PANDUAN PRAKTIKUM PENILAIAN, EVALUASI, DAN REMEDIASI HASIL BELAJAR BIOLOGI

Dr. Bambang Subali, M.S.

Jurusan Pendidikan Biologi

Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Negeri Yogyakarta

2010

KATA PENGANTAR

Puji dan syukur saya panjatkan ke hadirat Allah sehinga buku diktat Panduan Praktikum Penilaian, Evaluasi,dan Remediasi Pembelajaran Biologi dapat saya perbaiki disesuaikan kemajuan dan perkembangan ilmu dibidang penilaian, evaluasi dan remediasi.

Sebagaimana diketahui bersama bahwa dengan adanya stándar nasional pendidikan, maka stándar penilaian juga menjadi salah satu produk hukum yang arus dipatuhi oleh guru. dengan demikian, calon guru juga perlu memahaminya dan mampu melaksanakan prosedur pengembangan instrumen beserta pemenuhan bukti baik secara kualitatif maupun secara kuantitatif. Buku panduan praktikum ini diharapkan jadi pegangan bagi mahasiswa dalam menyusun instrumen, dan melakukan análisis ítem secara kualitatif dan secara empiris untuk mendukung validitas instrumen.

Kritik dan saran sangat saya perlukan untuk penyempurnaan buku ini ke depan.

Yogyakarta, 2010

Penulis

DAFTAR ISI

halaman

HALAMAN JUDUL …… 1

KATA PENGANTAR …… 2

DAFTAR ISI …… 3

BAB I. PENYIAPAN KISI-KISI ……...…… 4

BAB II. PENYUSUNAN INSTRUMEN HASIL BELAJAR …… 7

BAB III. ANALISIS ITEM SECARA KUALITATIF …….. 23

BAB IV. ANALISIS ITEM SECARA EMPIRIS SEBAGAI BUKTI VALIDITAS ……..…….. 25

BAB V. PENENTUAN INDEKS PERSETUJUAN DAN INDEKS KAPPA ……..……... 67

BAB I
PENYIAPAN KISI-KISI

Karena fungsi penilaian hasil belajar untuk mengetahui seberapa jauh “kompetensi yang ditargetkan” telah tercapai, maka kunci utama dalam melakukan penilaian adalah ketepatan dalam merumuskan indikator pencapaian kompetensi. Indikator tersebut menjadi kesatuan dengan Kompetensi Dasar yang ditargetkan untuk dicapai. Dengan kata lain, dalam melakukan penilaian harus diawali dengan perencanaan berupa menyusun kisi-kisi penilaian.

Dalam panduan penilaian yang diterbitkan oleh BSNP tahun 2007 dinyatakan bahwa kisi-kisi penilaian merupakan bagian yang tak terpisahkan dari kegiatan perencanaan pembelajaran dalam bentuk silabus dan rencana pelaksanaan pembelajaran (RPP). Di dalam silabus, pendidik menunjukkan keterkaitan antara SK, KD, materi pokok/materi pembelajaran, alokasi waktu, sumber belajar di satu sisi, dengan indikator pencapaian KD yang bersangkutan beserta teknik penilaian dan bentuk instrumen yang digunakan. Teknik penilaian dan bentuk instrumen dapat dituliskan dalam satu kolom, dan dapat pula dituliskan pada kolom yang berbeda. Berikut ini disajikan contoh format kisi-kisi penilaian yang menyatu dengan silabus.

Silabus Pembelajaran

Sekolah : .....
Mata Pelajaran : .....
Kelas/Semester : .....
Standar Kompetensi : .....

Kompetensi Dasar	Materi Pokok/ Materi Pembelajaran	Kegiatan Pembelajaran	Indikator Pencapaian	Penilaian		Alokasi Waktu	Sumber Belajar
				Teknik Penilaian	Bentuk Instrumen		

Perencanaan penilaian yang sudah dilengkapi dengan contoh instrumen disajikan secara menyatu dengan Rencana Pelaksanaan Pembelajaran (RPP). Berikut ini adalah contoh kisi-kisi penilaian yang menyatu dengan RPP.

Rencana Pelaksanaan Pembelajaran (RPP)

Sekolah : .....
Mata Pelajaran : .....
Kelas/Semester : .....
Alokasi Waktu : … jam pelajaran (… x pertemuan)
A. SK :.....
B. KD : .....
C. Materi Pembelajaran : .....
D. Model/Metode Pembelajaran : .....
E. Skenario/Langkah-langkah Kegiatan Pembelajaran
Pertemuan 1 : .....
Pertemuan 2 : .....
dst.
F. Sumber Belajar : .....
G. Penilaian

Indikator Pencapaian	Teknik Penilaian	Bentuk Instrumen	Contoh Instrumen
Berupa indikator yang ada di dalam rumusan silabus sesuai dengan KD yang bersangkutan	Dipilih sesuai dengan karakteristik indikator pencapaian, seperti tes tertulis, tes lisan, tes kinerja, dan portofolio.	Dipilih sesuai dengan teknik penilaian yang dipilih, misalnya memilih bentuk pilihan ganda untuk teknik penilaian tertulis atau memilih bentuk instrumen lembar penilaian portofolio untuk teknik penilaian portofolio.	Disusun sesuai dengan bentuk instrumen yang telah dipilih.

Format di atas sangat efektif bagi guru dalam merancang RPP karena guru tidak perlu meniliskan rumusan indikator dua kali. Ndikator cukur dipaparkan dalam kolom penilaian.

Untuk menilai pencapaian standar kompetensi dalam satu semester, pendidik merancang penilaian untuk semester yang bersangkutan. Kisi-kisi ulangan akhir semester memuat SK, KD, dan indikator pencapaiannya yang dapat dijadikan dasar penyusunan tes pada akhir semester. Kisi-kisi ulangan akhir semester dapat dirancang dengan memuat tes tertulis dan tes praktik yang formatnya disajikan sebagai berikut.

Kisi-Kisi Ulangan Akhir Semester

Sekolah : .....
Mata Pelajaran : .....
Kelas/Semester : .....
Alokasi waktu : .....

Standar Kompetensi	Kompetensi Dasar	Indikator Pencapaian	Teknik Penilaian	
			Tes Tertulis	Tes Praktik
Dituliskan seluruh SK dalam semester bersangkutan	Dituliskan KD yang esensial dari SK yang bersangkutan	Dituliskan indikator pencapaian yang esensial dari KD yang bersangkutan.	Dicantumkan bentuk butir tes yang dipilih, seperti benar-salah, menjodohkan, dan pilihan ganda	Dituliskan bentuk tes yang dipilih seperti tes keterampilan tertulis, tes identifikasi, tes simulasi, atau tes contoh kerja

Untuk tes tertulis, guru dapat membuat kisi-kisi tes tertulis untuk ulangan akhir semester seperti contoh berikut.

Kisi-Kisi Tes Tertulis Ulangan Akhir Semester

Sekolah : .....
Mata Pelajaran : .....
Kelas/Semester : .....
Alokasi waktu : .....

Standar Kompetensi	Kompetensi Dasar	Indikator Pencapaian	Bentuk Butir Tes			
			Pilihan Ganda	Uraian	.....	.....
Dituliskan seluruh SK dalam semester bersangkutan	Dituliskan KD yang esensial dari SK yang bersangkutan	Dituliskan indikator pencapaian yang esensial dari KD yang bersangkutan.	... butir	...butir	...butir	...butir

*) Keterangan: di bawah kolom bentuk butir tes diisi bentuk butir tes yang akan digunakan seperti pilihan ganda, uraian, dan menjodohkan

Hal yang penting untuk diperhatikan dalam mengembayusun kisi-kisi ujian/tes sumatif, yaitu perlunya memilih indicator yang paling esensial dari seluruh KD yang dibelajarkan dalam semester yang bersangkutan. Dengan demikian validitas isi dapat terpenuhi dengan baik.

BAB II
PENYUSUNAN INSTRUMEN HASIL BELAJAR

A. PENYUSUNAN INSTRUMEN PENILAIAN AFEKTIF

Dalam menyusun angket harus memperhatikan skala sikap yang digunakan. Pengukuran sikap yang dapat digunakan misalnya sebagai berikut.

1. Skala Likert

Skala Likert merupakan suatu skala penilaian untuk mengukur sikap dengan skala ordinal. Rentangan yang dipilih dari yang sangat positif sampai sangat negatif, misal dengan alternatif pilihan mulai dari sangat setuju (SS), setuju (S), ragu-ragu (R), tidak setuju (T), dan sangat tidak setuju (ST), dapat pula dari yang tidak pernah sampai yang selalu siswa lakukan sehingga rentangannya mulai dengan tidak pernah (TP), jarang (J), sering (S), hampir selalu (HS), dan selalu (S).

Dalam menyusun skala Likert sangat tergantung kemampuan penilai/penyusun angket dalam merumuskan indikator-indikator dari variabel yang akan diukur. Jika indikatornya sudah diperoleh baru disusun daftar pernyataan yang mencerminkan indikator-indikator tersebut. Misalnya, ingin diukur bagaimana persepsi siswa terhadap kemandirian dalam belajar di rumah, untuk itu harus dicari indikator- indikator yang relevan. Misal diperoleh indikator yang di antaranya tentang:

a. Keterlibatan orang lain dalam penyusunan jadwal belajar (1 item).
a. Keterlibatan orang lain membantu belajar (2 item).
b. Keterlibatan orang lain dalam menyiapkan alat bantu belajar (1 item).
c. dst.

Kemudian dibuat daftar pertanyaan sebagai berikut :

Nama: ..... Nomor presensi: ..... Kelas: .....

Setujukah kamu terhadap pernyataan di bawah ini?
Bila sangat setuju beri tanda V pada kolom pilihan SS, bila hanya setuju beri tanda V pada kolom pilihan S, bila netral atau ragu beri tanda V pada kolom pilihan R, bila tidak setuju beri tanda V pada kolom pilihan T, dan bila sangat tidak setuju beri tanda V pada kolom pilihan TS!

No.	Pernyataan	Pilihan				
		ST	T	R	S	SS
1	Dalam menyusun jadwal belajar sebaiknya minta bantuan orang tua					
2	Saat belajar dirumah lebih baik minta kakak, orang tua, atau orang lain untuk menjelaskan apa yang dipelajari.					
3	Sebaiknya dicoba lebih dahulu sebelum bertanya kepada kakak, orang tua, atau orang jika mengalami kesulitan dalam mengerjakan pekerjaan rumah.					
4	Sebaiknya ada orang lain yang ikut menyiapkan peralatan belajar saat saya belajar di rumah ataupun sebelum saya berangkat sekolah.					
5.	dst.					

Catatan:
Bila pernyataan bersifat positif maka untuk pilihan SS diberi skor 5 dan pilihan STS diberi skor 1. Sebaliknya jika pernyataan bersifat negatif (justru sepenuhnya melibatkan orang lain), maka pilihan SS diberi skor 1 sedangkan pilihan STS diberi skor 5.

1. Skala Perbedaan Semantik/Skala Berdeferensiasi Sematik

Skala perbedaan semantik/skala berdeferensiasi semantik merupakan suatu model skala dengan meletakkan suatu rentangan di antara dua kata atau ide yang berlawanan, sehingga berupa skala perbedaan sematik. Model skala yang bipolar ini sangat baik untuk mengungkap unsur evaluasi (baik/buruk, bersih/kotor, jujur/tidak jujur, menguntungkan/merugikan dan sejenisnya), atau untuk mengungkap unsur potensi (besar/kecil, kuat/lemah, berat/ringan, dan sejenisnya), ataupun unsur aktivitas (aktif/pasif, cepat/lambat, loyal/tak loyal, penuh perhatian/tak acuh).

Misal untuk mengetahui bagaimana persepsi siswa terhadap mata pelajaran yang diajarkan, guru menyuruh siswa untuk menyilang titik-titik yang tersedia di antara dua

ajektif sehingga akan diketahui ia bersikap positif ataukah negatif terhadap hal yang ditanyakan.

Pasangan ajektif tersebut harus dicari yang sesuai dengan konsep atau obyek yang akan diukur. Pasangan ajektif tersebut perlu diuji secara empiris, yaitu dengan terlebih dahulu mencari dua kelompok yang benar-benar pro dan yang benar-benar anti terhadap hal tersebut. Pasangan ajektif yang benar-benar dapat membedakan antara kedua kelompok tersebut dapat dipakai, dan yang tak dapat membedakan yakni baik kelompok pro dan anti sama-sama memilih titik tengah (ragu-ragu) dibuang. Dalam penskorannya, semakin ke arah yang positif semakin besar, dan skor total merupakan penjumlahan skor setiap pasangan ajektif.

Contoh:
Nama siswa: ..... No presensi: ..... Kelas: .....

Menurut Anda bagaimanakah pelajaran Matematika yang telah diselenggarakan selama 1 semester?

Perhatikan contoh berikut ini.

Menarik .____×____.____.____.____. ____. Membosankan

Karena Anda memberi tanda silang pada posisi di atas berarti menurut Anda pelajaran Matematika yang telah diselenggarakan selama satu semester menarik

Sebaliknya kalau Anda menyilang sebagai berikut.

Menarik .____.____.____.____×. ____. Membosankan

Berarti menurut Anda pelajaran Matematika yang telah diselenggarakan selama satu semester sangat membosankan.

Pembelajaran Matematika yang telah berjalan dalam satu semester

Menarik	.____.____.____.____.	Membosankan
Mudah	.____.____.____.____.	Sukar
Ringan	.____.____.____.____.	Berat
Menguntungkan	.____.____.____.____.	Merugikan
Bermanfaat	.____.____.____.____.	Merugikan
Menantang	.____.____.____.____.	Tidak menantang
Mengasyikkan	.____.____.____.____.	Menjemukan

2. Skala Thurstone

Tahapan dalam pengembangan instrumen skala sikap memakai skala Thurstone adalah sebagai berikut.

a. Pengembangan daftar pernyataan yang ditawarkan pada panelis yakni dengan menyusun minimal 50 pernyataan dari yang sangat positif sampai sangat negatif yang berkait dengan mata pelajaran Biologi.
b. Menyiapkan anggota panelis, misalnya dengan memilih sekurang-kurangnya 80 guru biologi atau mahasiswa yang menempuh program Pendidikan Biologi.
c. Meminta panelis untuk memberik-an skor terhadap setiap pernyataan yang ditawarkan. Meminta panelis untuk memberikan skor sangat rendah terhadap pernyataan yang bersifat negatif dan skor yang sangat tinggi untuk pernyataan yang sangat positif. Kisaran skor 1 sampai 11.

Contoh:
Berikut ini pernyataan-pernyataan yang berkait dengan bidang otomotif. Bila suatu pernyataan dinilai positif di beri skor besar, sedangkan bila negatif diberi skor kecil. Skor unuk pernyataan yang paling positif 11, sedangkan skor untuk penyataan yang paling negatif diberi skor 1.

No	Pernyataan	Pilihan skor										
		1	2	3	4	5	6	7	8	9	10	11
1	Menguasai bidang biologi dengan baik sangat mendukung wirausaha di masa depan											
2.	Berusaha di bidang yang ada hubungannya dengan biologi merupakan usaha yang sia-sia											
3.	Keahlian dalam bidang biologi memiliki prospek yang baik bila ditekuni dengan sungguh-sungguh											
4.	Biologi tidak bedanya dengan mata pelajaran yang lain di sekolah											
Dst												

d. Menghitung nilai median untuk tiap pernyataan berdasarkan skor yang diberikan panelis

Contoh:
Terhadap suatu butir pernyataan, setelah dihadapkan pada 80 panelis, sebanyak 4 panelis memberi skor 1, sebanyak 28 panelis memberi skor 2, sebanyak 31 panelis memberi skor 3, dan sebanyak 17 panelis memberi skor 4.

Skor	Banyaknya Panelis	Frekuensi Kumulatif
1	4	4
2	28	32
3	31	63
4	17	80

$$Md = L + \frac{\{N/2 - cfb\}}{fw} i$$

L = batas bawah nyata kelas interval yang mengandung median
N = banyaknya kasus (frekuensi kumulatif),
cfb = frekuensi kumulatif kelas interval di bawah kelas interval yang mengandung median (sebanyak 32)
fw = frekuensi kelas interval yang mengandung median (sebanyak 31)
i = panjang kelas interval (sebanyak 1)

$$Md = 2,5 + \frac{\{80/2 - 32\}}{31} \times 1$$

Md = 2,5 + (0,26) 1 = 2,76

e. Menghitung nilai kuartil (Q_3 dan Q_1) dan deviasi kuartil (Q_3-Q_1) untuk tiap pernyataan berdasarkan skor yang diberikan panelis

Contoh:

Skor	Banyaknya Panelis	Frekuensi Kumulatif
1	4	4
2	28	32
3	31	63
4	17	80

$$Q_3 = L_3 + \frac{\{3N/4 - cfb\}}{fw} i$$

$$Q_1 = L_1 + \frac{\{N/4 - cfb\}}{fw} i$$

Deviasi kuartil = $(Q_3 – Q_1)/2$

L_3 = batas bawah nyata kelas interval yang mengandung Q_3
L_1 = batas bawah nyata kelas interval yang mengandung Q_1
N = banyaknya kasus (frekuensi kumulatif),
cfb = frekuensi kumulatif kelas interval di bawah kelas interval yang mengandung median (sebanyak 32)
fw = frekuensi kelas interval yang mengandung median (sebanyak 31)
i = panjang kelas interval (sebanyak 1)

$$Q_3 = 2,5 + \left\{ \frac{3(80/4) – 4}{31} \right\} x\ 1$$

$Q_3 = 2,5 + 0,99 = 3,49$

$$Q_1 = 1,5 + \left\{ \frac{(80/4) – 32}{28} \right\} x\ 1$$

$Q_1 = 1,5 + 0,57 = 2,07$

Deviasi kuartil = $(Q_3 – Q_1)/2 = (3,49 – 2,07)/2 = 0,667$

f. Memilih pernyataan yang memenuhi syarat.
Menyisakan 15 sampai 20 pernyataan yang bersifat positif, netral, sampai yang negatif dengan cara membuang pernyataan yang memiliki nilai deviasi kuartil yang besar.

g. Menentukan besarnya skor untuk setiap pernyataan
Skor tiap pernyataan merupakan besarnya median yang diberikan oleh panelis

Contoh:

No.	Pernyataan	Skor
1	Menguasai bidang biologi dengan baik sangat mendukung wirausaha di masa depan	10,5
2	Berusaha di bidang yang ada hubungannya dengan biologi merupakan usaha yang sia-sia	1,2
3	Keahlian dalam bidang biologi memiliki prospek yang baik bila ditekuni dengan sungguh-sungguh	5,75
4	Biologi tidak bedanya dengan mata pelajaran yang lain di sekolah	6,5
dst.		

h. Menyiapkan angket siap dipakai

Contoh sajian

Pilihlah dengan cara membubuhkan tanda X pada kolom TS jika Anda tidak setuju dan pada kolom S jika setuju terhadap pernyataan-pernyataan di bawah ini!

Contoh:

No.	Pernyataan	Pilihan	
		TS	S
1	Bidang biologi menjamin masa depan		X

Artinya: Anda *setuju* bidang biologi *menjamin* masa depan wirausaha di bidang tersebut.

Berikut ini daftar pernyataan yang harus Anda pilih!

No.	Pernyataan	Pilihan	
		TS	S
1	Menguasai bidang biologi dengan baik sangat mendukung wirausaha di masa depan		
2	Berusaha di bidang yang ada hubungannya dengan biologi merupakan usaha yang sia-sia		
3	Keahlian dalam bidang biologi memiliki prospek yang baik bila ditekuni dengan sungguh-sungguh		
4	Biologi tidak bedanya dengan mata pelajaran yang lain di sekolah		
Dst.			

3. Lembar Observasi

Lembar observasi dapat digunakan untuk melihat sikap siswa saat berinteraksi sosial dengan orang lain, atau kalau dalam melakukan suatu pekerjaan ada norma atau sikap tertentu yang harus dipatuhi. Misalnya saat anak diminta bersimulasi bagaimana ia harus berpidato, maka dapat diobservasi bagaimana kesantunan saat ia berpidato. Dalam hal ini guru harus menyiapkan lembar observasi yang akan dipakai untuk mengamati sikap siswa. Misalnya saat berdiskusi dapat digunakan lembar observasi sebagai berikut.

LEMBAR PENGAMATAN ASPEK AFEKTIF

Nama: ..... Nomor presensi: ..... Kelas: .....

NO	ASPEK YANG DIAMATI	4	3	2	1
1a.	Kemauan mendengarkan dengan penuh perhatian				
b.	Kemaun mendengarkan nasehat orang lain				
c.	Kemauan dalam melibatkan diri dalam aktivitas di kelas dan/atau laboratorium				
d.	Kemauan menerima teman lain apa adanya				
e.	Kemauan untuk mencatat hal-hal yang penting				
f.	Kepedulian dengan persoalan yang dihadapi orang lain				
2a.	Kemauan dalam mengerjakan tugas rumah/laboratorium				
b.	Kesungguhan dalam menjawab pertanyaan				
c.	Kemauan berpartisipasi dalam diskusi kelompok/kelas				
d.	Kepatuhan dalam mengikuti kesepakatan bersama				
e.	Kemauan membantu/membimbing/menolong orang lain				
f.	Kemauan bergabung dengan kelompoknya dengan senang hati/sukarela				
3a.	Kemauan untuk mengambil keputusan bersama secara demokratis/tidak memaksakan kehendaknya pada orang lain				
b.	Kemauan menghargai hal-hal yang baik				
c.	Kemauan menerapkan pengetahuan dalam kehidupan sehari-hari (misalnya membuang sampah pada tempat yang telah tersedia)				
d.	Kemauan mendukung rencana yang telah dibuat kelompoknya				
e.	Kemauan mendukung pendapat orang lain yang benar dan memprotes pendapat orang lain yang salah				
f.	Kemauan untuk menyampaikan ide/pendapat dalam pemecahan masalah				
4a.	Kemauan untuk mendiskusikan standar/kriteria yang dipakai untuk menetapkan kebenaran suatu hal				
b.	Kemauan untuk menilai segala sesuatu dengan cara membandingkannya dengan standar/criteria				
c.	Kemauan mengakui kelebihan/kelemahan diri				
d.	Kemauan mendukung perencanaan yang baik yang dibuat kelompoknya				
e.	Kemauan menunjukkan peranserta dalam suatu perencanaan/kegiatan atas dasar minat, tanggungjawab, dan keyakinannya				
5a.	Kemauan berupaya menghindari kesalahan yang pernah ia lakukan				
b.	Kemauan memilih prosedur yang benar (tidak asal bertindak) untuk menyelesaikan masalah				
c.	Kemauan untuk mengatasi hal-hal yang tidak dikehendaki (kemauan tidak mendorong orang lain untuk berbuat salah)				

d.	Kemauan untuk menunjukkan kepercayaan diri bekerja secara individual (tidak bertanya pada orang lain saat ulangan, tidak minta bantuan orang lain padahal belum mencoba/mencarinya)				
e.	Kemauan untuk bekerja secara produktif (kemauan untuk menghasilkan karya)				
f.	Kemauan untuk menghindari konflik dan berusaha dinilai baik oleh orang lain				

Keterangan: 1: receiving; 2: responding; 3: valuing; 4: organization; 5: characterization by a value or value complex

4. Lembar Penilaian Antar Teman (*Peer Assessment*)

Lembar penilaian antar teman sangat baik untuk meningkatkan tanggung jawab dalam penyelesaian tugas kelompok atau penerimaan seseorang terhadap orang lain. Misalnya saat menyelesaikan tugas kelompok dalam bentuk tugas rumah, guru tidak dapat melakukan observasi. Oleh karena itu penilaian antar teman dapat digunakan.

a. Contoh Lembar penilaian antar teman dalam kerja kelompok

Nilailah setiap anggota dalam kelompokmu! Berilah nilai 10 bila sangat baik, sebaliknya berilah nilai 0 bila sangat jelek! Selanjutnya jumlahkan hasil penilaianmu untuk memperoleh nilai masing-masing anggota dalam kelompokmu!

No.	Nama siswa	Nomor Presensi	Hal yang dinilai					
			1	2	3	4	5	Jumlah
1								
2								
3								
dst								

Keterangan:

No.	Hal yang dinilai
1	Mendengarkan dengan perhatian penuh saat temannya berbicara,
2	mengajukan usul, atau memberikan arahan tentang tugas yang harus diselesaikan
3	Menyambut baik terhadap tugas yang diberikan kepadanya
4	Menyelesaikan tugas dengan baik dan sesuai waktu yang ditetapkan
5	Menegur dengan sopan bila ada temannya yang tidak serius dalam berkerja

B. PENGEMBANGAN INSTRUMEN PENILAIAN KINERJA

Penilaian kinerja adalah penilaian yang memfokuskan aspek keterampilan yang berkait dengan ranah psikomotor yang dapat didemonstrasikan oleh peserta didik. Dilihat dari kinerja atau kemampuan yang didemonstrasikan, kinerja dapat digradasi dari kinerja yang paling rendah sampai yang paling tinggi. Kinerja yang paling rendah misalnya kemampuan siswa mampu menjawab saat ditanya besarnya uang kembalian jika membayar dengan mata uang yang nilainya lebih besar dari harga barang. Kinerja agak tinggi misalnya siswa diminta mendemonstrasikan besarnya uang yang harus ia kembalikan menggunakan mata uang yang sesungguhnya. Kinerja yang lebih tinggi lagi misalnya siswa diminta bersimulasi dengan pasangannya mendemonstrasikan besarnya uang kembalian menggunakan mata uang yang sesungguhnya dengan nilai yang berbeda-beda, dan kinerja yang sangat tinggi jika siswa mampu berbelanja di toko dengan membawa sejumlah uang dan memperoleh uang kembalian/sisa uang sebesar nilai uang yang dibawa dikurangi harga barang yang dibelinya.

Dalam konteks di atas, maka kinerja mencakup ranah kognitif dan sekaligus mencerminkan ranah psikomotor. Ranah kognitif adalah tingkat kebenaran dari aspek berpikir yang mendasari tindakannya, dan keterampilan psikomotor yang didemonstrasikan berupa kemampuan membedakan mata uang sesuai dengan nilainya dan memilih mata uang yang sesuai/lebih besar dari harga barang saat ia sebagai pembeli, dan memilih mata uang yang nilainya sama dengan nilai pengembalian saat ia berperan sebagai penjual. Dengan kata lain aspek psikomotor menyangkut penguasan prosedur. Akan tetapi, dapat pula aspek psikomotor dapat dinilai dari produk yang dihasilkan oleh suatu tindakan tertentu yang dilakukan peserta didik. Penilaian terhadap prosedur berarti lebih mengarah kepada aktualisasi aspek psikomotor yang ditampilkan dalam suatu kinerja (*performance*).

Dari taksonomi ranah psikomotor dapat diidentifikasikan bahwa ada aspek dari ranah psikomotor yang murni sebagai gerak bagian tubuh dan kemampuan fisik tubuh dan ada pula gerak dari bagian tubuh yang berkait dengan pemakaian alat. Gerak yang dilakukan juga ada gerak yang tidak dipelajari yaitu gerak refleks, dan gerak yang dipelajari. Gerak yang dipelajari ada yang berupa gerak yang sederhana dan ada pula gerak yang kompleks.

Penguasaan teoretik tentang suatu prosedur pun oleh Simson dimasukkan sebagai aspek dari ranah psikomotor yakni termasuk dalam kesiapan untuk berperan aktif dalam melakukan aktivitas motorik. Oleh karena itu, dalam pembelajaran motorik, tahapan yang ditempuh adalah penguasaan teori tentang teknik/prosedur yang berupa tahapan-tahapan dalam melakukan aktivitas, dilanjutkan dengan artukulasi berupa latihan untuk menguasai suatu teknik/prosedur yang dipelajari, dan otomatisasi untuk menguasai teknik secara terlatih dan spontan.

Sebagai contoh agar seseorang mampu bermain tennis maka ia harus menguasai berbagai teori tentang teknik baik yang berkait dengan teknik memukul dan mengembalikan bola serta peraturan bermain tennis. Kemudian melalui tahapan artikulasi ia berlatih berlatih cara memukul, mengembalikan bola, dan menerapkan aturan bermain. Selanjutnya melalui tahapan otomatisasi ia harus berlatih berulang-ulang (*drilling*), termasuk uji coba, dan bertanding, sehingga ia dapat melakukan gerak-gerak otomatis dalam memukul dan mengembalikan bola, serta mampu bermain secara sportif.

Penilaian terhadap prosedur dilakukan dengan pertimbangan: (1) tidak ada produk yang bisa dinilai, (2) prosedurnya memiliki langkah-langkah yang urut dan dapat diamati, (3) langkah-langkah yang benar dari suatu prosedur menunjukkan suatu keberhasilan, dan/atau (4) analisis terhadap prossedur dapat meningkatkan mutu produk. Penilaian terhadap produk dilakukan dengan pertimbangan: (1) berbeda prosedur berbeda produk, (2) prosedur sudah dikuasai, (3) prosedurnya tidak dapat dinilai, (4) prosedur tidak perlu dinilai (misalnya pekerjaan rumah), dan/atau (5) produk memiliki kejelasan aspek yang dinilai

PENGEMBANGAN ITEM TES KINERJA

1. Pengembangan Item Tes Kinerja untuk Penguasaan Prosedur

Untuk mengembangkan item tes kinerja dalam bentuk prosedur harus memperhatikan hal-hal berikut.

a. Menyesuaikan dengan jenis kinerja/performance yang harus ditampilkan, apakah merupakan gerak dasar yang fundamental, kemampuan perceptual, kemampuan fisik, gerak terlatih ataukah gerak yang mengekspresikan komunikasi.
b. Menyesuaikan dengan tehnik penilaian yang dipilih, yaitu:

1) tes identifikasi untuk mengukur kinerja seseorang atas dasar tanda-tanda atau sinyal yang diberikan saat diberikan tes.
2) tes simulasi untuk mengukur kinerja dalam situasi yang mirip dengan situasi yang sebenarnya.
3) uji petik kerja (*work sampel test*) untuk mengukur kinerja dalam situasi yang sebenarnya.

c. Menyusun rubrik/pedoman penskoran

Di dalam penyusunan rubrik/pedoman penskoran ada beberapa hal yang perlu diperhatikan tergantung pada bentuk instrumen.

1) Tes identifikasi: (a) menentukan jenis kemampuan kinerja yang akan diidentifikasi, (b) menentukan banyaknya hal/aspek yang akan diidentifikasi, dan (c) membuat rubrik untuk penskoran yang dilengkap dengan kategorisasi keberhasilan identifikasi.
2) Uji petik kerja/simulasi: (a) mengidentifikasi aspek kinerja yang diskor, (b) menentukan model skala yang dipakai untuk menyekor, yakni skala penilaian (*rating scale*) atau daftar cek (*check list*), dan (c) membuat rubrik penskoran yang dilengkapi dengan kategorisasi keberhasilan kinerja.

2. Pengembangan Item Tes Kinerja untuk Penguasaan Produk

Untuk mengembangkan item tes kinerja dalam bentuk penguasan produk maka harus memperhatikan hal-hal berikut.

a. Menyesuaikan dengan jenis produk yang harus dihasilkan, apakah produk dua dimensi ataukah produk tiga dimensi.
b. Memperhatikan tehnik penilaian yang dipakai, yaitu
   1) Tes tulis (*paper and penci test*) untuk menilai produk dua dimensi yang diujudkan dalam bentuk sketsa, tulisan, gambar, lukisan, atau bentuk dua dimensi lainnya.
   2) Penugasan produk tiga dimensi untuk menilai produk tiga dimensi yang diujudkan dalam bentuk kerajinan, pahatan, dan produk tiga dimensi lainnya.
c. Menyusun rubrik/pedoman penskoran

   Di dalam penyusunan rubrik/pedoman penskoran ada beberapa hal yang perlu diperhatikan tergantung pada bentuk instrumen.

1) Tes *paper and pencil*: (a) menentukan cara penskoran secara holistik atau analitik, (b) menentukan aspek-aspek yang dinilai atau kata kunci, (c) menentukan bobot skor, dan (d) menentukan klasifikasi peringkat penilaian.
2) Penugasan produk tiga dimensi: (a) menentukan aspek produk yang akan dinilai, (b) menentukan bobot skor, dan (c) menentukan klasifikasi peringkat penilaian.

a. Contoh Instrumen Pengukuran Kinerja untuk ranah Kognitif

1) Pengukuran Aspek komunikasi

(a) Lembar Observasi untuk Kinerja Umum (Bentuk *Check List*)

Lembar observasi untuk kinerja umum dari aspek kognitif dalam bentuk *check list* berarti performan yang benar harus muncul dalam hal yang diamati.

No.	Nama Siswa	Nomor presensi	Aspek yang Diukur					
			1	2	3	4	5	Jmlh
1								
2								
3								
dst.								

Aspek yang dinilai:
1. Pertanyaan mengungkap kemampuan berpikir
2. Penjelasan lengkap dan jelas
3. Argumen logis dan kuat
4. Saran jelas dan logis
5. Bahasa baik/benar

Catatan: Beri skor 1 untuk setiap aspek jika sesuai

(b) Lembar Observasi untuk Kinerja Diskusi Kelompok Bentuk *Rating Scale*

Nama siswa: ……..... Kelas: …..... Tgl: …….....

No.	Aspek yang Diukur	Skala		
		1	2	3
1	Pola berpikir saat menyampaikan informasi/pendapat			
2	Pola berpikir saat memberikan argumentasi			
3	Pola berpikir saat memberikan kritikan			
4	Kejelasan fokus dan arah pertanyaan			
5	Kemampuan dalam berbicara (dalam memberikan informasi, berpendapat, berargumentasi)			
6	Penguasan bahasa (saat menyampaikan informasi, kritikan, ataupun argumentasi)			
	Jumlah skor			

Rubrik:

Aspek 1: pola berpikir saat menyampaikan informasi

1 = sulit dimengerti dan dipaparkan dengan tidak runtut/teratur

2 = dapat dimengerti tetapi tidak dipaparkan secara runtut/teratur

3 = dipaparkan secara runtut/teratur dan mudah dimengerti

Aspek 2: pola berpikir saat menyampaikan argumentasi

1 = sulit dimengerti dan dipaparkan dengan tidak runtut/teratur

2 = dapat dimengerti tetapi tidak dipaparkan secara runtut/teratur

3 = dipaparkan secara runtut/teratur dan mudah dimengerti

Aspek 3: pola berpikir saat menyampaikan kritik

1 = sulit dimengerti dan dipaparkan dengan tidak runtut/teratur

2 = dapat dimengerti tetapi tidak dipaparkan secara runtut/teratur

3 = dipaparkan secara runtut/teratur dan mudah dimengerti

Aspek 4: Fokus/arah pertanyaan

1 = tidak jelas fokus/arahnya sehingga tidak dimengerti apa yang ditanyakan

2 = dapat dimengerti pertanyaannya tetapi tidak langsung pada fokus permasalahannya/arahnya (berputar-putar)

3 = dapat dimengerti pertanyaannya dan terfokus/jelas arahnya

Aspek 5: kemampuan berbicara

1 = tergagap-gagap, sulit berbicara

2 = kalimat diganti/diulang

3 = lancar

Aspek 6: Penguasaan bahasa

1 = tidak menggunakan EYD

2 = menggunakan EYD tetapi tidak komunikatif (tidak jelas subjek predikatnya)

3 = menggunakan EYD dan komunikatif (jelas subjek predikatnya)

2. Contoh Instrumen untuk Pengukuran Kinerja Psikomotor

Misalnya guru akan mengukur kinerja keterampilan menggunakan neraca, maka langkah pertama invetarisasilah langkah-langkah yanag harus dikerjakan jika seseorang menggunakan neraca untuk menimbang suatu benda. Setelah diinventarisasi langkah/tahapannya misalnya diperoleh hasil sebagai betrikut.

a. Mengecek posisi kedua meja neraca
b. Mengatur kembali posisi kedua meja neraca jika tidak dalam posisi seimbang
c. Memilih anak neraca sesuai dengan hasil pengukuran yang diinginkan
d. Meletakkan anak neraca pada meja neraca pada tempatnya
e. Memasukkan bbarang yang ditimbang sedikit demi sedikit ke dalam wadah sampai ujung meja benda dan ujung meja anak neraca sama tingginya.
f. Menambah atau mengurangi sehingga ujung meja neraca yang berhadap-hadapan sama tingginya.
g. Mengecek kembali dengan ditunggu bahwa kedua ujung meja neraca yang berhadap-hadapan benar-benar sama tingginya
h. Menurunkan dengan hati-hati wadah neraca yang berisi benda yang ditimbaqng.
i. Menurunkan anak timbangan dari mja neraca.

a. Lembar observasi bentuk *check list*

Bila keterampilan menimbang menggunakan neraca akan dibuat dalam bentuk *check list* maka dapat dibuat sajian sebagai berikut.

Nama siswa: …….…..... Kelas: ..... Tgl: ….....…

..... a. Mengecek posisi kedua meja neraca dalam posisi seimbang
..... b. Mengatur kembali posisi kedua meja neraca jika tidak dalam posisi seimbang dengan menambah beban di salah satu meja sampai posisi seimbang
..... c. Memilih anak neraca seberat 1 kg
..... d. Meletakkan anak neraca pada meja neraca pada tempatnya
..... e. Memasukkan beras sedikit demi sedikit ke dalam wadah sampai ujung meja benda dan ujung meja anak neraca sama tingginya.
..... f. Menambah atau mengurangi dengan hati-hati beras pada wadah sampai ujung meja neraca yang berhadap-hadapan sama tingginya.
..... g. Mengecek kembali dengan ditunggu bahwa kedua ujung meja neraca yang berhadap-hadapan benar-benar sama tingginya
..... h. Menurunkan dengan hati-hati wadah neraca yang berisi beras, dan menuangkan beras kedalam kantung plastik yang telah disediakan.
..... i. Menurunkan anak timbangan dari meja neraca.

Catatan: Beri tanda V untuk setiap kinerja berikut ini! yang dinyatakan benar dari setiap tindakan yang dilakukan siswa untuk melakukan penimbangan beras seberat 1 kg!

b. Lembar Observasi Bentuk *Rating Scale*

Nama siswa: ……..... Kelas: …..... Tgl: ….....……

1	2	3	4	a.	Cara mengontrol bahwa posisi kedua meja neraca sudah seimbang
1	2	3	4	b.	Cara memilih anak timbangan sesuai dengan berat benda yang diinginkan.
1	2	3	4	c.	Cara meletakkan anak timbangan pada meja neraca.
1	2	3	4	d.	Cara memasukkan benda yang akan ditimbang ke dalam wadah
1	2	3	4	e.	Cara menambah atau mengurangi benda yang ditimbang agar kedua meja neraca pada posisi setimbang.
1	2	3	4	f.	Cara mengontrol posisi kedua meja/bagian neraca setelah diberi anak timbangan dan di sisi lain diberi benda yang ditimbang benar-benar sudah setimbang.
1	2	3	4	g.	Cara mengambil wadah untuk menuangkan/mengambil benda yang ditimbang
1	2	3	4	h.	Cara menuang/mewadahi benda yang telah selesai ditimbang dengan wadah yang telah disediakan.
1	2	3	4	i.	Cara mengambil anak neraca dari meja neraca.

Catatan: Lingkari angka 5 jika sangat tepat, angka 4 jika tepat, angka 3 jika agak tepat, angka 2 jika tidak tepat dan angka 1 jika sangat tidak tepat.

Agar lebih objektif dibuat kriteria untuk tiap butir yang direntang mulai dari skala 1 sampai 4 tersebut.

BAB III
ANALISIS ITEM SECARA KUALITATIF

Setelah item instrument baik instrument tes/ujian ataupun instrument nontes disusun, maka perlu diselidiki kualitasnya dengan cara ditelaah leh teman sejawat. Langkah ini dikenal dengan analisis instrument secara kualitatif. Dalam melakukan penyelidikan kualitas item tes hasil belajar, telaah difokuskan kepada pemenuhan aspek materi/substansi, aspek konstruksi, dan aspek bahasa. Dalam hal ini, bentuk instrument akan membedakan karakteristik terutama dari aspek konstruksinya.

Berikut disajikan lembar telaah instrument tes/ujian untuk bentuk soal pilihan ganda dan uraian. Untuk bentuk lainnya mahasiswa diharap dapat menyusunnya sendiri.

LEMBAR TELAAH ITEM BENTUK PILIHAN GANDA

ASPEK YANG DITELAAH	NOMOR ITEM				
	1	2	3	….	N
a. Aaspek materi:					
1) Butir soal sesuai indikator					
2) Hanya ada satu kunci atau jawaban yang benar					
3) Isi materi sesuai dengan tujuan pengukuran					
4) Isi materi sesuai dengan jenjang, jenis sekolah dan tingkatan kelas					
5) Pilihan benar-benar berfungsi, jika pilihan merupakan hasil perhitungan, maka pengecoh berupa pilihan yang salah rumus/salah hitung					
b. Aspek konstruksi:					
1) Pokok soal (stem) dirumuskan dengan jelas					
2) Rumusan soal dan pilihan dirumuskan dengan tegas					
3) Pokok soal tidak memberi petunjuk/mengarah kepada pilihan jawaban yang benar					
4) Pokok soal tidak mengandung pernyataan negatif genda					
5) Bila terpaksa menggunakan kata negatif, maka harus digarisbawahi atau dicetak lain					
6) Pilihan jawaban homogen					
7) Hindari adanya alternatif jawaban : "seluruh jawaban di atas benar" atau "tak satu jawaban di atas yang benar" dan yang sejenisnya					
8) Panjang alternatif /pilihan jawaban relatif sama, jangan ada yang sangat panjang dan ada yang sangat pendek					
9) Pilihan jawaban dalam bentuk angka/waktu diurutkan.					
10) Wacana, gambar, atau grafik benar-benar berfungsi					
11) Antar butir tidak bergantung satu sama lain					
c. Aspek bahasa:					
1) Rumusan kalimat komunikatif					
2) Kalimat menggunakan bahasa yang baik dan benar, sesuai dengan jenis bahasanya					
3) Rumusan kalimat tidak menimbulkan penafsiran ganda atau salah pengertian.					
4) Menggunakan bahasa/kata yang umum (bukan bahasa lokal)					
5) Rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan siswa.					

LEMBAR TELAAH ITEM BENTUK URAIAN

ASPEK YANG DITELAAH	NOMOR ITEM				
	1	2	3	….	N
a. Aspek materi:					
1) Butir soal sesuai indikator					
2) Batasan pertanyaan dan jawaban yang diharapkan jelas					
3) Isi materi sesuai dengan tujuan pengukuran					
4) Isi materi yang ditanyakan sesuai dengan jenjang, jenis sekolah, dan tingkat kelas					
b. Aspek konstruksi:					
1) Rumusan kalimat dalam bentuk kalimat tanya atau perintah yang menuntut jawaban terurai.					
2) Ada petunjuk yang jelas cara mengerjakan/ menyelesaikan soal					
3) Ada pedoman penskorannya					
4) Tabel, grafik, diagram, kasus, atau yang sejenisnya bermakna (jelas keterangannya atau ada hubungannya dengan masalah yang ditanyakan.					
5) Butir soal tidak bergantung pada butir soal sebelumnya					
c. Aspek bahasa:					
1) Rumusan kalimat komunikatif					
2) Kalimat menggunakan bahasa yang baik dan benar, sesuai dengan jenis bahasanya					
3) Rumusan kalimat tidak menimbulkan penafsiran ganda atau salah pengertian.					
4) Menggunakan bahasa/kata yang umum (bukan bahasa lokal)					
5) Rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan siswa.					
12) Antar butir tidak bergantung satu sama lain					
c. Aspek bahasa:					
6) Rumusan kalimat komunikatif					
7) Kalimat menggunakan bahasa yang baik dan benar, sesuai dengan jenis bahasanya					
8) Rumusan kalimat tidak menimbulkan penafsiran ganda atau salah pengertian.					
9) Menggunakan bahasa/kata yang umum (bukan bahasa lokal)					
10) Rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan siswa.					

BAB IV
ANALISIS ITEM SECARA EMPIRIS BUKTI PENDUKUNG VALIDITAS

PENDAHULUAN

Menurut Embretson & Gorin (2001) validitas konstrak adalah sentral untuk menetapkan mutu tes. Secara tradisional, spesifikasi item sering hanya samar-samar. Pengembang tes membuat spesifikasi item yang sering kali hanya berisi pertimbangan-pertimbangan dari segi isi secara umum (seperti penetapan ruang lingkup/topik area dan ringkasan dari materi/isi) atau samar-samar menggambarkan pengolahan derajat pemrosesan berpikir (seperti abstrak melawan konkrit). Begitu keseluruhan spesifikasi item dihasilkan, dilakukan review item oleh reviewer untuk meyakinkan justifiabilas dari kunci jawaban dan menguji konten/isi item untuk berbagai isu penyetaraan dan mutu tes. Metode-metode psikometrik diterapkan setelah item dikembangkan. Statistika item dari uji coba empiris menjadi sangat esensial untuk menentukan kualitas item, terutama untuk mengevaluasi bahwa proses-proses “konstrak-relevansi” telah terukur. Secara umum item-item yang tidak saling berkorelasi dengan item-item yang lain tidak dipilih. Prosedur tradisional yang standar tersebut sangat cocok dengan paradigma validitas konstrak menurut teori klasik.

Berdasarkan prinsip psikologi kognitif, hal pertama yang terpenting dalam pemenuhan validitas konstrak suatu tes kemampuan/abilitas adalah adanya dukungan satu set prinsip teoritis dari item-item tes yang akan disusun. Kedua, suatu set standar yang baru yang disusun harus didasarkan pada prinsip-prinsip psikologi kognitif. Mengutip pendapat Messick (1995) bahwa item-item yang disusun tidak sekedar cocok/sesuai dengan kriteria/ukuran tradisional, seperti tingkat kesulitan yang sesuai dan daya pembeda yang tinggi, tetapi item-item tersebut juga dibenarkan sebagai bagian juga relevan dengan konstrak dari proses-proses kognitif. Ketiga, bahwa tes yang disusun adalah untuk mengindikasikan kecakapan atau untuk mendiagnosis ketrampilan yang ada sehingga memerlukan bermacam informasi baru tentang item-item yang disusun.

Menurut Frisbie (2005) dengan mengacu pendapat Messick (1989) bahwa validitas dari ”tests standars” bukan terhadap instrumennya itu sendiri melainkan berkait dengan interpretasi skor dan penggunaan tes itu sendiri. Validitas suatu tes dinyatakan baik untuk tujuan yang satu tidak akan sebaik untuk tujuan yang lainnya. Pertanyaan mendasar

tentang validitas tes adalah: ”Can a valid test yield scores that should not be used in the way the maker orginaly intended?”

Berkait dengan tes hasil belajar di dalam glossary ”the Test Standars” 1999 menurut Frisbie dinyatakan bahwa: ” **Criterion-referenced interpretation** see criterion reference test (p.174). **Criterion-referenced test**: A test that allow its users to make score interpretations in relation to a functional performance level, as distinguished form those interpretations that are made in relation to the performance of others. Examples include comparison to cut scores ...(p.174). **Norm-referenced interpretation**: A score interpretations based on comparison of a test teaker’s performance to the performance of the other people in a spesific population. See criterion-reference test. (p 178)”.

Berkait dengan reliabilitas tes, Frisbie (2005) menyatakan bahwa reliabilitas tes asil belajar berbeda dengan reliabilitas tes untuk seleksi karena tes hasil belajar memiliki varians yang rendah manakala anak berhasil semua dalam belajarnya. Oleh karena itu secara praktis Friesbie membuat tabel pembandingan ideal antara interpretasi dalam situasi *Norm Reference* (NR) dan *Criterion Reference.*

Tabel 1. Comparative ideals for NR and CR interpretation situations

	Norm-Reference	Criterion-Reference
Item difficulty	Moderate	Easy-to-hard
Item discrimination	High positive	Nonnegative
Score variability	Maximize	Non-issue
Error estimate	High reliability coefficient	High decision consistency index

Berdasarkan informasi pada Tabel 1 maka tes untuk mengukur keberhasilan belajar merupakan tes yang skornya diinterpretasikan dalam situasi *criterion-reference*, sehingga item-itemnya memiliki tingkat kesulitan item bervariasi dari mudah sampai sukar (sebagai cerminan tingkat keberhasilan belajar) dan tidak boleh memiliki indeks daya beda yang negatif (sebagai cerminan bahwa tidak ada testi yang cerdas menjawab salah). Oleh karena itu estimasi error didasarkan pada tingginya indeks konsistensi (indeks yang tinggi menunjukkan semua testi pasti benar bila sudah belajar, semua testi salah bila belum belajar). Sebaliknya, tes untuk tujuan seleksi adalah tes yang dapat memisahkan kelompok yang lolos seleksi dan yang tidak lolos seleksi. Oleh karena itu, iterpretasinya dalam situasi *norm-reference*, sehingga item-itemnya memiliki indeks kesulitan yang harus

moderate (sebagai cerminan bahwa kelompok ataslah yang pasti dapat mengerjakan) dan indeks daya beda harus tinggi (sebagai cerminan yakin dapat membedakan kelompok atas dan bawah). Oleh karena itu, estimasi *error* didasarkan pada tingkat tingginya reliabilitas tes (indeks yang tes tinggi mencerminkan bahwa semakin cerdas testi di dalam kelompoknya semakin tinggi pula skor yang diperolehnya).

Menurut Stark *et. al.*(2001), pemilihan item tes dalam prosedur pengembangan tes menggunakan CTT umumnya didasarkan pada: (a) nilai kesukaran item, dan (b) korelasi skor item dan skor total atau disingkat korelasi item-total. Item yang memiliki korelasi item-total paling tinggi dipakai sebagai elemen suatu tes untuk membentuk suatu skala dengan konsistensi internal tinggi guna memperkecil sumbangan *error* acak skor-skor tes. Distribusi skor-skor tes total yang diperoleh dari lapangan dibandingkan dengan distribusi yang diinginkan oleh pengembang tes. Sejumlah item mungkin perlu diganti untuk memperoleh sedekat/semirip mungkin antara distribusi skor total yang diinginkan dan distribusi skor total yang diperoleh dari lapangan. Format-format paralel pada umumnya diciptakan untuk memperoleh distribusi-distribusi skor tes yang identik. Kesamaan dari nilai rata-rata, varians, dan *error* skor ditafsirkan sebagai bukti bahwa format tes-tes bersifat paralel.

Menurut Stark *et. al.*(2001), seharusnya langkah pertama sebelum penulisan item mulai, pengembang tes harus mempunyai suatu pemahaman yang baik tentang konstrak variabel (kemampuan) yang akan diukur. Mengacu pendapat Nunnally *et. al.*, berdasarkan ”*rule of thumb*” ia menyatakan bahwa lazimnya disepakati bahwa banyaknya item tes yang harus dibuat sedikitnya dua kali dari banyaknya item tes final yang diperlukan. Sejumlah besar item pilihan ganda diperlukan, jika format-format ganda harus dikembangkan. Item-item tersebut harus diteskan terlebih dahulu menggunakan suatu sampel yang serupa dengan populasi pelamar. Sampel ini, yang diacu selanjutnya sebagai suatu sampel yang dijadikan pedoman saat kalibrasi, harus besar, agar cukup untuk menyediakan statistika item CTT yang stabil. Item-item dengan korelasi item-total tinggi harus tercakup di dalam tes karena item-item tersebut meningkatkan konsistensi skala internal (reliabilitas), dan hal seperti itu akan mereduksi *standard error* pengukuran. Kesulitan item (nilai p) juga harus dipertimbangkan untuk membuat suatu tes dengan distribusi skor total yang diinginkan.

Untuk memperoleh distribusi skor skala yang diinginkan dilakukan penggantian item. Agar skalanya meningkat maka maka item dengan nilai p yang rendah harus

digantikan dengan nilai p yang tinggi. Untuk memperkecil dampak penggantian item terhadap reliabilitas skala, yakni dengan mencoba menggantikan item-item yang memiliki korelasi item-total yang rendah sebelum menghapus item-item yang memiliki daya pembeda yang lebih tinggi. Dapat pula dalam praktik, beberapa penyeimbangan konten/isi juga diperlukan. Setelah dilakukan penggantian kemudian dianalisis lagi.

Ada keterbatasan penggunaan pendekatan CTT (Stark *et. al.*, 2001). Pertama, statistika CTT bergantung kepada subpopulasi penempuh tes. Berbeda grup penempuh tes berbeda pula nilai rata-rata skor dari atribut variabel yang diukur. Dengan demikian, para pengembang tes harus hati-hati ketika memilih sampel untuk kalibrasi item. Jika sampel-sampel kalibrasi berbeda karakteristik/sifat dengan sampel operasional (sampel populasi yang sesungguhnya sebagai target), properti-properti psikometri hasil pengukuran akan berubah secara dramatis. Kedua, di dalam CTT, ketepatan pengukuran suatu tes (galat baku atau *standard error* pengukuran) secara implisit dirata-ratakan ke semua level kemampuan yang diukur. Dengan demikian, ketepatan pengukuran pada level-level skor yang tertentu tidak dikenal/tidak diketahui. Oleh karena itu, dikembangkan analisis item menggunakan teori respons item atau *item response theory* (IRT).

Kegiatan mengkonstruksi tes menggunakan pendekatan IRT, seperti halnya pada penggunaan pendekatan CTT, penulis harus membuat dua sampai tiga kali banyaknya item seperti yang diinginkan di dalam format final. Dalam IRT diperlukan sampel kalibrasi heterogen yang besar. Model IRT yang lebih kompleks, seperti model IRT untuk skala politomus, memerlukan sampel lebih besar untuk mengestimasi parameter. Sebelum mengestimasi parameter item, perlu untuk melakukan suatu analisis item menurut teori tes klasik untuk menghapuskan item-item yang mempunyai skor mendekati nihil (tidak atau sedikit sekali yang dapat mengerjakan), tentu saja item yang demikian akan memiliki korelasi-korelasi item-total negatif. Item ini akan menyebabkan permasalahan konvergensi/pemusatan. Demikian pula item yang mempunyai skor prefect, dimana untuk tes pilihan ganda skor prefect adalah 1 untuk setiap testi atau *person/case*.

ANALISIS ITEM MENGGUNAKAN PROGRAM QUEST

Analisis item menggunakan program QUEST memberikan informasi hasil analisis item menurut teori tes klasik (*classical test theory* atau CTT) dan menurut teori tes modern atau teori respons item (*item response theory* atau IRT). IRT hasil program QUEST mengacu kepada model logistic satu parameter (1-parameter logistic) atau disingkat model

1-PL. Dalam hal ini parameter yang dimaksud adalah tingkat kesulitan item. Model ini dikenal dengan model Rasch untuk data dengan skala dikotomus (kategori-1 bila memiliki skor 0 dan kategori-2 bila memiliki skor 1). Untuk data dengan skala politomus (lebih dari dua kategori (misalnya kategori-1bila memiliki skor 0, kategori-2 bila memiliki skor 1, kategori-3 bila memiliki skor 2, dan dapat ditambah kategori selanjutnya sesuai dengan penambahan skor yang dimiliki). Program QUEST dapat menganalisis data skala politomus sampai 10 kategori (kategori terendah yakni kategori-1 yakni berskor 0 karena salah atau melewatinya, dan kategori tertinggi adalah kategori-10 yakni kategori berskor 9 karena tidak mengerjakan akibat kehabisan waktu. Program QUEST menganalisis data politomus dengan mengacu kepada model kredit parsial (Partial Credit Model) yang merupakan perluasan model Rasch, sehingga tetap menggunakan model 1-PL.

Sebagaimana prinsip IRT, bahwa syarat yang dikenakan adalah bahwa seluruh item bersifat unidimensi, artinya mengukur aspek yang sama dan indeks kemampuan atau abilitas (*ability*) testi diplot pada satu garis yang sama dengan tingkat kesulitan item. Dengan demikian, bila ada testi yang tidak *fit* dengan model dapat dimaknai bahwa pola respons abilitas testi yang bersangkutan tidak sesuai/sejalan dengan pola respons item yang dikerjakannya. Bila item tidak fit dengan model maka item yang bersangkutan tidak satu dimensi dengan item-item yang lainnya, dengan kata lain tidak mengukur dimensi yang sama.

Program analisis atau disebut program kalibrasi menggunakan IRT mendasarkan pda distribusi logistik, yakni distribusi yang menyerupai distribusi normal dengan nilai logistik D sebesar 1,7. Analisis item menggunakan IRT ada yang melakukan kalibrasi berdasar berdasar satu parameter yakni hanya didasaran pada tingkat kesulitan (diberi simbol β atau b) sehingga disebut model satu paramemeter logistik tau model 1-PL atau disebut Model Rasch (*Rasch Model*). Ada yang mendasarkan pada dua parameter, yakni daya beda (diberi simbol a) dan tingkat kesulitan (b) sehingga disebut model 2-PL. Ada pula yang mendasarkan pada tiga parmeter, yakni daya pembeda, tngkat kesukaran, dan *guessing* (diberi simbol c), sehingga disebut Model 3-PL.

UKURAN SAMPEL

Semakin bertambah banyak parameter di dalam model politomus sebagai lawan model dikotomus, akan semakin bertambah pula informasi di dalam data. Namun,

diperlukan estimasi yang stabil di dalam ukuran sampel yang sama. Ukuran sampel untuk data politomus menggunakan *Graded Model* (GM) yang merupakan model 2-PL sekitar 250 dapat diterima untuk aplikasi dalam penelitian, sedangkan 500 sampai 1000 untuk penggunaan operasional (Muraki & Bock, 1998: 35). Ahli lain ada yang menyatakan bahwa untuk keperluan kalibrasi dalam IRT ukuran sampel antara 200 sampai 1000 tergantung model yang dipilih. Penelitian disertasi dapat menggunakan sampel yang kecil (Crocker & Algina, 1986: 322). Sebagian ahli menyatakan bahwa ukuran sampel khusus untuk model 1-PL berupa *Rasch Model* (RM) antara 30 sampai 300 dengan batas INFIT t sebesar -2 sampai +2 (Bond & Fox, 2007: 43). Jadi dalam hal ini menggunakan batas kesalahan 5%, sehingga besarnya nilai INFIT t ±1,96 atau dibulatkan menjadi ±2,0. Dengan demikian, suatu item menjadi tidak *fit* menurut Model Rasch bila memiliki nilai <-2,0 atau > +2,0 (probability atau peluang <0,05).

PROSEDUR ANALISIS ITEM MENGGUNAKAN PROGRAM QUEST

Hasil tes uraian dapat dianalisis menurut model kredit parsial (*Partial Credit Model* atau PCM) sedangkan hasil tes pilihan dianalisis dianalisis menurut model Rasch (*Rasch Model* atau RM).

Beberapa pertimbangan dalam pemakaian PCM sebagai perluasan RM yang merupakan model 1-PL, dapat menggunakan sampel yang tidak sebesar kalau melakukan kalibrasi data politomus menggunakan model 2-PL atau 3-PL (Keeves & Masters, 1999: 12-13). Kedua, bahwa karakteristik respons terhadap setiap item mengikuti PCM yakni bahwa tingkat kesulitan dari suatu tahapan kategori di bawahnya ke kategori di atasnya tidak sama antaritem satu dan yang lain, sehingga besarnya *delta* untuk suatu tahapan kategori di bawahnya dan *delta* untuk tahapan kategori di atasnya tidak sama antaritem satu dengan item lainnya.

PCM adalah perluasan dari model 1-PL/RM. Persamaan RM menurut Han & Hambleton (2007: 15) dituliskan sebagai berikut.

$$P_i(\theta) = \frac{1}{1 + e^{-D(\theta - b_i)}} \qquad (1)$$

e adalah konstanta eksponensial yang nilainya kira-kira 2,718, dan *D* adalah faktor penskalaan yang nilainya 1,7. Nilai kemungkinan setiap testi berhasil mengerjakan item i

dipahami sebagai fungsi logistik perbedaan dua parameter, yakni parameter kemampuan/kecakapan sebesar θ dan parameter tingkat kesulitan item sebesar b_i. Persamaan nomor satu dapat ditulis kembali sebagai berikut.

$$P_i(\theta) = \frac{1}{1+e^{-D(\theta-b_i)}} = \frac{\exp(D(\theta-b_i))}{1+\exp(D(\theta-b_i))} = \frac{P_{i1}(\theta)}{P_{i0}(\theta)+P_{i1}(\theta)} \qquad (2)$$

$P_{i1}(\theta)$ adalah peluang testi yang dipilih acak, yang memiliki tingkat kemampuan/kecakapan (*proficiency level*) sebesar θ untuk meraih skor 1 pada item *i*, sedangkan $P_{i0}(\theta)$ adalah peluang testi yang dipilih acak, yang memiliki tingkat kemampuan sebesar θ untuk memperoleh skor 0 pada item *i*.

Persamaan RM dalam bentuk persamaan untuk testi (*case/person*) n dan item i dengan skor x sebesar 0 atau 1 dengan kemampuan sebesar β dan tingkat kesulitan item sebesar δ dituliskan sebagai berikut (Masters, 1999: 101; Wright & Masters, 1982: 39-40).

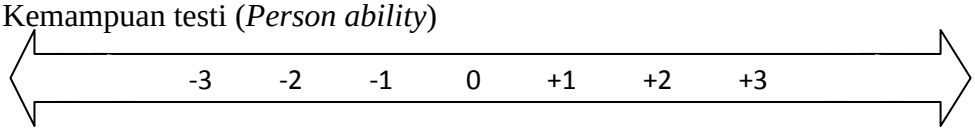
$$P_{nix} = \frac{1}{1+exp(\beta_n - \delta_{i1})} \quad \text{untuk x = 0} \qquad (3)$$

dan

$$P_{nix} = \frac{exp(\beta_n - \delta_{i1})}{1+exp(\beta_n - \delta_{i1})} \quad \text{untuk x = 1} \qquad (4)$$

Tingkat kemampuan testi maupun tingkat kesulitan item dalam *Rasch Model* (RM) diekspresikan pada satu garis berupa absis pada grafik dengan satuan berupa logit (*logg-odd unit*). Garis tersebut terbentang dari -∞ sampai dengan +∞, bila digambar akan tersaji pada Gambar 1 (Keeves & Alagumalai, 1999: 27).

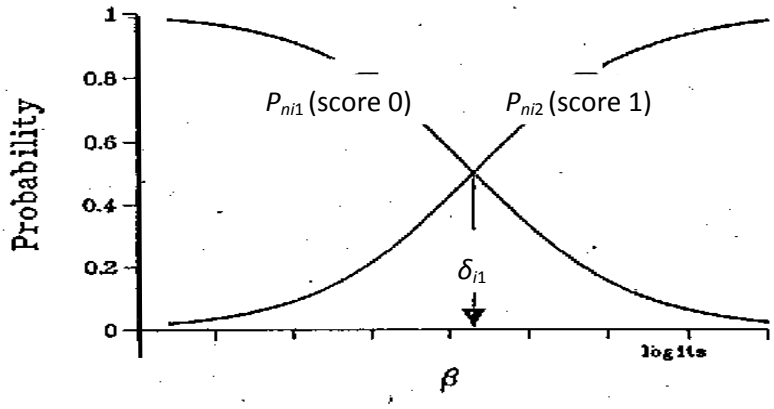
Kemampuan testi (*Person ability*)



Tingkat kesulitan item (*Item difficulty*)

Gambar 1. Skala Rasch
(Sumber: Keeves & Alagumalai, 1999: 27).

Lokasi tingkat kesulitan sebesar δ_{i1} pada grafik merupakan perpotongan antara curve respons P_{ni1} berskor 0 dengan kurve respons P_{ni2}. ber skor 1, dan pada grafik akan terlihat sebagaimaa tersaji pada Gambar 2.



Catatan: Titik potong antara kurve peluang skor 0 (kategori-1) dan skor 1 (kategori-2) menunjukkan lokasi *Delta*-1 sebesar δ_{i1}

Gambar 2. Kurve Peluang pada Skala Dikotomus Menurut *Rasch Model*
(Sumber: Wright & Masters, 1982: 40)

Skala politomus memiliki skor x sebesar 0, 1, 2, 3 ....., m_i. Peluang seorang testi (*case/person*) pada tingkat kemampuan θ meraih skor sebesar *x* di atas *x*-1 dapat dihitung dengan persamaan sebagai berikut (Han & Hambleton, 2007: 15).

$$\frac{P_{ix}(\theta)}{P_{ix-1}(\theta)+P_{ix}(\theta)} = \frac{\exp(D(\theta - b_{ix}))}{1+\exp(D(\theta - b_{ix}))} \text{ untuk x = 0, 1, 2, ..., } m_i \qquad (5)$$

$P_{ix}(\theta)$ dan $P_{ix-1}(\theta)$ mengacu pada peluang seorang testi (*case/person*) sebesar θ, meraih skor *x* dan *x*-1. Hal yang perlu diperhatikan bahwa pada persamaan nomor delapan, jumlah parameter kesukaran item kini menjadi m_i (jumlah kategori respons dikurangi satu). Peluang seorang testi (*case/person*) yang dipilih acak, dengan tingkat kemampuan sebesar

θ, untuk memperoleh skor x pada item i dapat dituliskan dengan persamaan sebagai berikut.

$$P_{ix}(\theta) = \frac{\exp \sum_{k=0}^{x} (D(\theta - b_{ik}))}{\sum_{h}^{m_i} \exp \sum_{k=0}^{h} (D(\theta - b_{ik}))} \quad \text{untuk x = 1, 2, 3, ….., } m_i \qquad (6)$$

Fungsi Persamaan nomor sembilan sering disebut fungsi respons kategori skor (*score category response function* atau SCRF).

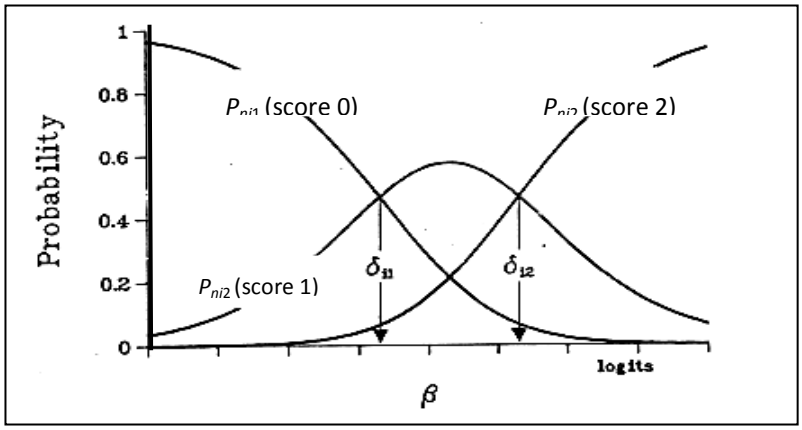
Masters (1999: 101) dan Wright & Masters, (1982: 39) menuliskan persamaan nomor untuk testi (*case/person*) n dan item i dengan skor x sebesar 0, 1, 2, …., m_i dengan kemampuan sebesar β dan tingkat kesulitan item sebesar δ yang dituliskan dalam PCM dengan rumus sebagai berikut.

$$P_{nix} = \frac{1}{1 + exp(\beta_n - \delta_{ij})} \quad \text{untuk x = 0} \qquad (7)$$

dan

$$P_{nix} = \frac{exp\ (\beta_n - \delta_{ij})}{1 + exp\ (\beta_n - \delta_{ij})} \quad \text{untuk x = 1, 2, 3, ..., } m_i \qquad (8)$$

Dengan demikian, tingkat kesulitan item (*difficulty*) untuk item i sebesar δ akan terurai menjadi nilai *delta* sebesar δ_{ij} untuk x = 1, 2, 3, m_i. Item nomor 1 yang memiliki tiga kategori atau diskor secara politomus tiga kategori, memiliki δ_{11} dan δ_{12}, item nomor 2 memiliki δ_{21} dan δ_{22}. Besarnya nilai *delta*-1 menunjukkan nilai yang diperlukan testi (*case/person*) untuk berpindah dari kategori-1 (skor 0) ke kategori-2 (skor 1) dan nilai *delta*-2 menunjukkan nilai yang diperlukan untuk berpindah dari kategori-2 (skor 1) ke kategori-3 (skor 2). Besarnya *delta*-1 dapat lebih kecil, sama, atau lebih besar dari *delta*-2. *Delta*-1 dalam grafik menunjukkan perpotongan antara kurve respons P_{ni1} (skor 0) dengan kurve respons P_{ni2} (skor 1) dan *delta*-2 menunjukkan perpotongan kurve respons kurve respons P_{ni2} (skor 2) dengan kurve respons P_{ni3} (skor 3), dan pada grafik akan terlihat pada Gambar 3.



Catatan: Titik potong antara kurve peluang skor 0 (kategori-1) dan skor 1 (kategori-2) menunjukkan lokasi *Delta*-1 sebesar δ_{i1}, titik potong kurve peluang skor (kategori-2) dan skor 2 (kategori-3) menunjukkan lokasi *Delta*-2 sebesar δ_{i2}

Gambar 3. Kurve Peluang Skala Politomus Tiga Kategori Menurut *Partial Credit Model* (PCM)
(Sumber: Wright & Masters, 1982: 44).

Kategori pada *Rating Scale Model* atau RSM, merupakan kategori yang berjenjang (*ordered category*). Sebagai contoh, respons item 1, 2, 3 dengan kategori-1 "tidak setuju", kategori-2 "setuju", dan kategori-3 "sangat setuju merupakan kategori yang berjenjang. Oleh karena itu, nilai *delta* sebesar δ_{ij} (yang menunjukkan karakteristik spesifik tingkat kesulitan item i pada tahapan/step/kategori j) dipecah menjadi tingkat kesulitan (*difficulty*) sebesar δ_i ditambah nilai *tau* sebesar τ_{ij} (yang menunjukkan karakteristik spesifik tingkat kesulitan suatu tahapan/step/kategori j dari item i) mengikuti pada formula yang diajukan Andrich (1978). Oleh karena itu, persamaannya dapat ditulis sebagai berikut (Masters, 1999: 101 dan Wright & Masters, 1982: 39).

$$P_{nix} = \frac{exp(\beta_n - \delta_i - \tau_{ij})}{1 + exp(\beta_n - \delta_i - \tau_{ij})} \quad \text{untuk x = 1, 2, 3, ..., } m_i \qquad (9)$$

Elemen sentral dari program QUEST adalah IRT mengikuti *Rasch Model* (RM). Dalam hal ini, dapat pula digunakan pada data respons yang diskor secara politomus.

Program QUEST dalam melakukan estimasi parameter, baik untuk item maupun untuk testi (*case*/*person*) menggunakan *unconditional* (UCON) atau *joint maximum likelihood* (Adam & Khoo, 1996: 89).

Skor mentah seorang testi dalam penskalaan sebesar *r* dikonversi menjadi skala logit yang menunjukkan n kemampuan sebesar b = log[(*r*/(*L*-*r*)], di mana *L* adalah banyaknya *activities* (item). Sementara, nilai r dapat dikonversi menjadi skala logit yang menunjukkan tingkat kesulitan sebesar d = log[(*N*-*S*)/*S*)], di mana *N* adalah banyaknya testi (*case*/*person*) dan *S* adalah skor suatu item (Wright & Masters, 1982: 28-31). Besarnya *S* untuk data pengukuran yang diskor secara politomus dalam program QUEST diubah menjadi w_{ij} dan tingkat kesulitan sebesar d akan diubah menjadi nilai δ_{ij}. Persamaan untuk RSM dalam program QUEST dituliskan sebagai berikut.

$$P(X_{ni} = x_{ni}) = \frac{\exp \sum_{j=0}^{X_n} w_{ij}(\beta_n - \delta_i - \tau_{ij})}{\sum_{k=0}^{ni} \exp \sum_{j=0}^{k} w_{ij}(\beta_n - \delta_i - \tau_{ij})} \qquad (10)$$

β_n adalah komponen tingkat kemampuan (*ability*) dari testi (*case*/*person*) n, w_{ij} adalah skor yang ditetapkan untuk step j dalam suatu item i, sedangkan *difficulty* sebesar δ_i serta *tau* sebesar τ_{ij} adalah karakteristik spesifik tingkat kesulitan item dan tingkat kesulitan kategori j dari item i mengikuti formula yang diajukan Andrich, 1978 (Wright & Masters, 1982: 28-31; Swaminathan, 1999: 50). Item i yang diskor secara politomus tiga kategori menurut RSM memiliki satu nilai kesukaran item atau *item difficulty* sebesar δ_{ij} dan dua buah nilai parameter *tau* berupa *tau*-1 dan *tau*-2 sebesar τ_{i1} dan τ_{i2}.

Selain menyajikan nilai *tau*, program QUEST menyajikan tingkat kesulitan dalam bentuk nilai *threshold* (ambang batas) untuk RSM. Nilai *threshold* yang dihitung berdasarkan nilai *tau* sebagaimana yang diperkenalkan oleh Masters (1988) mengikuti kaidah Thurstone (Adam & Khoo, 1996: 90). Dengan demikian, suatu *threshold* (yang juga diberi simbol τ_{ij} oleh Wright & Master, 1982) merepresentasikan atau keberfungsiannya identik dengan nilai *tau*. Nilai *threshold* untuk suatu tahapan dari item i adalah tingkat kemampuan (*ability level*) yang dibutuhkan oleh testi (*case/person*) untuk melewatinya dengan peluang 0,50 (peluang tertinggi). Nilai *threshold* mengikuti kaidah Thorstone yang diasumsikan bahwa setiap respons tunggal berada pada lokasi μ (baik

respons kelompok maupun individual). Proses respons yang melalui teknik penskalaan diubah ke dalam fungsi logistik (Andrich, 1999: 113-114).

Besarnya menjadi w_{ij} sebagai skor yang ditetapkan untuk step j suatu item i sesuai dengan banyaknya kategori, dan besarnya menjadi w_{ij} adalah 0, 1, 2, ... m. Bila besarnya menjadi w_{ij} = 0, maka persamaannya dapat dituliskan dalam ekspresi tunggal dalam formula untuk menjadi *w* :

$$\exp \sum_{j=0}^{0} w_{ij}(\beta_n - \delta_i - \tau_{ij}) \equiv 1 \qquad (11)$$

Untuk kepastian identifikasi digunakan dua *cosntraint*, yakni:

$$\sum_{j=0}^{ni} \tau_{ij} \equiv 0 \text{ dan } \sum_{i=1}^{l} \delta_i \equiv 0.$$

Besarnya δ_i ditambah τ_{ij} pada RSM sama dengan δ_{ij} pada *Partial Credit Model* (PCM) (Wright & Masters, 1982: 56; Swaminathan, 1999: 51). Dengan demikian, persamaan untuk RSM dapat dituliskan menjadi persamaan untuk PCM sebagai berikut.

$$P(X_{ni} = x_{ni}) = \frac{\exp \sum_{j=0}^{x_n} w_{ij}(\beta_n - \delta_{ij})}{\sum_{k=0}^{ni} \exp \sum_{j=0}^{k} w_{ij}(\beta_n - \delta_{ij})} \qquad (12)$$

Penerapan pada data pengukuran yang skor secara dikotomus akan direduksi sehingga formula *Rasch Model* (RM) atau disebut model parameter logistik dalam program QUEST dituliskan dengan persamaan:

$$P(X_{ni} = x_{ni}) = \frac{\exp(x_{ni} w_{ij}(\beta_n - \delta_i))}{1 + \exp(x_{ni} w_{ij}(\beta_n - \delta_i))} \qquad (13)$$

a. Perhitungan Estimasi untuk Item

Penetapan *fit* item secara keseluruhan dengan model dalam program QUEST (Adam & Kho, 1996) didasarkan pada besarnya nilai rata-rata INFIT *Mean of Square* (INFIT MNSQ) beserta simpangan bakunya atau nilai rata-rata INFIT *Mean of* INFIT *t*.

Penetapan *fit* tiap item dengan model dalam program QUEST didasarkan pada besarnya nilai INFIT MNSQ atau nilai INFIT *t* item yang bersangkutan.

Besarnya kuadrat tengah yang tertimbang (*Wighted Mean Square*)—dalam program QUEST disingkat INFIT MNSQ)—adalah dengan ekspektasi sebesar 1 dan varians sebesar 0. Sementara besarnya kuadrat tengah tertimbang terstandar (*Standardized Weighted Mean Square*) atau t_i dengan ekspektasi sebesar 0 dan varians sebesar 1.

b. Perhitungan Estimasi untuk Testi

Penetapan *fit* testi (*case/person*) secara keseluruhan dengan model dalam program QUEST (Adam & Kho, 1996) juga didasarkan pada besarnya nilai rata-rata INFIT *Mean of Square* (INFIT MNSQ) beserta simpangan bakunya. Dapat pula didasarkan pada besarnya nilai rata-rata INFIT *Mean of* INFIT *t*. Penetapan *fit* tiap testi (*case/person*) dengan model dalam program QUEST didasarkan pada besarnya nilai INFIT MNSQ atau nilai INFIT *t* item yang bersangkutan (Wright & Masters, 1982: 108-109).

Besarnya jumlah kuadrat tertimbang (*Weighted Sum of Square*) untuk setiap testi (*case/person*) dengan ekspektasi sebesar 1 dan varians sebesar 0. (atau mau menggunakan kriteria menurut

c. Pengujian Validitas untuk Mengetahui *Fit* Item dan Testi terhadap Model

Item characteritic curve (ICC) akan mendatar (*flat*) bila besarnya INFIT MNSQ untuk item atau *e* lebih besar dari satuan logit > 1,30 atau <0,77. Akibatnya membentuk *platokurtic curve* dan tidak lagi membentuk *leptokurtic curve* (Keeves & Alagumalai 1999: 36). Oleh karena itu, dalam program QUEST ditetapkan bahwa suatu item atau testi/*case*/*person* dinyatakan *fit* dengan model dengan batas kisaran INFIT MNSQ dari 0,77 sampai 1,30 (Adam & Khoo, 1996:30 & 90). Ada pula peneliti yang menggunakan batas yang lebih ketat, yakni dengan kisaran 0,83 sampai dengan 1,20 dan ada yang menggunakan pengujian berdasarkan besarnya nilai INFIT t. Dalam hal ini menggunakan kisaran nilai t adalah ± 2,0 (pembulatan ± 1,96) jika taraf kesalahan atau *alpha* sebesar 5% (Keeves & Alagumalai 1999: 34-36; Bond & Fox, 2007: 43). Dengan demikian, suatu item menjadi tidak *fit* menurut Model Rasch bila memiliki nilai <-2,0 atau > +2,0 (probability atau peluang <0,05).

d. Estimasi Reliabilitas

Etimasi reliabilitas menurut IRT dihitung berdasarkan item disebut indeks sparasi item dan berdasarkan testi (*case/person*) dan disebut dengan indeks sparasi person. Semakin tinggi estimasi ideks sparasi item semakin tepat keseluruhan item dianalisis menurut model yang digunakan (apakah menurut RM, PCM, atau RSM). Semakin tinggi indeks sparasi *person* semakin konsisten setiap item pengukur digunakan untuk mengukur testi yang bersangkutan. Estimasi reliabilitas berdasarkan testi (*case/person*) sama kedudukannya dengan reliabilitas menurut CTT—yakni reliabilitas menurut alpha Cronbach untuk data politomus dan reliabilitas menurut Kuder-Richardson-20 untuk data dikotomus. Indeks separasi item (*item separation index* atau R_I) oleh Wright & Master (1999: 96) disebutnya dengan istilah ”reliabilitas sampel”, sedangkan indeks sparasi person disebut dengan ”reliabilitas tes”.

Program QUEST juga menyajikan hasil realiabilitas tes menurut CTT, yakni berupa indeks konsistensi internal, yang untuk penskoran politomus merupakan indeks alpha Cronbach dan untuk penskoran dikotomus merupakan indeks KR-20 (Adam & Khoo, 1996: 93). Dalam hal ini, reliabilitas yang berlaku adalah untuk tes yang berfungsi seleksi, bukan untuk pengukuran prestasi. Untuk pengukuran prestasi maka perlu dikonversi menjadi indeks Kappa atau indeks persetujuan (*agreement index*).

I. Penyiapan file Perintah dan File Data

A. Untuk Pilihan Ganda dengan Data ditulis Menggunakan Huruf

1. File Perintah

```
title PRESTASI (50 ITEM PG dengan 4 alternatif)
data_file prest.txt
codes 0ABCD9
format id 1-4 items 5-50 items 52-54
key CCBABCCBBACCBBABACBAAACDBBCDCBDABDDBBADBAACDCCCBCD
set width=107 ! page
estimate
show >> prestsh.out
show items >> prestit.out
show cases >> prestca.out
itanal >> presttn.out
logit >> prestlo.out
quit
```

Keterangan

a. Title PRESTASI (50 ITEM PG dengan 4 alternatif) → menunjukkan nama identitas file
b. data_file prest.txt → menunjukkan nama file data. Dalam hal ini dapat pula diberi nama dengan ekstensi .dat bila komputer tidak berisi program macromedia
c. codes 0ABCD9 → kode bahwa data ditulis dalam bentuk huruf A, B, C, D dengan 0 bila dilewati dan 9 bila tidak dikerjakan (omit)
d. format id 1-4 items 5-54 → spasi 1 sampai 4 untuk identitas testi (dalam hal ini hanya menggunakan nomor), dan spasi 5 sampai 54 adalah untuk data sebanyak 50 item
e. key CCBABCCBBACCBBABACBAAACDBBCDCBDABDDBBADBAACDCCCBCD → kunci jawaban
f. set width=107 ! page → lebar halaman kertas
g. estimate → diestimasi secara otomatis menurut program QUEST
h. show ! scale=all >> prestsh.out → hasil analisis secara simultan
i. show items >> prestit.out → hasil analisis menyajikan informasi tentang item secara singkat (estimasi tingkat kesukaran, nilai INFIT MNSQ, nilai INFIT t)
j. show cases >> prestca.out → hasil analisis menyajikan informasi testi (skor mentah, estimasi skor kalibrasi, nilai INFIT MNSQ, nilai INFIT t)
k. itanal ! scale=all >> presttn.out → hasil analisis menyajikan informasi tentang item secara lengkap hasil analisis menurut CTT dan IRT
l. quit → kode perintah diakhiri

Catatan:
Dengan menuliskan angka 0 dan 9 pada code di file perintah maka testi yang tidak mengerjakan dengan cara melompati soal yang bersangkutan diberi skor 0, sedangkan yang tidak mengerjakan atau omit diberi skor 9

Setelah selesai simpan dengan extensi .CTL → beri nama prest.ctl (*jangan lupa gunakan menu all file saat menyimpan supaya tidak ganda ekstensinya*).
Catatan: beri nama dengan nama depan yang konsisten agar tidak bermasalah ketika diesekusi. Misalnya, dengan nama file perintah prest.ctl maka file data diberi nama prest.txt dan hasil diawali pula dengan prest sehingga menjadi prestsh.out kemudian prestit.out dan seterusnya seperti contoh di atas.

2. File data

```
001 CCBABCCBBACCBBABACBAAACDBBCDCBDABDDBBADBAACDCCCBCD
002 CCBABCCBBACCBBABACBAAACDBBCDCBDABDDBBADBAACDCCCBCD
003 CCBABCCBBACCBBABACBAAACDBBCDCBDABDDBBADBAACDCCCBCD
004 CCBABCCBBACCBBABACBAAACDBBCDCBDABDDBBADBAACDCCCBCD
```

B. Untuk Pilihan Ganda dengan Data Ditulis Menggunakan Angka

1. File Perintah

```
title PRESTASI (50 ITEM PG dengan 4 alternatif)
data_file prest.txt
codes 012349
format id 1-4 items 5-54
key 12134234432123234342 13432341234323444432123223344221133123 44
set width=107 ! page
estimate
show >> prestsh.out
show items >> prestit.out
show cases >> prestca.out
itanal >> presttn.out
quit
```

Catatan:

Dengan menuliskan angka 0 dan 9 pada code di file perintah maka testi yang tidak mengerjakan dengan cara melompati soal yang bersangkutan diberi skor 0, sedangkan yang tidak mengerjakan atau omit diberi skor 9

2. File data

```
001 121342344321232343421343234123432344443212322334422113312344
002 321342343321232343421343234123432344443212322334422313312342
003 131322344321232343421343234123432344443212322334422113312344
004 121342344321232343423343234123432344443212322334422143312344
```

C. File Perintah dan File Data untuk Analisis Data Politomus

1. Untuk penyiapan hasil angket dengan pilihan:

1= tidak pernah
2= jarang
3= sering
4= selalu

a. File perintah

```
title AKTIVASI BELAJAR OLEH GURU (14 ITEM)
data_file GRBLJ.DAT
codes 1234
format id 1-12 items 14-27
set width=107 ! page
estimate
show ! scale=all >> 5bljgrsh.out
show items >> bljgrit.out
show cases >> bljgrca.out
itanal ! scale=all >> bljgrtn.out
quit
```

title AKTIVASI BELAJAR OLEH GURU (14 ITEM) → nama file
data_file GRBLJ.DAT → nama file data GRBLJ.DAT harus diganti dengan ektensi .txt bula komputer berisi program makromedia.

codes 1234 → kode bahwa data ditulis dalam bentuk angka 1, 2, 3, dan 4
format id 1-12 items 14-27 → spasi 1 sampai 12 spasi untuk identitas testi
spasi 14 sampai 27 untuk data 14 item
set width=107 ! page → lebar halaman kertas
estimate → diestimasi secara otomatis menurut program QUEST
show ! scale=all >> bljgrsh.out → hasil analisis secara simultan
show items >> bljgrit.out → hasil analisis menyajikan informasi tentang item secara singkat (estimasi tingkat kesukaran, nilai INFIT MNSQ, nilai INFIT t)
show cases >> bljgrca.out → hasil analisis menyajikan informasi testi (skor mentah, estimasi skor kalibrasi, nilai INFIT MNSQ, nilai INFIT t)
itanal ! scale=all >> bljgrtn.out → hasil analisis menyajikan informasi tentang item secara lengkap hasil analisis menurut CTT dan IRT
quit → kode perintah diakhiri

Setelah selesai simpan dengan extensi **.CTL** misalnya **GRBLJ.CTL (*jangan lupa gunakan menu all file saat menyimpan supaya tidak ganda ekstensinya*).**

b. File data

```
Jogonalan204 44433334233322
Jogonalan185 31314344442244
Jogonalan287 23232222333233
Jogonalan070 14433324332333
Jogonalan062 24433334333332
Jogonalan061 23433334333322
```

Setelah selesai simpan dengan extensi **.DAT** misalnya **GRBLJ.DAT (*Jangan lupa ganti ekstensi dengan* .txt *bila ada program multimedia dalam komputer yang digunakan!*)**

2. Untuk tes uraian dengan kunci yang seragam

Misal setiap item diberi skor maksmum 3, maka

alternatif jawaban siswa score 0 = ketegori 1 → dikerjakan dan salah
score 1 = kategori 2 → ikerjakan dan benar 1
score 2 = kategori 3 → dikerjakan dan benar 2
score 3 = kategori 4 → dikerjakan dengan sempurna (skor maksimum)

a. File Perintah

```
title tes prestasi (uraian 18 item)
data_file PRES.dat
codes 012349
format id 1-13 items 14-32 grP 33
key 0000000000000000000 ! score=0
key 1111111111111111111 ! score=1
key 2222222222222222222 ! score=2
key 3333333333333333333 ! score=3
key 4444444444444444444 ! score=4
set width=107 ! page
estimate
show >> PRESsh.out
show items ! stat=tau >> PRESit.out
show cases >> PRESca.out
itanal >> PREStn.out
logit >> PRESlo.out
quit
```

Catatan:
Kode 012349 artinya skor dari 0 sampai 4 (jadi ada lima kaegori) dan diberi 9 jika omit

Setelah selesai simpan dengan extensi **.CTL** misalnya **PRESU.CTL**

b. File data

```
Jogonalan204 44033334233322
Jogonalan185 31314344042244
Jogonalan287 23232222333233
Jogonalan070 14430324339333
Jogonalan062 24433334333332
Jogonalan061 23433334333322
... dan seterusnya
```

Setelah selesai simpan dengan extensi **.DAT** misalnya **PRESU.DAT**
Atau diberi ekstensi**.TXT**

3. Untuk Item Bentuk Benar-Salah atau Isian Singkat

Misalnya setiap item diberi skor 0 bila salan dan 1 bila benar.

a. File Perintah

```
EKSAMINI 165 ITEM 47 (KOLOM 6-52)
data_file FA.dat
codes 01
format id 1-5 items 6-52
set width=107 ! page
estimate!ITER=20
show >> PFAsh.out
show items ! stat=DELTA >> PFAiD.out
show items ! stat=TAU >> PFAit.out
show cases >> PFAca.out
show cases ! form=export >> PFAsc.out
```

b. File data

```
001 10101111000011100001111000010101010000111111111
002 11101111000011100101111000010101010000111101111
003 10101111000011100001111000010101010000111111111
004 11101111000011100101111000010101010000110101111
005 00101011000011100001111000010101010000111111111
```

C. Untuk Data untuk Analisis Data Kombinasi Dikotomus dan Politomus

1. Kombinasi PG dan uraian

a. File Perintah

```
title INDONESIAN LANGUAGE KELAS 3 (25 PG DAN 5 URAIAN)
data_file ind1.dat
codes 012349
format id 1-50 items 51-90
key 212414234323221341432134311111 ! score=1
key xxxxxxxxxxxxxxxxxxxxxxxxx22222 ! score=2
key xxxxxxxxxxxxxxxxxxxxxxxxxxxx3x ! score=3
key xxxxxxxxxxxxxxxxxxxxxxxxxxxx4x ! score=4
set width=107 ! page
estimate
show >> ind1sh.out
show items ! stat=tau >> ind1it.out
show cases >> ind1ca.out
itanal >> ind1tn.out
```

b. File data

```
001 212211231143423431412144300020
002 313412224144234331434412220100
003 444211431312244423332111400012
004 413221221142421434242331200000
005 113431121132423331113311301030
006 344431231242222123321231201100
007 213344111342234222343331202141
```

2. Kombinasi PG, Isian singkat, dan Uraian

a. File Perintah

```
title  INDONESIAN LANGUAGE KELAS 3 (25 PG, 10 ISISNGKT, 5 URAIAN)
data_file ind1.dat
codes 012349
format id 1-50 items 51-90
key 2124142343232213414321343111111111111111 ! score=1
key xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx22222 ! score=2
key xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx3x ! score=3
key xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx4x ! score=4
set width=107 ! page
estimate
show >> ind1sh.out
show items ! stat=tau >> ind1it.out
show cases >> ind1ca.out
itanal >> ind1tn.out
```

b. File data

```
DENI         B1 212211231143423431412144301010000000020
SUKRO        B1 313412224144234331434412201000000000100
ANTAKA       B1 444211431312244423332111411000090000000
TATU         B1 413221221142421434242331200010000000000
SUPI         B1 113431121132423331113311300000000000000
.... dan seterusnya
```

II. PERINTAH ANALISIS

Langkah untuk analisis sebagai berikut.

1. Klik **QUEST**

2. Ketik **SUBMIT** spasi kemudian NAMA FILE PERINTAH LENGKAP kode extensinya. Jika nama File Perintahnya **prest.ctl** maka perintahnya sebagai berikut.

>SUBMIT **PREST.CTL**
Atau
>submit prest.ctl

3. Kemudian **tekan tombol ENTER**

   Janga lupa, ada jarak **1 spasi** antara tulisan **submit** dan **nama file perintah!**

III. MEMBACA HASIL ANALISIS

A. Hasil analisis Data Dikotomus soal bentuk PG

Out put yang diberi kode dengan akhiran file sh.out misalnya prestsh.out

Berikut ini disajikan hasil analisis data menggunakan program QUEST. Sekali lagi agar mudah mengingat maka penamaan file out put dinamai dengan akhirtan sh.out seperti pada file **prestsh.out** (sh dari singkata show) untuk menampilkan hasil dalam bentuk informasi global beserta grafiknya. Pemakaian akhiran it.out seperti pada file **prestit.out** (it singkata item) untuk memperoleh informasi item secara global, sementara file dengan akhiran tn.out seperti pada nama file **presttn.out** adalah file yang memberikan informasi detail analisis item (tn dari singkatan item analysis). Pemakaian akhiran ca seperti pada nama file prestca.out (ca singkatan case) untuk memperoleh informasi skor case/testi/person.

```
PRESTASI
-------------------------------------------------------------------------------
Current System Settings                                          9/ 2/ 8 12:25
all on all (N =16699 L = 50 Probability Level= .50)
-------------------------------------------------------------------------------
Data File     =  prest.txt
Data Format   =  id 1-20 items 21-70

Log file      =  LOG not on

Page Width    =  107
Page Length   =   65
Screen Width  =   78
Screen Length =   24

Probability level =   .50

Maximum number of cases set at 60000

VALID DATA CODES      A B C D

GROUPS
1 all                 (16699 cases ) : All cases

SCALES
1 all                 (   50 items ) : All items

DELETED AND ANCHORED CASES:

No case deletes or anchors

DELETED AND ANCHORED ITEMS:

No item deletes or anchors

RECODES

SCORING KEYS

Score = 1     CCBABCCBBACCBBABACBAAACDBBCDCBDABDDBBADBAACDCCCBCD
===============================================================================
```

Artinya ada data sebanyak 16699 testi yang dianalisis dengan item sebanyak 50 dengan peluang 0,5 sesai dengan prinsip *Likelihood Maximum.* Tidak ada case (testi), item maupun anchor yang dihapus atau tidak disertakan dalam analisis. Anchor atau common item adalah item yang ada pada dua set yang hasilnya dianalsis secara bersamaandalam sekali analisis agar diperoleh hasil estimasi kemampuan testi dan tingkat kesulitan item kedua pengukuran tersebu menjadi satu skala), shingga hasil kedua tes dapat diperbandingkan, baik dalam hal tingkat kesulitan item, maupun kemampuan testi.

```
PRESTASI
-------------------------------------------------------------------------------
Item Estimates (Thresholds)                                        9/ 2/ 8 12:25
all on all (N =**** L = 50 Probability Level= .50)
-------------------------------------------------------------------------------
 Summary of item Estimates
=========================
 Mean                           .00
SD                             .69
SD (adjusted)                  .69
Reliability of estimate       1.00

 Fit Statistics
===============

 Infit Mean Square         Outfit Mean Square

    Mean    1.00             Mean    1.02
    SD       .06             SD       .08

      Infit t                  Outfit t

    Mean   -1.54             Mean     .45
    SD      9.83             SD      6.44

   0 items with zero scores
   0 items with perfect scores
===============================================================================
```

Nilai reliabilitas tes (untuk ***Norm-Reference***) berdasarkan estimasi item Wrigh & Master (1982) disebut dengan reliabitas sampel. Semakin tinggi nilainya semakin meyakinkan bahwa sampel uji coba sesuai dengan item yang diujikan. Semakin rendah semakin

banyak sampel untuk uji coba yang tidak memberikan informasi yang diharapkan (tidak mengerjakan, atau mengerjakan secara asal-asalan). **Atau justru mengerjakan tetapi sebagian besar testi benar semua atau salah semua, karena dengan mengikuti kurve logistik yang identik dengan kurve normal maka testi yang memiliki skor sempurna dan yang memiliki skor nol tidak dimasukkan dalam analisis.**

Dengan mean INFIT MNSQ 1,0 dan SD 0,6 artinya secara keseluruhan item sesuai dengan model Rasch, karena ini hasil tes pilihan ganda, jadi berupa data dengan skala dikotomus.

```
PRESTASI
-------------------------------------------------------------------------------
Case Estimates                                                   9/ 2/ 8 12:25
all on all (N =**** L = 50 Probability Level= .50)
-------------------------------------------------------------------------------

Summary of case Estimates
=========================

Mean                          -.67
SD                             .64
SD (adjusted)                  .55
Reliability of estimate        .74

 Fit Statistics
===============
 Infit Mean Square         Outfit Mean Square

    Mean    1.00             Mean    1.02
    SD       .11             SD       .18

      Infit t                  Outfit t

    Mean    -.06             Mean     .05
    SD       .90             SD       .71

   0 cases with zero scores
   1 cases with perfect scores
===============================================================================
```

Nilai reliabilitas berdasarkan estimasi case atau testi oleh Wrigh & Master (1982) disebut dengan reliabitas tes. Semakin tinggi nilainya semakin meyakinkan bahwa pengukuran memberikan hasil yang konsisten. Hasil ini juga ditentukan oleh karakteristik sampel.

Semakin rendah berarti juga semakin banyak sampel untuk uji coba yang tidak memberikan informasi yang diharapkan. (tidak mengerjakan, atau mengerjakan secara asal-asalan). **Atau malah mengerjakan tetapi sebagian besar testi benar semua atau salah semua, karena dengan mengikuti kurve logistic yang identik dengan kurve normal maka testi yang memiliki skor sempurna dan yang memiliki skor nol tidak dimasukkan dalam analisis.**

Dengan mean INFIT MNSQ 1,0 dan SD 0,11 artinya secara keseluruhan testi sesuai dengan model Rasch, karena ini hasil tes pilihan ganda, jadi berupa data dengan skala dikotomus.

```
PRESTASI
-------------------------------------------------------------------------------
Item Estimates (Thresholds)                                        9/ 2/ 8 12:25
all on all (N =**** L = 50 Probability Level= .50)
-------------------------------------------------------------------------------
  3.0                             |
                                  |
                                  |
                                  |
                                  |
                                  |
                                  |
                                  |
  2.0                             |
                                  |
                                  |
                                  |
                                  |
                                  |
                              X   |      38     39
                                  |      34
                              X   |
  1.0                         X   |
                              X   |      35
                              X   |
                              X   |       9     20     25     31
                             XX   |      32     40
                             XX   |       1     11     13     19    21
                             XX   |      10     12     43
                            XXX   |       2      3     44     45
                            XXX   |      36     41
   .0                     XXXXXXX   |       8     18     42     46
                            XXXXX   |      14     24     49
                           XXXXXX   |      30
                        XXXXXXXXX   |
                        XXXXXXXXX   |
           XXXXXXXXXXXXXXXXXXXXX    |       5      7     17     22    23   27   29   47   48
                   XXXXXXXXXXXXX    |      33     37
                  XXXXXXXXXXXXXX    |
                 XXXXXXXXXXXXXXX    |      28
                 XXXXXXXXXXXXXXX    |      16     26
 -1.0             XXXXXXXXXXXXXX    |
                  XXXXXXXXXXXXXX    |       4
                    XXXXXXXXXXXX    |      15
                        XXXXXXXXX   |
                          XXXXXXX   |       6
                                  |
                             XXXX   |      50
                              XXX   |
                                  |
 -2.0                         X   |
                              X   |
                                  |
                                  |
                                  |
                                  |
                                  |
                                  |
 -3.0                             |
-------------------------------------------------------------------------------
  Each X represents   85 students
```

Grafik diatas menunjukkan nilai threshold. Khusus skala dikotomus sama besarnya dengan tingkat kesulitan item dalam pengertian sebagai *difficulties index*. Dari grafik di atas dapat diperoleh informasi bahwa yang paling sukar adalah item nomor 38 dan 39, dan yang paling mudah item nomor 50. Setiap tanda X mewakili 85 testi/person.

```
PRESTASI
-------------------------------------------------------------------------------
Item Fit                                                          9/ 2/ 8 12:25
all on all (N =**** L = 50 Probability Level= .50)
-------------------------------------------------------------------------------
INFIT
 MNSQ           .56       .63       .71       .83      1.00      1.20      1.40
-----------------+---------+---------+---------+---------+---------+---------+--
  1 item 1                                 .              |*             .
  2 item 2                                 .              *              .
  3 item 3                                 .              *              .
  4 item 4                                 .             *|              .
  5 item 5                                 .           *  |              .
  6 item 6                                 .            * |              .
  7 item 7                                 .              |*             .
  8 item 8                                 .              |*             .
  9 item 9                                 .              |*             .
 10 item 10                                .              |*             .
 11 item 11                                .             * |              .
 12 item 12                                .              *              .
 13 item 13                                .              *              .
 14 item 14                                .              | *            .
 15 item 15                                .             *|              .
 16 item 16                                .            * |              .
 17 item 17                                .              |   *          .
 18 item 18                                .            * |              .
 19 item 19                                .              |  *           .
 20 item 20                                .             *|              .
 21 item 21                                .              |*             .
 22 item 22                                .            * |              .
 23 item 23                                .            * |              .
 24 item 24                                .              *              .
 25 item 25                                .             *|              .
 26 item 26                                .        *     |              .
 27 item 27                                .              | *            .
 28 item 28                                .            * |              .
 29 item 29                                .              *              .
 30 item 30                                .             * |              .
 31 item 31                                .              |  *           .
 32 item 32                                .             *|              .
 33 item 33                                .              |   *          .
 34 item 34                                .              |    *         .
 35 item 35                                .              | *            .
 36 item 36                                .             *|              .
 37 item 37                                .            * |              .
 38 item 38                                .              |*             .
 39 item 39                                .              *              .
 40 item 40                                .             *|              .
 41 item 41                                .              *              .
 42 item 42                                .             *|              .
 43 item 43                                .              |*             .
 44 item 44                                .              |   *          .
 45 item 45                                .              | *            .
 46 item 46                                .              |         *    .
 47 item 47                                .          *   |              .
 48 item 48                                .          *   |              .
 49 item 49                                .             * |              .
 50 item 50                                .             *|              .
```

Artinya 50 item *fit* atau cocok dengan model Rasch atau model 1-PL dengan batas penerimaan ≥0,77 sampai ≤1,30.

Hasil analisis dengan akhiran it.out misalnya prestit.out

PRESTASI

Item Estimates (Thresholds) In input Order 9/ 2/ 8 12:25
all on all (N =** L = 50 Probability Level= .50)**

	ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1	item 1	409316631		.52 .02	1.03	1.07	3.0	4.7
2	item 2	491216553		.25 .02	1.00	1.02	-.1	1.4
3	item 3	472616561		.31 .02	1.00	1.01	-.4	.5
4	item 4	994016645		-1.09 .02	.98	.98	-3.4	-2.1
.	……	.	…….……	. .	…..	.	….	…..
…..	…….							
.	……	.	…….……	. .	…..	.	….	…..
…..	…….							
.	……	.	…….……	. .	…..	.	….	…..
…..	…….							
46	item 46	542716571		.09 .02	1.20	1.24	26.2	20.6
47	item 47	744516579		-.46 .02	.90	.89	-20.5	-11.3
48	item 48	760316561		-.50 .02	.91	.90	-22.1	-10.2
49	item 49	608116539		-.10 .02	.95	.95	-7.6	-4.6
50	item 50	1194416529		-1.68 .02	.97	.97	-3.7	-2.3
Mean				.00	1.00	1.02	-1.5	.5
SD				.69	.06	.08	9.8	6.4

SCORE dan MAXSCR (maximum score) tergabung menjadi satu karena banyaknya testi, tetapi harus diingat bahwa data yang dianalisis berasal dari 16699 testi, sehingga angka 409316631 adalah 4093 dan 16631. Artinya, skor untuk item 1 sebanyak 4093 dari skor maksimum 16631, jadi ada 4093 testi yang mengerjakan dengan benar dari 16631 siswa yang mengerjakan.

Catatan: Bila menganalisisnya dengan program QUEST yang baru, yang bersimbul Q, maka hasilnya sudah terpisah dengan baik, karena memang kemampuannya lebih besar, yakni menganalisis 100.000 case untuk 1200 item. Sementara program Quest berkode QUEST hanya mampu menganalisis 60.000 testi untuk maksimum 400 item.

Berikut hasil program QUEST berkode Q**.**

PRESTASI (50 ITEM PILIHAN GANDA)

Item Estimates (Thresholds) In input Order 3/12/2009 19:58

all on all (N = 16699 L = 50 Probability Level=0.50)

	ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1	item 1	4093	16631	0.52 .02	1.03	1.07	3.0	4.7
2	item 2	4912	16553	0.25 .02	1.00	1.02	-0.1	1.4
3	item 3	4726	16561	0.31 .02	1.00	1.01	-0.4	0.5
4	item 4	9940	16645	-1.09 .02	0.98	0.98	-3.4	-2.1
.	……	. ……	.	….. . ….	…..	…..	…….	
.	……	. ……	.	….. . ….	…..	…..	…….	
.	……	. ……	.	….. . ….	…..	…..	…….	
46	item 46	5427	16571	0.10 .02	1.20	1.24	26.2	20.6
47	item 47	7445	16579	-0.46 .02	0.90	0.89	-20.5	-11.3
48	item 48	7603	16561	-0.50 .02	0.91	0.90	-22.1	-10.2
49	item 49	6081	16539	-0.10 .02	0.96	0.95	-7.6	-4.6
50	item 50	11944	16529	-1.68 .02	0.97	0.97	-3.7	-2.3
Mean				0.00	1.00	1.02	-1.5	0.5
SD				0.69	0.06	0.08	9.8	6.4

Jika menggunakan batas penerimaan item menggunakan INFIT MNSQ maka item nomor 1 diterima, atau fit menurut model, tetapi bila menggunakan INFIT t dengan batas ±2,0 maka item 1 ditolak atau tidak *fit* karena besarnya INFIT t 3,0, hal yang sama berlaku untuk item nomor 4 dengan INFIT t sebesar -3,4 juga untuk item nomor 46 sampai 50.

Berikut adalah hasil analisis dengan akhiran tn.out misalnya presttn.out. Hasil ini adalah hasil analisis detail untuk setiap item, yang menyajikan informasi baik hasil analisis menurut CTT maupun IRT. CTT berupa percent (%) yang merupakan indkes kesukaran item dan pt-biserial menunjukkan indeks daya beda atau ada yang menyebutnya indeks daya pembeda.

Berikut hasil dari program QUEST berkode Q

PRESTASI (50 ITEM PILIHAN GANDA)

```
-------------------------------------------------------------------------------
Item Analysis Results for Observed Responses                    3/12/2009 19:58
all on all (N = 16698 L = 50 Probability Level=0.50)
-------------------------------------------------------------------------------
Item    1: item 1                                Infit MNSQ = 1.03
                                                       Disc = 0.22

Categories        A [0]     B [0]     C [1]     D [0]    missing

Count             5884      4485      4093      2169        67
Percent (%)       35.4      27.0      24.6      13.0
Pt-Biserial      -0.11     -0.08      0.22     -0.03
Mean Ability     -0.77     -0.75     -0.42     -0.72     -0.49
StDev Ability     0.57      0.54      0.78      0.56      0.62

                       01

Delta                 0.52
Gamma                 0.52
Error                 0.02
...............................................................................
Item    2: item 2                                Infit MNSQ = 1.00
                                                       Disc = 0.28

Categories        A [0]     B [0]     C [1]     D [0]    missing

Count             4061      5854      4912      1726       145
Percent (%)       24.5      35.4      29.7      10.4
Pt-Biserial      -0.08     -0.17      0.28     -0.05
Mean Ability     -0.76     -0.81     -0.40     -0.77     -0.44
StDev Ability     0.57      0.50      0.73      0.64      0.93

                       01

Delta                 0.25
Gamma                 0.25
Error                 0.02
...............................................................................
```

dan seterusnya sampai dengan

```
...............................................................................
```

```
Item   50: item 50                              Infit MNSQ = 0.97
                                                      Disc = 0.28

Categories        A [0]     B [0]     C [0]     D [1]    missing

Count             1654      1221      1710     11944       169
Percent (%)       10.0       7.4      10.3      72.3
Pt-Biserial      -0.14     -0.13     -0.15      0.28
Mean Ability     -0.95     -0.98     -0.97     -0.56     -0.70
StDev Ability     0.52      0.50      0.54      0.64      0.70

                       01

Delta                 -1.68
Gamma                 -1.68
Error                  0.02
.............................................................................


Mean test score       17.82
Standard deviation     6.36
Internal Consistency   0.76

The individual item statistics are calculated
using all available data.

The overall mean, standard deviation and internal
consistency indices assume that missing responses
are incorrect.  They should only be considered useful when
there is a limited amount of missing data.
=============================================================================
```

Hasil analisis menurut teori tes klasik menunjukkan untuk item nomor 1 dengan kunci jawaban C memiliki indeks kesulitan 0,25 (atau 24,6%) dan indke daya beda atau ada yang menyebut daya pembeda sebesar 0,22.

Jika item ini untuk tujuan seleksi maka item ini dinyatakan terlalu sukar dan tidak memiliki daya pembeda yang baik.

Jika item ini untuk tujuan pencapaian hasil belajar, dan ini adalah hasil posttest, JIKA ITEM SPEC dan hasil analisis item secara kualitatif memenuhi syarat, maka dapat diartikan bahwa item ini belum menggambarkan hasil belajar yang diharapkan, tetapi dengan daya beda yang tidak negatif berarti tidak ada siswa yang lebih cerdas yang lebih berpeluang gagal mengerjakan item ini.

Pada bagian akhir akhir analisis disajikan nilai internal consistency sebesar 0.76. Nilai ini adalah nilai reliabilitas tes menurut teori tes klasik yang dihitung menurut **Indeks Reliabilitas Kuder-Richardson-20 (reliabilitas untuk Norm-Reference). (Jika data politomus, kaka menunjukkan indeks alpha Cranbach).**
Informasi IRT berupa INFIT MNSQ sebagai bukti fit atau tidaknya item menurut model Rasch, Kemudian delta yang menunjukkan estimasi indeks tingkat kesulitan item pada skala *logit*.

Berikut disajikan hasil menurut program QUEST dengan kode QUEST

PRESTASI (50 ITEM PILIHAN GANDA)

```
-------------------------------------------------------------------------------
Item Analysis Results for Observed Responses                      3/12/ 9 19:57
all on all (N =**** L = 50 Probability Level= .50)
-------------------------------------------------------------------------------
Item    1: item 1                              Infit MNSQ = 1.03
                                                     Disc =  .22

Categories          A         B         C*        D      missing

Count             5884      4485      4093      2169        67
Percent (%)       35.4      27.0      24.6      13.0
Pt-Biserial       -.11      -.08       .22      -.03
p-value           .000      .000      .000      .000
Mean Ability      -.77      -.75      -.42      -.72      -.49

Step Labels                  1
Thresholds                  .52
Error                       .02
...............................................................................


Item    2: item 2                              Infit MNSQ = 1.00
                                                     Disc =  .28

Categories          A         B         C*        D      missing

Count             4061      5854      4912      1726       145
Percent (%)       24.5      35.4      29.7      10.4
Pt-Biserial       -.08      -.17       .28      -.05
p-value           .000      .000      .000      .000
Mean Ability      -.76      -.81      -.40      -.77      -.44

Step Labels                  1
Thresholds                  .25
Error                       .02
...............................................................................
.....  dan seterusnya

...............................................................................
```

```
Item   50: item 50                              Infit MNSQ =  .97
                                                      Disc =  .28

Categories          A         B         C         D*     missing

Count             1654      1221      1710     11944       169
Percent (%)       10.0       7.4      10.3      72.3
Pt-Biserial       -.14      -.13      -.15       .28
p-value           .000      .000      .000      .000
Mean Ability      -.95      -.98      -.97      -.56      -.70

Step Labels                  1
Thresholds                -1.68
Error                       .02
.............................................................................

Mean test score       17.82
Standard deviation     6.36
Internal Consistency    .76

The individual item statistics are calculated
using all available data.

The overall mean, standard deviation and internal
consistency indices assume that missing responses
are incorrect.  They should only be considered useful when
there is a limited amount of missing data.
==============================================================================
```

Pada sajian yang terakhir tidak menggunakan istilah delta tetapi menggunakan istilah threshold. **Istilah threshold dipakai untuk skala sikap**. Istilah delta atau threshold pada skala dikotomus yang dimaksud adalah estimasi tingkat kesulitan menurut IRT.

Step labels hanya ditulis dengan kode 1, artinya tingkat kesulitan untuk testi menjawab dengan benar (skor 1). Pada sajian menurut program QUEST dengan kode Q ditulis step labels 01 artinya untuk meningkat dari skor 0 (dari kategori-1) ke skor 1 (kategori-2) diperlukan kemampuan sebesar delta.

B. Hasil Analisis Item Bentuk Uraian Menggunakan Item *Anchor*

1. File perintah

```
EKSAMINI 537 ITEM 47 (KOLOM 7-53)
data_file dataequ.dat
codes 123
format id 1-6 items 7-53
set width=107 ! page
estimate
show >> QEFAsh.out
show items ! stat=DELTA >> QEFAiD.out
show items ! stat=TAU >> QEFAit.out
show cases >> QEFAca.out
show cases ! form=export >> QEFAsc.out
itanal >> QEFAtn.out
```

```
0011  321221233  22221223213221119999999999999999999
0022  31111123311332122321111111199999999999999999999
0033  32232321112221223323123231299999999999999999999
0044  33113123312232323111121211199999999999999999999
0055  11113313211211312111122121199999999999999999999
0066  31311223111321221112332111199999999999999999999

…… dan seterusnya

1564  33311112323221322133233332399999999999999999999
1577  21231113313131223133332322399999999999999999999
1588  33312113313223222233333332399999999999999999999
1599  32233113311213233333332331199999999999999999999
1600  32223113312322323133333322399999999999999999999
0011  331112999999999999999999999311 1212213211133321
0022  311123999999999999999999999323222221133111223 21
0033  33311399999999999999999999932211332313312311111
0044  23111399999999999999999999933  3333233311111321
0055  33111399999999999999999999312 3232222213311221

…. dan seterusnya
```

Dari data yang ada tampak bahwa **ada dua te**s yang diujikan pada dua kelompok testi tetapi **kedua tes memiliki *anchor/common item*** sebanyak 6 item. Dengan demikian**, hasil tes baik item maupun kemampuan testi kedua kelompok tersebut dapat diplotkan ke dalam satu skala.**

Hasil estimasi *logit* (permintaan QEFAlo.out sebagai berikut

Score Equivalence Table 4/12/ 9 7:14
all on all (N = 365 L = 47 Probability Level= .50)

Score	Estimate (logits)	Error	(unit= 1.00 , origin= .00) Transformed Estimate	Transformed Error
(max= 93)				
92	4.60	1.00	4.60	1.00
91	3.90	.71	3.90	.71
90	3.49	.58	3.49	.58
89	3.19	.51	3.19	.51
88	2.96	.46	2.96	.46
87	2.77	.42	2.77	.42
86	2.61	.39	2.61	.39
85	2.47	.37	2.47	.37
84	2.34	.35	2.34	.35
83	2.22	.34	2.22	.34
82	2.11	.32	2.11	.32
……	……	……	……	……
……	……	……	……	……
……	……	……	……	……
8	-2.50	.38	-2.50	.38
7	-2.65	.40	-2.65	.40
6	-2.82	.43	-2.82	.43
5	-3.02	.47	-3.02	.47
4	-3.26	.52	-3.26	.52
3	-3.57	.60	-3.57	.60
2	-4.00	.72	-4.00	.72
1	-4.72	1.01	-4.72	1.01

Artinya skor mentah maksimum 2 dan setelah dikonversi menjadi 4,6 pada skala logit. Skor minium 1 diubah ke dalam skala logit sebesar -4,72. Ingat secara teoretik skala logit terbentang dari -∞ sampai +∞.

Hasil permintaan QEFAit.out

```
QUEST: The Interactive Test Analysis System
-----------------------------------------------------------------------------------------
Item Estimates (Difficulty and Taus) In input Order                        4/12/ 9 7:14
all on all (N = 365 L = 47 Probability Level= .50)
-----------------------------------------------------------------------------------------
    ITEM NAME      |SCORE MAXSCR| DIFFCLTY  TAU/S                |  INFT  OUTFT INFT  OUTFT
                   |            |            1       2       3   |  MNSQ  MNSQ   t     t
-----------------------------------------------------------------------------------------
1   item 1         |   609  730 |  -1.44    -.44     .44         |   1.08  1.09    .9    .7
                   |            |     .10     .26     .12         |
                   |            |                                 |
2   item 2         |   269  730 |    .64    -.32     .32         |    .94   .92  -1.0   -.9
                   |            |     .08     .11     .14         |
                   |            |                                 |
3   item 3         |   314  730 |    .39    -.20     .20         |    .89   .87  -2.2  -1.8
                   |            |     .07     .12     .13         |
                   |            |                                 |
…… dan seterusnya
                   |            |                                 |
30  item 30        |   165  406 |    .68   -1.14    1.14         |   1.14  1.27   1.6   2.3
                   |            |     .12     .16     .22         |
                   |            |                                 |
31  item 31        |    69  179 |    .68                         |   1.12  1.15   2.4   1.4
                   |            |     .16                         |
                   |            |                                 |
32  item 32        |   185  386 |    .28    -.58     .58         |    .97   .97   -.3   -.2
                   |            |     .10     .17     .17         |
…… dan seterusnya
                   |            |                                 |
46  item 46        |   233  410 |   -.17   -1.16    1.16         |    .88   .88  -1.5  -1.1
                   |            |     .12     .21     .16         |
                   |            |                                 |
47  item 47        |   159  410 |    .55     .28    -.28         |    .90   .87  -1.5  -1.2
                   |            |     .09     .15     .17         |
                   |            |                                 |
-----------------------------------------------------------------------------------------
Mean               |            |    .00                         |   1.00  1.02   -.1    .1
SD                 |            |    .93                         |    .12   .19   1.5   1.3
=========================================================================================
```

Jika dilihat dari besarnya nilai tingkat kesulitan yang berupa difficulty, maka item nmor 1 paling mudah di antara item yang ada pada tabel di atas(-1,44), kemudian item nomor 30 dan 31 paling sukar (+0,68). Item nomor 31 tidak memilki nilai tau-2 artinya tidak ada yang berhasil mengerjakansampai skor maksimum 3.

Hasil permintaan QEFAtn.out

```
QUEST: The Interactive Test Analysis System
--------------------------------------------------------------------------------------------
Item Analysis Results for Observed Responses                                  4/12/ 9 7:14
all on all (N = 365 L = 47 Probability Level= .50)
--------------------------------------------------------------------------------------------

Item    1: item 1                                Infit MNSQ = 1.08
                                                       Disc =  .23

Categories          1         2         3      missing

Count              17        87       261         0
Percent (%)       4.7      23.8      71.5
Pt-Biserial     -.17      -.13       .21
p-value         .001      .005      .000
Mean Ability    -.40       .00       .21        NA

Step Labels                  1         2         3

Thresholds                          -2.16      -.73
Error                                 .34       .26
.............................................................................................

Item    2: item 2                                Infit MNSQ =  .94
                                                       Disc =  .43

Categories          1         2         3      missing

Count             166       129        70         0
Percent (%)      45.5      35.3      19.2
Pt-Biserial     -.39       .12       .34
p-value         .000      .011      .000
Mean Ability    -.14       .24       .57        NA

Step Labels                  1         2         3

Thresholds                                .00      1.28
Error                                     .19       .22
.............................................................................................
```

Hasil analisis menurut **CTT untuk item tes sebagai alat pengukur hasil belajar**, memberikan infomasi dapat diinterpretasi sebagai berikut.

Item 1:

Seluruh testi mengerjakan (365 testi) dengan skor terendah 1 dan skor teringgi 3.

Tingkat kesulitan untuk memperoleh skor 1 sebesar 0,047 (17 testi atau 4,7% dari 365), tingkat kesulitan untuk memperoleh skor 2 sebesar 0,230 (87 testi atau 23,8% dari 365), dan tingkat kesulitan untuk memperoleh skor 3 sebesar 0,715 (261 testi atau 71,5% dari 365).

Dengan melihat nilai point biserial (indeks daya beda) untuk skor 1 yang negatif, yakni sebesar -,17, berarti **testi yang memperoleh skor 1, testi yang cerdas < testi yang tidak cerdas**. Hal yang sama berlaku untuk skor 2, karena hasilnya juga negatif. Untuk skor 3,

hasilnya positif, yakni +0,21, sehingga **yang memperoleh skor 3, testi yang cerdas lebih banyak dibanding testi yang tidak cerdas**

Hasil analisis menurut **CTT untuk item tes sebagai alat pengukur untuk tujuan seleksi**, memberikan infomasi dapat diinterpretasi sebagai berikut. Tingkat kesulitan item tergolong sangat mudah karena sebanyak 71,5% testi berhasil mengerjakan. Dengan tingkat kesulitan 0,715, maka item tidak memenuhi syrat sebagai item untuk keperluan seleksi. Dilihat dari nilai daya beda (point biserial) sebesar 0,21, item ini memiliki daya pembeda yang rendah, sehingga tidak memenuhi syarat sebagai item untuk keperluan seleksi. Kesimpulan, baik dari tingkat kesulitan maupun daya beda item ini tidak memenuhi syarat sebagai item untuk seleksi.

BAB V

PENENTUAN INDEKS PERSETUJUAN DAN INDEKS KAPPA

Ddiangkat dari tulisan Mikhael J.Subkoviak, Universitas Wisconsin-Madison)
Journal of Educational Measurement Spring 1988. Vol.25. No. 1. pp. 47-.55

Pendahuluan

Ada kendala bagi guru dan praktisi dalam menerapkan metode untuk memperoleh indeks keandalan untuk tes penguasaan (*mastery test*) seperti koefisien persetujuan (*agreement coefficient*) dan koefisien kappa.

1. Memerlukan banyak tenaga untuk menghitungnya karena memerlukan dua pengadministrasian tes
2. Melibatkan prosedur-prosedur statistis yang kompleks dan memerlukan akses ke *software* komputer yang sesuai jika menggunakan pengadminitrasian tunggal.

Artikel ini menawarkan tabel praktis dari koefisien persetujuan dan koefisien kappa dimana koefisien tersebut dapat dibaca secara langsung.

Karena koefisien indeks keandalan yang dihitung secara tradisional berbeda dengan prosedur ini, maka sekaligus dibahas tentang seberapa jauh nilai koefisien persetujuan dan koefisien kappa dapat diterima. Perhitungan indeks keandalan tes yang mengacu pada acuan kriteria, baik dari segi metode untuk menghitung maupun cara menginterpretasikan koesisien keandalan untuk tes penguasaan menjadi sangat kompleks menurut pandangan praktisi. Sebagai contoh, metoda yang diusulkan oleh Swaminathan, Hambleton, dan Algina (1974) memerlukan dua pengadministrasian tes yang sama atau uji paralel, dengan pengklasifikasian berhasil-gagal (*mastery-nonmastery*) bagi penempuh ujian di kedua pengadministrasian tes tersebut.

Dua indeks keandalan yang telah direkomendasikan adalah (a) koefisien persetujuan (*agreement coefficient*), (b) koefisien kappa. Dua koefisien ini mengukur konsistensi dari klasifikasi berhasil-gagal (*mastery-nonmastery*) antarkedua pengadministrasian tes, dan kedua-duanya sedikit banyak memerlukan interpretasi yang berbeda dibanding koefisien keterandalan tradisional, yang menggunakan koefisien korelasi dari kedua pengadministrasian tes.

Ketika pengklasifikasian berhasil-gagal pada kedua pengadministrasian tes diringkas maka hasilnya seperti yang tersaji pada Tabel 1 berikut

Tabel 1

Klasifikasi Penempuh Ujian di Dua Pengdministrasian Tes

		Pengadministrasian 2		
		Berhasil	Gagal	
Pengadministrasian 1	Berhasil	*A*	*b*	*(a + b)*
	Gagal	*C*	*d*	*(c + d)*
		(a + c)	*(b + d)*	*N*

Koefisien persetujuan adalah proporsi penempuh tes yang secara konsisten tergolongkan ke dalam kelompok yang berhasil dan yang gagal dari kedua-dua pengadministrasian pengujian tes. Koefisien persetujuan menunjuk nilai p_o yang diperoleh dengan rumus:

$$p_o = (a + d)/N \qquad (1)$$

di mana *a* dan *d* berturut-turut menunjukkan banyaknya penempuh tes yang digolongkan menjadi kelompok yang berhasil dan yang gagal di kedua-dua pengadministrasi tes dan *N* menunjukkan banyaknya peserta tes di kedua pengadministrasian tes yang bersangkutan.

Batas atas dari koefisien persetujuan adalah 1,00, yang tercapai atau terjadi jika klasifikasi-klasifikasi di kedua-dua pengadministrasian tes bersifat konsisten untuk semua penempuh tes dimasing-masing kelompok. Batas bawah dari koefisien persetujuan diberi oleh:

$$p_{chance} = [(a + b)(a + c) + (c + d)(b + d)]/N^2 \qquad (2)$$

Batas bawah, p_{chance}, menunjukkan proporsi klasifikasi-klasifikasi yang konsisten yang diharapkan secara kebetulan jika hasil 'mastery-nonmastery' pada pengadministrasian tes yang kedua dengan sepenuhnya tidak terikat pada hasil pada pengadministrasin tes yang pertama. Di dalam konteks seperti itu besarnya $p_{chance} \geq 0{,}50$.

Koefisien kappa, K , diperoleh dengan rumus:

$$K = (\boldsymbol{p}_o - p_{chance})/(1 - p_{chance}) \qquad (3)$$

di mana p_o dan p_{chance} diperoleh dari (1) dan (2).

Koefisien kappa mencerminkan proporsi klasifikasi-klasifikasi yang konsisten yang sesuai dengan harapan dan yang secara kebetulan. Batas atas dan batas bawah dari koeifisen kappa adalah 1,00 dan 0,00, yang terjadi ketika hasil-hasil pada kedua pengadministrasian tes bertutur-turut di dalam persetujuan yang sempurna atau bebas secara penuh.

Metode untuk menaksir koefisien persetujuan dan koefisien kappa dari satu pengadministrasian tes sudah diusulkan, dengan demikian tidak memerlukan lagi pengadministrasian tes yang kedua (Huynh, 1976; Marshall & Haertel, 1976; Subkoviak, 1976), namun metoda tersebut juga sulit untuk diterapkan para praktisi karena memerlukan akses fasilitas komputer dan perangkat lunak yang sesuai dan mereka juga sedikit banyak harus memahami latar belakang lanjut dari teori tes.

Metoda prakiraan yang disertai kalkulasi manual untuk menghitung besarnya koefisien persetujuan dan koefisien kappa dari suatu pengadministrasian tes telah pula diusulkan (Huynh, 1976. p.258; Peng & Subkoviak, 1980. p.363). Meskipun metode-metode ini merupakan metode yang paling sederhana yang sampai sekarang diusulkan, diperlukan pemakaian tabel statistika bivariat dan univariat agihan/distribusi normal, yang bagi para guru sama sekali tidak akan terbiasa menggunakan, terutama para guru kelas atau praktisi-praktisi pengukuran di bagian distrik.

Artikel ini menyajikan tabel koefisien persetujuan dan koefisien kappa yang dapat terbaca secara lngsung. Artikel ini juga menjawab pertanyaan para praktisi yang kurang memiliki naluri membaca literatur dimana perhitungan-perhitungan yang ada sampai saat ini mendasarkan pada metode statistika dari estimasi/perkiraan.

Tabel Koefisien Persetujuan dan Koefisien Kappa

Tabel 2 berisi nilai-nilai prakiraan/aproksimasi dari koefiein persetujuan, dan Table 3 berisi nilai-nilai prakiraan/aproksimasi dari koefisien kappa.

Untuk menggunakan kedua table tersebut diperlukan dua nilai yaitu:

(1) skor penggalan (*cutoff score*) dari tes yang dinyatakan sebagai suatu skor patokan (*standard score* atau *z-score*) dan

(b) keandalan tradisional dari skor tes (*r*).

Taksiran *z* dan *r* dapat diperoleh dari data untuk suatu pengadministrasian tes dengan menggunakan formula yang dapat ditemukan di dalam setiap teks pengantar pengukuran.

Tabel 2

Nilai Taksiran (Approximate Values) dari Koefisien Persetujuan (Agreement Coefficient)

r

IzI	.10	.20	.30	.40	.50	.60	.70	80	.90
.00	.53	.56	.60	.63	.67	.70	.75	.80	.86
.10	.53	.57	.60	.63	.67	.71	.75	.80	.86
.20	.54	.57	.61	.64	.67	.71	.75	.80	.86
.30	.56	.59	.62	.65	.68	.72	.76	.80	.86
.40	.58	.60	.63	.66	.69	.73	.77	.81	.87
.50	.60	.62	.65	.68	.71	.74	.78	.82	.87
.60	.62	.65	.67	.70	.73	.76	.79	.83	.88
.70	65	.67	.70	.72	.75	.77	.80	.84	.89
.80	.68	.70	.72	.74	.77	.79	.82	.85	.90
.90	.7I	.73	.75	.77	.79	.81	.84	.87	.90
1.00	.75	.76	.77	.77	.81	.83	.85	.80	.91
1.10	.78	.79	.80	.81	.83	.85	.87	.89	.92
1.20	.80	.81	.62	.84	.85	.86	.88	.90	93
1.30	.83	.84	85	.86	.67	.88	.90	.91	.94
1.40	.86	.86	.87	.88	.89	.90	.91	.93	.95
1.50	.88	.88	.89	.90	.90	.91	.97	.94	.95
1.60	.90	.90	.90	.91	.92	.93	.93	.95	.96
1.70	.92	.92	.92	.93	.93	.94	95	.95	.97
1.80	.93	.93	.94	.94	.94	.95	.95	.96	.97
1.90	.95	.95	.95	.95	.95	.96	.96	.97	.98
2.00	.96	.96	.96	.96	.96	.97	.97	.97	.98

Pemakaian Tables 2 dan 3 dapat digambarkan, seandainya satu set data riil dari suatu ulangan bentuk pilihan sebanyak 10 item, dengan suatu skor penggalan (*cutoff score*) sebesar 8, yang diujikan kepada 30 siswa memperoleh skor rata-rata M = 4,63, dan deviasi standar S = 1,81.

Besarnya z standar yang muncul pada Tabel 2 dan 3, dapat diperoleh sebagai berikut:

$$z = \frac{(c - 0{,}5 - M)}{S} \qquad (4)$$

di mana c adalah skor penggalan mentah dari tes, M adalah skor rata-rata yang diperoleh siswa.

Dari data diatas dapat dihitung besarnya harga
skor z = (8 – 0,5 - 4.63)/1.81 = 1,59, atau kira-kira = 1,60, yang akan dibawa ke Tabel 2 atau Tabel 3.

Nilai 0,5 pada persamaan (4) adalah suatu koreksi kekontinuan yang dibangun dari fakta dalam Tables 2 dan 3 yang diperoleh dengan memperkirakan skor tes mengikuti distribusi diskret untuk diubah menjadi distribusi normal yang kontinum, yang akan dibahas kemudian.

Nilai z yang dihitung menggunakan persamaan (4) dapat menghasilkan harga positif ataupun negatif. Oleh karena distribusi z adalah distribusi yang simeteris maka besarnya koefisien persetujuan atau koefisien kappa untuk nilai z = -1,0 sama saja untuk z = +1,0. Jadi, dengan memberikan tanda mutlak untuk z dapat menggunakan Tabel 2 dan 3. Keandalan skor tes *r*, yang muncul di dalam Tabel 2 dan 3, dapat diperoleh dengan menggunakan manapun indeks reliabilitas tradisional seperti koefisien keandalan Kuder-Richardson (K-R) atau alfa-Cronbach.

Secara khusus, Koefiesien kenadalan menggunakan rumus Kuder-Richardson dengan Formula 20 atau 21 dapat dihitung berdasarkan besarnya nilai rata-rata M dan simpangan baku S, beserta banyaknya item dari item yang diskor dengan cara benar atau salah. Formula K-R 21 di bawah ini dirumuskan dengan anggapan yang tak realistis yaitu bahwa semua item memiliki tingkat kesulitan yang sama. Formula K-R 21 adalah:

$$r = \frac{nS^2 - M/(n - M)}{(n - 1)S^2}$$

di mana *n* adalah banyaknya item dan *M* dan *S* adalah skor rata-rata dan simpangan baku seperti yang sudah digambarkan sebelumnya.

Dari contoh sebelumnya dapat diperoleh besarnya *r*:

r = [(I0)(1,81)^2 - (4.63)]/[(10-1)(1,81)2] = 0,27,

atau kira-kira r = 0,30 dalam tabel Tabel 2 dan 3.

Tabel 3

Nilai Taksiran (Approximate Value) dari Koefisien Kappa (Kappa Coefficient)

r

|z|	.10	.20	.30	.40	.50	.60	.70	.8O	.90
.00	.06	.13	.19	.26	.33	.41	.49	.59	.71
.10	.06	.13	.19	.26	.33	.41	.49	.59	*.71*
.20	.06	.13	.19	.26	.33	41	.49	.59	.71
.30	.06	.12	.19	.26	.33	.40	.49	.59	.71
.40	.06	.12	.19	.25	.32	.40	.48	.58	.71
.50	.06	.12	.18	.25	.32	.40	.48	.58	.70
.60	.06	.12	.18	.24	.31	.39	.47	.57	.70
.70	.05	.11	.17	.24	.31	.38	.47	.57	.70
.60	.05	.11	.17	.23	.30	.37	.46	.56	.69
.90	.05	.10	.16	.22	.29	.36	.45	.55	.68
1.00	.05	.10	.15	.21	.28	.35	.44	.54	.68
1.10	.04	.09	.14	.20	.27	.34	.43	.53	.67
1.20	.04	.08	.14	.19	.26	.33	.42	.52	.66
1.10	.04	.08	.13	.18	.25	.32	.41	.51	.65
1.40	.03	.07	.12	.17	.23	.31	.39	.50	.64
1.50	.03	.07	.11	.16	.22	.29	.30	.49	.63
1.60	.03	.06	.10	.15	.21	.28	.37	.47	.62
1.70	.02	.05	.09	.14	.20	.27	.35	.46	.61
1.90	.02	.05	.08	.13	.16	.25	.34	.45	.60
1.90	.02	.04	.08	.12	.17	.24	*.32*	.43	.59
2.00	.02	.04	.07	.11	.16	.22	.31	.42	.58

Formula K-R 20 tidak mengasumsikan semua item memiliki tingkat kesulitan yang sama, sehingga hasil perhitungan menggunakan formula K-R 20 menghasilkan taksiran *r* = 0,47 untuk data ini, atau kira-kira *r* = 0,50 pada Tabel 2 dan 3.

Jika item-item tersebut tidak dikategorikan benar atau salah dapat menggunakan perhitungan koefisien alfa-Cronbach untuk perhitungannya. Memasuki Tabel 2 dengan nilai |z| = 1,60 dan *r* = 0,50, mendekati nilai dari koefisien persetujuan adalah p_o = 0,92, yang menunjukkan bahwa 92% dari kelompok secara konsisten dapat digolongkan sebagai kelompok master dan nonmaster jika dilakukan dua pengadministrasian tes yang

setara. Dengan cara yang sama, memasuki Table 3 dengan nilai-nilai |z| = 1,60 dan r = 0,50, mendekati nilai dari koefisien kappa atau K = 0,21. Dengan demikian, besarnya koefisien persetujuan dan koefisien kappa sungguh-sungguh berbeda sebagaimana yang tersaji di dalam contoh tersebut(p_o = 0,92 melawan K = 0,21), karena keduakoefisien menggunakan ukuran yang terpisahmaka jelaslah bahwa diperlukan interpretasi yang berbeda dari keduanya.

Tujuan yang diharapkan dari pembuatan Tabel 2 dan 3 menyediakan bagi para praktisi untuk memperoleh perkiraan besarnya koefisien persetujuan dan koefisien kappa dengan melibatkan komputasi seminimal mungkin. Seperti yang digambarkan di atas dengan dengan menghitung besarnya |z| = 0,60 dan r = 0,50 dapat menggunakan kedua tabel tersebut. Meskipunsedikit banyak tidak konsisten dengan tujuan ini, namun dapat diperoleh besarnya nilai p_o dan K dari kedua tabel. Dengan intrapolasi akan dapat diperoleh nilai yang lebih eksak berdasar nilai |z| = 1,59 dan r = 0,47 yang ada di dalam contoh (Subkoviak: 1980. pp. 141-142). Bagaimanapun, untuk tujuan-tujuan praktis, cukup dengan menggunakan nilai yang paling mendekati yang tersedia di dalam tabel tersebut.

Menginterpretasikan Koefisien Persetujuan dan Koefisien Kappa

Mungkin saja hasil yang diperoleh sesuai yang dicatat pada Tabele 2 dan 3 secara umum sungguh yang berbeda, seperti di contoh di atas, dimana nilai p_o = 0,92 sementara nilai K = 0,21. Perbedaan seperti itu terjadi karena kedua koefisien tersebut dihitung berdasarkan ukuran-ukuran yang terpisah, oleh karenanya dalam menginterpretasikannya juga harus berbeda (lihat Berk. 1984, pp. 211-242; Subkoviak, 1984, pp. 286-287).

Adanya satu kesadaran terhadap perbedaan antara kedua koefisien tersebut sangat penting adanya ketika mengevaluasi nilai numerik dari suatu laporan. Dengan demikian, penting adanya diskusi untuk meninjau ulang perbedaan-perbedaan dasar antara kedua koefisien tersebut sehingga dalam menilai menilai tiap koefisien tersebut dapat diterima.

Koefisien Persetujuan

Seperti sebelumnya telah diuraikan, koefisien persetujuan (p_o) menunjukkan proporsi penempuh ujian yang secara konsisten dikelompokkan dari dua

pengadministrasian tes penguasaan. Sebagaimana digambarkan pada Tabel 2, magnitud dari nilai tersebut dipengaruhi oleh nilai keandalan dari skor tes (r) dan jarak skor penggalan dari skor rata-rata menurut distribusi standar (|z|).

Pertama, meningkatnya keandalan tes sejalan dengan panjang tes, seperti yang terlihat pada deret-deret angka dalam Tabel 2. Dengan kata lain, tes yang lebih panjang akan memiliki kenadalan yang lebih baik di dalam mengklasifikasi peserta tes menjadi peserta menjadi yang master (menguasai) dan yang nonmaster (tidak menguasai). Kedua, untuk distribusi skor tes yang unimodal (hanya memiliki satu modus), p_o akan meningkat ketika jarak antara skor penggalan dan skor rata-rata juga meningkat, seperti yang digambarkan di dalam kolom-kolom dari Tabel 2.

Dalam terminologi yang lain dapat dikatakan bahwa klasifikasi akan lebih tidak konsisten bila skor penggalan sangat dekat dengan skor rata-ratanya. Hal tersebut membawa kepada suatu pertanyaan umum yang tidak dijawab di dalam literatur yakni: apakah nilai koefisien persetujuan memuaskan? Dengan kata lain, bilamana proporsi (p_o) dari suatu kelompok dapat secara konsisten digolongkan di dalam dua pengulangan tes? Apakah jawaban mengenai seberapa besar proporsi tersebut bergantung kepada banyaknya pengulangan?

Keputusan yang diambil untuk menyatakan bahwa suatu tes memiliki keandalan yang baik adalah jika memiliki koefisien persetujuan $p_o \geq 0{,}86$ juga tergantung kepada penempatan skor penggalan sehubungan dengan skor rata-rata yang diperoleh dari tes yang bersangkutan. Nilai $p_o = 0{,}86$ berpasangan dengan situasi di mana harga $|z| = 0{,}00$, yang menyiratkan bahwa separuh dari penempuh ujian adalah menguasai dan separuh lainnya tidak menguasai.

Pada umumnya, banyaknya jawaban tebakan pada sekolah-sekolah distrik yang unggul dari siswa kelas tiga sebanyak 10% tidak menguasai ketrampilan dasar pada nilai $|z| \geq 1{,}30$. Dengan kata lain nilai p_o akan menjadi dekat dengan angka 0,95. Memperpendek diskusi di atas petunjuk berikut dapat dijadikan pegangan: ”tes-tes yang digunakan dalam pengambilan keputusan yang serius akan menggunakan batas koefisien persetujuan sebesar 0,85.

Batas yang lebih tinggi dapat diterapkan misalnya dengan mematok batas 0,95 pada umumnya berharap hanya kurang dari 10% siswa yang akan tergolong tidak menguasai, seperti yang sering digunakan pada ujian sertifikasi. Tes buatan guru

digunakan secara rutin dan digunakan untuk mengetahui apakah siswa sudah menguasai materi yang diajarkan.

Diederich (1973, p.10) menyatakan umumnya guru membuat tes yang dirancang untuk kelas yang tunggal yang biasanya diharapkan dapat mencapai koefisien keandalan (*r*) antara 0,60 sampai 0,80, dan hal ini juga sesuai dengan pengalaman penulis yang menyediakan analisis tes untuk 11 departemen di universitas. Biasanya diasumsikan bahwa r minimal yang bisa diterima adalah sebesar 0,70.

Dalam tabel 2 untuk r = 0,70 akan menghasilkan koefisien persetujuan (p_o) sebesar ≥ 0,75 tergantung pada lokasi skor penggalan. Nilai minimal dari p_o = 0,75 berpasangan dengan situasi di mana |z| = 0.00, yang dalam keadaan seperti itu separoh penempuh ujian itu akan tergolong menguasai dan separoh yang lain tergolong tidak menguasai. Nilai yang lebih besar dapat ditetapkan manakala antara kelompok yang menguasai dan tidak menguasai proporsinya tidak sama.

Sebagai contoh, suatu kelas dengan koefisien keandalan (*r*) = .70) diatur mengikuti ketentuan dimana yang tidak menguasai sebanyak ≤ 15% (|z| =1,00), maka sebagai sebagai konsekwensinya nilai p_o yang diharapkan akan mencapai ≥ 0,85. Secara ringkas ukuran berikut boleh jadi dapat dijadikan pegangan: Suatu tes buatan guru untuk satu periode pembelajaran penuh (katakanlah 1 semester) akan dijamin keandalannya jika memiliki koefisien persetujuan sedikitnya 0,75. Jika harapan guru kelompok yang tidak menguasai hanya sebanyak 15 % maka tes dinyatakan memiliki keandalan jika besarnya koefisien persetujuan ≥ 0,85.

Koefisien Kappa (*K*)

Sebagaimana dapat dilihat di dalam deret-deret angka pada Table 3, koefisien kappa (K) juga meningkat sejalan dengan meningkatnya keandalan skor tes (r). Sebenarnya, jika dibandingkan dengan koefisien persetujuan, koefisien kappa meningkat secara lebih dramatis. Seperti dapat dilihat di dalam kolom-kolom dari Table 3, koefisien kappa berkurang (sedangkan koefisien persetujuan meningkat) ketika jarak antara skor penggalan dan skor rata-rata (|z|) meningkat. Dengan demikian, perlu interpretasi yang berbeda antara koefisien kappa dan koefisien persetujuan.

Pertanyaan tentang apakah nilai koefisien kappa dapat diterima dijawab dengan meninjau ulang bagaimana koefisien ini diperoleh. Seperti yang disajikan pada persamaan (3), koefisien kappa melibatkan

a) nilai p_o—the proporsi klasifikasi amatan yang konsisten dari tes yang dilakukan,
b) p_{chance}—proporsi klasifikasi harapan yang konsisten jika skor total dari tes yang dilakukan tidak dapat dipercaya , dan
c) 1—proporsi atau klasifikasi yang diharapkan dari keandalan skor tes benar-benar konsisten secara sepurna.

Oleh karena itu, pembilang dari koefisien kappa (p_o - p_{chance}) adalah suatu ukuran keuntungan di dalam konsistensi yang disediakan oleh suatu tes yang dilakukan di atas suatu secara total tak dapat dipercaya keandalannya. Dengan demikian, $K = (p_o - p_{chance})/(1 - p_{chance})$ adalah rasio keuntungan nyata terhadap keuntungan maksimum.

Di dalam terminologi yang sederhana, koefisien persetujuan mengukur konsistensi secara menyeluruh, sedangkan koefisien kappa mengukur keuntungan di dalam konsistensi yang direalisir dengan menggunakan tes yang bersangkutan. Pertanyaan tentang berapa banyak keuntungan di dalam konsistensi yang realistis tang diharapkan dari suatu pengujian (yaitu berapa besarnya koefisien kappa yang seharusnya) lagi-lagi kembali bergantung pada keseriusan keputusan yang sedang dibuat.

Jika suatu distrik membuat keputusan-keputusan sekitar siapa yang akan atau tidak akan lulus dari sekolah menengah, maka koefisien keandalan yang digunakan dapat mencapai 0,90. Seperti dapat dilihat pada kolom yang terakhir dari Table 3, tes seperti itu bisa diharapkan akan memiliki koefisien kappa di dalam mendekati daerah 0,60 sampai 0,70, tergantung di mana skor penggalan ditempatkan.

Nilai kappa dekat dengan angka 0,70 dapat diharapkan jika ukuran proporsi yang menguasai dan yang tidak menguasai adalah sama; nilai kappa medekat 0,60 boleh jadi diharapkan jika ukuran ini adalah sungguh yang berbeda. Sebagai contoh, jika 10% dari siswa kelas tiga tidak menguasai ketrampilan dasar (yaitu., $|z| = 1,30$), nilai dekat 0,65 dapat diantisipasi.

Menyederhanakan diskusi di atas aturan berikut ini dapat dijadikan pegangan: Tes yang digunakan untuk membuat keputusan-keputusan penting dapat menggunakan koefisien kappa pada daerah 0,60 sampai 0,70, nilai kappa yang lebih tingi dapat diterapkan manakala ukuran nisbi antara kelompok yang menguasai dan tidak menguasai

sama. Nilai kappa mendekati angka 0,65 bisa diterapkan manakala sekitar 10% dari penempuh tes tidak menguasai.

Seperti pembahasan sebelumnya, tes buatan guru yang digunakan untuk membuat keputusan-keputusan yang rutin biasanya menggunakan batas nilai koefisien keandalan sebesar 0,60 sampai 0,80, dan biasanya umumnya menggunakan batas minimal nilai r yang dapat diteraima sebesar 0,70. Seperti yang tersaji pada Tabel 3, suatu tes yang memiliki nilai r 0,70 akan memiliki nilai koefisien kappa mendekati daerah 0,35 sampai 0,50 tergantung pada lokasi skor penggalan terhadap skor rata-rata atau ukuran yang nisbi antara kelompok yang menguasai dan yang tidak menguasai.

Sebagai contoh, jika koefisien keandalan tes r = 0,70 dan suatu kelas diharapkan hanya memiliki 15 % siswa yang tidak menguasai (|z| =1,00) maka besarnya koefisien kappa yang diharapkan sekitar 0,45. Tentu saja, nilai kappa yang lebih tinggi dapat diharapkan bila suatu tes diharapkan memiliki koefisien keandalan yang lebih tinggi pula.

Dari diskusi di atas aturan berikut ini dapat dijadikan pegangan: Suatu tes yang diterapkan dalam suatu kelas pada periode yang penuh (katakanlah 1 semester) dapat menggunakan nilai koefisien kappa sebesar 0,35 sampai 0,50. Nilai yang lebih tinggi dapat diterapkan disesuaikan dengan proporsi kelompok yang menguasai dan yang gagal. Nilai kappa mendekati 0,45 untuk tes yang terdiri dari 10 item jika siswa yang gagal diharapkan hanya 15%.

Di dalam menerapkan kriteria seperti yang diusulkan di atas, harus melihat kembali bagaimana perhitungan kedua koefisien itu diperoleh. Sebagai contoh, tes yang terdiri dari 10 item dengan koefisien keandalan r = 0,50 dan penggalan |z| =1,60, akan menghasilkan koefisien persetujuan p_o = 0,92 berarti melebihi batas minimal p_o = 0.85 karena lokasi skor penggalan menghasilkan proporsi antara siswa yang berhasil dan yang gagal cukup besar. Namun demikian besarnya koefisien kappa k = .21 untuk tes yang sependek itu dibawah patokan minimal yang yang diusulkan (yakni = 0,35). Dalam kasus tertentu guru dimungkinkan membuat perjanjian yang berkebalikan. Artinya patokan minimal untuk koefisien kappa ditetapkan terlebih dahulu.

Ketelitian Tabel

Tabel 2 dan 3 dibangun dengan suatu prosedur yang diusulkan oleh Peng dan Subkoviak (1980. p.363) untuk menaksir besarnya koefisien persetujuan dan koefisien

kappa. Prosedur tersebut didasarkan pada asumsi jika dua pengadministrasian tes benar-benar diselenggarakan, maka sebaran bersama dari skor-skor pada kedua tes tersebut dapat didekati atau diprakirakan distribusi normal (bivariat). Bagaimanapun, Peng dan Subkoviak menemukan prosedur yang hasilnya disajikan pada Tabel 2 dan 3 yang secara umum menyediakan perkiraan-perkiraan yang dapat dipakai, bahkan ketika data ujian tidak terdistribusi secara normal. Mereka melakukan simulasi untuk 125 kondisi yang berbeda, dan kemudian mereka bandingkan dengan koefisien persetujuan dan koefisien kappa yang eksak untuk data tersebut untuk membuat perkiraan-perkiraan dua koefisien itu seperti yang tersaji pada Tabel 2 atau 3. Ternyata nilai rata-rata pertentangan (*average discrepancy*) antara yang eksak dan nilai perkiraan untuk 125 kondisi secara keseluruhan mendekati adalah 0,013 untuk koefisien persetujuan dan 0,037 untuk koefisien kappa.

Sebagaimana yang diharapkan, pertentangan-pertentangan yang terbesar terjadi untuk sebaran/distribusi yang tidak normal, dan membentuk distribusi bentuk U, dan besarnya nilai rata-rata pertentangan untuk di atas 25 kasus seperti itu sebesar 0,019 untuk koefisien persetujuan dan 0,011 untuk koefisien kappa. Untuk hasil simulasi skor tes yang terdistribusi hampir normal angka pertentangan antara nilai eksak dan nilai perkiraan semakin menurun. Dari 25 distribusi yang menceng yang muncul dari tes ini, besarnya nilai rata-rata pertentangan adalah 0,015 untuk koefisien persetujuan dan 0,032 untuk koefisien kappa. Dengan demikian, itu Tables itu 2 dan 3 secara umum menyediakan perkiraan-perkiraan koefisien persetujuan dan koefisien kappa bagi praktisi-praktisi yang sangat bermanfaat karena didasarkan pada bermacam kondisi-kondisi data yang realistis.

Huynh (1976, p.258) juga telah mengusulkan suatu prosedur untuk penaksiran koefisien kappa dengan menggunakan suatu transformasi arcsin atau data skor dengan distribusi yang diperkirakan normal. Peng dan Subkoviak (1980) pada dasarnya menggunakan pendekatan yang sama, tetapi tidak melibatkan transformasi arcsin.

DAFTAR PUSTAKA

Adams, R.J. & Kho, Seik-Tom. (1996). *Acer quest version 2.1*. Camberwell, Victoria: The Australian Council for Educational Research.

Andrich, D. (1999). Rating scale analysis. In: Masters, G.N. & Keeves, J.P. (1999). *Advances in measurement in educational research and assessment.* Amasterdam: Pergamon, An imprint of Elsevier Science.

Bond, T.G. & Fox, Ch.M. (2007). *Applying the rasch model*: *Fundamental measurement in the human sciences*. 2-nd ed. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Djemari Mardapi. (2007). *Teknik penyusunan instrumen tes dan non tes*. Yogyakarta: Mitra Cendekia Press.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory*. Boston: Kluwer Nijjhoff Publiser.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Foundamentals of item responses theory*. Newbury Park: Sage Publications.

Han, Kyung T. & Hambleton, R.K. (2007). *User's manual for WinGen2*: *Windows software that generates IRT model parameters and item response*. (Media elektronik]. Massachusetts: Center for Educational Assessment.

Keeves, J.P. & Masters, G.N. (1999). Introduction. In: Masters, G.N. & Keeves, J.P. (1999). *Advances in measurement in educational research and assessment.* Amasterdam: Pergamon, An imprint of Elsevier Science.

Kolen, M.J. (1999). Equating of test. In: Masters, G.N. & Keeves, J.P. (1999). *Advances in measurement in educational research and assessment.* Amasterdam: Pergamon, An imprint of Elsevier Science.

_____ & Brennan, R.L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag New York Inc.

Masters, G.N. (1999). Partial credit model. In: Masters, G.N. & Keeves, J.P. (1999). *Advances in measurement in educational research and assessment.* Amasterdam: Pergamon, An imprint of Elsevier Science.

Messick, S. (1988). The one and future issues of validity: Assessing the meanng and consequences of measurement. In: Waine, H. & Braun, H.I. (1988). *Test validity*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Muraki, E. & Bock, R.D. (1998) Parscale: *IRT item analysis and test scoring for rating scale data*. Chicago: Scientific Software Internatinal, Inc.

Smith, J.K. (2003). Reconsidering reliability in classroom assessment and grading [Versi elektronik]. *Educational Measurement, Issues and Practice*, 22, 4, 26-33.

Stark, S., Chernyshenko, S., Chuah, D.,Wayne Lee, & Wilington, P. (2001). *IRT modeling lab: IRT tutorial* [Versi elektronik]. Urbana: University of Illinois.

Stark, S., Chernyshenko, S., Chuah, D.,Wayne Lee, & Wilington, P. (2001). *IRT modeling lab: Test Development Using Classical Test Theory* [Versi elektronik]. Urbana: University of Illinois.

_____, (2001). *IRT modeling lab: IRT tutorial* [Versi elektronik]. Urbana: University of Illinois.

Thissen, D., Nelson, L, & Surygert, K.A. (2001). Item response theory applied to to combination of multiple-choise and constructed response items—Approximation methods for scale score. In: Thissen, D. & Wainer, H. (2001). *Test Scoring*. Mahwah, New Jerrsey: Lawrence Erlbraum Associates, Publishers.

Wright, B.D. (1999). Rasch measurement model. In: Masters, G.N. & Keeves, J.P. (1999). *Advances in measurement in educational research and assessment.* Amasterdam: Pergamon, An imprint of Elsevier Science.

_____ & Masters, G.N. (1982). *Rating scale analsis*. Chicago: Mesa Press.