dan valid lazimnya menghasilkan skor dengan distribusi yang simetris atau normal (Kline, 1986).

*Keempat, u*ntuk memastikan daya diskriminasi skala, hitunglah koefisien delta Ferguson dengan rumus seperti yang sudah kita bahas di bagian yang menguraikan langkah-langkah analisis item dalam penyusunan tes abilitas. Sekadar mengingatkan, skala yang berdaya diskriminasi baik lazimnya memiliki koefisien delta Ferguson ≥ 0,90 (Kline, 1986).

E. Merevisi Item

Lazimnya, hasil uji coba dan analisis item pertama belum akan menghasilkan item-item yang memuaskan sehingga menghasilkan skala yang juga memuaskan. Idealnya, langkah-langkah yang harus ditempuh sesudah memperoleh hasil uji coba dan analisis item pertama adalah sebagai berikut:

1. Masukkanlah kembali data parameter item (r_{it}) beserta identitas itemnya ke dalam tabel spesifikasi.
2. Dengan memperhatikan item-item yang memiliki parameter item yang memenuhi syarat, periksalah apakah struktur inventori kepribadian yang Anda peroleh sudah baik, dalam arti:

a. Semua komponen atribut psikologis yang menjadi objek atau sasaran pengukuran terepresentasikan atau terwakili dengan memadai, seperti Anda rencanakan.
b. Jumlah item pada masing-masing komponen kurang lebih seimbang.
c. Jumlah item *Favorable* dan *Unfavorable* pada masing-masing komponen kurang lebih seimbang.
d. Jumlah total item sebagai kesatuan skala mencapai minimal 20-30 item.

3. Jika semua syarat di atas terpenuhi, pekerjaan kita mengonstruksi sebuah skala pada dasarnya sudah selesai. Selanjutnya kita periksa reliabilitas dan koefisien delta Ferguson dari keseluruhan item sebagai skala seperti sudah disinggung.
4. Jika jumlah item kurang memadai dan struktur skala juga kurang baik, maka harus dilakukan revisi terhadap item-item yang kita perlukan sesuai jumlah item dan struktur skala ideal seperti yang kita harapkan.
5. Idealnya, skala baru yang mencakup item-item terevisi tersebut harus kita uji cobakan kembali untuk kemudian kita analisis item kembali. Langkah revisi-uji coba kembali-analisis item kembali ini harus kita ulang secukupnya sampai kita peroleh bentuk final skala yang sungguh-sungguh memuaskan, yaitu memiliki jumlah item memadai dengan struktur skala yang memenuhi harapan pula. Ψ

Bab 12
Penggunaan Hasil Tes

Hasil tes berupa *raw scores* atau skor kasar, atau dalam bahasa teori tes klasik adalah *observed score* atau skor tampak. Skor kasar atau skor tampak merupakan taraf informasi paling dasar yang dihasilkan oleh sebuah tes psikologis. Skor kasar lazim merupakan *tally* atau jumlah, yaitu: (a) jumlah pertanyaan atau soal yang dijawab secara benar atau berhasil pada *maximal performance tests*; atau (b) jumlah pertanyaan atau pernyataan yang dijawab sesuai *kunci* yang mengarah pada penafsiran tertentu pada *typical performance tests*.

Bisa dikatakan, sebagai informasi paling dasar skor kasar atau skor tampak praktis belum bermakna apa pun. Pada dirinya skor kasar atau skor tampak belum dapat ditafsirkan, dalam arti dua hal (Allen & Yen, 1979). Pertama, skor kasar yang dicapai oleh seorang testi dalam sebuah tes belum memberikan gambaran tentang bagaimana kinerja testi tersebut dibandingkan dengan kinerja testi lain. Kedua, skor kasar yang dicapai oleh seorang testi dalam sebuah tes tidak bisa dibandingkan dengan skor kasar yang dicapai oleh testi yang sama dalam satu atau lebih tes lain.

Namun cara memaknai skor kasar semacam itu sangat bergantung pada keperluan untuk apa skor kasar sebagai hasil tes tersebut akan digunakan. Berdasarkan uraian tentang aneka penggunaan tes yang disajikan di Bab 4, penggunaan hasil tes pada dasarnya dapat digolongkan menjadi dua kategori besar: (a) **keperluan ilmiah,** yaitu penggunaan hasil tes untuk penelitian ilmiah tentang atau yang melibatkan pengukuran aneka atribut psikologis; dan (b) **keperluan praktis,** yaitu penggunaan hasil tes untuk aneka layanan praktis pengguna dalam bidang psikologi terapan (Kline, 1986).

Untuk kategori pertama, yaitu penggunaan hasil tes untuk penelitian ilmiah tentang atribut psikologis tertentu atau tentang tema lain namun melibatkan pengukuran kekhususan individual

terkait atribut psikologis tertentu, skor kasar dapat bahkan lebih tepat langsung digunakan sebagai data penelitian. Sebagaimana sudah disinggung, kualitas skor kasar sebagai data penelitian semacam ini ditentukan antara lain oleh taraf pengukurannya. Semakin data berupa skor kasar hasil pengukuran tersebut mendekati taraf pengukuran rasio atau paling tidak mencapai taraf pengukuran interval, semakin terbuka data tersebut untuk diolah dengan aneka tehnik statistik yang *robust* atau handal sehingga menghasilkan kesimpulan atau informasi yang semakin bisa dipercaya.

Untuk kategori kedua, yaitu penggunaan hasil tes untuk aneka layanan praktis pengguna dalam bidang psikologi terapan di berbagai sektor kehidupan, seperti layanan kesehatan mental masyarakat, pendidikan sekolah, pembinaan pegawai di lingkungan industri, atau penjaminan mutu dalam pengelolaan aneka profesi, skor kasar hasil pengukuran aneka atribut psikologis lazim perlu dikonversikan atau ditransformasikan dulu menjadi kategori atau bilangan lain yang lebih bermakna dengan cara membandingkannya dengan standar tertentu. Cara konversi atau transformasi skor kasar beserta standar yang dipakai sebagai patokan pembanding bisa bermacam-macam namun secara garis besar bisa digolongkan dalam dua kategori besar: (a) transformasi beracuan **kriteria** atau **patokan;** dan (b) transformasi beracuan **norma.** Marilah kita bahas kedua jenis transformasi tersebut satu demi satu.

A. Transformasi Skor Beracuan Patokan

Jenis transformasi ini lazim dipraktekkan di lingkungan pendidikan sekolah, khususnya untuk keperluan menginterpretasikan atau memaknai skor kasar hasil berbagai jenis tes prestasi yang disusun sendiri oleh guru kelas atau guru pengampu mata pelajaran yang bersangkutan. Jenis transformasi ini pada dasarnya hanya

melibatkan satu langkah penting yaitu *setting standard* atau **penentuan standar** atau **patokan.**

Inti penentuan standar atau patokan adalah penentuan *cutoff scores* atau **skor pemisah** (Crocker & Algina, 2008). Salah satu cara sederhana yang paling lazim dalam penggunaan skor pemisah semacam ini adalah menerapkannya secara langsung pada skor kasar hasil pengukuran atribut psikologis tertentu, khususnya hasil belajar dalam suatu mata pelajaran di sekolah, untuk menentukan kelulusan atau ketidaklulusan. Sebagai contoh, jika sekelompok guru mata pelajaran Bahasa Indonesia di sebuah SMA bersepakat menentukan skor pemisah hasil ujian sekolah untuk mata pelajaran Bahasa Indonesia bagi siswa kelas XII sebesar 0.80, berarti setiap siswa peserta ujian sekolah harus menjawab dengan benar 80% atau lebih dari keseluruhan soal-soal Bahasa Indonesia yang diujikan untuk dinyatakan lulus dalam mata pelajaran yang bersangkutan. Siswa yang menjawab benar kurang dari 80% dari soal-soal yang diujikan dinyatakan tidak lulus.

Salah satu pendekatan dalam penentuan patokan yang cukup sederhana dalam arti tidak melibatkan tehnik psikometrik adalah yang disebut *consensus judgment based on holistic inspection* atau **musyawarah berdasarkan pemeriksaan menyeluruh** (Crocker & Algina, 2008). Pendekatan ini meliputi sejumlah langkah sebagai berikut. *Pertama,* sekelompok pakar terdiri dari sekitar lima orang diminta memeriksa secara cermat isi tes. Masing-masing diminta mengusulkan persentase dari jumlah soal yang harus dijawab benar oleh seorang testi untuk dinyatakan menguasai kompetensi minimum atas isi atau materi yang diujikan. *Kedua,* patokan yang diusulkan oleh masing-masing pakar tersebut dikumpulkan dan dihitung reratanya secara musyawarah untuk akhirnya disepakati sebuah patokan akhir.

Selanjutnya, patokan berupa skor pemisah tersebut dapat digunakan untuk mentransformasikan skor kasar masing-masing testi atau peserta tes ke dalam salah satu dari dua kategori, **lulus** dan **tidak lulus** atau nama kategori biner lainnya seperti **diterima** dan **tidak diterima**. Kategori yang terakhir tersebut juga lazim dipakai sebab

penerapan skor pemisah semacam ini konon juga jamak dilakukan dalam menafsirkan hasil tes dalam rangka admisi atau penerimaan calon siswa atau calon pegawai. Di sini hasil tes dipakai sebagai dasar untuk memprediksikan taraf penguasaan kriteria eksternal tertentu berupa keberhasilan studi di jenjang pendidikan tertentu atau kinerja yang baik dalam berbagai tugas atau jabatan dalam bidang pekerjaan tertentu (Crocker & Algina, 2008).

Ada yang berpendapat kelompok pakar yang diminta mengusulkan patokan tersebut sebaiknya diperluas mencakup berbagai unsur pemangku kepentingan terkait (Shepard, 1976; dalam Crocker & Algina, 2008). Misal dalam konteks pendidikan sekolah kita, dalam penentuan patokan kelulusan ujian sekolah untuk mata pelajaran tertentu, selain kelompok guru mata pelajaran yang bersangkutan pihak lain yang juga pantas diminta pendapatnya adalah Komite Sekolah, kelompok pakar terkait dari perguruan tinggi terdekat atau mitra, dan Dewan Pendidikan setempat.

Menurut Crocker dan Algina (2008), kendati sangat lazim diterapkan karena sifatnya yang sederhana, pendekatan **musyawarah berdasarkan pemeriksaan menyeluruh** dalam penentuan skor pemisah atau patokan ini memiliki kelemahan, khususnya tidak bisa dijamin atau dipastikan bahwa kelompok-kelompok penilai tersebut akan mendasarkan penilaian mereka pada aspek-aspek tes yang sama atau bahwa mereka memiliki pandangan yang sama tentang cakupan isi yang semestinya dari atribut psikologis yang diukur atau diujikan. Akibatnya, skor pemisah atau patokan yang diusulkan oleh masing-masing kelompok pakar bisa sangat berlainan. Penyusun tes dituntut bersikap tegas namun bijak menghadapi situasi semacam ini.

B.Transformasi Skor Beracuan Norma

Norma adalah sejenis patokan juga, namun yang ditetapkan berdasarkan distribusi skor yang dikumpulkan dari sampel testi

yang representatif atau mewakili populasi testi yang menjadi sasaran tes. Norma semacam ini lazim berupa *average performance* atau kinerja rerata sampel yang dipakai sebagai dasar penentuan norma, khususnya berupa *Mean* disertai informasi tentang variabilitas skor-skor di sekitar *Mean* berupa *SD* (*standard deviation*) atau deviasi standar. Dengan kata lain, dalam transformasi skor beracuan norma skor kasar seorang testi ditafsirkan dengan cara membandingkannya dengan norma berupa kinerja rerata dari kelompok sebaya yang dipakai sebagai referensi atau acuan. Cara ini lazim diterapkan dalam menafsirkan hasil pengukuran dengan aneka tes yang bersifat standar atau baku (*standardized tests)*, baik berupa *maximal performance tests* maupun *typical performance tests.* Dalam konstruksi atau penyusunan tes, langkah penentuan norma ini merupakan salah satu aspek dari standarisasi atau pembakuan tes khususnya pembakuan cara menafsirkan skor kasar tes.

Ada dua langkah penting dalam penentuan norma, yaitu: (1) pelaksanaan *norming study* yaitu penelitian empiris dengan cara mengadministrasikan tes pada kelompok testi tertentu sebagai *standardization sample*; dan (1) pemilihan jenis norma yang dipakai sebagai acuan dalam menafsirkan skor kasar. Marilah kita bahas kedua langkah tersebut satu demi satu.

1. Pelaksanaan *Norming Study*

Kompleksitas pelaksanaan *norming study* atau penelitian dalam rangka penyusunan norma sebuah tes ditentukan oleh tujuan penggunaan tesnya. Sebuah tes yang dimaksudkan untuk digunakan sebagai *standardized test* atau tes baku bagi populasi testi berskala nasional atau bahkan berskala lintas bangsa tentu menuntut *norming study* yang lebih kompleks khususnya terkait jenis dan jumlah sampel yang dibutuhkan, dibandingkan sebuah tes buatan guru yang dimaksudkan untuk digunakan sebagai ujian sekolah atau bahkan ulangan umum untuk jenjang kelas tertentu di suatu sekolah. Kendati demikian, semua jenis penelitian dalam rangka penyusunan

norma sebuah tes akan mencakup langkah-langkah dasar studi atau penelitian sebagai berikut:

a. **Mengidentifikasikan Populasi Sasaran Tes.** Seperti sudah disinggung, populasi sasaran tes bisa memiliki cakupan luas seperti orang dewasa berusia 18 sampai dengan 56 tahun di sebuah negara atau bahkan lintas negara, atau cakupan menengah seperti semua pelamar calon pegawai negeri sipil (PNS), atau cakupan sempit seperti siswa kelas V SD tertentu di kota tertentu.

b. **Mengidentifikasikan Jenis Statistik yang Diperlukan.** Maksudnya, jenis statistik penting yang akan dihitung berdasarkan data yang diperoleh dari sampel dan yang akan diperlukan untuk mengembangkan norma. Ada dua jenis statistik yang lazim diperlukan yaitu *mean* dan deviasi standar atau *SD*. Sebagai tambahan juga dapat dihitung jenis statistik lain yang kiranya akan diperlukan dalam penyusunan norma, seperti jenjang persentil.

c. **Menetapkan Besar *Sampling Error* yang Dipandang Bisa Ditolerir.** *Sampling error* atau kesalahan akibat pengambilan sampel secara random merupakan ukuran diskrepansi atau kesenjangan atau perbedaan antara nilai murni suatu parameter dengan estimasi parameter tersebut yang didasarkan pada sampel. Jenis kesalahan ini perlu diperiksa untuk memastikan bahwa besarnya masih berada dalam batas yang bisa ditolerir. Salah satu ukuran yang lazim dipakai untuk menetapkan batas toleransi *sampling error* semacam ini adalah *standard error of measurement* atau kesalahan baku pengukuran, disingkat *SEM* mengikuti istilah bahasa Inggrisnya.

   Besar atau nilai *SEM* ditentukan oleh dua karakteristik tes, yaitu reliabilitas dan deviasi standar atau *SD*-nya sehingga cara menghitung *SEM* pun mengandalkan dua statistik tersebut, sebagaimana sudah disinggung di Bab 6 (Rumus 6.2.). *SEM* bermanfaat untuk menetapkan besar *sampling error* yang dipandang bisa ditolerir dengan cara menentukan apa yang disebut *confidence interval* atau interval kepercayaan. Interval

kepercayaan adalah *range* nilai yang mengandung kemungkinan letak skor murni dari suatu skor tes. Dengan menggunakan skor tes seorang testi sebagai estimasi skor murninya serta berpedoman pada logika distribusi normal, maka kita bisa 68% yakin bahwa skor murni testi tersebut terletak dalam *range X*±1 *SEM,* atau kita bisa 95% yakin bahwa skor murni testi tersebut terletak dalam *range X*±1,96 *SEM.* Artinya, pada kasus pertama dengan 1 *SEM* kita menentukan bisa menolerir *sampling error* sebesar 32%, sedangkan pada kasus kedua dengan 1,96 *SEM* kita menentukan bisa menolerir *sampling error* sebesar 5% saja.

d. Menetapkan Prosedur Memilih Sampel dari Populasi. Sebelum menetapkan prosedur untuk memilih sampel dari populasi, tentu harus diputuskan dulu apakah akan menggunakan sampel atau menggunakan seluruh populasi. Jika cakupan populasinya sempit misal dua atau tiga kelompok siswa kelas tertentu di sebuah SD seperti sudah disinggung, kiranya lebih tepat menggunakan data dari seluruh populasi untuk menentukan norma. Baru jika cakupan populasinya luas misal seluruh siswa SD kelas tertentu di sebuah kabupaten atau bahkan provinsi, penyusun tes perlu memutuskan menggunakan sampel dan memilih prosedur untuk mengambil sampel dari populasi.

Secara garis besar ada dua kategori *sampling* atau cara pengambilan sampel, yaitu *probability sampling* atau pengambilan sampel dengan menerapkan prinsip probalilitas; dan *nonprobability sampling* atau pengambilan sampel tanpa menerapkan prinsip probabilitas. Pada cara yang pertama, pada dasarnya setiap anggota populasi memiliki kesempatan sama untuk terpilih menjadi anggota sampel dengan mengandalkan tehnik yang disebut *random sampling* atau pengambilan sampel secara random. Pada cara kedua, tidak ada jaminan bahwa setiap anggota populasi memiliki kesempatan sama untuk terpilih menjadi anggota sampel. Cara kedua akan menghasilkan jenis sampel yang disebut *samples of convenience,* lazimnya berupa kelompok-kelompok testi yang kebetulan bisa diakses oleh penyusun tes.

Tentu saja, penentuan norma berdasarkan *probability sampling* akan menghasilkan norma yang lebih baik dalam arti lebih baku dibandingkan norma yang dihasilkan lewat *nonprobability sampling*.

e. **Menetapkan Besar Minimal Sampel yang Akan Menghasilkan *Sampling Error* dalam Batas yang Bisa Ditolerir.** Berbagai pakar mengemukakan pedoman yang sedikit berlainan. Leedy dan Ormrod (2005) memberikan pedoman sebagai berikut: (1) jika populasi terdiri dari 100 orang atau satuan atau kurang, sebaiknya tidak mengambil sampel atau seluruh populasi digunakan sebagai subjek atau partisipan studi; (2) jika populasi sebesar lebih kurang 500, perlu diambil sampel sebesar 50% dari populasi; (3) jika populasi sebesar sekitar 1.500, perlu diambil sampel sebesar 20% dari populasi; dan (4) jika populasi sebesar 5.000 atau lebih, sampel sebesar 400 dipandang memadai.

f. **Mengambil Sampel dan Mengadministrasikan Tes untuk Memperoleh Data.** Apa yang sudah ditetapkan terkait jenis dan besar sampel harus ditaati dalam pengambilan sampelnya. Seandainya sasaran jumlah sampel tidak tercapai, harus dicermati apakah pengurangan jumlah sampel ini bisa diterima atau harus diganti dengan mengambil sampel baru dalam jumlah yang diperlukan untuk mengisi kekurangan yang terjadi.

g. **Menghitung Nilai-nilai statistik yang Sudah Ditetapkan dan Nilai Kesalahan Standarnya, Khususnya *SME*.** Seperti sudah disinggung, jenis statistik yang lazim digunakan dalam penyusunan norma meliputi *mean, SD, persentil,* dan *SME* untuk menentukan batas *sampling error* yang bisa ditolerir.

h. **Menetapkan Jenis Norma yang Akan Digunakan dan Menyusun Tabel Konversi Skornya.** Tergantung dari jenis statistik yang digunakan sebagai dasar penyusunan norma, norma yang dihasilkan bisa bertaraf ordinal atau interval. Jenis norma yang dipilih perlu disesuaikan dengan keperluan penggunaan hasil tes.

i. **Menyusun Penjelasan Tertulis tentang Proses Penyusunan Norma serta Petunjuk Penggunaan Norma dalam Menafsirkan Hasil Tes.** Informasi tentang proses penyusunan norma meliputi antara lain (Crocker & Algina, 2008): (a) populasi yang menjadi khalayak sasaran tes serta prosedur pengambilan sampel yang digunakan dalam penelitian dalam rangka penyusunan norma; (b) masa saat penelitian dalam rangka penyusunan sampel dilaksanakan serta komposisi sampel dari segi gender, latar belakang etnik, status sosio-ekonomik, tempat tinggal, dan jenis kelompoknya seperti pelajar, pekerja, ibu rumah tangga, dan sebagainya; (c) *standard error of the mean* (*SME)* serta interval skor yang mengandung *mean* populasi pada taraf kepercayaan yang berbeda-beda; dan (d) penjelasan tentang makna dan cara menafsirkan hasil konversi skor berdasarkan norma.

2. Jenis Norma

Hal penting yang perlu diperhatikan dalam melakukan transformasi skor kasar dalam rangka penyusunan norma adalah pembedaan transformasi skor ke dalam dua jenis, yaitu *transformasi linear* dan *transformasi nonlinear* (Allen & Yen, 1979; Friedenberg, 1995). Sebagaimana sudah kita lihat, transformasi linear hanya memanfaatkan empat jenis operasi dasar matematika yaitu perkalian, pembagian, penambahan, dan pengurangan. Bentuk persamaan dasar transformasi linear adalah $Y = aX + b$, di mana a dan b merupakan bilangan konstan, X merupakan skor kasar yang ditransformasikan, dan Y merupakan skor baru hasil transformasi skor kasar. Transformasi linear hanya berdampak mengubah satuan pengukuran dan tidak berdampak mengubah skala pengukuran, hubungan antara masing-masing skor kasar, maupun bentuk distribusi secara keseluruhan.

Sebaliknya, transformasi nonlinear atau juga disebut *area transformation* atau transformasi wilayah (Friedenberg, 1995) memanfaatkan jenis-jenis operasi matematika lain seperti akar

atau pangkat. Transformasi semacam ini berdampak tidak hanya mengubah satuan pengukuran melainkan juga mengubah skala pengukuran, hubungan antar skor, dan bentuk distribusi. Kita hanya akan membahas beberapa jenis transformasi skor yang secara bersama-sama akan memenuhi dua syarat: (a) mewakili jenis transformasi linear dan nonlinear, serta (b) lazim diterapkan dalam bidang pendidikan sekolah maupun dalam bidang layanan psikologi secara umum.

a. Norma Persentil

Transformasi skor nonlinear yang banyak digunakan dalam menyusun norma penilaian di bidang pendidikan sekolah adalah ***percentile rank*** atau jenjang persentil. Nama lain adalah **persentil** atau **skor persentil** (Allen & Yen, 1979). Sebagai norma, jenjang persentil menunjukkan jenjang atau kedudukan masing-masing skor yang dicapai oleh sekelompok testi berdasarkan frekuensi atau seberapa sering aneka skor tersebut muncul. Karena yang menentukan kedudukan suatu skor adalah frekuensi dan bukan nilainya, maka hubungan antar skor berubah (Friedenberg, 1995).

Langkah pertama untuk menghitung jenjang persentil adalah menyusun tabel frekuensi skor yang menunjukkan distribusi atau sebaran frekuensi masing-masing skor atau kelompok skor yang diurutkan dari skor tertinggi sampai dengan skor terendah. Selanjutnya salah satu jenis jenjang persentil yang lazim digunakan sebagai norma penilaian adalah jenjang persentil yang menunjukkan proporsi orang yang mencapai skor **di bawah** atau **lebih rendah** dibandingkan masing-masing skor yang sedang kita tentukan jenjang persentilnya (Friedenberg, 1995). Dengan kata lain, jenjang persentil sebuah nilai hasil pengukuran atribut psikologis tertentu setara dengan persentase subjek atau testi dalam *norm group* atau kelompok norma, yaitu sampel yang digunakan untuk melakukan *norming study* atau penelitian dalam rangka penyusunan norma, yang memiliki atau memperoleh nilai-nilai kurang atau sama dengan nilai tertentu tersebut (Allen & Yen, 1979).

Cara menghitung jenis jenjang persentil seperti diuraikan di atas adalah sebagai berikut. Mula-mula perlu dihitung **frekuensi kumulatif di bawah** (fk_b) masing-masing skor, yaitu jumlah frekuensi skor-skor yang berada di bawah atau lebih rendah dari masing-masing skor. Kemudian frekuensi kumulatif di bawah masing-masing skor tersebut dibagi dengan jumlah subjek atau testi (N), dan akhirnya dikalikan 100. Bila dinyatakan dalam sebuah formula, diperoleh rumus sebagai berikut (Friedenberg, 1995):

$$JP_b = (fk_b / N) \times 100$$ **Rumus 12.1.**

JP_b = Jenjang persentil skor tertentu, disebut skor sasaran
fk_b = frekuensi kumulatif di bawah skor sasaran
N = jumlah subjek yang digunakan dalam penelitian dalam rangka penyusunan norma

Kelebihan utama jenjang persentil, persentil, atau skor persentil sebagai norma penilaian ada dua, yaitu langsung bisa dihitung tanpa perlu mempertimbangkan bentuk distribusinya dan mudah dipahami. Namun kelemahannya adalah: (1) skala pengukuran yang dihasilkan hanya bertaraf ordinal sehingga kurang leluasa untuk dianalisis lebih lanjut dengan tehnik statistik; (2) bentuk distribusi persentil adalah rektangular atau segi empat dan bukan normal sehingga tidak bisa dianalisis dengan menggunakan tehnik statistik parametrik; dan (3) jenjang persentil bisa menghasilkan penafsiran yang berbeda terhadap perbedaan kecil yang sama antar skor yang terletak di tengah dan di ujung distribusi (Allen & Yen, 1979; Kline, 1986; Friedenberg, 1995).

Khususnya untuk mengatasi kelemahan kedua, jenjang persentil lazim ditransformasikan terlebih dulu menjadi *normalized scores* atau skor yang dinormalisasikan sehingga distribusinya pun menjadi normal. Salah satu bentuk normalisasi jenjang persentil adalah ***stanines*** kependekan dari ***standard nines,*** berupa rangkaian skor satu digit berdistribusi normal yang berkisar antara 1 sampai dengan 9 dan yang sudah dinormalkan. Normalisasi yang dimaksud

diperoleh dengan cara membagi distribusi jenjang persentil menjadi 9 kategori dengan pembagian *ranges* atau wilayah sebagai berikut: kategori 1 dan 9 masing-masing meliputi 4% dari keseluruhan wilayah distribusi, kategori 2 dan 8 masing-masing meliputi 7% wilayah, kategori 3 dan 7 masing-masing meliputi 12% wilayah, kategori 4 dan 6 masing-masing meliputi 17% wilayah, dan kategori 5 meliputi 20% wilayah. Hasilnya adalah sebuah norma penilaian dengan kisaran nilai 1-9, *mean* = 5 dan *SD* kurang lebih = 2, serta batas-batas dan *range* jenjang persentil seperti disajikan dalam Tabel 12.1. (Allen & Yen, 1979; Kline, 1986; Friedenberg, 1995; Crocker & Algina, 2008).

Tabel 12.1.
Norma Penilaian *Stanines* Beserta Jenjang dan *Range* Persentilnya

Stanine	Jenjang Persentil	*Range* Persentil
9	98	96-100
8	94,5	89-96
7	83	77-89
6	68,5	*60-77*
5	50	40-60
4	31,5	23-40
3	17	11-23
2	5,5	4-11
1	2	0-4

Adaptasi dari Allen & Yen (1979), h. 165.

Untuk mentransformasikan sebuah skor kasar menjadi *Stanine,* pertama-tama harus dihitung dulu jenjang persentil skor kasar tersebut kemudian ditentukan *Stanine*-nya berpedoman pada norma seperti disajikan dalam Tabel 12.1.

b. Norma *Standard Score* atau Skor Baku

Transformasi linear yang lazim digunakan sebagai dasar untuk menyusun norma adalah transformasi *standard score* atau skor baku atau *z-score* atau skor-*z* atau cukup disebut *z,* dengan rumus perhitungan sebagai berikut (Allen & Yen, 1979):

$$z = (X - \mu_x)/\sigma_x$$ **Rumus 12.2.**

X = skor kasar
μ_x = *mean* distribusi skor kasar
σ_x = deviasi standar distribusi skor kasar

Skor-z merupakan salah satu jenis skor baku yang menempatkan skor kasar dalam sebuah skala baru dengan *mean* = 0 dan σ = 1. Skor-z menunjukkan letak skor kasar dari *mean* dalam satuan deviasi standar. Jika z dari sebuah skor kasar = +1 berarti skor kasar tersebut terletak 1 σ di atas μ; jika z dari sebuah skor kasar lain = -2 berarti skor kasar tersebut terletak 2 σ di bawah μ. Transformasi linear skor-z semacam ini tidak mengubah bentuk distribusi skor kasarnya serta menghasilkan data pengukuran baru pada taraf interval. Kendati demikian, skor-z memiliki dua kelemahan yang menjadikannya kurang disukai orang, yaitu: (1) kira-kira setengahnya akan berupa bilangan negatif, dan (2) tidak jarang berupa bilangan pecahan.

Untuk mengatasi kelemahan-kelemahan di atas, orang lazim menerapkan *standardized scores* atau skor yang dibakukan. *Standardized scores* merupakan transformasi linear terhadap skor kasar atau skor-z yang merupakan ekuivalennya untuk menghilangkan khususnya bilangan negatif. Rumus umum perhitungannya adalah sebagai berikut (Allen & Yen, 1979):

$$Y = \sigma^* z + \mu^*$$ **Rumus 12.3.**

Y = *standardized score*
z = skor baku yang ditransformasikan
σ^* dan μ^* = *SD* dan *mean standardized score* yang dipilih

Ada beberapa pilihan *standardized scores* dengan *mean* dan *SD* yang berbeda-beda. Salah satu jenis *standardized scores* yang cukup popular adalah *T-score* atau skor-T atau disingkat T. T memiliki μ = 50

dan $\sigma = 10$. *Mean* distribusi skor kasar yang setara $z = 0$ menjadi T = 50, sehingga skor kasar yang terletak di atas *mean* akan memiliki T > 50 sedangkan skor kasar yang terletak di bawah *mean* akan memiliki T , 50. Untuk menghindari pecahan, nilai T selalu dibulatkan ke bilangan bulat terdekat (Friedenberg, 1995). Ψ

Bab 13
Penutup

Selama sekitar satu abad sejak kelahirannya di awal abad ke-20 hingga kini, pengukuran psikologis beserta perwujudan konkretnya tes psikologis atau psikotes telah menikmati sejenis 'hegemoni' sebagai salah satu sarana andalan dalam rekayasa sosial di berbagai bidang kehidupan masyarakat, baik di Inggris dan Amerika Serikat yang merupakan tanah kelahirannya, maupun praktis di semua negara di dunia. Dalam bab-bab sebelumnya, kita telah membahas berbagai hal esensial terkait pengukuran psikologis dan tes psikologis baik yang bersifat konseptual maupun yang bersifat praktis. Agar memiliki wawasan yang proporsional tentang tes psikologis sehingga mampu menyikapinya secara tepat dalam pengembangan maupun penerapannya dalam kehidupan sehari-hari di berbagai wilayah kehidupan, pada bab penutup buku ini akan disajikan sejumlah catatan kritis sekitar tiga hal: (1) esensialisme sebagai ontologi yang mendasari psikologi modern pada umumnya maupun pengukuran psikologis khususnya; (2) konteks sosio-historis lahirnya tes psikologis baik di Inggris maupun di Amerika Serikat dan dampaknya terhadap perkembangan psikologi dan tes psikologis di berbagai negara lain termasuk Indonesia; dan (3) beberapa asumsi dasar penerapan tes psikologis khususnya dalam bidang pendidikan sekolah sebagai sejenis *caveat* bagi siapa pun yang bersinggungan dengan penggunaan tes psikologis. Marilah kita bahas ketiga hal tersebut satu demi satu.

A. Pengukuran Psikologis dan Esensialisme

Pengukuran psikologis atau psikometri merupakan salah satu subdisiplin psikologi modern. Sudah menjadi pengetahuan umum,

psikologi modern lahir di Jerman pada sekitar akhir abad ke-19 dan menjelma menjadi psikologi arus utama di Amerika Serikat sejak awal abad ke-20. Di Bab I sudah dicoba diuraikan adanya dua arus besar paradigma yang mempengaruhi cara kerja ilmu-ilmu sosial termasuk Psikologi, yaitu realisme yang diwakili oleh positivisme dan relativisme yang diwakili oleh konstruktivisme (Guba & Lincoln, 1994). *Realisme* merupakan pandangan dunia yang meyakini bahwa realitas bersifat *real* atau nyata atau benar-benar ada dan tunggal. Tugas ilmu pengetahuan termasuk Psikologi adalah mengungkap hakikat dan cara kerja aneka benda-gejala yang terdapat di dalam realitas faktual tersebut (Guba & Lincoln, 1994). Sebaliknya, *relativisme* merupakan pandangan dunia yang meyakini bahwa realitas bersifat relatif dalam arti tidak bersifat tunggal melainkan jamak bahkan kadang-kadang saling bertentangan, sebab merupakan hasil konstruksi mental yang bersifat lokal dan spesifik dalam arti bahwa bentuk dan isinya ditentukan oleh pengalaman masing-masing orang atau kelompok orang yang meyakini atau mendukungnya. Maka, tugas ilmu pengetahuan adalah memahami aneka konstruksi tentang realitas tersebut serta mendialogkannya secara dialektis satu sama lain agar aneka konstruksi tentang realitas tersebut mengalami revisi dan rekonstruksi secara terus-menerus. Dalam bab tersebut juga sudah dicoba ditunjukkan bahwa psikologi modern cenderung berada dalam arus paradigma realisme. Secara lebih spesifik, di sini akan ditunjukkan bahwa Psikologi dalam arti *mainstream psychology* atau Psikologi arus utama seperti yang mendominasi wacana psikologi di banyak negara termasuk di Tanah Air dewasa ini khususnya dalam bidang pengukuran psikologis, cenderung menganut esensialime atau secara lebih khusus *real essentialism* atau esensialisme realis atau nyata (Oderberg, 2007).

Secara garis besar, esensialisme sebagai pandangan dunia bertumpu pada lima keyakinan dasar sebagai berikut (Oderberg, 2007). *Pertama,* dunia nyata memang benar-benar ada dan semua hal yang ada di dalamnya adalah benar-benar nyata berupa jenis *beings* atau makhluk tertentu yang keberadaannya tidak ditentukan oleh

pendapat atau dugaan orang. Memang, orang harus merumuskan atau memberi sebutan pada aneka jenis 'makhluk' tersebut seperti pada contoh aneka atribut psikologis yang dirumuskan oleh para pakar psikologi sebagai konstruk teoretis. Namun, kalau pun belum atau tidak pernah dirumuskan dan diberi sebutan, aneka atribut psikologis sebagai jenis 'makhluk' tertentu menurut keyakinan para esensialis tetaplah ada di antara aneka jenis 'mahkluk' lain di dunia ini (Oderberg, 2007).

Kedua, alam materi benar-benar memiliki keberadaan otonom. Esensialisme tentu saja tidak menyangkal atau mengakui keberadaan 'makhluk' yang bersifat imaterial. Namun, esensialisme tidak mengurangi nilai keberadaan alam materi dalam rupa realitas fisik konkret dengan mengacu pada alam imaterial tertentu sebagai landasan atau dasar terakhir keberadaannya (Oderberg, 2007).

Ketiga, esensi atau hakikat dapat dipahami oleh akal manusia. Pikiran atau akal budi manusia mampu menangkap atau memahami esensi atau hakikat segala jenis makhluk. Pengetahuan tentang kebenaran merupakan hasil persinggungan atau kesejalanan (*conformity*) antara akal budi manusia dengan hakikat aneka makhluk. Memang seringkali pengetahuan tersebut masih bersifat parsial atau tidak sempurna, namun kaum esensialis yakin bahwa manusia senantiasa mampu mencapai pengetahuan yang memadai dan lengkap atau penuh tentang hakikat aneka makhluk yang ada di dalam dunia nyata (Oderberg, 2007).

Keempat, pengetahuan tentang esensi atau hakikat dicapai lewat perumusan *real definition* atau definisi nyata. Mendefinisikan sesuatu berarti menetapkan batas-batas sesuatu tersebut sedemikian rupa sehingga kita bisa membedakannya dari segala sesuatu lain dari jenis yang berbeda. Mendefinisikan sesuatu berarti menyatakan sesuatu itu apa. Mengutip pernyataan seorang pakar lain, "just as we may define a word, or say what it means, so we may define an object, or say what it is" (Fine, dalam Oderberg, 2007, h. 19). Artinya, sama seperti kita bisa mendefinisikan sebuah kata, atau menyatakan makna kata tersebut, begitu pula kita bisa mendefinisikan sebuah objek, atau

menyatakan apa gerangan objek tersebut. Intinya, kaum esensialis berkeyakinan bahwa kita bisa menyatakan secara tepat apa gerangan hal-hal yang kita temui di sekitar kita (Oderberg, 2007).

Kelima, kaum esensialis meyakini bahwa dunia ini tertib-teratur maka segala sesuatu di dunia ini bersifat *classifiable* atau bisa diklasifikasikan atau bisa digolong-golongkan menurut suatu sistem kategorisasi tertentu. Lebih lanjut menurut keyakinan mereka, klasifikasi atau penggolongan tersebut tidak didasarkan pada dimensi nyata dari objek atau gejala, melainkan berdasarkan keberadaan objek atau gejala tersebut dalam keseluruhan, kesatuan, atau keutuhannya. Kaum esensialis juga yakin bahwa klasifikasi atau penggolongan tersebut sungguh-sungguh mungkin atau ada, dengan demikian mereka juga yakin bahwa ada struktur penggolongan tertentu sebab makna hakikat itu sendiri mengandaikan *inclusion* atau penggabungan dan *exclusion* atau pemisahan objek-objek dan gejala-gejala, sehingga juga terbuka kemungkinan terbentuknya hirarki (Oderberg, 2007).

Paham atau pandangan di atas sangat berseberangan dengan tiga corak teori sosial interpretif mewakili paradigma relativis-konstruktivis yang lazim diterapkan dalam penelitian di bidang ilmu sosial-humaniora pada umumnya dan kajian budaya pada khususnya (Lewandowski, 2001). Corak yang pertama menggunakan apa yang oleh Lewandowski (2001) disebut *the logic of textuality and deconstruction* atau logika tekstualitas dan dekonstruksi. Teori ini didasarkan pada pandangan bahwa kebudayaan atau realitas sosial merupakan sejenis teks. Maka tugas ilmu sosial adalah *reading* atau membaca, atau *evoking* atau memunculkan dalam arti *writing* atau menuliskan teks tersebut.

Menurut Lewandowski (2001) ada dua cabang dalam logika tekstualitas dan dekonstruksi ini. Yang pertama memandang kebudayaan atau realitas sosial sebagai teks yang *thick* atau tebal, atau *deep* atau dalam, yang harus dibaca. Pendekatan ini didasarkan pada metode *hermeneutika* atau interpretasi yang dikembangkan oleh *Paul Ricoeur* yang dikenal sebagai *a theory of textual depth and character-readability of human action* atau teori tentang tindakan manusia sebagai

teks yang dalam dan dapat dibaca, yang memandang tindakan manusia sebagai teks sosial. Menurut Ricoeur (dalam Lewandowski, 2001), sebagai teks sosial tindakan manusia memiliki empat aspek yang saling terjalin. Yang *pertama* adalah sifat dalam atau kedalamannya. Maka penelitian sosial harus bergerak dari yang di permukaan menuju kedalaman tindakan manusia yang menjadi objek penelitiannya. Namun tujuannya bukan menemukan makna orisinal atau maksud pengarang atau pelaku sebagai sesuatu yang tersembunyi di dalam kedalaman itu, melainkan mengungkapkan secara kritis sesuatu atau makna yang dimunculkan oleh tindakan itu. Aspek *kedua* adalah sifat konstitutifnya sebagai bahasa (*linguistic constitution*). Artinya, teks pada dasarnya merupakan diskursus atau wacana atau peristiwa bahasa, sehingga teks senantiasa bersifat *referentially deep* atau mengacu sesuatu sebab teks senantiasa merupakan diskursus *tentang* sesuatu. Aspek *ketiga,* analisis sosial pada dasarnya merupakan metode membaca, yaitu membaca teks sosial yang lebih mirip membaca sebuah karya seni. Artinya, yang dibutuhkan dalam analisis sosial bukan sebuah sains yang bertujuan menjelaskan melainkan sebuah "art of deciphering" atau seni mengupas yang bertujuan menguliti lapis demi lapis lapisan makna teks sosial yang memiliki kedalaman tanpa batas. Aspek *keempat,* hubungan antara teks berisi makna yang dalam dan dunia kehidupan. Menurut Ricoeur, dunia kehidupan merupakan kumpulan makna yang diungkap melalui teks. Artinya, pembacaan atas teks-teks sosial akan melahirkan aneka makna baru, menciptakan aneka kemungkinan baru, menyingkapkan aneka potensi baru yang akan mendobrak atau membuyarkan aneka kerangka acuan atau sudut pandang yang sudah ada (Lewandowski, 2001).

Cabang kedua dari logika tekstualitas dan dekonstruksi memandang kebudayaan atau realitas sosial sebagai sebuah *planar text* atau teks yang membentang, sebuah permainan penanda yang menunggu untuk disingkap atau dimunculkan. Pendekatan ini didasarkan pada teori *dekonstruksi* yang dikembangkan oleh *Jacques Derrida* (Lewandowski, 2001). Menurut Derrida (dalam Lewandowski,

2001), permainan atau main-main pertama-tama merupakan peristiwa atau kejadian kebahasaan melibatkan proses substitusi atau penggantian atau penambahan tanpa batas. Menginterpretasikan sebuah teks bukan berarti menggandakan atau mempenetrasinya, melainkan membentangkannya dalam sebuah rangkaian *signification* atau pemaknaan melalui sejenis proses *writing* atau menulis. Bagi Derrida, teks bukan lagi merupakan naskah tebal melainkan situs atau medan tempat berlangsungnya permainan aneka penanda, menulis, aneka suara, aneka *evocations* atau penyingkapan. Karena kebudayaan atau realitas sosial dipandang sebagai arena penulisan tempat seorang pengarang ditulis dan menulis, maka tugas ilmu sosial bukan pertama-tama membuat penafsiran melainkan menyajikan apa yang disebut *a linguistic reflexivity* atau sejenis pertukaran bahasa tempat sebuah teks dipertemukan dengan teks lain, tempat sebuah *gaze* atau tatapan dipertemukan dengan tatapan lain (Lewandowski, 2001).

Corak kedua teori sosial interpretif mewakili paradigma relativis-konstruktivis yang lazim diterapkan dalam penelitian di bidang ilmu sosial-humaniora pada umumnya dan kajian budaya pada khususnya adalah *the logic of rationality and reconstruction* atau logika rasionalitas dan rekonstruksi sebagaimana ditemukan dalam karya *Jurgen Habermas* (Lewandowski, 2001). Pendekatan ini memandang kebudayaan dan kehidupan sosial sebagai praktek komunikasi. Bagi Habermas (dalam Lewandowski, 2001), pandangan ini melahirkan tiga implikasi penting. Pertama, bahasa merupakan penentu praktek sosial. Artinya, bahasa merupakan medium atau sarana praktek kekuasaan dan pemaksaan atau penindasan, sarana pembagian pekerjaan dan dominasi, sekaligus merupakan sarana pemahaman sejarah. Kedua, karena data bagi penelitian sosial berupa pengalaman komunikasi maka aneka metode hermeneutika merupakan bagian yang tidak bisa dihindari dalam penelitian sosial. Ketiga, hermeneutika cocok diterapkan untuk menafsirkan masyarakat modern, dan secara khusus cocok untuk melakukan kritik ideologi melalui rekonstruksi komunikasi (Lewandowski, 2001).

Corak ketiga teori sosial interpretif mewakili paradigma relativis-konstruktivis yang lazim diterapkan dalam penelitian di bidang ilmu sosial-humaniora pada umumnya dan kajian budaya pada khususnya adalah *the logic of constructing constellations* atau logika konstelasi yang mengkonstruksikan atau membangun (Lewandowski, 2001). Pendekatan ini memandang kehidupan sosial sebagai konstruksi yang bersifat materialis atas aneka praktek dan produksi yang dilakukan oleh manusia yang tidak harus bercorak rasional atau tekstual melainkan *relasional*. Artinya, masyarakat merupakan hasil konstruksi relasional antara berbagai unsur dan praktek sosial yang terpisah-pisah berupa aneka konteks atau *fields* atau wilayah dan bukan berupa teks-teks bahasa atau tindakan-tindakan komunikatif. Menurut pendekatan ini, corak atau sifat relasional inilah yang menjadikan eksistensi dan praktek kehidupan manusia bersifat *sosial*. Selain bersifat relasional, pendekatan ini juga memandang kehidupan sosial sebagai hasil konstruksi. Maka, menurut pendekatan ini tugas ilmu sosial pertama-tama bukanlah memunculkan makna atau merekonstruksikan norma-norma sebuah konteks sosial melainkan mengubah konstitusi atau bangunan praktis konteks sosial tersebut (Lewandowski, 2001).

Sekali lagi, psikologi arus utama yaitu psikologi sebagaimana dikenal umum di berbagai negara termasuk di Tanah air termasuk pengukuran psikologis sebagai salah satu bidang kajian di dalamnya, cenderung menganut pandangan positivis-realis-esensialis. Psikologi sosial sebagai salah satu bidang psikologi khususnya berkat keterlibatannya yang lebih intens dalam kajian budaya yang bersifat lintas disiplin, cenderung lebih condong pada pandangan relativis-konstruktivis baik yang bercorak strukturalis maupun paskastrukturalis. Masing-masing memiliki landasan filosofis-konseptual yang dalam konteks dan sudut pandang tertentu dipandang bisa dipertanggungjawabkan, antara lain terbukti dari keberlangsungan keberadaannya. Para pendukung kedua paradigma keilmuan dalam komunitas psikologi termasuk di kampus-kampus di Tanah Air seringkali terlibat dalam sejenis persaingan bahkan

cenderung menjadi semacam permusuhan, tanpa menyadari posisi pendirian masing-masing terbukti dari munculnya pertikaian itu sendiri. Lewat catatan kecil ini diharapkan semua pihak terbantu untuk menyadari secara proporsional posisi masing-masing, kelebihan dan keterbatasan dari posisinya, demi mampu memberikan kontribusi yang pas dalam rangka pembentukan masyarakat kita yang semakin cerdas.

B. Konteks Sosio-Historis Lahirnya Tes Psikologis

Pemahaman secara sekilas tentang perkembangan pengukuran psikologis atau lebih khusus tes psikologis sejak kelahiran hingga keberadaannya di masa kini kiranya juga diperlukan agar bisa menilai kelebihan-kekurangannya secara lebih proporsional dan akurat sehingga mampu menimbang dan menempatkan secara pas kontribusinya dalam pembentukan masyarakat zaman sekarang.

Pengukuran psikologis lahir seiring dengan munculnya gerakan *testing* dalam rangka pembentukan masyarakat di Inggris dan Amerika Serikat sekitar awal abad ke-20. Di Inggris di masa pra Perang Dunia, pembentukan sosial atau lebih tepat *social engineering* sangat dipengaruhi oleh pandangan tentang pentingnya inteligensi sebagai penentu keberhasilan hidup seseorang di tengah masyarakat. Kala itu inteligensi dipandang sebagai kemampuan tunggal khususnya berupa intelek dan bersifat *innate* atau bawaan. Pendidikan pada umumnya maupun pendidikan sekolah pada khususnya dilaksanakan dalam rangka mewujudkan cita-cita *eugenetika Francis Galton,* yaitu perbaikan atau peningkatan kualitas masyarakat melalui seleksi orang-orang yang memiliki inteligensi bawaan yang tinggi untuk diberi kesempatan mengembangkan diri secara optimal melalui kesempatan studi lanjut ke jenjang pendidikan yang semakin tinggi yang kala itu lazimnya memang masih terbatas (Torrance, 1981). Jadi, semacam Darwinisme sosial yaitu penerapan prinsip seleksi *survival*

of the fittest ala Darwin dalam formasi sosial masyarakat. Maka, di Prancis *Alfred Binet* mengembangkan tes inteligensi untuk memilah anak-anak yang terbelakang secara mental dari anak-anak normal di sekolah-sekolah di Paris. Di Amerika Serikat kala itu tes inteligensi dikembangkan antara lain untuk membatasi arus imigran dari negara-negara Eropa timur dan selatan yang secara umum dipandang kurang berkualitas dibandingkan imigran dari Eropa barat dan utara, dengan asumsi "To him that hath a superior intellect is given also on the average a superior character" (Thorndike, seorang pakar testing Amerika Serikat, dalam Torrance, 1981). Ringkas kata, pada awalnya dan baik di Inggris maupun Amerika Serikat, gerakan testing yang mendorong lahir dan mekarnya pengukuran psikologis sebagai salah satu bidang penting dalam psikologi merupakan sarana pembentukan masyarakat atau tepatnya perekayasaan sosial berdasarkan prinsip eugenetika.

Dalam perkembangan selanjutnya di Eropa maupun khususnya Amerika Serikat paska Perang Dunia seiring dengan semangat demokratisasi, gerakan testing mengalami pergeseran tujuan dari semula sebagai sarana perekayasaan sosial berdasarkan prinsip eugenetika yang berbau elitis dan diskriminatif menjadi sarana mewujudkan efisiensi dalam melayani kebutuhan pendidikan yang beragam dalam masyarakat (Torrance, 1981). Salah satu pemicu pergeserean ini adalah perubahan pandangan tentang inteligensi yang kemudian dipandang bersifat ganda dan bukan tunggal, serta bukan merupakan satu-satunya penentu keberhasilan hidup manusia. Gerakan testing kemudian lebih bertujuan akomodatif, yaitu sebagai sarana manajerial untuk melakukan diferensiasi dan klasifikasi anak-anak yang memiliki latar belakang dan kemampuan yang berlainan, agar bisa diberikan jenis pengalaman pendidikan yang sesuai dengan kemampuan maupun minat mereka. Di Amerika Serikat semangat ini melahirkan konsep *elektif* yaitu diperkenalkannya aneka pilihan kurikulum dalam sistem pendidikan agar murid-murid bisa memilih jalur yang sesuai arah minat dan kemampuan masing-masing (Torrance, 1981). Dengan kata lain, pada fase ini manajemen

masih dilaksanakan sebagai sarana untuk mewujudkan efisiensi dalam arti positif dalam memberikan pelayanan terhadap kebutuhan masyarakat yang semakin demokratis dan beragam oleh negara. Dalam semangat ini secara teoretis gerakan testing sebagai salah satu sarana manajemen diharapkan mendorong demokratisasi dan pemerataan pelayanan khususnya dalam pemenuhan salah satu hak dan kebutuhan ekonomi-sosial masyarakat, yaitu pendidikan.

Globalisasi dan internasionalisasi juga dalam bidang pendidikan yang berlangsung masif sejak dasawarsa 1980-an mengubah wajah manajerialisme yang semula berwatak positif menjadi manajerialisme baru sebagai momok yang menakutkan. Neoliberalisme telah mengubah dunia menjadi sebuah pasar global dan semua barang serta seluruh aktivitas manusia menjadi komoditas. Dalam ekonomi global berbasis pengetahuan, pendidikan pada umumnya maupun pendidikan tinggi pada khususnya dipandang sebagai kunci atau jaminan untuk meraih status sosial yang lebih tinggi melalui pemerolehan pekerjaan profesional yang lebih mapan. Pendidikan pun menjadi komoditas. Maka, kerja sama antar negara di bidang pendidikan yang semula dilaksanakan dalam kerangka *aid* bergeser menjadi *trade,* yang semula berupa *cooperation* bergeser menjadi *competition*. Perdagangan pendidikan antar negara dalam berbagai bentuknya (studi di luar negeri, penerapan kurikulum sekolah luar negeri di sekolah dalam negeri, pembukaan kampus asing di dalam negeri, dsb.) menjadi semakin lazim. "Perdagangan" tenaga kerja profesional antar negara akan menjadi masa depan yang tak terelakkan (Ninnes, & Hellsten, 2005).

Untuk memuluskan transaksi tersebut semua negara harus menerapkan **standar mutu yang sama**, baik dalam mengelola diri ke dalam maupun dalam menjalankan aneka bentuk "kerjasama" dengan lembaga atau negara lain. *Mutual recognition of qualifications framework* (Departement of International Research and Cooperation, NIER, 2001) dikembangkan untuk menjamin kesetaraan mutu kurikulum pendidikan di berbagai jenjang pendidikan dan mutu lulusan yang dihasilkan antar negara, begitu pula sistem akreditasi nasional,

regional maupun internasional dikembangkan untuk menjamin kesetaraan mutu pendidikan untuk pada akhirnya mencapai tingkat *world class school* dan *world class university.* Untuk itu, sekali lagi standar mutu yang sama juga wajib diterapkan dalam pengelolaan seluruh sumber daya di dalam suatu negara, termasuk dalam bidang pendidikan. *Testing* dan tes psikologis menjadi sarana penting dalam rangka mengevaluasi dan menciptakan kesetaraan mutu proses dan hasil pendidikan antar negara.

Namun di samping arus *mainstream* penerapan standar tunggal dalam berbagai bidang kehidupan termasuk pendidikan sekolah seperti diuraikan di atas, sebenarnya juga tengah berlangsung arus lain yang lebih menghargai keragaman khususnya terkait penilaian hasil belajar di dunia pendidikan sekolah. Dalam kesadaran bahwa pendidikan merupakan hak setiap orang serta dilandasi semangat menghargai keragaman latar belakang dan kemampuan manusia seperti yang terjadi pada masa ketika manajemen dimaknai secara positif sebagai sarana untuk mengakomodasi keragaman kemampuan dan latar belakang peserta didik, dalam dunia pendidikan sekolah di banyak negara kini ada gerakan yang mengarahkan penilaian hasil belajar lebih untuk tujuan *sertifikasi,* yaitu memberikan *feedback* baik kepada peserta didik maupun kepada pihak lain terkait taraf penguasaan berbagai kemampuan pada aneka konteks kehidupan dan tingkat perkembangan (Broadfoot, 2009). Sesuai tujuannya untuk memberikan apa yang sering disebut *accreditation of real-life practice* alias pengakuan atas kemampuan melakukan praktek dalam kehidupan nyata, maka modus atau cara penilaian yang ditempuh pun tidak lagi mengandalkan teknik tes tulis objektif tradisional seperti yang berlangsung dalam gerakan *testing* di awal perkembangan pengukuran psikologis maupun dalam demam standarisasi di zaman sekarang, melainkan lebih mengakomodasi khususnya teknik nontes yang mampu mengungkap beraneka ragam kemampuan-ketrampilan yang dituntut oleh aneka konteks kehidupan nyata secara lebih otentik (Supratiknya, 2012). Pelaksanaan penilaian hasil belajar semacam ini juga tidak menuntut usaha formal yang dilakukan secara sentralistik

oleh sebuah lembaga atau negara seperti pusat tes ditingkat nasional, regional, maupun global, melainkan lebih mengandalkan upaya *on-demand, personalized 'micro-assessments'* yang dilaksanakan sendiri oleh guru atau sekolah. Hasil-hasil asesmen yang kendati cenderung informal namun tetap serius dan diperoleh di berbagai konteks belajar semacam ini, secara kumulatif akan menjadi sejenis *personal portofolio,* yaitu rekaman prestasi yang menorehkan narasi pribadi tentang pengalaman belajar seseorang sepanjang hayatnya. Spirit semacam ini ternyata juga dicanangkan dan dicoba diimplementasikan dalam Kurikulum 2013 pada jenjang pendidikan dasar dan pendidikan menengah kita, maupun dalam gerakan *student centered learning* pada jenjang pendidikan tinggi kita. Lewat catatan kecil ini diharapkan para pihak terkait di lingkungan komunitas psikologi terbantu untuk menyadari perkembangan mutakhir bidang asesmen khususnya dalam pendidikan sekolah, sehingga mampu menempatkan secara proporsional dan akurat peran dan sumbangan pengukuran psikologis khususnya dalam upaya bersama meningkatkan kualitas proses dan hasil pendidikan sekolah di Tanah Air.

C. Asumsi yang Mendasari Penerapan Tes Psikologis

Menyadari salah satu keunggulan tes psikologis sebagai sarana yang efektif khususnya dalam pemberian layanan asesmen untuk berbagai keperluan di berbagai bidang kehidupan, pada akhir Bab sekaligus akhir buku ini kiranya perlu dikemukakan sejumlah asumsi dasar yang melatari sekaligus menyertai penerapan tes psikologis sebagai sejenis *caveat* atau peringatan untuk berhati-hati agar pihak-pihak terkait dalam komunitas psikologis bisa lebih proporsional dan akurat dalam menggunakan dan menafsirkan hasil pengukuran psikologis. Dari uraian Newland (1973), bisa disarikan lima asumsi dasar yang perlu diperhatikan dalam penggunaan tes psikologis di berbagai bidang kehidupan sebagaimana dipaparkan di bawah ini.

Pertama, orang yang mengadministrasikan tes harus benar-benar mendapatkan pendidikan dan latihan yang memadai untuk menjalankan tugas tersebut. Artinya, dia harus memiliki pemahaman konseptual yang kuat tentang seluk-beluk pengukuran psikologis pada umumnya maupun tes psikologis khususnya, serta memiliki pengalaman yang memadai dalam mengadministrasikan berbagai jenis tes pada berbagai kelompok usia maupun kelompok khusus, mulai dari menjalin *rapport,* menyampaikan instruksi, melakukan *probing* pada jenis-jenis tes tertentu yang lazim bersifat individual sampai ke penskoran dan interpretasinya. Asumsi atau persyaratan ini semakin mendesak untuk dipenuhi dalam pengadministrasian tes dalam rangka membuat diagnosis, khususnya terkait jenis-jenis tes yang bersifat tidak terstruktur atau proyektif.

Kedua dan sebagaimana sudah diuraikan, setiap tes hanyalah merupakan sampel tingkah laku yang memenuhi kriteria memadai secara statistik pada taraf tertentu. Sebagai sampel tingkah laku setiap tes harus memenuhi dua kriteria. Pertama, tes harus memadai dalam hal jumlahnya, baik dalam arti jumlah tugas atau pertanyaan yang dipakai sebagai item-itemnya maupun jumlah subjek yang dipakai dalam pengujian kualitas psikometrik maupun dalam rangka penyusunan normanya. Aspek ini akan menentukan kualitas tes khususnya dari segi reliabilitasnya. Kedua, tes harus memadai dalam merepresentasikan ranah isi dari atribut psikologis yang diukur. Jumlah item yang banyak tidak secara otomatis menjamin terwakilinya seluruh cakupan isi atribut yang diukur secara memadai. Diperlukan *expert judgment* yang cermat serta didukung pembuktian psikometris yang memadai untuk memperoleh tes yang benar-benar mencerminkan populasi isi atribut psikologis yang diukur. Aspek ini akan menentukan kualitas tes khususnya dari segi validitasnya.

Ketiga dan juga sebagaimana sudah diuraikan, senantiasa terdapat kesalahan pengukuran dalam pengukuran psikologis baik yang bersifat sistematis maupun yang bersifat random. Penyusun tes wajib berupaya meminimalisasikan kesalahan pengukuran ini. Untuk jenis kesalahan yang bersifat sistematis, kewajiban tersebut bersifat

mutlak karena ini terkait dengan validitas atau keabsahan penafsiran hasil tes sesuai tujuan tes itu disusun. Untuk jenis kesalahan yang bersifat random pemilihan aneka strategi yang tepat dalam berbagai tahap penyusunan tes maupun penyusunan normanya dapat meminimalkan besar kesalahan random tersebut. Namun, sampai batas tertentu hadirnya jenis kesalahan tersebut memang tak terelakkan dan justru hal inilah yang mendorong perkembangan psikometri. Maka, setiap tes harus disertai dengan informasi yang lengkap terkait berbagai aspek dalam proses penyusunannya baik terkait rancangan maupun kualitas psikometriknya, agar tidak menimbulkan kesesatan dalam penggunaannya.

Keempat, dalam pengetesan yang langsung bisa diamati adalah tingkah laku atau kinerja testi pada masa kini yaitu saat mengerjakan tes, sedangkan tingkah laku sebenarnya yang dicoba diungkap baru akan terjadi kemudian atau di masa mendatang dan hanya baru bisa diinferensikan berdasarkan tingkah laku saat mengerjakan tes di masa kini. Tingkah laku testi saat mengerjakan tes dipengaruhi oleh banyak faktor meliputi antara lain kondisi pribadinya, aparatus atau instrumen yang dipakai, lingkungan fisik tempat dilaksanakannya pengetesan, dan orang yang mengadministrasikan tes. Dengan kata lain, sesungguhnya ada kemungkinan jarak yang lebar antara tingkah laku testi yang bisa diamati pada masa kini dan tingkah lakunya di masa mendatang yang hanya bisa diinferensikan pada masa kini. Administrator tes yang baik harus mempertimbangkan kemungkinan jarak ini saat menginterpretasikan hasil tes dan menyusun inferensi. Meminjam contoh yang dikemukakan oleh Newland (1973), pernyataan bahwa seorang anak tidak melihat kesamaan di antara sejumlah objek adalah observasi, sedangkan pernyataan tentang anak yang sama bahwa dia *tidak mampu* melihat kesamaan di antara sejumlah objek adalah inferensi. Dua pernyataan berdasarkan fakta yang sama tersebut bisa dipisahkan oleh jarak yang sangat lebar bahkan mungkin tidak akan pernah bertemu. Dengan kata lain, inferensi tersebut bisa keliru.

Kelima, karena kebanyakan tes baku yang digunakan di Indonesia merupakan hasil adaptasi dari tes yang dikembangkan di negara lain, kendati terdengar klise namun tetap harus diakui bahwa perbedaan atau kesenjangan antara konteks budaya tempat tes-tes itu dikembangkan di negara asalnya dan konteks budaya setempat di Indonesia tempat tes-tes itu diadaptasikan dan diadministrasikan bisa memperlebar jarak antara observasi dan inferensi sebagaimana dibahas dalam asumsi yang keempat. Ψ

Daftar Acuan

Abdi, H. (2010). Guttman scaling. Dalam Neil Salkind (Ed.), *Encyclopedia of research design.* Thousand Oaks, CA: Sage.

Allen, M.J., & Yen, W.M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: MacMillan.

Anderson, L.W. (1990a). Guttman scales. *The international encyclopedia of educational evaluation* (333-334). Oxford: Pergamon.

Anderson, L.W. (1990b). Likert scales. Dalam H.J. Walberg, & G.D. Haertel (Eds.), *The international encyclopedia of educational evaluation* (334-335). Oxford: Pergamon.

Anderson, L.W., & Krathwohl, D.R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing. A revision of Bloom's taxonomy of educational objectives.* New York: Longman.

Andrich, D. (1990). Thurstone scales. Dalam H.J. Walberg & G.D. Haertel (Eds.), *The international encyclopedia of educational evaluation* (h. 329-332). Oxford: Pergamon.

Azwar, S. (1997). *Reliabilitas dan validitas.* Yogyakarta: Pustaka Pelajar.

Azwar, S. (1998). *Tes prestasi, fungsi dan pengembangan pengukuran prestasi belajar* (ed. ke-2). Yogyakarta: Pustaka Pelajar.

Bakker, A., & Zubair, A.C. (1990). *Metodologi penelitian filsafat.* Yogyakarta: Kanisius.

Blalock, Jr., H.M. (1979). *Social statistics* (Rev. Ed.). New York: McGraw-Hill.

Bloom, B.S., Englehart,M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives. Handbook I: The cognitive domain.* New York: David McKay.

Bray, J.H. (2010). The future of psychology practice and science. *American Psychologist, 65,* 355-369.

Broadfoot, Patricia. (2009). Foreword. Signs of change: Assessment past, present and future. Dalam Claire Wyatt-Smith & J. Joy Cumming (Eds.), *Educational assessment in the 21st century. Connecting theory and practice* (v-xi). New York: Springer.

Browne, M.W. (2000). Psychometrics. *Journal of the American Statistical Association, 95*(450), 661-665.

Burisch, M. (1984). Approaches to personality inventory construction. *American Psychologist*, *39*(3), 214-227.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Carroll, J.B. (1992). Cognitive abilities: The state of the art. *Psychological Science, 3*(5), 266-270.

Crocker, Linda, & Algina, James. (2008). *Introduction to classical and modern test theory.* Mason, OH: Cengage Learning.

Cronbach, Lee J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671-684.

Cronbach, Lee J., & Meehl, Paul E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

De Champlain, Andre F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44,* 109-117.

Departement of International Research and Cooperation, NIER. (2001). *Mutual recognition of qualifications. Practices, challenges, and prospects in university mobility.* Tokyo: Pengarang.

Edwards, A.L. (1959). *Edwards Personal Preference Schedule.* New York: The Psychological Corporation.

Embretson, S.E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher, 36*(8), 449-455.

Friedenberg, L. (1995). *Psychological testing. Design, analysis, and use.* Boston: Allyn & Bacon.

Gardner, H. (1999). *Intelligence reframed: Multiple intelligences for the 21st century.* New York: Basic Books.

Gergen, K.J. (2001). Psychological science in a postmodern context. *American Psychologist, 56,* 803-813.

Gregory, R.J. (2007). *Psychological testing. History, principles, and applications* (5th ed.). Boston: Pearson.

Goodwin, L.D., & Leech, N.L. (2003, October). The meaning of validity in the new standards for educational and psychological testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36,* 181-191.

Guba, E.G., & Lincoln, Y.S. (1994). Competing paradigms in qualitative research. Dalam N.K. Denzin & Y.S. Lincoln (Eds.), *Handbook of qualitative research* (105-117). Thousand Oaks: Sage.

Guilford, J.P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Gulliksen, H. (1974). Looking back and ahead in psychometrics. *American Psychologist, 29,* 251-261.

Haladyna, Thomas M., Downing, Steven M., & Rodriguez, Michael C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Hall, C.S., & Lindzey, G. (1993). *Teori-teori sifat dan behavioristik.* Yogyakarta: Kanisius.

Hambleton, Ronald K., & Jones, Russell W. (1993, Fall). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice,* 38-47.

Krathwohl, D.R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Pratice, 41*(4), 212-218.

Leedy, Paul D., & Ormrod, Jeanne Ellis. (2005). *Practical research. Planning and design* (8th ed.). Upper Saddle River, NJ: Pearson.

Lewandowski, Joseph D. (2001). *Interpreting culture. Rethinking method and truth in social theory.* Lincoln: University of Nebraska Press.

Lilienfeld, S.O., Wood, J.M., & Garb., H.N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, (1)*2, November, 27-66.

Lord, F.M. (1954). Scaling. *Review of Educational Research, 24*(5), 375-392.

Marzano, R.J., & Kendall, J.S. (2007). *The new taxonomy of educational objectives* (2nd ed.). Thousand Oaks, CA: Corwin.

McCormick, E.J., & Ilgen, D. (1980). *Industrial psychology* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

McGrew, K.S. (2009). CHC theory of intelligence and its impact on contemporary intelligence test batteries. Diunduh dari http://www.iapsych.com/articles/chcbrief.pdf

Meehl, P.E. (1945, 1971). The dynamics of "structured" personality tests. Dalam L.D. Goodstein & R.I. Lanyon (Eds.), *Readings inf personality assessment* (245-253). New York: Wiley.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*(3), 215-237.

National Institute for Educational Policy Research. (2001). *Mutual recognition of qualifications: Practices, challenges and prospects in university mobility.* Tokyo: Author.

Newland, T. Ernest. (1973). Assumptions underlying psychological testing. *Journal of School Psychology, 11*(4), 316-322.

Ninnes, Peter, & Hellsten, Meeri. (Eds., 2005). *Internalizing higher education. Critical explorations of pedagogy and policy.* Dordrecht: Springer.

Novick, Melvin R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3,* 1-18.

Nunnally, Jr., J.C. (1970). *Introduction to psychological measurement.* Tokyo: Kogakusha.

Oderberg, David S. (2007). *Real essentialism.* New York: Routledge.

Shermis, M.D., & Di Vesta, F.J. (2011). *Classroom assessment in action.* Lanham: Rowman & Littlefield.

Stevens, S.S. (1935). The operational basis of psychology. *The American Journal of Psychology, 47*(2), 323-330.

Stevens, S.S. (1936). Psychology: The propaedeutic science. *Philosophy of Science, 3*(1), 90-103.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677-680.

Supratiknya, A. (2006). Penyusunan tes kompetensi bidang Psikologi Kepribadian (Sebuah langkah awal). *Widya Dharma, 17*(1), 28-46.

Supratiknya, A. (2012). *Penilaian hasil belajar dengan teknik nontes.* Yogyakarta: Penerbit Universitas Sanata Dharma.

Tarrant, Marie; Knierim, Aimee; Hayes, Sasha K.; & Ware, James (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today,* doi:10.1016/j.nedt.2006.07.006

Teo, T. (2005). *The critique of psychology. From Kant to postcolonial theory.* New York: Springer.

Torrance, Harry. (1981). The origins and development of mental testing in England and the United States. *British Journal of Sociology of Education, 2*(1), 45-59.

Traub, Ross E. (1997, Winter). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice,* 8-14.

Woodcock, R.W. (2002). New looks in the assessment of cognitive ability. *Peabody Journal of Education, 77*(2), 6-22.

Yaremko, R.M., Harari, H., Harrison, R.C., & Lynn, E. (1982). *Reference handbook of research and statistical methods in psyhcology. For students and professionals.* New York: Harper & Row. Ψ

INDEKS

E

F

G

H

I

J

K

L

M

N

O

P

R

S

T

U

V

W

Tentang Penulis

A. Supratiknya, guru besar Psikologi pada Fakultas Psikologi Universitas Sanata Dharma, Yogyakarta. Tamat dari Fakultas Psikologi Universitas Gadjah Mada (B.A., 1977; Drs., 1980) dan dari *Department of Psychology, College of Social Sciences and Philosophy, University of the Philippines,* Diliman (Ph.D., 1992). Pernah mengikuti *Fulbright Visiting Scholar Program* di *Center for Cross-Cultural Research, Department of Psychology, Western Washington University,* Bellingham, Washington, dan *School of Psychology, Florida Institute of Technology,* Melbourne, Florida, Amerika Serikat (2003-2004). Menjadi anggota Himpunan Psikologi Indonesia (HIMPSI) dan *American Psychological Association* (*APA*), serta anggota Masyarakat Karawitan Jawa (*Maskarja*). Menulis dan menerjemahkan sejumlah buku, menulis artikel, dan melakukan penelitian tentang psikologi, dengan perhatian khusus pada psikologi budaya dan pendidikan.