

KOLEKSI BUKU
 TIDAK DIPINJAM
 MHSUS BINA

di FPK IKIP Padang

Pada Tanggal 19 Juni - 12 Agustus 1989

Ujungpandang, dan Padang

Staf Pengajar FPK IKIP Jakarta, Surabaya,

Metode Mengajar/Media Pendidikan Tingkat Lanjutan

Disampaikan pada Penataran:

CALL NO	371.06 Kum PD
NO. INVENTARIS	1219 (No. 191-10/89)
KOLEKSI	KCI
SUMBER HARSA	TRAJAH
DITEMPAI	Juni 1991
MILIK PBT PERPUSTAKAAN IKIP PADANG	

Kurnaldi, M., Phd
 FPK IKIP Padang

Dth:

IKIP PADANG

PERPUSTAKAAN

PENGANTAR PENILAIAN PENDIDIKAN

PENGANTAR EVALUASI PENDIDIKAN

Oleh: Kumaidi
FPTK IKIP Padang

UNIT PERPUSTAKAAN
IKIP PADANG

PENDAHULUAN

Berbagai kritik sering kita dengar melalui media masa terhadap testing atau pengukuran dan penilaian pendidikan yang ada saat ini. Kritik tersebut di samping ditujukan kepada bentuk item yang dipakai, sampai dengan tidak adanya standar atau acuan yang pasti dalam menetapkan angka indeks (nilai) yang dipakai. Kritik yang lain dapat pula dilontarkan terhadap kenyataan bahwa tingkat kesukaran masing-masing set butir soal (tes) dari tempat ke tempat lainnya tidaklah sama. Kritik-kritik ini menunjukkan bahwa teknologi testing yang ada dan kita kuasai saat ini masih perlu ditingkatkan lagi.

Kritik terhadap bentuk butir soal (item) yang dipakai biasanya ditujukan terhadap pemakaian bentuk Benar-Salah dan Pilihan-Berganda (atau yang dikenal dengan bentuk tes Objektif). Pemakaian bentuk butir soal ini, di samping keunggulannya terhadap bentuk lain---esei---, memiliki berbagai kelemahan sebagaimana yang di "over" emphasized" oleh para pengkritik. Padahal dari berbagai penelitian di USA (tempat berkembangnya teknologi testing ini) ditemukan bahwa hasil pengesanan dengan bentuk item pilihan berganda dan esei sangat tinggi, yang dapat disimpulkan bahwa apabila pemakaian bentuk tes tersebut sesuai dengan tujuan pengtesannya maka pilihan berganda dan esei menunjukkan hasil yang tidak jauh berbeda. Problem yang sering dijumpai biasanya terletak pada

kemampuan kita untuk menulis butir soal yang "baik" masih perlu ditingkatkan. Para guru kita masih belum menguasai skill atau ketrampilan menulis item sebagaimana yang diharapkan. Akibatnya banyak item bentuk pilihan berganda yang hanya mengukur tingkat "know ledge" (pengetahuan) dan comprehension (pemahaman) dalam klasifikasi tujuan pengajaran menurut Bloom. Dua tingkatan pemahaman ini merupakan dua tingkatan terendah dari klasifikasi Bloom.

Dua kritik terhadap pemakaian testing di dunia pendidikan yang berkaitan dengan standar yang mantap dan tingkat kesulitan yang sebanding dari daerah ke daerah lainnya ini perlu penanganan yang tidak kalah seriusnya dengan isue pertama. Disini diperlukan suatu badan yang dapat mengembangkan butir soal yang terkalibrasikan, sebagaimana alat ukur yang kita kenal sehari-hari pun perlu dikalibrasi. Dengan demikian setiap set item atau setiap tes dapat diskalakan kepada referensi metrik yang sama. Teknologi ini sudah cukup berkembang, mulai dari teori klasik sampai dengan teori modern yang dikenal dengan teori response (item response theory), dan teknologi ini dikenal sebagai "equating" dalam ilmu pengukuran. Secara sederhana "equating" atau penskalaan ini dapat dicontohkan dengan pengukuran temperatur yang dapat dilakukan dengan skala Celcius dan/atau skala Fahrenheit. Pengukuran suhu dengan dua macam skala ini walaupun medium yang diukur sama akan menghasilkan angka (skala) yang berbeda. Akan tetapi dengan diketahuinya hubungan atau pola transformasi skalanya maka dua hasil pengukuran dengan skala yang berbeda ini dapat dikonformasikan kepada skala standar

(baik menurut Celcius maupun Fahrenheit, sebagai skala standarnya). Apalagi dengan pengukuran fisika maka pengukuran kemampuan dapat pula dilaksanakan.

Makalah ini tidak akan membicarakan berbagai macam teknologi yang diceritakan di atas, akan tetapi hanya sekedar menunjukkan bahwa pengukuran pendidikan telah sampai pada suatu tahapan yang dapat mengatasi berbagai kesulitan. Ini tidak berarti semua kesulitan telah dapat diatasi. Makalah ini akan mencoba mereview konsep-konsep yang telah dibicarakan, pada penataran tingkat pertama yang lalu, menambah dengan beberapa isu lain yang relevan, dan memberi berbagai "flavor" atau rasa tentang dasar psikometrika yang melandasi berkembangnya pengukuran pendidikan. Dasar psikometrika ini akan dilanjutkan ketika kita akan bertemu lagi dalam topik analisa tes hasil belajar.

REVIEW KONSEP PENILAIAN PENDIDIKAN

Pada awal pembicaraan kita terdahulu, kita cenderung membedakan pengukuran dari penilaian. Pengukuran merupakan suatu usaha pengumpulan informasi tentang individu dalam bentuk kuantitatif (mengikuti aturan main tertentu). Sedangkan penilaian (evaluasi) merupakan usaha pemberian arti atau makna dari informasi yang terkumpul. Informasi yang digunakan untuk penilaian tidak selamanya dikumpulkan melalui pengukuran akan tetapi dapat pula bersumber dari usaha lain. Dengan demikian penilaian akan memanfaatkan semua sumber informasi baik yang kuantitatif maupun yang kualitatif.

Pemberian makna dalam tahap evaluasi memerlukan acuan

tertentu. Dalam penataran tingkat pertama telah dibicarakan dua macam prosedur penilaian (interpretasi skor) melalui apa yang dikenal sebagai Penilaian Acuan Norma (PAN) dan Penilaian Acuan Patokan (PAP). PAN memakai acuan kelompok (misalnya kelompok pengambil tes bidang studi tertentu); sedangkan PAP memakai acuan patokan yang ditetapkan sebelum pengetesan dilaksanakan. PAP biasanya memakai acuan obyektif atau tujuan (sasaran) pengajaran untuk menetapkan tingkat keberhasilan seorang murid, contohnya anak yang dapat menguasai minimal 90% tujuan pengajaran (bahan ajar) akan diberi predikat A (sangat baik).

Penilaian hasil belajar pada dasarnya memiliki berbagai fungsi yang dapat diraih (capitalized) dari diselenggarakannya testing di sekolah/kelas. Salah satu fungsi yang sangat penting adalah untuk evaluasi formatif, pemantauan keberhasilan membelajarkan murid dalam rangka mencapai tujuan pengajaran. Guru tentu sangat berkepentingan dengan pemantauan proses pembelajaran tersebut, karena guru memang diangkat dan digaji untuk melaksanakan proses pembelajaran bagi muridnya. Dengan usaha ini maka profesionalitas guru akan dapat terus ditingkatkan. Fungsi lainnya yang tak kalah pentingnya dengan fungsi yang disebutkan di atas adalah fungsi diagnostik dari tes hasil belajar. Fungsi diagnostik dari penilaian hasil belajar ini akan sangat bermanfaat untuk menggalakkan usaha peningkatan mutu yang saat ini sedang hangat-hangatnya dibicarakan di negara kita. Fungsi diagnostik pada dasarnya berusaha untuk mengenali kelemahan dan keunggulan masing-masing individu murid agar pencapaian

tujuan belajar dapat dioptimalkan.

Di samping kedua fungsi tersebut masih terdapat berbagai fungsi penilaian hasil belajar yaitu grading, penempatan, dan seleksi. Masing-masing fungsi ini telah kita bicarakan pada kesempatan pertama penataran kita ini.

Ada beberapa konsep yang juga disinggung dalam penataran terdahulu adalah konsep penilaian formatif dan sumatif. Di kalangan guru SMTA evaluasi formatif selalu diartikan sebagai penilaian atau tes yang dilaksanakan sebelum semesteran berakhir. Dengan kata lain, penilaian formatif dikaitkan dengan waktu pelaksanaan ujian yaitu tengah atau mid-semester. Dalam kajian kita yang terdahulu, pengertian penilaian formatif tidak dikaitkan dengan waktu pelaksanaannya akan tetapi lebih kepada fungsi yang akan dicapai oleh proses evaluasi tersebut. Penilaian formatif ditujukan untuk mengetahui hambatan-hambatan dalam pelaksanaan proses pembelajaran dalam kelas agar dapat dicapai hasil (belajar) yang optimal. Penilaian sumatif biasanya dikaitkan dengan keperluan untuk penentuan hasil akhir dari suatu proses pembelajaran. Dalam kaitannya dengan penilaian hasil belajar penentuan hasil akhir ini biasanya ditandai dengan penentuan nilai akhir atau "grading" bagi murid. Oleh sebab itu, penilaian sumatif sebenarnya tidak selalu dilaksanakan di akhir semester (karena nilai akhir murid tidak sekedar ditentukan oleh satu kali ujian, yaitu ujian semesteran), akan tetapi kapan saja ujian itu dilaksanakan yang tujuannya adalah meng"grade" keberhasilan belajar murid.

Konsep terakhir yang dijelaskan secara singkat dalam penataran yang lalu (pertama) adalah konsep reliabilitas (keterangan) dan validitas (kesahihan) skor suatu tes. Posisi yang diambil oleh penulis makalah ini adalah reliabilitas skor dan validitas skor, sebelum kita bisa menarik "inferences" bahwa tes dimaksud menghasilkan skor atau instrumen yang reliabel (terandalkan) dan valid (sahih) untuk tujuan tertentu. Hal ini dikaitkan dengan asumsi bahwa tes (hasil belajar) hanya merupakan sampel (contoh) kemampuan yang dimiliki oleh siswa. Oleh sebab itu, tes biasanya mengacu kepada suatu populasi tingkah laku atau domain tertentu, yang dalam dunia pertestingan populasi item ini dikenal dengan "universe of behaviors".

Reliabilitas skor suatu tes berkaitan dengan ketelitian skor yang dihasilkan, sedangkan validitas skor kesesuaian dengan yang ingin dipenuhi oleh hasil pengujian tersebut. Reliabilitas skor ini sejalan dengan konsep kesalahan atau kemelesetan pengukuran (error of measurement). Reliabilitas dalam praktek tidak akan dapat diukur akan tetapi dapat diprakirakan (estimated). Berbagai model estimasi (prakiraan) reliabilitas akan dibicarakan pada waktu kita akan bertemu kembali dalam analisis tes. Model yang sangat banyak dikenal oleh kalangan pendidikan di Indonesia adalah model split-halves (belah dua) dan model Kuder-Richardson (KR) formula 20 dan 21. Validitas skor suatu tes, ini disesuaikan dengan tujuan pemakaian hasil tesnya. Validitas ini kita definisikan mengikuti Test Standards 1985 yaitu "the appropriateness, meaningfulness, and usefulness of the specific in-

ferences made from test scores"--- kesesuaian, keberartian, dan kemanfaatan dalam pengacuan yang ingin dibuat dari skor-tes yang dihasilkan --- (APA, NCME, dan AERA, 1985, halaman 9). Definisi diatas dapat dengan mudah dikaitkan dengan model pemalidasian test yang bisa kita kenal yaitu validitas isi, konstruk, dan prediksi (atau berdasar kriteria dari luar tes skor). Untuk tes hasil belajar mungkin kita dapat mengikuti pendapat Ebel (1983) yang menyatakan bahwa, untuk tes hasil belajar, validitas isi (content validity) yang lebih mendasarkan pada logical aspects of validation process merupakan model pemvalidasian yang terpenting (walaupun pendapat ini mendapat kritik dari ahli lain, misalnya Flanagan). Berbagai perbincangan yang lebih rinci akan kita temui dalam analisis tes nanti.

Isue lainnya

Isue lainnya yang belum dibicarakan dalam penataran tingkat awal lalu adalah isue-isue yang dapat ditunjukkan dalam berbagai pertanyaan berikut. Apa yang akan diukur dalam tes hasil belajar? Apakah prestasi belajar dapat diukur? Dan bagaimana guru dapat mempertahankan obyektivitas suatu tes hasil belajar? Isue ini saya rasa tidak kalah pentingnya dengan berbagai isue yang telah disampaikan di muka.

Isue apa yang akan diukur oleh suatu tes hasil belajar, tentu dihadapi oleh setiap guru (dosen). Isue ini terutama muncul pada waktu seorang guru akan membuat tes hasil belajar. Seorang guru yang sedang mempersiapkan pembuatan tes hasil belajar (baik untuk ujian mid-semester atau ujian

akhir semester) harus menentukan apa yang akan diukurnya. Bagi seorang dosen/guru masalah kemampuan atau tingkah laku apa yang harus diukur tentu dikaitkan dengan penguasaan materi ajar yang diberikan selama selang waktu pembelajaran yang telah dilalui bersama muridnya. Sebagai awal dari jawaban apa yang harus diukur oleh guru, maka kurikulum serta silabus yang dikembangkannya (yang biasanya berisikan TIU dan TIK) merupakan pengangan yang dapat dipedomani. Kurikulum dan silabus dengan TIU dan TIKnya merupakan sumber penjabaran tujuan yang ingin dicapai oleh proses pembelajaran dalam kelas. Mengingat dalam sistem yang kita anut, yang terdiri dari kegiatan tatap muka, belajar terstruktur, dan belajar mandiri, maka mungkin saja aspek tingkah laku atau penguasaan materi ajar ada yang harus dikuasai murid, seharusnya dikuasai murid, dan mungkin dapat dikuasai murid. Aspek apa yang akan diukur dalam tes hasil belajar sebaiknya mengacu kepada ketiga aspek/ciri sistem semester tersebut. Karena penjabaran materi ajar dapat saja sangat luas, maka sang guru/dosen sebaiknya memilih konsep utama yang terkandung dalam unit pengajaran atau pokok bahasan. Penguasaan konsep utama dalam unit pengajaran atau pokok bahasan tersebut harus dituangkan dalam tingkatan pemahaman dan/atau keterampilan yang seharusnya dimiliki oleh murid/mahasiswa yang telah sukses dalam proses pembelajarannya. Dengan demikian terdapat relevansi antara apa yang akan diukur oleh suatu tes hasil belajar dengan apa yang diajarkan dalam kelas.

Apabila kita ingin meningkatkan relevansi antara apa yang akan diukur dengan kebutuhan tenaga kerja di luar

sekolah, maka penyermatan terhadap kurikulum dan silabus (serta TIU dan TIKnya) setiap kali ingin mengembangkan tes hasil belajar sangatlah penting. Kadang kala kurikulum dan silabus yang dikembangkan oleh guru sudah ketinggalan jaman, oleh sebab itu kerjasama dengan pihak industri tentu sangat membantu, terutama untuk pendidikan kejuruan dan teknologi yang kita tekuni ini. Di sinilah perlunya dilibatkan pihak sekolah kejuruan tingkat atas (misalnya STM, SMKK, dsb) dalam pengujian calon guru yang FPTK hasilkan mungkin perlu dipikirkan agar tidak hanya terbatas pada praktek lapangan kependidikan dan keguruan saja. Ada kalanya, kurikulum dan silabus tidak begitu jelas jabarannya. Ini tentu akan banyak ditemui oleh guru/dosen muda (asisten dosen) yang masih perlu supervisi oleh dosen seniornya. Disinilah "moderating" (atau pemvalidasian isi tes) perlu dilaksanakan. Mungkin kita akan melaksanakan simulasi pemvalidasian isi lewat prosedur moderating ini nanti ketika kita bertemu lagi dalam analisis tes hasil belajar.

Kemampuan yang mungkin diukur sebagai manifestasi dari hasil pembelajaran dapat meliputi aspek kognisi, psikomotor, dan afeksi serta ketrampilan mengkomunikasikan gagasan (lewat gambar atau sketsa, misalnya). Aspek kognisi sudah banyak kita kenal dan sudah demikian berkembangnya, dengan model klasifikasi (taxonomy) tujuan pengajaran Bloom yang mendarah-daging pada guru-guru kita. Aspek psikomotor jelas bukan barang baru lagi bagi kita semua, yang biasanya ditunjukkan dengan penguasaan ketrampilan motorik atau manual yang biasanya diajarkan oleh sekolah kejuruan dan teknologi.

Aspek afeksi yang biasanya berkaitan dengan sikap, keyakinan, dan watak murid belum terlalu banyak mendapat perhatian. Aspek afeksi ini mungkin hanya dapat diobservasi dari tingkah laku murid. Aspek ketrampilan berkomunikasi, baik lisan maupun lewat media lain, sering diuji sesuai dengan tujuan penilaiannya. Misalnya, kemampuan berkomunikasi seorang perancang kepada juru teknik biasa diwujudkan dalam bentuk gambar sketsa atau gambar kerja dengan segala tanda gambarnya menurut sistem tertentu.

Isue dapat tidaknya prestasi belajar diukur menurut penulis makalah ini merupakan isue psikologi pendidikan yang menerobos kepada keahlian pengukuran pendidikan. Kenapa isue ini saya nyatakan sebagai isue psikologi pendidikan? Alasannya adalah prestasi belajar merupakan suatu konstruk (construct) atau konsep yang levelnya sama dengan pertanyaan tentang 'apa yang dimaksud dengan belajar' itu sendiri. Jawaban terhadap isue seperti ini sangat tergantung pada pemahaman kita tentang proses pendidikan itu sendiri, yang pada saat ini pemahaman tersebut masih jauh dari sempurna. Oleh sebab itu kemampuan kita mengukur berbagai aspek (sebagai hasil) proses pembelajaran juga masih sangat terbatas. Prestasi belajar hanya dapat diukur apabila kita dapat mengetahui semua aspek penting yang ditimbulkan oleh adanya usaha pendidikan itu. Apabila kita yakin dapat menunjukkan adanya pengaruh atau dampak proses pendidikan (pembelajaran), maka pada kondisi itu kita harus yakin akan dapat mengukur prestasi belajar. Seandainya kita sepakat untuk mengatakan bahwa terdapat perbedaan antara orang yang telah belajar

(suatu keahlian) dengan orang yang belum belajar (suatu keahlian), dan timbulnya perbedaan ini diakibatkan oleh proses pembelajaran, maka prestasi belajar akan dapat diukur. Pengukuran (dan akhirnya penilaian) prestasi belajar ini mungkin tidak selamanya menggunakan bentuk ujian tertulis, akan tetapi dapat juga memakai bentuk ujian lainnya, termasuk bentuk ujian yang mungkin saat ini belum diketahui oleh ahli-ahli pengukuran dan evaluasi pendidikan ataupun ahli pendidikan lainnya.

Isue obyektivitas dalam pengukuran dan penilaian pendidikan sering disinggung sebagai sesuatu yang sulit (kalau tidak mungkin) dicapai. Pendapat penulis makalah ini tidaklah demikian, sebab menurut penulis makalah, semua tes (penilaian) hasil belajar haruslah obyektif (tidak identik dengan harus memakai tes yang kita kenal sebagai tes obyektif). Isue obyektivitas dalam evaluasi hasil belajar mungkin dapat disejajarkan dengan isue keadilan dalam kehidupan sosial sehari-hari. Obyektivitas penilaian hasil belajar dapat dicapai apabila kita sepakat bahwa kurikulum dan silabus pengajaran merupakan satu-satunya sumber pengembangan tes hasil belajar. Kurikulum dan silabus ini biasanya dikembangkan dan dijabarkan lagi menjadi TIK yang ingin dicapai dalam proses pembelajaran dalam kelas. Setiap TIK harus mendapat perhatian yang memadai dalam pengembangan tes hasil belajar, sehingga level pemahaman yang tertinggi dari aspek-aspek yang berbeda dapat dibuatkan butir soal (item)nya. Paling tidak contoh atau sampel tingkah laku yang harus diperlihatkan oleh murid yang sukses diacu dalam pengembangan tes

hasil belajar dimaksud. Dari uraian ini tergambar ide atau konsep obyektivitas suatu tes hasil belajar seharusnya dicermati dari unsur apa yang diteskan sehingga ditemukan apakah materi tes sesuai dengan materi ajar yang diberikan di dalam kelas. Materi ajar yang diberikan dalam kelas mestinya sesuai dengan isi kurikulum dan silabus pengajarannya. Isue ini oleh Calfee (1983) disebut sebagai isue validitas kepengajaran (instructional validity) atau oleh Linn (1983) juga disebut sebagai validitas kurikuler (curricular validity).

Pengantar Dasar Teori Psikometrika

Untuk dapat memahami konsep analisis item dan tes hasil belajar dengan baik, kita dituntut untuk memahami teori yang dipakai untuk dasar analisis tersebut. Teori testing atau pengukuran hasil belajar ini disebut sebagai teori psikometrika, salah satu cabang teori ini disebut sebagai teori klasik pengukuran (classical test theory). Teori ini disebut teori klasik karena memang sudah cukup tua umurnya. Cabang teori terbaru dalam bidang psikometrika biasanya dikenal sebagai teori modern atau teori response (item response theory), salah satu modelnya dikenal sebagai "Rasch model".

Untuk keperluan kita di sini, hanya akan diperkenalkan sedikit konsep pemikiran teori klasik, yang karena umurnya (mungkin) yang paling mudah dipahami. Dasar teori pengukuran ini adalah statistik terpakai (applied statistics) dalam bidang psikologi dan pendidikan. Bunyi teori ini sangat sederhana yaitu hasil pengukuran yang teramati (observed scores) selalu terdiri dari dimensi (skor) sebenarnya (true

scores) ditambah dengan kemelesetan/kesalahan pengukuran (error of measurement). Secara matematik teori ini dinotasikan sebagai :

$$X = T + E$$

dimana : X = skor teramati

T = skor sebenarnya

E = kemelesetan/kesalahan pengukuran

Dari teori ini dapat dipahami bahwa skor teramati dipengaruhi oleh kesalahan pengukuran (apapun sumber kesalahan tersebut). Untuk memberi gambaran sederhana terhadap teori dasar ini dapat dicontohkan apabila kita ingin mengukur benda tertentu, misalnya panjang meja makan dengan memakai rol meter (meter gulung). Apabila kita mengukur meja tersebut berulang-ulang, maka hasil ukur yang tepat sama hampir pasti tidak ditemukan. Hasil pengukuran yang teramati ini menunjukkan adanya unsur kesalahan dalam pengukuran, baik yang disebabkan oleh kecerobohan orang yang mengukur dalam membaca skala meterannya ataupun faktor lainnya. Oleh sebab itu ukuran yang sebenarnya mungkin tidak pernah ditemukan, tetapi dapat diprakirakan.

Untuk memperkirakan skor sebenarnya dari kemampuan seorang murid menurut pemahaman teori klasik ini diasumsikan bahwa rata-rata skor teramati, apabila dilakukan pengukuran berulang-ulang dengan instrumen yang sama dan setiap kali habis mentes kemampuan murid tersebut di "kembalikan" kepada keadaan awalnya. Oleh sebab itu, T ini diartikan sebagai rata-rata hitung dari distribusi teoritis X yang dihasilkan dari pengukuran yang berulang-ulang dengan instrumen (tes)

yang sama kepada murid yang sama pula. Dalam praktek tentu ini hampir tidak mungkin dilaksanakan.

Selanjutnya, asumsi yang harus diterapkan adalah tidak terdapat korelasi antara T dan E. Artinya kesalahan pengukuran yang terjadi independen dari pengaruh skor sebenarnya (atau kemampuan yang diukur) pada tingkat populasi. Apabila dalam praktek terlihat adanya korelasi (yang sangat rendah), maka korelasi yang terjadi haruslah dipahami sebagai "chance error" dari sampling responden. Dapat pula dipahami bahwa dalam pengukuran ini tidak terjadi kesalahan terstruktur, yang ada adalah kesalahan acak dari proses pengukurannya. Implikasi dari asumsi ini adalah variance (variasi) masing-masing komponen yang membentuk rumusan teori klasik adalah independen. Implikasi terakhir ini dapat dituliskan secara matematis sebagai :

$$\text{Var}(X) = \text{Var}(T) + \text{Var}(E)$$

dimana : $\text{Var}(X)$ = variasi skor ternati

$\text{Var}(T)$ = variasi skor sesungguhnya

$\text{Var}(E)$ = variasi kesalahan pengukuran.

Teori variasi ini akan merupakan dasar untuk analisis tes hasil belajar, terutama dalam memprakirakan (mengestimasi) besarnya koefisien reliabilitas dan kesalahan pengukuran.

Asumsi selanjutnya menyebutkan bahwa apabila dua tes hasil belajar yang ditunjukkan untuk mengukur kemampuan yang berbeda (misalnya mengukur kemampuan matematik dan berbahasa) maka korelasi kesalahan pengukuran dari kedua tes hasil belajar tersebut sama dengan nol. Dengan kata lain, distribusi kesalahan pengukuran untuk dua tes hasil belajar yang

berbeda haruslah independen. Asumsi ini juga menempatkan kita pada pemahaman bahwa apabila seorang murid memiliki kesalahan pengukuran positif pada tes pertama, maka kita tidak mempunyai alasan untuk memprediksikan bahwa kesalahan pengukuran murid tersebut pada tes kedua akan positif atau negatif. Asumsi ini tentu tidak berlaku atau mungkin kurang tepat dipakai untuk tes hasil belajar yang menghasilkan skor yang dipengaruhi oleh faktor-faktor kesalahan, hasil latihan mengerjakan tes pertama (practice effects), mood (kestabilan) pengambil tes, dan kondisi lingkungan tempat pengujian dilaksanakan. Apabila berbagai faktor ini ikut mempengaruhi skor yang dihasilkan oleh suatu tes hasil belajar, kita harus mencoba untuk menyebarkan pengaruh tersebut secara merata kepada setiap murid andaikata memang tidak dapat dikontrol atau dihilangkan keberadaannya. Asumsi selanjutnya yang juga mendasari berlakunya teori klasik adalah tidak berkorelasinya kesalahan pengukuran pada tes yang satu dengan skor sesungguhnya dari tes yang lain. Asumsi ini kelihatannya akan dilanggar bila tes kedua merupakan tes ranah efektif (misalnya sikap atau keyakinan) yang kadangkala mempengaruhi hasil tes pertama (misalnya tes pertama ini tentang minat kejuruan ataupun mungkin tes kemampuan matematik).

Apabila kita berfikir dalam konteks tes paralel (yaitu dua tes yang sama tetapi memakai item yang berbeda) maka kita perlu juga menambah asumsi yang berkaitan dengan paralelisme dua tes tersebut. Asumsi ini dapat dinyatakan sebagai berikut:

$$X = T + E \quad (\text{tes pertama})$$

$$X' = T' + E' \quad (\text{tes paralel terhadap tes pertama})$$

maka: $T = T'$ dan $\text{Var}(E) = \text{Var}(E')$ untuk semua murid yang mengambil kedua tes tersebut. Dari definisi tes paralel ini juga tergambar bahwa kedua tes yang paralel tersebut memiliki rata-rata hitung (mean) skor teramati yang sama, variasi skor yang sama, dan korelasi dengan skor kriteria yang sama pula. Dua tes yang paralel ini tidak harus memiliki korelasi yang sempurna.

Asumsi terakhir yang sering dikaitkan dengan teori klasik adalah berkenaan dengan dua tes yang setara. Konsep ini dimunculkan oleh para ahli psikometrika untuk mengatasi kesulitan mereka mengembangkan dua tes yang benar-benar paralel, karena beratnya memenuhi asumsi yang berlaku untuk kondisi tes paralel ini. Tes setara pada dasarnya tes paralel yang dimodifikasi dengan melepaskan beberapa asumsi dasar keparalelan tes paralel. Salah satu asumsi yang di "perlunak" kan dalam konsep tes setara ini adalah tidak diperlukannya keharusan dua tes tersebut memiliki rata-rata hitung (mean) skor teramati yang sama, akan tetapi diganti dengan suatu konstante yang additive (menambah) sifatnya terhadap mean tes pertama. Di samping itu, tes setara juga tidak memiliki keharusan bervariasi yang sama dan tes yang satu dapat saja memiliki ketelitian yang lebih tinggi dalam mengukur skor sebenarnya dari tes lainnya.

Dengan bekal penambahan ini kita akan melangkah lebih lanjut kepada analisis tes hasil belajar. Konsep-konsep dasar psikometrika sederhana ini akan melandasi pengembangan

teori dan rumusan analisis tes hasil belajar dengan segala kompleksitasnya. Di samping itu, pemahaman kita tentang statistik elementer juga diperlukan. Oleh sebab itu, bagi peserta penataran yang memerlukan persiapan penguasaan dasar statistik elementer dianjurkan untuk mereview beberapa konsep dasar statistik misalnya rata-rata hitung, deviasi standar, dan korelasi antara dua variabel, di samping juga perlu dipahami konsep varian suatu variabel dan gabungan variabel (composite variables).

Konsep dasar teori klasik ternyata merupakan model statistik (matematis) dalam usaha kita mengenal proses pengukuran hasil pendidikan. Model tersebut merupakan model linier bentuk additive (tambah) antara skor sebenarnya dan skor kesalahan/kemelesetan pengukuran yang terlibat dalam bentuk rancu pada pengukuran teramati (observed). Model ini menurut penulis makalah merupakan model yang paling sederhana dan paling mudah dipahami. Model lain tentu memiliki kekuatan dan kelemahan masing-masing yang tidak selayaknya dibicarakan di sini. Bagi yang berminat untuk mendalami teori ini bisa konsultasi dengan bukunya Lord dan Novick (1968) untuk level lanjutan atau bukunya Crocker dan Algina (1986) untuk tingkat pemula.

KEPUSTAKAAN

- APA, NCME & AERA. (1985). Standards for educational and psychological testing. Author.
- Allen, M.J. & Yen, W.M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole Publishing Company.
- Calfee, R. (1983). Establishing instructional validity for minimum competency programs. Dalam George F. Madaus (Ed.), The courts, validity, and minimum competency testing, halaman 95-113. Boston, MA: Kluwer-Nijhoff.
- Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston.
- Ebel, R. L. (1983). The practical validation of tests of ability. Educational Measurement: Issues and Practice, 2, 7-10.
- Ebel, R. L. & Frisbie, D. A. (1986). Essentials of educational measurement (4th. edition). Englewood Cliffs, NJ: Prentice Hall.
- Linn, R. L. (1983). Curricular validity: Convincing the courts that it was taught without precluding the possibility of measuring it. Dalam George F. Madaus (Ed.), The courts, validity, and minimum competency testing, halaman 115-132. Boston, MA: Kluwer-Nijhoff.
- Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.
- Ward, C. (1980). Designing a scheme of assessment. Cheltenham, UK: Stanley Thornes.

