ANALISIS SOAL BAIK KUALITATIF MAUPUN KUANTITATIF*

Prof. Dr. Bambang Subali, M.S.
Universitas Negeri Yogyakarta

(* Disajikan pada Kegiatan Workshop *Item Development* Dosen Poltekes Kebidanan Politeknik Kesehatan Surakarta tanggal 18-19 Agustus di Griya Persada Conventional Hotel & Resort, jl Boyong Kaliurang Barat)

Pedahuluan

Tes merupakan salah satu bentuk instrumen dalam penilaian penguasaan kompetensi baik berkaitan dengan kinerja atau keterampilan dan pemahaman konsep. Kinerja atau keterampilan dapat berupa keterampilan yang murni psikomotor, gabungan psikomotor dan kognitif. Kinerja dalam aspek kognitif yang tertinggi adalah kemampuan berkreasi. Penyusunan instrumen tes melalui prosedur yang berurutan. Prosedur penyusunan instrumen tes untuk tes hasil belajar mencakup tahapan (1) menentukan standar kompetensi dan Kompetensi Dasar, (2) menentukan indikator pencapaian Kompetensi Dasar, (3) menyusun kisi-kisi tes, (4) menulis item tes berdasarkan kisi-kisi, (5) melakukan telaah item tes, (6) mengujikan soal pada peserta didik, (7) analisis hasil tes, dan (8) merevisi soal tes.

a. Penyusunan Kisi-kisi Tes

Kisi-kisi tes disusun dahulu baru menulis tes, bukan sebaliknya. Pada kisi-kisi ada (1) kompetensi dasar, (2) materi pokok, (3) indikator pencapaian, (4) teknik dan bentuk penilaian, dan (5) nomor item. Kisi-kisi perlu (1) mewakili isi kurikulum secara tepat, (2) komponennya jelas dan mudah dipahami, (3) dapat ditulis item tesnya. Kisi-kisi tes untuk pengukuran hasil belajar harus mengacu kepada *learning continuum* (periksa lampiran). *Learning continuum* tersebut merupakan garis yang berupa *abstract continuum*, karena mengukur berarti meletakkan posisi subjek yang diukur pada garis tersebut (Keeves & Alagumalai, 1999).

Di dalam kisi-kisi ada indikator, yaitu ciri yang dapat diamati sebagai petunjuk bahwa kompetensi dasar telah tercapai. Indikator yang baik adalah : (1) memuat ciri kompetensi dasar yang hendak diukur, (2) mengandung kata kerja operasional, (3) terkait dengan isi/materi, dan (4) dapat ditulis item tesnya.

b. Penulisan Item Tes

Item tes ditulis berdasarkan kisi-kisi. Ada rambu-rambu tertentu dalam penulisan tes objektif. Kualitas butir soal sangat ditentukan, antara lain oleh (1) kejelasan rumusan pertanyaan, yakni tidak meragukan dan komunikatif, (2) meyakinkan bahwa hanya ada satu jawaban benar, (3) tidak menggunakan pernyataan/kalimat negatif ganda. (4) pilihan jawaban homogen dan logis, (5) pengecoh (*distracter*) jangan memberikan petunjuk jawaban benar, dan (6) menggunakan bahasa Indonesia yang baik dan benar.

Untuk soal uraian, rambu-rambu penulisan butir soal adalah sebagai berikut (Hopkins, 1981).

1) Soal ditulis secara spesifik dan dapat ditangkap jelas oleh siswa. Soal seperti : *"Apa yang kau ketahui tentang bacaan di atas?"* adalah soal yang kurang spesifik. Soal

tersebut terbuka dan akan menghasilkan jawaban yang sangat bervariasi, sehingga menyulitkan penskoran.

2) Pertanyaan diawali dengan kata: *bandingkan, berikan alasan, jelaskan, uraikan, mengapa, tafsirkan, hitunglah, simpulkan, buktikan*, dan semacam itu.
3) Menghindari awal pertanyaan dengan kata : siapa, apa, kapan, atau bila, sebab kata-kata tersebut hanya akan memancing jawaban yang berupa reproduksi informasi belaka.
4) Beberapa butir soal dengan jawaban pendek lebih baik daripada satu soal dengan jawaban panjang. Hal ini terkait dengan reliabilitas tes, makin banyak jumlah butir, makin tinggi koefisien reliabilitas tes tersebut.

Berikut disajikan persyaratan baik dari aspek materi/substansi, konstruksi, dan aspek bahasa untuk tiap bentuk item.

1) Syarat item tes uraian

Item bentuk uraian atau disingkat dengan item uraian terdiri atas butir pertanyaan disertai dengan rubrik dan/atau pedoman penskoran. Item tes uraian dapat berpola memusat/konvergen (*convergent*) dan menyebar/divergen (*divergent*). Item uraian pola konvergen hanya ada satu jawaban yang benar yang sama untuk seluruh testi. Item uraian pola divergen memiliki banyak jawaban sehingga berpeluang akan terjadi perbedaan jawaban yang benar yang berbeda antartesti.

Pemberian skor untuk item tes uraian dapat mengikuti penskalaan dikotomus atau politomus. Jika pemberian skor mengikuti penskalaan dikotomus, hanya terdiri atas dua kategori yaitu kategori-1 dan kategori-2. Kategori-1 jika jawaban salah sehingga memiliki skor 0 dan kategori-2 jika jawaban benar sehingga memiliki skor 1. Jika pemberian skor item tes uraian mengikuti penskalaan politomus, dapat dibuat tiga, empat, sampai sebanyak-banyaknya 10 kategori. Kategori-1 dengan skor 0 jika semua jawaban salah, kategori-2 dengan skor 1 jika satu jawaban benar, kategori-3 dengan skor 2 jika dua jawaban benar, dan seterusnya sampai kategori-10 jika semua jawaban (yang merupakan seluruh jawaban) benar. Penskalaan item tes uraian ini sangat penting kaitannya dengan analisis item secara kuantitatif menggunakan program analisis berbantuan komputer. Penyusunan rubrik item tes uraian pola konvergen dengan penskalaan politomus untuk kepentingan analisis kualitas item secara empiris menggunakan analisis berbantuan komputer berupa model analitik.

Berikut ini disajikan mengenai persyaratan item bentuk uraian dari aspek materi/substansi, konstruksi, dan bahasa.

a) Aspek materi/substansi

(1) item sesuai indikator,
(2) pertanyaan dan rubrik dan/atau pedoman penskoran terumuskan dengan benar,
(3) materi/substansi sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status), serta
(4) materi/substansi 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) rubrik dan/atau pedoman penskorannya jelas/operasional,
(4) tabel, grafik, diagram, kasus, atau yang sejenisnya bermakna (jelas keterangannya atau ada hubungannya dengan masalah yang ditanyakan, serta
(5) antaritem tidak bergantung satu sama lain.

c) Aspek bahasa

(1) rumusan kalimat soal 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 atau bahasa serapan baru yang belum dikenal oleh seluruh testi), dan
(5) rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi.

2) Syarat item melengkapi (*completion*)

Item bentuk melengkapi atau disingkat menjadi item melengkapi, biasa disebut dengan item bentuk isian singkat atau item isian singkat karena yang digunakan untuk melengkapi dapat berupa kata atau frase/penggalan kalimat yang singkat. Item bentuk melengkapi terdiri atas pernyataan yang tidak lengkap yang bersifat terbuka. Testi/peserta ujian diminta untuk melengkapinya dengan membubuhkan kata atau frase di ujung kalimat tersebut sehingga membentuk suatu kalimat yang lengkap yang mengandung arti/benar. Jawaban kunci berupa kata/frase yang digunakan untuk melengkapinya. Oleh karena itu, dalam penyusunannya harus diusahakan agar jawabannya kunci bersifat tunggal, jangan sampai muncul jawaban alternatif. Persyaratan item bentuk melengkapi dari aspek materi/substansi, konstruksi, dan bahasa adalah sebagai berikut.

a) Aspek materi/substansi

(1) item sesuai indikator.
(2) pertanyaan jelas batasannya,
(3) rumusan kalimat yang lengkap sebagai jawaban kunci terumuskan dengan benar
(4) materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status), serta
(5) materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah, dan tingkat kelas.

b) Aspek konstruksi

(1) rumusan kalimat dalam bentuk kalimat terbuka (yang belum lengkap) yang hanya memerlukan tambahan kata atau frase pendek yang merupakan jawaban/kunci dan
(2) antaritem tidak bergantung satu sama lain.

c) Aspek bahasa

(1) rumusan kalimat soal 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* atau bahasa serapan baru yang belum dikenal oleh seluruh testi), dan
(5) rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi.

3) Syarat item benar-salah (*true-false*)

Item bentuk ini berupa suatu pernyataan yang ditawarkan kepada peserta ujian untuk diputuskan benar atau salah. Item bentuk benar-salah akan berbobot jika pernyataan mengandung unsur analisis atau sintesis. Untuk menyatakan benar salahnya pernyataan yang diajukan dapat dengan melingkari/menyilang huruf B jika benar atau S jika salah, dapat pula dengan membubuhkan tanda positif (+) jika benar dan tanda negatif (-) jika salah, atau menulis kata *ya* jika benar dan *tidak* jika salah, dan sebagainya. Oleh karena itu, harus dilengkapi dengan petunjuk cara mengerjakannya.

Item bentuk benar-salah dapat disajikan dengan lebih kompleks susunannya dengan cara, jika isi kalimat dibalik apakah tetap benar (B), tetap salah (S) menjadi benar (MB) atau menjadi salah (MS).

Contoh:

1. B S MB MS Semua pohon adalah tumbuhan
2. B S MB MS Glukose dapat berupa karbohidrat
3. B S MB MS *Monocotiledoneae* termasuk ke dalam *Oryza*
4. B S MB MS Manusia termasuk Mammalia sebab manusia tergolong hewan menyusu

Penjelasan:

1. Jawaban dinyatakan benar jika testi menyilang MS karena pernyataan "Semua pohon adalah tumbuhan" merupakan pernyataan yang benar dan pernyataan "Semua tumbuhan adalah pohon" adalah pernyataan yang salah.
2. Jawaban dinyatakan benar jika testi menyilang huruf MB karena pernyataan "Glukose dapat berupa karbohidrat" merupakan pernyataan yang salah dan pernyataan "Karbohidrat dapat berupa glukose" pernyataan yang benar.
3. Jawaban dinyatakan benar jika testi menyilang huruf MB karena pernyataan "Monocotiledoneae termasuk ke dalam Oryza" merupakan pernyataan yang salah dan pernyataan "Oryza termasuk ke dalam Monocotiledoneae" adalah pernyataan yang benar.
4. Jawaban dinyatakan benar jika testi menyilang huruf B karena pernyataan "Manusia termasuk Mammalia sebab manusia tergolong hewan menyusu" merupakan pernyataan yang benar dan pernyataan "Manusia tergolong hewan menyusu sebab manusia termasuk Mammalia" juga pernyataan yang benar.

Item bentuk benar-salah dapat disajikan dengan lebih kompleks susunannya dengan cara testi diminta menuliskan yang benar jika ada bagian pernyataan yang salah.

Beberapa hal yang perlu diperhatikan dalam menyusun item bentuk benar-salah dari aspek materi/substansi, konstruksi dan bahasa sebagai berikut.

a) Aspek materi/substansi

(1) item sesuai indikator,
(2) pernyataan sebagai pedoman jawaban kunci terumuskan dengan benar,
(3) materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status), serta
(4) materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah, dan tingkat kelas.

b) Aspek konstruksi

(1) hindarkan penggunaan kata negatif,
(2) hindarkan pernyataan yang setengah-setengah, jadi harus tegas benar atau salah,
(3) banyaknya item yang benar harus seimbang dengan banyaknya item yang salah,
(4) usahakan panjang tiap pernyataan hampir sama satu dengan yang lain, dan
(5) hindarkan pernyataan yang sekedar kutipan dari buku, dan
(6) antaritem tidak bergantung satu sama lain.

c) Aspek bahasa

(1) rumusan kalimat soal 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* atau bahasa serapan baru yang belum dikenal oleh seluruh testi), dan
(5) rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi.

4) Syarat item pilihan ganda (*multiple choice*)

Item bentuk pilihan yang berupa bentuk pilihan ganda tersusun atas butir soal dan jawaban kunci. Butir soal terdiri atas pernyataan yang belum lengkap disebut pokok soal (*stem*) yang untuk melengkapinya dengan cara memilihnya di antara beberapa alternatif/pilihan yang telah disediakan. Alternatif jawaban yang benar disebut kunci, dan yang salah atau yang kurang benar disebut pengecoh (*distracter*). Selain berupa pernyataan yang belum lengkap, pokok soal dapat berupa suatu pertanyaan. Ada yang hanya berupa pertanyaan dan ada yang diawali dengan uraian kasus, tabel, grafik, atau gambar. Jika item tes bertujuan untuk seleksi, harus ada yang terseleksi masuk kelompok atas dan kelompok bawah. Oleh karena itu, pengecoh harus berfungsi. Setiap pengecoh dipersyaratkan dapat dipilih minimal oleh 5% testi. Jika suatu item memiliki lima alternatif/pilihan jawaban maka yang terkecoh diharapkan minimal 20%. Diharapkan pula yang terkecoh adalah testi/peserta ujian yang berasal dari kelompok bawah. Jika untuk tujuan mengukur prestasi atau pencapaian hasil belajar maka pengecoh tidak akan ada yang memilih jika seluruh testi/peserta ujian benar-benar telah menguasai kompetensinya. Dengan demikian, semuanya akan memilih jawaban kunci. Sebaliknya jika testi/peserta ujian benar-benar belum menguasai kompetensi maka tidak akan ada yang memilih jawaban kunci. Beberapa hal yang perlu diperhatikan dalam menyusun item bentuk pilihan ganda dari aspek materi/substansi, konstruksi, dan bahasa adalah sebagai berikut.

a) Aspek materi/substansi

(1) item sesuai indikator,

(2) pernyataan/pertanyaan dan jawaban kunci terumuskan dengan benar,
(3) materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status), serta
(4) materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah dan tingkatan kelas.

b) Aspek konstruksi

(1) pokok soal (*stem)* dirumuskan dengan jelas,
(2) pokok 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 ganda,
(5) jika terpaksa menggunakan kata negatif, 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) uraian kasus/wacana, gambar, tabel atau grafik benar-benar berfungsi,
(11) hanya ada satu jawaban kunci yang benar, dan
(12) antaritem tidak bergantung satu sama lain.

c) Aspek bahasa

1) rumusan kalimat soal 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* atau bahasa serapan baru yang belum dikenal oleh seluruh testi), dan
5) rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi.

5) Syarat item bentuk klasifikasi atau penggolongan (*Classification*)

Item bentuk klasifikasi menggiring testi/peserta ujian untuk mengelompokkan objek atau hal berdasarkan kesamaan ciri yang dimiliki. Jadi dasarnya adalah prinsip kesamaan dan perbedaan.

Beberapa hal yang perlu diperhatikan dalam menyusun item bentuk klasifikasi dari aspek materi/substansi, konstruksi, dan bahasa adalah sebagai berikut.

a) Aspek materi/substansi

(1) item sesuai indikator,
(2) hanya ada satu jawaban kunci yang benar,
(3) materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status), serta
(4) materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah dan tingkatan kelas.

b) Aspek konstruksi

(1) di antara seluruh alternatif/pilihan hanya ada satu saja yang tidak sejenis/segolongan,
(2) dasar penggolongannya hanya satu macam,
(3) memilih alternatif jawaban sebagai pengecoh berupa objek/hal yang mirip dengan jawaban kunci, dan
(4) antaritem tidak bergantung satu sama lain.

c) Aspek bahasa

(1) rumusan kalimat soal 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* atau bahasa serapan baru yang belum dikenal oleh seluruh testi), dan
(5) rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi.

6) Syarat item bentuk pasangan (*matching*)

Item bentuk pasangan bukan merupakan item tunggal tetapi disusun seri, yakni satu seri pertanyaan dan satu seri jawaban, peserta ujian harus mencari atau menempatkan jawaban sehingga sesuai atau cocok dengan pertanyaannya. Cara menjawabnya dapat dengan menyilang atau melingkari nomor/ tanda jawaban yang sesuai atau disuruh menulis secara lengkap.

Beberapa hal yang perlu diperhatikan dalam menyusun item bentuk pasangan dari aspek materi/substansi, konstruksi, dan bahasa adalah sebagai berikut.

a) Aspek materi/substansi

(1) item sesuai indikator,
(2) hanya ada satu kunci atau jawaban yang benar untuk setiap pernyataan yang akan ditetapkan pasangannya,
(3) materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status), serta
(4) materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah dan tingkatan kelas.

b) Aspek konstruksi

(1) seri pernyataan dan seri alternatif pasangan masing-masing harus homogen,
(2) banyaknya alternatif pasangannya harus lebih banyak daripada banyaknya pernyataan yang diberi pasangan jika setiap pilihan pasangan hanya dipilih satu kali, sebaliknya pilihan pasangan harus lebih sedikit jika pilihan pasangan dapat dipilih berulang,
(3) seri item jangan terlalu banyak supaya tidak membingungkan testi/peserta ujian jika banyaknya pilihan pasangan hanya boleh dipilih satu kali, serta
(4) seri item harus homogen.

c. Aspek bahasa

(1) rumusan kalimat soal 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* atau bahasa serapan baru yang belum dikenal oleh seluruh testi), dan
(5) rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi.

7) Syarat item bentuk sebab-akibat

Item bentuk sebab-akibat memiliki tingkat kesukaran yang relatif tinggi dibandingkan dengan bentuk yang lain sehingga untuk ujian masuk perguruan tinggi misalnya, masih tetap dipertahankan penggunaannya. Pada dasarnya bentuk sebab-akibat terdiri atas dua pernyataan. Pernyataan pertama berupa akibat dan pernyataan kedua berperan sebagai penyebab. Oleh karena itu, antara pernyataan pertama dan kedua selalu disisipi/dihubungkan dengan kata **sebab** atau **karena**. Testi/peserta ujian disuruh untuk memutuskan apakah memang kedua pernyataan tersebut menunjukkan hubungan sebab-akibat, kedua pernyataan benar tetapi tak menunjukkan hubungan sebab-akibat; kedua pernyataan tersebut salah satu di antaranya salah atau kedua pernyataan salah. Oleh karena itu, banyaknya alternatif jawabannya dapat empat atau lima. Alternatif jawaban sebanyak lima jika alternatif "salah satu pernyataan salah" dipisah menjadi "pernyataan pertama benar dan pernyataan kedua (alasan) salah" serta "pernyataan pertama salah dan pernyataan kedua (alasan) benar".

Beberapa hal yang perlu diperhatikan dalam menyusun item bentuk sebab-akibat dari aspek materi/substansi, konstruksi, dan bahasa adalah sebagai berikut.

a) Aspek materi/substansi

(1) item sesuai indikator,
(2) pernyataan dalam bentuk hubungan sebab-akibat sebagai jawaban kunci terumuskan dengan benar,
(3) materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status); serta
(4) materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah dan tingkatan kelas.

b) Aspek konstruksi

(1) sifat hubungan harus tegas, apakah memang menunjukkan hubungan sebab akibat atau tidak
(2) setiap pernyataan harus tegas, apakah pernyataan tersebut termasuk kategori benar ataukah termasuk kategori salah,
(3) dalam menyusun suatu pernyataan hindari kata-kata "selalu", "pasti", "tidak", "tidak mungkin", dan sejenisnya, serta
(4) antaritem tidak bergantung satu sama lain.

c) Aspek bahasa

(1) rumusan kalimat soal 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* atau bahasa serapan baru yang belum dikenal oleh seluruh testi), dan
(5) rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi.

8) Item bentuk asosiasi pilihan ganda

Item bentuk asosiasi pilihan ganda atau bentuk pilihan berganda ini meminta setiap testi/peserta ujian disuruh mencari alternatif jawaban yang benar. Mungkin semua alternatif jawaban benar, mungkin hanya sebagian yang benar, dan mungkin tak ada yang benar. Bentuk pilihan yang biasa digunakan adalah sebagai berikut.

a. jawaban (1), (2), dan (3) benar
b. jawaban (1) dan (3) benar
c. jawaban (2) dan (4) benar
d. jawaban (4) benar

Bentuk pilihan seperti di atas memiliki kelemahan yang mendasar karena konsep asosiasi dalam arti pilihan berganda tidak terpenuhi. Testi/peserta ujian akan dapat menyiasati dengan hanya menetapkan apakah pernyataan nomor 1 salah. Jika salah, otomatis tinggal pilihan c dan d. kemudian jika pernyataan nomor 2 salah maka otomatis jatuh pada pilihan d. Jika pernyataan nomor 1 benar, testi/peserta ujian juga tinggal menetapkan apakah pernyataan nomor 2 benar atau salah. Jika pernyataan nomor 2 benar maka otomatis pilihan jatuh pada alternatif a, jika salah akan jatuh pada alternatif b. Oleh karena itu, akan lebih baik jika dibuat model pilihan sebagai berikut.

Beberapa hal yang perlu diperhatikan dalam penyusunan item bentuk asosiasi ganda dari aspek materi/substansi, konstruksi, dan bahasa adalah sebagai berikut.

a) Aspek materi/substansi

(1) item sesuai indikator,
(2) pernyataan/pertanyaan dan jawaban kunci terumuskan dengan benar,
(3) materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status), serta
(4) materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah, dan tingkat kelas.

b) Aspek konstruksi

(1) pokok soal harus dirumuskan dengan jelas dan tegas,
(2) semua pernyataan yang dijadikan pilihan dirumuskan secara lugas sehingga jelas dapat dibedakan apakah setiap tersebut benar atau salah,
(3) pokok soal tidak mengarah ke jawaban yang benar, misalnya menggunakan kata salah satu, tiga hal atau yang sejenisnya,
(4) seluruh pernyataan pada alternif jawaban harus homogen, baik dalam bentuk bunyi kata ataupun istilah,
(5) gambar, grafik, tabel ataupun wacana benar-benar berfungsi; serta
(6) antaritem tidak bergantung satu sama lain.

c) Aspek bahasa

(1) rumusan kalimat soal 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* atau bahasa serapan baru yang belum dikenal oleh seluruh testi), dan
(5) rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi.

9) Syarat item bentuk ikhtisar

Item bentuk ikhtisar menggiring testi/peserta ujian untuk menyusun ikhtisar dengan benar. Beberapa hal yang perlu diperhatikan dalam menyusun item bentuk ikhtisar dari aspek materi/substansi, konstruksi, dan bahasa adalah sebagai berikut.

a) Aspek materi/substansi

(1) item sesuai indikator,
(2) Ikhtisar/penggolongan sebagai jawaban kunci terumuskan dengan benar,
(3) materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status);serta
(4) materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah, dan tingkat kelas.

b) Aspek konstruksi

(1) jawaban yang diminta hanya ada satu macam ikhtisar,
(2) rambu-rambu dalam membuat ikhtisar harus tegas,
(3) ikhtisar yang diminta jangan terlalu panjang, dan
(4) antaritem tidak bergantung satu sama lain.

c) Aspek bahasa

(1) rumusan kalimat soal 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* atau bahasa serapan baru yang belum dikenal oleh seluruh testi), dan
(5) rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi.

10) Syarat item bentuk hubungan konteks

Item bentuk hubungan konteks terdiri atas teks/wacana/uraian kasus dan sejumlah pernyataan. Testi/peserta ujian diminta menetapkan terlebih dahulu apakah setiap pernyataan benar dan bagaimana hubungan dengan teks/uraian kasusnya. Beberapa hal yang perlu diperhatikan dalam menyusun item bentuk mencari hubungan konteks dari aspek materi/substansi, konstruksi, dan bahasa adalah sebagai berikut.

a. Aspek materi/substansi

(1) item sesuai indikator,
(2) Teks/wacana dan tiap-tiap pertanyaan sebagai jawaban kunci terumuskan dengan benar,

(3) materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status); serta
(4) materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah, dan tingkat kelas.

b) Aspek konstruksi

(1) sifat hubungan antara pernyataan dengan teks/wacana harus tegas, apakah memang menunjukkan ada hubungan atau tidak;
(2) setiap pernyataan harus tegas, apakah pernyataan tersebut benar atau salah; serta
(3) menghindari pernyataan yang menggunakan kata "selalu", "pasti", "tidak", "tidak mungkin", dan sejenisnya serta
(4) antaritem tidak bergantung satu sama lain.

c) Aspek bahasa

(1) rumusan kalimat soal 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* atau bahasa serapan baru yang belum dikenal oleh seluruh testi), dan
(5) rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi.

Pengembangan Item Tes Pola Divergen

Untuk item tes pola konvergen dan bentuk divergen yang sederhana perlu disiapkan rubrik disiapkan rubrik sebagai jawaban dan sekaligus sebagai pedoman penskoran. Untuk item tes bentuk divergen yang kompleks, selain ada rubrik juga ada pedoman penskorannya Rubrik memuat semua alternatif jawaban yang benar sedangkan pedoman penskorannya untuk menetapkan berapa skor maksimum yang dicapai.

Berikut disajikan contoh butir tes yang termasuk pola divergen untuk Biologi

1. Apa yang menyebabkan Anda salah ukur saat Anda mengukur panjang tubuh hewan menggunakan pita ukur? **Tuliskan lebih dari satu jawaban!**

   Rubrik (Jawaban-jawaban yang benar)
   a. pita ukur **tidak standar**/jarak antargaris batas tidak sama
   b. angka dalam pita ukur sudah **tidak jelas**
   c. **tidak cermat/teliti** membaca angka dalam pita ukur
   d. mata **rabun/sakit**
   e. **tidak** tepat **memasangkannya** pada tubuh hewan/pita **terlipat**
   f. hewan **terus menerus bergerak**
   g. pita ukur **tidak sebanding** dengan tubuh hewan
   h. **posisi mata** tidak tegak lurus pada skala
   i. **salah mencatat** hasil pengamatan

Pedoman penskoran:

Kategori-1	=	tidak ada jawaban yang benar
Kategori-2	=	jawaban yang benar hanya 1
Kategori-3	=	jawaban yang benar >1

2. Apa yang Anda lakukan agar Anda dapat bekerja menggunakan termometer dengan aman juga tidak merusakkannya? **Tuliskan lebih dari satu cara!**

   Rubrik (Jawaban-jawaban yang benar)
   a. menjaga **jangan sampai pecah** (**jatuh/terbentur**)
   b. **kaca dikumpulkan** bila sampai pecah
   c. **memilih rentang skala** suhu sesuai dengan taksiran suhu benda yang diukur
   d. yang diukur **zat beracun**, setelah selesai **termometer juga tangan dicuci bersih**
   e. **tidak meletakkannya** di **sembarang tempat**
   f. bila telah **selesai** masukkan **ke tempatnya**
   g. **tidak memegangnya** dengan **tangan licin/basah** agar tidak lepas
   h. **bertanya** pada guru bila belum tahu cara menggunakan
   i. **memegang erat** saat menggoyang termometer agar **posisi air raksa ke titik nol**
   j. **menggunakan** sesuai prosedur/sebagaimana mestinya**/hati-hati/serius**
   k. **mengecek** termometer tidak **pecah/rusak** sebelum dipakai
   l. **membaca petunjuknya** sebelum menggunakannya
   m. meminta guru/ahli untuk **membimbing/memperagakannya**

Pedoman penskoran:		
Kategori-1	=	tidak ada jawaban yang benar
Kategori-2	=	jawaban yang benar hanya 1
Kategori-3	=	jawaban yang benar >1

Analisis Kualitatif/Telaah Item Tes

Telaah item tes atau analisis kualitatif sering disebut analisis teoretik. Item tes yang telah ditulis diperiksa kesesuaiannya dengan kisi-kisi yang diacunya dengan memperhatikan substansi/isi materi, konstruksi, dan bahasa. Telaah item tes dilakukan oleh (1) bukan oleh penulis item tes, dan (2) dilakukan oleh pakar yang menguasai isi/materi yang diujikan. Dari kegiatan tersebut, penulis item tes dapat mengetahui validitas isi dari item tes yang disusun.

Dalam melakukan analisis kualitatif, setiap item ditelaah sesuai dengan persyaratan item yang bersangkutan. Berikut disajikan tabel analisis kualitatif sesuai dengan bentuk tes. Cara penggunaannya, penelaah memberi tanda silang (X) jika suatu persyaratan dari suatu item dapat terpenuhi.

Telaah item tes bentuk klasifikasi atau penggolongan (*Classification*)

No	Aspek	Nomor item				
a.	**Aspek materi/substansi**	1	2	3	4	dst
1.	Item sesuai dengan indikator					
2.	Hanya ada satu jawaban kunci yang benar					
3.	Materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status),					
4.	Materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah dan tingkatan kelas					
b.	**Aspek konstruksi**					
1.	Di antara seluruh alternatif/pilihan hanya ada satu saja yang tidak sejenis/segolongan					
	Dasar penggolongannya hanya satu macam					
2.	Memilih alternatif jawaban sebagai pengecoh berupa objek/hal yang mirip dengan jawaban kunci					
3.	Antaritem tidak bergantung satu sama lain					
c.	**Aspek bahasa**					
1.	Rumusan kalimat soal 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 atau bahasa serapan baru yang belum dikenal oleh seluruh testi)					
5.	Rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi					

Telaah item tes bentuk uraian

No	Aspek	No item				
a.	**Aspek materi/substansi**	1	2	3	4	dst
1.	Item sesuai dengan indikator					
2.	Pertanyaan dan rubrik dan/atau pedoman penskoran terumuskan dengan benar					
3.	Materi/substansi sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status),					
4.	Materi/substansi 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.	Rubrik dan/atau pedoman penskorannya jelas/operasional					
4.	Tabel, grafik, diagram, kasus, atau yang sejenisnya bermakna (jelas keterangannya atau ada hubungannya dengan masalah yang ditanyakan)					
5.	Antaritem tidak bergantung satu sama lain					
c.	**Aspek bahasa**					
1.	Rumusan kalimat soal komunikatif					
2.	Kalimat menggunakan bahasa yang baik dan benar, sesuai dengan jenis bahasanya					
3.	Rumusan kalimat tidak menimbulkan penafsiran ganda					
4.	Menggunakan bahasa/kata yang umum (bukan bahasa lokal atau bahasa serapan baru yang belum dikenal oleh seluruh testi)					
5.	Rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi					

Telaah item tes bentuk melengkapi (*completion*)

No	Aspek	Nomor item				
a.	**Aspek materi/substansi**	1	2	3	4	dst
1.	Item sesuai dengan indikator					
2.	Pertanyaan jelas batasannya					
3.	Rumusan kalimat yang lengkap sebagai jawaban kunci terumuskan dengan benar					
4.	Materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status),					
5.	Materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah, dan tingkat kelas					
b.	**Aspek konstruksi**					
1.	Rumusan kalimat dalam bentuk kalimat terbuka (yang belum lengkap) yang hanya memerlukan tambahan kata atau frase pendek yang merupakan jawaban/kunci					
2.	Antaritem tidak bergantung satu sama lain					
c.	**Aspek bahasa**					
1.	Rumusan kalimat soal 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 atau bahasa serapan baru yang belum dikenal oleh seluruh testi),					
	Rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi					

Telaah item tes bentuk benar-salah (*true-false*)

No	Aspek	Nomor item				
a.	**Aspek materi/substansi**	1	2	3	4	dst
1.	Item sesuai dengan indikator					
2.	Pernyataan sebagai pedoman jawaban kunci terumuskan dengan benar					
3.	Materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status)					
4.	Materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah, dan tingkat kelas					
b.	**Aspek konstruksi**					
1.	Hindarkan penggunaan kata negatif					
2.	Hindarkan pernyataan yang setengah-setengah, jadi harus tegas benar atau salah					
3.	Banyaknya item yang benar harus seimbang dengan banyaknya item yang salah					
4.	Usahakan panjang tiap pernyataan hampir sama satu dengan yang lain					
5.	Hindarkan pernyataan yang sekedar kutipan dari buku					
6.	Antaritem tidak bergantung satu sama lain					
c.	**Aspek bahasa**					
1.	Rumusan kalimat soal 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 atau bahasa serapan baru yang belum dikenal oleh seluruh testi)					
5.	Rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi					

Telaah item tes bentuk pilihan ganda (*multiple choice*)

No	Aspek	Nomor item				
a.	**Aspek materi/substansi**	1	2	3	4	dst
1.	Item sesuai dengan indikator					
2.	Pernyataan/pertanyaan dan jawaban kunci terumuskan dengan benar					
3.	Materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status),					
4.	Materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah dan tingkatan kelas					
b.	**Aspek konstruksi**					
1.	Pokok soal (*stem)* dirumuskan dengan jelas					
2.	Pokok 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 ganda jika terpaksa menggunakan kata negatif, harus digarisbawahi atau dicetak lain					
5.	Pilihan jawaban homogen					
6.	Hindari adanya alternatif jawaban: "seluruh jawaban di atas benar" atau "tak satu jawaban di atas yang benar" dan yang sejenisnya					
7.	Panjang alternatif /pilihan jawaban relatif sama (jangan ada yang sangat panjang dan ada yang sangat pendek)					
8.	Pilihan jawaban dalam bentuk angka/waktu diurutkan					
9.	Uraian kasus/wacana, gambar, tabel atau grafik benar-benar berfungsi					
10.	Hanya ada satu jawaban kunci yang benar					
11.	Antaritem tidak bergantung satu sama lain					
c.	**Aspek bahasa**					
1.	Rumusan kalimat soal 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 atau bahasa serapan baru yang belum dikenal oleh seluruh testi)					
5.	Rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi					

Lembar telaah item tes bentuk pasangan (*matching*)

No	Aspek	Nomor item				
a.	**Aspek materi/substansi**	1	2	3	4	dst
1.	Item sesuai dengan indikator					
2.	Hanya ada satu kunci atau jawaban yang benar untuk setiap pernyataan yang akan ditetapkan pasangannya					
3.	Materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status),					
4.	Materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah dan tingkatan kelas					
b.	**Aspek konstruksi**					
1.	Seri pernyataan dan seri alternatif pasangan masing-masing harus homogen					
2.	Banyaknya alternatif pasangannya harus lebih banyak daripada banyaknya pernyataan yang diberi pasangan jika setiap pilihan pasangan hanya dipilih satu kali, sebaliknya pilihan pasangan harus lebih sedikit jika pilihan pasangan dapat dipilih berulang					
3.	Seri item jangan terlalu banyak supaya tidak membingungkan testi jika banyaknya pilihan pasangan hanya boleh dipilih satu kali					
4.	Seri item harus homogen					
5.	Antaritem tidak bergantung satu sama lain					
c.	**Aspek bahasa**					
1.	Rumusan kalimat soal 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* atau bahasa serapan baru yang belum dikenal oleh seluruh testi)					
5.	Rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi					

Lembar telaah item tes bentuk sebab akibat

No	Aspek	Nomor item				
a.	**Aspek materi/substansi**	1	2	3	4	dst
1.	Item sesuai dengan indikator					
2.	Pernyataan dalam bentuk hubungan sebab-akibat sebagai jawaban kunci terumuskan dengan benar					
3.	Materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status);					
4.	Materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah dan tingkatan kelas					
b.	**Aspek konstruksi**					
1.	Sifat hubungan harus tegas, apakah memang menunjukkan hubungan sebab akibat atau tidak					
2.	Setiap pernyataan harus tegas, apakah pernyataan tersebut termasuk kategori benar ataukah termasuk kategori salah					
3.	Dalam menyusun suatu pernyataan hindari kata-kata "selalu", "pasti", "tidak", "tidak mungkin", dan sejenisnya					
4.	Antaritem tidak bergantung satu sama lain.					
c.	**Aspek bahasa**					
1.	Rumusan kalimat soal 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* atau bahasa serapan baru yang belum dikenal oleh seluruh testi)					
5.	Rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi					

Lembar telaah item tes bentuk asosiasi pilihan ganda

No	Aspek	Nomor item				
a.	**Aspek materi/substansi**	1	2	3	4	dst
1.	Item sesuai dengan indikator					
2.	Pernyataan/pertanyaan dan jawaban kunci terumuskan dengan benar					
3.	Materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status),					
4.	Materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah, dan tingkat kelas					
b.	**Aspek konstruksi**					
1.	Pokok soal harus dirumuskan dengan jelas dan tegas					
2.	Semua pernyataan yang dijadikan pilihan dirumuskan secara lugas sehingga jelas dapat dibedakan apakah setiap tersebut benar atau salah					
3.	Pokok soal tidak mengarah ke jawaban yang benar, misalnya menggunakan kata salah satu, tiga hal atau yang sejenisnya					
4.	Seluruh pernyataan pada alternatif jawaban harus homogen, baik dalam bentuk bunyi kata ataupun istilah					
5.	Gambar, grafik, tabel ataupun wacana benar-benar berfungsi					
6.	Antaritem tidak bergantung satu sama lain					
c.	**Aspek bahasa**					
1.	Rumusan kalimat soal 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* atau bahasa serapan baru yang belum dikenal oleh seluruh testi)					
5.	Rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi					

Lembar telaah item tes bentuk ikhtisar

No	Aspek	Nomor item				
a.	**Aspek materi/substansi**	1	2	3	4	dst
1.	Item sesuai dengan indikator					
2.	Ikhtisar/penggolongan sebagai jawaban kunci terumuskan dengan benar					
3.	Materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status);					
4.	Materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah, dan tingkat kelas					
b.	**Aspek konstruksi**					
1.	Jawaban yang diminta hanya ada satu macam ikhtisar					
2.	Rambu-rambu dalam membuat ikhtisar harus tegas					
3.	Ikhtisar yang diminta jangan terlalu panjang					
4.	Antaritem tidak bergantung satu sama lain					
c.	**Aspek bahasa**					
1.	Rumusan kalimat soal 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* atau bahasa serapan baru yang belum dikenal oleh seluruh testi)					
5.	Rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi					

Lembar telaah item tes bentuk hubungan konteks

No	Aspek	Nomor item				
a.	**Aspek materi/substansi**	1	2	3	4	dst
1.	Item sesuai dengan indikator					
2.	Teks/wacana dan tiap-tiap pertanyaan sebagai jawaban kunci terumuskan dengan benar					
3.	Materi/substansi yang ditanyakan sesuai dengan tujuan pengukuran (untuk tujuan pengukuran hasil belajar, tujuan pengukuran untuk seleksi, atau tujuan pengukuran untuk konfirmatori/mengukur status);					
4.	Materi/substansi yang ditanyakan sesuai dengan jenjang, jenis sekolah, dan tingkat kelas					
b.	**Aspek konstruksi**					
1.	Sifat hubungan antara pernyataan dengan teks/wacana harus tegas, apakah memang menunjukkan ada hubungan atau tidak					
	Setiap pernyataan harus tegas, apakah pernyataan tersebut benar atau salah					
	Menghindari pernyataan yang menggunakan kata "selalu", "pasti", "tidak", "tidak mungkin", dan sejenisnya					
	Antaritem tidak bergantung satu sama lain					
c.	**Aspek bahasa**					
1.	Rumusan kalimat soal 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* atau bahasa serapan baru yang belum dikenal oleh seluruh testi)					
5.	Rumusan soal tidak mengandung kata-kata yang dapat menyinggung perasaan testi					

B. Analisis kuantitatif

Setelah dilakukan telaah item tes, item yang kurang/tidak memenuhi persyaratan direvisi oleh penulis item tes yang bersangkutan. Selanjutnya, item dirangkai menjadi seperangkat tes yang siap diujikan pada siswa. Selama pengujian, perlu diciptakan suasana ujian yang kondusif/optimal agar hasil yang diperoleh dapat benar-benar dapat dipercaya.

Data yang masuk kemudian dianalisis untuk memperoleh kejelasan tentang kualitas tes yang disusun. Analisis dilakukan baik terhadap tes maupun terhadap item tes. Analisis tes untuk mengetahui reliabilitas tes, sedangkan analisis item tes untuk mengetahui kualitas setiap item tes yang bersangkutan. Analisis dapat dilakukan dengan metode yang hanya berupa pengukuran klasik, seperti program komputer ITEMAN, dapat pula menggunakan

kombinasi pengukuran klasik dan modern (*Item Respons Theory*) dengan program komputer QUEST.

Agar hasil pengukuran baik, tes yang digunakan harus baik pula. Tes yang berkualitas tinggi apabila tes itu mampu mengukur secara tepat dan hasil pengukurannya dapat diandalkan. Dengan kata lain, tes perlu memenuhi syarat validitas dan reliabilitas. Suatu tes mempunyai validitas tinggi apabila tes tersebut dapat menjalankan fungsi ukurnya. Sisi lain dari konsep validitas yang penting adalah kecermatan pengukuran, yakni kemampuan untuk mendeteksi perbedaan-perbedaan kecil yang ada pada atribut yang diukurnya.

1. Validitas Tes

Suatu alat ukur dinyatakan sahih (valid), jika alat ukur tersebut benar-benar mampu memberikan informasi empirik sesuai dengan apa yang diukur. Suatu neraca pengukur massa dinyatakan sahih jika suatu benda yang memiliki massa 1 kg akan terukur 1 kg pula karena neraca tersebut memiliki akurasi/ keakuratan/ketelitian yang tinggi.

Hakekat kesahihan dalam ilmu sosial selain terkait dengan kelayakan suatu alat ukur juga berkait dengan masalah derajat. Jadi bukan masalah sahih dan tidak sahih. Kesahihan selalu bersifat spesifik kaitannya dengan tujuan dan interpretasi. Kesahihan juga mencerminkan ketunggalan (unidimensionalitas) konsep yang akan diukur. Selain itu alat ukur juga harus memiliki sifat andal (*reliable*). Artinya jika dipakai untuk mengukur secara berulang-ulang selalu tetap/konsisten/stabil hasilnya. Jadi, jika penimbangan pertama 50 kg, penimbangan kedua juga menunjukkan berat 50 kg, demikian pula pada ulangan yang selanjutnya. Alat ukur yang andal memiliki presisi/ketepatan/kecermatan yang tinggi.

Suatu alat ukur dapat saja memiliki keandalan yang tinggi namun tidak sahih. Misalnya, suatu benda yang memiliki massa 50 kg dan setelah diukur massanya menggunakan suatu alat pengukur massa secara berulang-ulang hasilnya selalu menunjukkan bahwa massa benda tersebut 49 kg. Artinya bahwa alat pengukur massa tersebut andal karena dapat mengukur secara ajeg/konsisten yaitu sebesar 49 kg tetapi tidak sahih/akurat karena tidak menunjukkan besar massa benda tersebut yang sesungguhnya. Kecuali itu, hakekat keandalan suatu alat ukur adalah mengacu kepada hasil yang diperoleh bukan alat ukurnya itu sendiri. Keandalan alat ukur juga tidak berlaku secara umum. Ganti waktu atau ganti sampel mungkin sudah dapat berubah keandalannya. Suatu alat ukur yang andal tidak otomatis sahih dan keandalan suatu alat ukur juga sangat statis.

Menurut Gronlund (1985) hakekat validitas tes dapat ditinjau dari beberapa aspek sebagai berikut.

a. Validitas tes mengacu kepada kepantasan/ketercukupan (*appropriateness*) interpretasi hasil pengukuran yang dikenakan pada kelompok atau individu testi/peserta ujian terhadap instrumen tes yang digunakan. Berbicara tentang validitas adalah Berbicara tentang interpretasi yang dapat dibuat dari hasil pengukuran. Semakin valid suatu instrumen tes semakin tepat interpretasi hasil pengukuran yang diperoleh.

b. Validitas adalah sesuatu hal yang memiliki derajat sehingga dapat dibedakan antara yang benar-benar sahih (*valid*) dan yang benar-benar tidak sahih (*invalid*) atau antara yang rendah, sedang, dan tinggi validitasnya.
c. Validitas adalah mengacu kepada suatu tujuan yang spesifik. Misalnya, suatu tes aritmatik dinyatakan memiliki derajat validitas yang tinggi jika dapat menunjukkan keterampilan dalam komputasi, dinyatakan memiliki derajat validitas yang rendah jika hanya untuk menunjukkan kemampuan berpikir aritmatik, dan memiliki derajat validitas yang sedang jika dapat digunakan untuk memprediksi keberhasilan belajar aritmatik ke depan. Akan tetapi, interpretasi tersebut tidak mampu untuk menilai atau menggambarkan validitas guna mempertimbangkan penggunaan hasil tes yang diperoleh.
d. Validitas adalah konsep kesatuan (*unitary*). Hakekat konsepsual validitas tes untuk bidang pendidikan dan psikologi kini didasarkan pada tiga aspek yang ditinjau berdasarkan bukti-bukti empiris yaitu aspek isi (*content*), aspek hubungannya dengan kriteria (*criterion related*) dan aspek konstruk (*construct*).

2. Pendekatan Untuk Menguji Validitas

a. Bukti yang berkaitan dengan isi

Hal yang menjadi pertanyaan pokok yang berkait dengan kesahihan isi (*content validity*) adalah "apakah sampel gugus alat ukur yang digunakan sudah representatif" atau bila berupa tes "seberapa baguskah sampel tes yang digunakan untuk mengukur gugus tugas merepresentasikan ranah (*domain*) gugus tugas yang akan diukur". Dengan kata lain, secara sederhana dapat dipertanyakan adalah apakah semua item dalam suatu perangkat tes yang diujikan sudah mampu mencerminkan semua komponen ranah kemampuan tertentu dari subjek yang diukur. .

Karena tujuan pengukuran untuk mengetahui sejauh mana kemampuan seseorang ditinjau dari segi domain ranah, tinjauan representasi dapat diartikan dari sisi kelengkapan komponen ranah yang akan diukur. Misalnya, jika yang akan diukur kemampuan kognitif peserta didik dan berdasarkan teori-teori yang ada peserta didik yang bersangkutan harus sudah memiliki kemampuan kognitif mulai dari mengingat (*to know*) sampai dengan berkreasi (*to create*) pada taksonomi Bloom yang terbaru, item yang digunakan dinyatakan terpenuhi kesahihan isinya alat uji yang digunakan meliputi item jenjang dari mengingat, memahami, mengaplikasikan, menganalisis, mengevaluasi, menyintesis, berimajinasi, dan berkreasi. Oleh karena dalam ranah kognitif antartingkatan berpikir sudah berbeda dimensinya, pemenuhan validitas dapat dibuat rentang kemampuan sesuai dengan tingkatan berpikir dan cakupan konsep keilmuan yang harus dikuasai. Dengan demikian, jika pengukuran itu dikaitkan dengan hasil belajar, maka akan dapat dibuat rentang kemampuan berdasarkan tingkatan berpikir dan keluasan serta kedalaman aspek keilmuannya dan akan menjadi suatu garis abstrak yang menunjukkan suatu kontinum kemampuan dari yang terendah sampai yang tertinggi yang akan menunjukkan capaian hasil belajar. Itulah yang kemudian disebut dengan *learning continuum*.

Dalam proses pembelajaran, tujuan yang ditargetkan sudah dirumuskan dalam bentuk kompetensi (SK dan KD). Oleh karena itu, persoalan validitas isi bergeser menjadi "Apakah item instrumen yang digunakan sudah mencerminkan seluruh KD yang ditargetkan?". Jika seluruh KD tidak dapat diukur, pertanyaannya adalah "Apakah item instrumen yang digunakan sudah mewakili seluruh KD yang akan diukur?". Jadi, dalam hal ini berbicara tentang representasi butir soal. Karena tinjauannya dikembalikan kepada KD yang ditargetkan, maka kesahihan isi juga disebut dengan kesahihan kurikuler (*Curricular validity*). Pertanyaan yang menarik yaitu apakah KD dalam target pembelajaran hanya spesifik ranah kognitif. Jika hanya ranah kognitif, bagaimana dengan tujuan pendidikan yang diharapkan mencakup seluruh ranah/domain. Bahkan, sebagai hasil pembelajaran idealnya menjadi satu kesatuan/*unity*.

b. Bukti yang berkaitan dengan konstruk atau *construct validity*

Persoalan pokok yang dikaji dalam rangka pemenuhan validitas konstruk adalah pertanyaan mengenai "Apakah hasil pengukuran benar-benar bermakna atau benar-benar dapat diinterpretasikan. Dengan kata lain, apakah alat ukur yang digunakan benar-benar mampu mengukur hal atau variabel yang diukur. Apakah jika setiap butir soal sudah mencerminkan indikator variabel yang hendak diukur sudah dapat dinyatakan bahwa pengukuran sudah mengukur kompetensi yang dimaksud.

Persoalan yang dihadapi dalam pemenuhan validitas konstruk dalam ranah kognitif bukan hanya terbatas pada kesesuaian item dengan indikator dengan pencapaian kompetensi. Persoalan yang mendasar adalah apakah sejumlah kompetensi yang diukur berada dalam satu dimensi. Jika tidak dalam satu dimensi, akan dipertanyakan bagaimana pemenuhan validitas konstruknya. Oleh karena tidak mudah untuk memenuhi validitas konstruk, dalam pengukuran penguasaan ranah kognitif akan lebih mudah dipahami jika pemenuhan validitas dibatasi dalam pemenuhan validitas isi yang tergambar dalam *learning continuum*.

Dalam bidang psikologi, termasuk di dalamnya psikologi pendidikan, pemenuhan kesahihan konstruk dikaitkan dengan sifat variabel yang akan diukur. Dalam hal ini adalah aspek psikologi meskipun variabel-variabel yang akan diukur bersifat abstrak. Dengan demikian, fenomenanya tidak dapat ditangkap secara langsung melalui panca indera. Oleh karena itu, perlu dicari indikator/indikan yang tepat agar variabel tersebut benar-benar dapat diukur. Hal ini akan mudah ditangkap jika dibandingkan dengan Ilmu Kimia. Fenomena reaksi kimia sulit ditangkap dengan panca indera. Oleh karena itu, cara yang dapat ditempuh adalah membubuhkan zat yang berfungsi sebagai indikator. Zat yang berfungsi sebagai indikator tersebut tidak akan ikut bereaksi, namun akan berubah warnanya jika reaksi kimia sudah berlangsung. Dengan demikian, dapat dikatakan bahwa zat A sudah bereaksi dengan zat B karena warna zat yang berfungsi sebagai indikator sudah berubah warnanya.

Cara untuk memenuhi validitas konstruk adalah dengan membuat definisi operasional variabel yang akan diukur. Jika akan mengukur minat, maka dibuat terlebih dahulu apa yang dimaksud dengan minat secara operasional. Dari definisi operasional yang sudah dirumuskan selanjutnya dicari indikator-indikatornya. Dengan cara demikian, pemenuhan unidimensionalitas variabel yang akan diukur berpeluang dapat dipenuhi.

Setelah variabel yang akan diukur dijabarkan ke dalam indikator-indikatornya barulah disusun pertanyaan-pertanyaan yang mencerminkan masing-masing indikator tersebut. Dengan demikian, alat uji akan memiliki kesahihan konstruk jika item-itemnya mencerminkan indikator-indikator dari variabel yang akan diukur, juga ketepatan memilih indikator-indikator dari variabel yang diukur.

Menurut Embretson & Gorin (2001) validitas konstruk adalah sentral untuk menetapkan mutu tes. Secara tradisional, spesifikasi item sering hanya samar-samar. Pengembang tes membuat spesifikasi item yang sering kali 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 konkret). Begitu keseluruhan spesifikasi item dihasilkan kemudian dilakukan telaah (*review*) item oleh penelaah (*reviewer*) untuk meyakinkan ketepatan 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 "konstruk-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 konstruk menurut teori klasik.

Berdasarkan prinsip psikologi kognitif, bahwa yang pertama yang terpenting dalam pemenuhan validitas konstruk suatu tes kemampuan/abilitas adalah adanya dukungan satu set prinsip teoretis dari item-item tes yang akan disusun. Kemudian, suatu set standar yang baru yang disusun harus didasarkan pada prinsip-prinsip psikologi kognitif. Mengutip pendapat Messick (1995) bahwa item-item yang disusun tidak sekadar 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 konstruk dari proses-proses kognitif. Terakhir, bahwa tes yang disusun adalah untuk mengindikasikan kecakapan atau untuk mendiagnosis ketrampilan yang ada sehingga memerlukan bermacam informasi baru tentang item-item yang disusun.

Pemenuhan validitas konstruk secara empiris akan dapat diperoleh jika aspek yang diukur bersifat spesifik. Misalnya, akan lebih mudah memenuhi bukti empiris mengembangkan tes untuk mengukur kemampuan berkreasi daripada mengembangkan untuk mengukur seluruh aspek kognitif mulai dari mengingat sampai berkreasi. Pemenuhan validitas konstruk secara empiris dapat didekati dengan membandingkan skor yang diperoleh dari kelompok-kelompok yang telah diketahui karakteristiknya. Misalnya, secara umum kemampuan seseorang akan bertambah meningkat sejalan dengan pertambahan umur. Artinya, bahwa suatu tes yang dikenakan pada kelompok yang lebih tua usianya akan menunjukkan skor rata-rata yang lebih tinggi dibanding kelompok yang lebih muda. Skor tes prestasi akan dicapai kelompok yang pernah ikut pelatihan dibandingkan yang belum mengikutinya. Skor rata-rata kemampuan *reading comprehension* pada perempuan lebih tinggi dibandingkan dengan laki-laki. Jadi, jika hasil tes yang dikembangkan setelah diujicobakan pada kelompok yang telah diketahui spesifikasinya menunjukkan hasil yang berkesesuaian berarti terpenuhi bukti empirisnya.

Hasil skor jenis tes tertentu, seperti tes intelegensi, tidak terpengaruh dengan adanya tambahan pengalaman pelatihan. Namun, tes prestasi belajar akan menunjukkan skor yang berbeda antara peserta didik sebelum dan sesudah belajar. Dalam eksperimen pun, akan ada perbedaan skor rata-rata antara kelompok perlakuan dan kelompok pembanding. Bila perlakuan memberikan dampak yang positif, skor rata-rata kelompok perlakuan akan menunjukkan hasil yang lebih tinggi secara nyata (*significant*). Hal ini akan menjadi bukti empiris terhadap tes yang dikembangkan.

Membandingkan skor hasil tes yang dikembangkan dengan skor tes lain yang sama/se bidang juga dapat untuk menunjukkan bukti empiris. Misalnya, hasil suatu tes bakat (*aptitude test*) yang satu akan memiliki korelasi dengan hasil tes bakat yang pernah dikembangkan sebelumnya. Cara ini merupakan cara yang dilakukan dengan menghubungkan hasil tes dengan kriteria (*criterion-related study*).

Cara yang tepat untuk memperoleh bukti empiris adalah menghubungkan dengan penguasaan testi terhadap tes yang ditempuhnya. Di dalam tes beracuan norma sulit untuk mengetahui bagaimana hasil yang dicapai testi orang per orang karena testi yang satu dibandingkan dengan testi lain di dalam kelompoknya. Sementara itu, di dalam tes beracuan kriteria, skor suatu item tes dapat menunjukkan *prefect score* (*zero variability*) jika semua testi benar-benar menguasai aspek yang diujikan melalui item yang bersangkutan. Tes prestasi yang diwujudkan dalam untuk pengujian guna mengetahui penguasaan testi (*mastery testing*) umumnya memiliki variabilitas yang kecil, bahkan variabilitasnya dapat sebesar 0.

c. **Bukti yang berkaitan dengan kriteria** (***criterion-related validity***)

Dalam hal ini untuk melihat kesahihan suatu alat ukur justru dibandingkan dengan alat ukur yang lain yang standar atau yang baku sebagai patokan (kriteria). Alat ukur yang kita buat dinyatakan memiliki kesahihan yang tinggi jika hasil pengukurannya menunjukkan korelasi yang sangat positif dengan hasil pengukuran yang menggunakan alat ukur baku/standar. Kesahihan yang demikian disebut kesahihan *concurrent validity*. Pemenuhan validitas berdasarkan bukti empiris yang dikaitkan dengan kriteria dapat dilakukan pula dengan mengorelasikan hasil tes yang telah dikembangkan dengan tes lain yang diterapkan pada waktu mendatang yang berdekatan waktunya. Pemenuhan validitas ini disebut pemenuhan prediktif (*predictive validity*). Kedua hal ini relatif sulit sekali dipenuhi mengingat tidak tersedianya tes standar maupun tes lain yang dapat digunakan untuk menunjukkan daya prediktif dari tes yang telah dikembangkan.

d. Kesahihan yang lain selain yang dilihat dari tiga aspek

1) Kesahihan muka (*face validity*)

Kesahihan ini dibedakan menjadi dua. Pertama adalah kesahihan muka yang hanya dilihat berdasarkan tampak luarnya. Misalnya, pada suatu item mata pelajaran Biologi tentang klasifikasi mikroorganisme, item tersebut dinyatakan memenuhi kesahihan muka jika yang ditanyakan adalah hal-hal yang menyangkut klasifikasi organisme. Kedua, kesahihan ditinjau berdasarkan pandangan orang yang lebih ahli, dapat berarti ahli dalam

ilmunya ataupun ahli dari segi evaluasi atau dari segi bahasa. Oleh karena itu, minta pendapat orang yang lebih ahli akan meningkatkan kesahihan item yang telah disusun.

2) Kesahihan antarbudaya (*cultural validity*)

Kesahihan alat ukur juga tercermin dari keseragaman daya tangkap subjek terhadap alat ukur yang bersangkutan. Suatu butir tes model kasus harus menyajikan kasus yang dipahami secara merata oleh para subjek yang diukur. Seorang yang memahami Biologi akan lebih terpandu mengerjakan soal bahasa Inggris dengan kasus Biologi daripada dengan kasus Sastra misalnya. Jenis-jenis tanaman tumpangsari sudah tidak perlu dihafal oleh anak-anak yang hidup di pedesaan, namun untuk anak kota menjadi pekerjaan rumah tersendiri.

Alat-alat ukur yang diadopsi dari luar negeri harus dikaji benar apakah pertanyaan-pertanyaan yang ada di dalamnya bebas dari faktor budaya. Tidak otomatis bahwa alat ukur yang sudah teruji kesahihannya di luar negeri dipertanyakan kembali jika digunakan di dalam negeri.

d. Hal-Hal yang Diperhatikan Untuk Meningkatkan Validitas

Kesahihan alat ukur juga dapat dipenuhi dengan membuat alat ukur sebaik mungkin antara lain dengan cara berikut ini.

1) Petunjuk mengerjakan harus jelas. Jika alat ukurnya berupa perangkat tes, perintahnya harus jelas agar testi/peserta ujian mengerti apa yang harus dikerjakan. Demikian pula, jika berupa instrumennya berupa angket, usahakan agar responden mengerti cara pengisiannya.
2) Struktur kalimat harus benar, lugas, komunikatif, dan kata-kata yang digunakan sesuai dengan jenjang kemampuan testi/peserta ujian.
3) Konstruksi alat ukur harus memenuhi persyaratan sesuai dengan jenis alat ukurnya. Misalnya alat ukur berupa tes objektif/pilihan ganda harus memenuhi kaidah pembuatan item jenis tersebut. Maksudnya, persyaratan stem, kunci dan pengecoh harus dipenuhi.
4) Pernyataan-pernyataannya jangan mengandung arti ganda karena testi yang lebih cerdas cenderung tidak akan menjawab/mengerjakannya.
5) Cukup waktu untuk mengerjakan.
6) Alat ukur tidak terlalu panjang atau terlalu pendek.
7) Khusus alat ukur berupa alat tes juga memperhatikan hal berikut ini.
   (a) Tingkat kesukaran item hendaknya disesuaikan dengan kemampuan subjek uji.
   (b) Setiap item tes seleksi harus dapat membedakan mana kelompok atas dan mana kelompok bawah. Item tes seleksi yang baik adalah tidak terlalu mudah dan tidak terlalu sukar (indeks kesukaran item idealnya sebesar 0,5) dan memiliki daya pembeda yang tinggi (indeks daya pembeda idealnya 1,0) sehingga mampu membedakan mana yang pandai dan yang tidak.
   (c) Setiap item tes prestasi benar-benar untuk mengukur keberhasilan belajar, jika tesnya untuk mengukur keberhasilan belajar.

(d) Item tes prestasi hasil belajar yang baik adalah yang berhasil dikerjakan oleh peserta didik yang sungguh-sungguh berhasil dalam belajar dan yang pasti gagal dikerjakan oleh peserta didik yang belum terlibat dalam proses pembelajaran. Hal tersebut akan ditunjukkan oleh besarnya indeks sensitivitas item sebesar 1,0.
(e) Item diurutkan dari yang termudah ke yang paling sukar.
(f) Jawaban jangan sampai ditemukenali oleh subjek uji, misalnya karena urutan jawabannya terpola (tidak acak).
(g) Jangan sampai terjadi penyimpangan saat dilaksanakan tes dan jangan sampai salah ketika melakukan penskoran hasil pengukuran.

3. Reliabilitas

Pengertian reliabilitas berkaitan dengan konsistensi. Suatu alat ukur dinyatakan *reliable*/andal jika memberikan hasil yang sama pada berkali-kali pengulangan pengukuran. Reliabilitas berlaku pada tingkat suatu perangkat tes. Jadi tidak berlaku untuk masing-masing item tes penyusun suatu perangkat alat ukur.

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

Tabel 2. *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 tersebut, tes untuk mengukur keberhasilan belajar menggunakan tes yang skornya diinterpretasikan dengan mengacu kepada kriteria (*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 pembeda yang negatif (sebagai cerminan tidak ada testi/peserta ujian yang cerdas menjawab salah). Oleh karena itu, estimasi *error* didasarkan pada tingginya indeks konsistensi (indeks yang tinggi menunjukkan semua testi/peserta ujian yang sudah belajar pasti dapat mengerjakan dengan benar, sementara semua testi/peserta ujian yang belum belajar pasti tidak dapat mengerjakan dengan benar). Sebaliknya, tes untuk tujuan seleksi adalah tes yang dapat memisahkan kelompok yang lolos seleksi dan yang tidak lolos seleksi.

Oleh karena itu, interpretasinya dalam dengan mengacu kepada distribusi normal (*norm-reference*) sehingga item-itemnya memiliki indeks kesulitan yang harus moderat/sedang. Artinya, item-itemnya tidak terlalu sukar dan tidak terlalu mudah (sebagai cerminan bahwa kelompok atas yang pasti dapat mengerjakan) dan indeks daya pembeda harus tinggi (sebagai cerminan dapat membedakan kelompok atas dan bawah dengan meyakinkan). Oleh karena itu, estimasi *error* didasarkan pada tingkat tingginya keandalan/reliabilitas tes (indeks yang tes tinggi mencerminkan bahwa semakin cerdas testi/peserta ujian di dalam kelompoknya semakin tinggi pula skor yang diperolehnya).

Hasil analisis item untuk tujuan seleksi yang diterapkan pada item untuk tujuan pengukuran prestasi akan menunjukkan hasil yang tidak sesuai harapan. Berikut disajikan ilustrasinya.

Tabel 3. Analisis Item dengan mencari Daya Beda dan Indeks Kesukaran Item untuk 15 Item

Peserta Ujian	Nomor item															Skor	Nilai (0 -100)	Capaian KKM 75
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15			
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15	100	Berhasil
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15	100	Berhasil
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15	100	Berhasil
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15	100	Berhasil
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15	100	Berhasil
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15	100	Berhasil
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15	100	Berhasil
8	1	1	1	0	1	1	0	0	1	1	1	1	1	1	1	12	80	Berhasil
9	1	0	0	1	1	1	1	1	0	0	1	1	1	1	1	11	73	Gagal
10	1	0	0	1	0	1	1	1	1	1	0	1	1	1	1	11	73	Gagal
11	0	0	1	0	1	1	1	1	1	1	0	1	1	1	1	11	73	Gagal
12	0	1	1	1	0	1	0	1	0	1	1	0	1	0	1	9	60	Gagal
13	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	7	47	Gagal
14	1	1	1	0	1	1	0	0	0	0	1	0	0	0	0	6	40	Gagal
15	0	0	0	0	0	1	1	1	0	0	1	1	1	0	0	6	40	Gagal
16	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	3	20	Gagal
17	1	1	0	0	0	0	1	1	1	0	0	1	1	1	0	8	53	Gagal
18	1	0	1	0	0	0	0	1	1	1	0	0	1	1	1	8	53	Gagal
19	0	1	1	0	0	0	1	0	0	1	0	1	0	0	1	6	40	Gagal
20	0	0	0	0	1	0	1	0	1	0	0	1	0	1	0	5	33	Gagal
21	0	0	1	1	1	0	0	0	1	0	0	0	0	1	0	5	33	Gagal
22	0	0	1	0	1	0	1	0	0	0	0	1	1	0	0	5	33	Gagal
23	0	0	0	1	1	0	0	0	0	1	0	0	1	0	1	5	33	Gagal
24	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	3	20	Gagal
25	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	2	13	Gagal
26	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	3	20	Gagal
27	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	3	20	Gagal
28	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	2	13	Gagal
29	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	2	13	Gagal
30	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	7	Gagal
Kel atas	12	11	12	11	12	15	11	12	10	11	13	12	13	11	12			
Kel bawah	3	3	5	4	6	0	5	4	4	3	6	6	5	4	3			
Daya beda	0.6	0.5	0.5	0.5	0.4	1.0	0.4	0.5	0.4	0.5	0.5	0.4	0.5	0.5	0.6			
Tingkat kesukaran	0.5	0.5	0.6	0.5	0.6	0.5	0.5	0.5	0.5	0.5	0.6	0.6	0.6	0.5	0.5			

Jika dilihat daya beda minimal 0,4 berarti seluruh item memenuhi syarat tes standar. Tingkat kesukaran berkisar antara 0,5 sampai 0,6 juga hampir ideal. Kenyataan hanya 8 dari 30 peserta didik yang mencapai KKM. Belum lagi jika memperhatikan syarat item NRT di mana agar upaya muncul kelompok bawah, masih ada syarat bahwa minimal setiap pilihan pengecoh pada setiap soal pilihan ganda harus ada 5% peserta yang terkecoh. Jika ada 4 pengecoh/distraktor dan satu kunci, dan yang terkecoh peserta yang sama, dalam keadaan ekstrim akan ada 20% peserta didik yang memperoleh skor 0 dan ada 80% peserta didik yang memperoleh skor 100. Jika dikombinasi dengan syarat daya beda dan indeks kesukaran, maka tidak mungkin ada testi yang memperoleh skor mutlak karena indeks tingkat kesukaran minimal 0,3. Dengan kata lain, harus ada peserta didik sebanyak 30% yang masuk dalam kategori gagal.

Penggunaan Program Quest: Skala Dikotomus, Skala Politomus, Dan Kombinasinya

Istilah validitas dan reliabilitas adalah istilah teori tes klasik. Penyelidikan validitas berdasarkan kriteria (harus ada instrumen standar) maupun prediktif (harus ada instrumen lain yang mengukur hal yang sama) sukar dilakukan akibat tidak tersedianya instrumen pembanding. Demikian pula penyelidikan vaaliditas item yang harus dilakukan dengan cara mengorelasikan skor setiap item dengan skor instrumen lain yang standar.

Dalam melakukan analisis item, dilakukan dengan pendekatan teori tes klasik (*Classical Test Theory* atau CTT) atau teori tes modern yang dikenal dengan teori respons item (*Item Respons Theory* atau IRT). Pemilihan item tes dalam prosedur pengembangan tes menggunakan CTT didasarkan pada: (a) nilai/indeks 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. Tujuannya adalah untuk 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. Tes dengan format-format paralel untuk pencarian bukti empiris pada umumnya diciptakan untuk memperoleh distribusi-distribusi skor tes yang identik. Kesamaan nilai rata-rata, varians, dan *error* skor ditafsirkan sebagai bukti empiris tes dengan format paralel (Stark *et. al.*, 2001).

Langkah pertama dalam CTT adalah sebelum penulisan item mulai, pengembang tes harus mempunyai suatu pemahaman yang baik terhadap konstruk variabel (kemampuan) yang akan diukur. Menurut Stark *et. al.* dengan mengacu pendapat Nunnally *et. al.*, berdasarkan "*rule of thumb*" lazimnya disepakati bahwa banyaknya item tes yang harus dibuat sedikitnya dua kali dari banyaknya item tes yang akan dikemas sebagai perangkat tes yang siap digunakan.

Jika bentuk itemnya berupa pilihan ganda maka sejumlah besar item harus disusun untuk menghasilkan format-format ganda yang dikembangkan. Item-item tersebut harus diujikan terlebih dahulu menggunakan suatu sampel yang serupa dengan populasi yang dituju. Sampel ini, yang selanjutnya diacu sebagai sampel untuk pedoman saat dilakukan kalibrasi, haruslah besar. Tujuannya adalah agar cukup untuk menyediakan statistika item CTT yang stabil. Item-item dengan korelasi item-total tinggi harus ada di dalam tes karena item-item tersebut meningkatkan konsistensi skala internal (reliabilitas). Hal seperti itu juga akan mereduksi/memperkecil kesalahan baku (*standard error*) pengukuran.

Kesukaran item (dinyatakan sebagai nilai p) juga harus dipertimbangkan untuk membuat suatu tes dengan distribusi skor total yang diinginkan (Stark *et. al.*, 2001). Selain itu, untuk memperoleh distribusi skor skala yang diinginkan maka harus dilakukan penggantian item. Agar skalanya meningkat maka item dengan nilai p yang rendah harus diganti dengan nilai p yang tinggi. Untuk memperkecil dampak penggantian item terhadap reliabilitas maka item-item yang memiliki korelasi item-total yang rendah diganti dengan item-item yang memiliki korelasi item-total yang tinggi. Di dalam praktik, beberapa penyeimbangan konten/isi juga diperlukan. Oleh karena itu, setelah dilakukan penggantian kemudian item-item tersebut dianalisis lagi.

Ada keterbatasan penggunaan pendekatan CTT untuk analisis item (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 karena tidak ada informasi tentang kemampuan testi (*ability*). Akibatnya ketepatan pengukuran pada level-level skor yang tertentu tidak diketahui sehingga dikembangkan analisis item menurut teori modern.

Suatu instrumen, baik tes maupun non tes, dinyatakan layak sebagai suatu alat ukur menurut *item response theory* (IRT) jika memenuhi persyaratan unidimensi dan setiap item penyusunnya cocok/sesuai (*fit)* dengan model pengukuran. Dengan syarat tersebut setara dengan pemenuhan validitas dalam teori tes klasik. Dalam hal ini, dikenal ada model 1-PL (satu parameter logistik yang didasarkan pada tingkat kesularan/*difficulty*), model 2-PL (yang didasarkan pada *difficulty* dan daya pembeda/*discriminant*), dan model 3-PL (yang mendasarkan pada *difficulty, discriminanty,* dan menjawab dengan asal-asalan/*guessing).*

Langkah kegiatan mengonstruksi tes menggunakan pendekatan IRT, seperti halnya pada penggunaan pendekatan CTT. Pertama membuat dua sampai tiga kali banyaknya item dalam perangkat tes yang final akan digunakan. Skor mentah akan dikalibrasi menurut IRT menjadi skor skala *logit*.

Jika menggunakan program tertentu dengan pendekatan berdasarkan teori kemungkinan yang maksimum (*maximum likelihood*) seperti program QUEST, maka 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 (Stark *et. al.* (2001). Demikian pula item yang mempunyai skor prefect, di mana untuk tes pilihan ganda skor *prefect* adalah 1 untuk setiap testi atau *person/case*.

Program analisis atau disebut program kalibrasi menggunakan IRT adalah mendasarkan pada distribusi logistik, yakni distribusi yang menyerupai distribusi normal dengan nilai logistik D sebesar 1,7. Analisis item menggunakan IRT ada yang melakukan kalibrasi berdasarkan satu parameter yakni hanya didasarkan pada tingkat kesukaran (diberi simbol β atau b) sehingga disebut model satu parameter 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 kesukaran (b) sehingga disebut model 2-PL. Ada pula yang mendasarkan pada tiga parameter, yakni daya pembeda, tingkat kesukaran, dan *guessing* (diberi simbol c), sehingga disebut Model 3-PL.

Berdasarkan skala yang digunakan ada yang hanya dibuat skala dikotomus, yakni hanya dibedakan menjadi dua kategori yakni kategori 1 untuk skor 0 dan kategori 2 untuk skor 1, seperti skor tes pilihan yang pada akhirnya hanya ditetapkan benar atau salah. Ada pula yang dapat dibuat dalam bentuk skala politomus, yakni hanya dibedakan lebih dari dua kategori. Misalnya, testi diminta mengemukakan dua faktor yang menjadi penyebab terjadinya suatu peristiwa. Dengan demikian, jawaban testi ada tiga kategori, yakni kategori

1 untuk skor 0 (testi tidak menjawab), kategori 2 untuk skor 1 (testi hanya mengemukakan dengan benar satu faktor penyebab), dan kategori 3 untuk skor 2 (testi dapat menyebutkan dengan benar dua faktor penyebab yang dimaksudkan).

BAHAN RUUJUKAN UNTUK PENDALAMAN LEBIH LANJUT

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.

Atkinson, R.L., Atkinson, R.C., & Hilgard, E.R. (1983). *Introduction to psychology.* Terjemahan Nurdjannah Taufik dan Rukmini Barhana. 1997. Jakarta: Erlangga.

Badan Standar Nasional Pendidikan. (2007). *Panduan penilaian kelompok mata pelajaran IPTEK*. Jakarta: Badan Standar Nasional Pendidikan, Depdiknas.

Bambang Subali. (2010). *Pengukuran kreativitas keterampilan proses sains dalam mata pelajaran biologi sma di provinsi DIY dan Jawa Tengah beserta deteksi biasnya dalam konteks assessment for learning*. Laporan penelitian. Tidak dipublikasikan.

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.

Brown, D.H. (2004). *Language assessment: Principle and classroom practice*. San Fransisco: Pearson Education, Inc.

Carin, A.A. dan Sund, R.B. (1989). *Teaching science through discovery*. Columbus: Merrill Publishing Company

Dettmer, P. (2006). New Blooms in established fields: Four domains of learning and doing. Roeper Review, *ProQuest Education Journal*. Vol 28. No. 2. 70-78

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

Frisbie, D.A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice*. Fall 2005. Vol. 24. No.3. pp.21 28.

Gronlund, N.E. (1998). *Assessment of student achievement*, 9-th ed. Boston: Allyn and Bacon.

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.

Hopkins, K.D & Stanley, J.C. 1981. *Educational and psychological measurement and evaluastion*. NewYersey: prentice Hall, Inc.

Johnson. E.B. (2002). *Contextual teaching and learning: What it is and why it's here to stay*. Thousand Oaks, California: Corwin Press, Inc.

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.

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

Kunnan, A.J. 1998. *Validation in language assessment*. New Jersey: Lawrence Elbaum Associates, Publisher.

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

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

Subcoviak, M.J. (1988). A Practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement Spring 1988. Vol.25. No. 1. pp. 47-.55.*

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.

Wright, B.D. & Masters, G.N. (1982). Rating scae analsis. Chicago: Mesa Press.