

PENULISAN BUTIR SOAL*)

Oleh :

Badrun Kartowagiran)**

**UNIVERSITAS NEGERI YOGYAKARTA
2012**



*) Makalah disampaikan pada Pelatihan penulisan dan analisis butir soal bagi Sumber daya PNS Dik-Rekinpeg, di Hotel Kawanua Aerotel, Jakarta pada tanggal 10 Oktober 2012

***) Dosen Universitas Negeri Yogyakarta

PENULISAN DAN ANALISIS BUTIR SOAL

Oleh: Badrun Kartowagiran

PENDAHULUAN

Setiap kegiatan pasti memiliki tujuan, demikian halnya dengan kegiatan pembelajaran juga memiliki kegiatan. Tujuan pembelajaran adalah untuk meningkatkan kualitas perilaku peserta didik, termasuk prestasi belajarnya. Untuk mengetahui peningkatan ini, guru melakukan penilaian. Penilaian itu sendiri didefinisikan sebagai proses pengumpulan dan pengolahan informasi untuk menentukan pencapaian hasil belajar peserta didik.

Agar informasi yang diperoleh tepat maka instrumen yang digunakan harus dipersiapkan dengan baik. Menurut Tim Puspisjan (1997/1998), langkah-langkah pengembangan suatu tes prestasi belajar adalah : (1) penentuan tujuan tes, (2) penyusunan kisi-kisi, (3) penulisan soal, (4) penelaahan soal (review dan revisi soal), (5) uji coba soal, termasuk analisis dan perbaikan, dan (6) perakitan soal menjadi perangkat tes.

PENULISAN BUTIR SOAL

Pada pelatihan ini hanya difokuskan pada penyusunan dan analisis butir yang digunakan untuk mengungkap aspek kognitif. Untuk itu, sebelumnya akan disampaikan peringkat kognitif menurut Bloom. Menurut Moore, B dan Stanley T (2010), taksonomi Bloom yang mencakup: mengingat, memahami, menerapkan, menganalisis, mengevaluasi, dan melakukan kreasi merupakan urutan, dari yang paling rendah (peringkat 1) ke yang paling tinggi (peringkat 6). Selanjutnya, Moore, B dan Stanley T (2010), menambahkan bahwa urutan nomor 1 – 3 dikategorikan *the lower level of thinking* dan 4 -6 *the higher level of thinking (HOT)*. Hal ini senada dengan pendapat Thomas, A. dan Thorne, G. (2007) yang mengatakan *HOT is thinking on a higher level than memorizing facts or telling something back to someone exactly the way the it was told to you. When a person memorizes and gives back the information without having to think about it, we call it rote memory. That's because it's much like a robot; it does what it's programmed to do, but it doesn't think for itself.* Kata kerja yang dapat digunakan pada peringkat kognitif Bloom dapat dilihat pada Tabel 1.

Tabel 1. Kata Kerja Dalam Peringkat Kognitif Bloom

C1	C2	C3
Menyebutkan	Membedakan	Menggunakan
Menghafal	Membandingkan	Menerapkan
Mengidentifikasi	Menjelaskan	Membentuk
Menuliskan	Mengilustrasikan	Mengatur
Menunjukkan	Menduga	Menghitung
	Menguraikan	Menentukan
	Menghubungkan	Menyelesaikan
	Menterjemahkan	Melakukan eksperimen
	Meringkas	Mendemonstrasikan
	Menggolongkan	

C4	C5	C6
Menganalisis	Mengkritik	Mengkreasi
Membandingkan konsep	Menentukan	Merancang
Memprediksikan	Memberi keputusan	Meyintesis
Mengorganisasikan informasi	Merekomendasikan	Membuat proposal
Mengupas	Menyetujui pendapat	Membangun
Memeriksa hubungan	Mendukung	Mengkompilasi
Menyimpulkan	Memilih	Membuat estimasi
	Membenarkan	Menciptakan
	Menyangkal	Memodifikasi
	Memprioritaskan	Mengembangkan
	Menilai	
	Mengevaluasi	

1. Langkah-langkah Penulisan Butir Soal

Seperti yang dijelaskan di atas bahwa langkah-langkah pengembangan suatu tes prestasi belajar adalah : (1) penentuan tujuan tes, (2) penyusunan kisi-kisi, (3) penulisan soal, (4) penelaahan soal (review dan revisi soal), (5) uji coba soal, termasuk analisis dan perbaikan, dan (6) perakitan soal menjadi perangkat tes.

a. Penentuan tujuan/penyusunan *blueprint*

Dalam melakukan pengetesan pasti ada tujuan yang ingin dicapai. Tujuan ini dapat berupa tujuan khusus, misal untuk mengetahui penguasaan materi, tes diagnostik, atau tes seleksi; dan tujuan umum, misal untuk mengetahui pengetahuan umum dari sekelompok responden atau sekelompok orang. Dalam kesempatan ini, tujuan pemberian tes adalah untuk mengetahui penguasaan peserta didik pada kompetensi/sub kompetensi tertentu setelah diajarkan. Penguasaan ini dapat diartikan, sejauh mana peserta didik memahami atau mungkin menganalisis materi tertentu yang telah dibahas di ruang kelas. Dengan kata lain, pada tingkat kognitif mana mereka menguasai materi yang telah diberikan, ditugaskan, atau dibahas, yang biasanya direncanakan dalam bentuk *blue print*. Tujuan tes harus jelas agar arah dan ruang lingkup pengembangan tes selanjutnya juga jelas.

b. Penyusunan Kisi-kisi

Kisi-kisi adalah panduan atau acuan dalam menyiapkan bahan ajar, menyelenggarakan pembelajaran, dan mengembangkan butir-butir soal uji. Kisi-kisi soal tes yang merupakan bagian dari silabus ini biasanya berisi standar kompetensi, kompetensi dasar, materi pokok, kegiatan pembelajaran, penilaian, waktu, dan sumber belajar. Hal yang harus diperhatikan dalam menyusun kisi-kisi adalah indikator jabaran dari kompetensi dasar (KD), kompetensi dasar jabaran dari standar kompetensi (SK), standar kompetensi jabaran dari standar kompetensi lulusan mata pelajaran (SKL-MP), dan standar kompetensi lulusan mata pelajaran jabaran dari standar kompetensi lulusan satuan pendidikan (SKL-P), dan standar kompetensi lulusan satuan pendidikan jabaran dari Tujuan Pendidikan Nasional.

Kompetensi lulusan dijabarkan ke dalam subkompetensi, selanjutnya subkompetensi dijabarkan menjadi indikator esensial dan deskriptor. Sama halnya pada kompetensi dan subkompetensi, kata utama dalam indikator esensial dan deskriptor juga kata kerja, hanya saja

skopanya sama atau lebih sempit dan peringkat kognitifnya sama atau lebih rendah. Contoh format Kisi-kisi dapat dilihat pada Lampiran 1.

c. Penulisan butir-butir soal/tes

Penulisan butir-butir soal merupakan langkah penting dalam upaya pengembangan alat ukur kemampuan atau tes yang baik. Penulisan soal adalah penjabaran indikator jenis dan tingkat perilaku yang hendak diukur menjadi pertanyaan-pertanyaan yang karakteristiknya sesuai dengan perinciannya dalam kisi-kisi. Butir soal merupakan jabaran atau dapat juga ujud dari indikator, Dengan demikian setiap pernyataan atau butir soal perlu dibuat sedemikian rupa sehingga jelas apa yang ditanyakan dan jelas pula jawaban yang diminta. Mutu setiap butir soal akan menentukan mutu soal tes secara keseluruhan. Butir-butir soal harus memiliki tingkat penalaran tinggi atau memiliki *Higher Order Thinking (HOT)*.

d. Telaah Soal atau Analisis Kualitatif Soal

Telaah soal atau analisis kualitatif soal adalah mengkaji secara teoritik soal tes yang telah tersusun. Telaah ini dilakukan dengan memperhatikan tiga aspek, yaitu aspek materi, aspek konstruksi, dan aspek bahasa. Tabel telaah butir dapat dilihat pada Lampiran 2.

e. Ujicoba Soal

Ujicoba soal pada dasarnya adalah upaya untuk mengetahui kualitas soal tes berdasarkan pada empirik atau respon dari peserta tes. Hal ini dapat terwujud manakala dilakukan analisis empirik atau analisis kuantitatif, baik menggunakan teori klasik maupun teori modern.

f. Analisis Empirik

Untuk mengetahui kualitas butir soal, maka hasil uji coba harus dianalisis secara empirik. Ada dua pendekatan yang digunakan untuk melakukan analisis empirik ini, yaitu: teori klasik dan teori respon. Masing-masing pendekatan ada kelebihan dan kekurangannya. Untuk

responden yang kecil (kurang dari 100) lebih cocok menggunakan teori klasik, sebaliknya untuk responden yang besar (lebih besar dari 200) lebih cocok menggunakan teori respon butir.

g. Perakitan Soal Tes

Agar skor tes yang diperoleh tepat dan dapat dipercaya maka soal tes harus valid dan reliabel. Butir-butir soal perlu dirakit menjadi alat ukur yang yang terpadu. Hal-hal yang dapat mempengaruhi validitas skor tes adalah urutan nomor soal, pengelompokan bentuk-bentuk soal, tata letak soal, dan sebagainya. Untuk itu, ada baiknya soal tes disajikan mulai dari butir mudah ke yang susah, pengelompokan rapi, tata letak bagus dan tidak terpotong-potong kalimatnya, dan kemasannya menarik.

2. Penulisan Butir Soal yang HOT

Sebelum penulisan butir-butir soal dimulai, terlebih dulu perlu dicermati peringkat kognitif atau Taksonomi Bloom yang mencakup mengingat, memahami, mengaplikasikan, menganalisis, menyintesis, dan mengevaluasi. Selanjutnya, taksonomi ini direvisi (Anderson, L.W., dan Krathwoh, D.R , 2001), yakni evaluasi berada pada tingkat 5, sedangkan sintesis ditiadakan diganti dengan kreasi yang ditempatkan pada peringkat 6. Contoh soal pada masing-masing tingkat kognitif Bloom dapat dilihat pada Tabel 2.

Tabel 2. Contoh butir soal pada masing-masing tingkat kognitif Bloom

BUTIR SOAL URAIAN	BUTIR SOAL PILIHAN GANDA
Mengingat	
Sebutkan Ibu Kota Republik Indonesia	1. Ibu Kota Republik Indonesia adalah... . A. Bandung B. Surabaya C. Jakarta D. Medan

Memahami	
Berilah contoh binatang kelas herbivora	Berikut ini termasuk binatang kelas herbivora, kecuali: a. sapi b. kambing c. harimau d. kerbau
Aplikasi	
Sebuah pensil diameter 1,5 Cm, panjang 10 Cm dicelupkan seluruhnya ke sebuah gelas yang penuh air. Berapa CC air yang tumpah? Atau: Hitunglah pengurangan berikut. $\begin{array}{r} 643 \\ \underline{278} \\ \hline \end{array}$	Sebuah pensil diameter 1,5 Cm, panjang 10 Cm dicelupkan seluruhnya ke sebuah gelas yang penuh air. Jumlah air yang tumpah sebanyak... . a. 15, 675 CC b. 150,674 CC c. 441,964 CC d. 492,705 CC
Analisis	
Bandingkan kelebihan dan kelemahan buku-buku biologi SMP yang ada di pasaran, terutama buku karangan Osama, Obama, Ogama, dan Oalahmak. Bagian mana dari ke empat buku itu yang memiliki kesamaan dan bagian mana yang memiliki perbedaan? Atau Mengapa buku Biologi SMP karangan Obama dapat digantikan buku karangan Ogama?	Untuk mata pelajaran Biologi SMP, dari berbagai buku yang ada, buku yang dapat menggantikan buku karangan Obama adalah buku karangan: a. Osama b. Otama c. Ogama d. Oalahmak
Evaluasi	
Dari sekian siswa Bapak/Ibu, siapakah yang paling pintar?	Dari sekian banyak siswa Bapak/Ibu, siswa yang paling pintar adalah: a. Badu b. Beda c. Budi d. Bada

Kreasi	
Orang yang cinta perdamaian merupakan warga negara yang yang baik. Kebanyakan orang terdidik lebih cinta perdamaian karena tertarik pada kemajuan. Kesimpulan dari alinea ini adalah...	<p>Orang yang cinta perdamaian merupakan warga negara yang yang baik. Kebanyakan orang terdidik lebih cinta perdamaian karena tertarik pada kemajuan. Kesimpulan dari alinea ini adalah... .</p> <ul style="list-style-type: none"> A. banyak orang yang cinta perdamaian tetapi tidak terdidik B. kebanyakan orang yang terdidik bukan warga negara yang baik C. warga negara yang baik pada umumnya orang yang terdidik* D. orang yang cinta perdamaian belum tentu orang yang terdidik E. warga negara yang baik hanyalah orang-orang yang terdidik

3. Syarat Soal Tes yang baik

Seperti instrumen lainnya, soal tes juga harus baik, yakni memiliki validitas dan reliabilitas. Adapun penjelasan validitas dan reliabilitas adalah sebagai berikut.

a. Validitas

Validitas suatu alat ukur adalah sejauhmana alat ukur itu mampu mengukur apa yang seharusnya diukur (Nunnally, 1978). Sementara itu, Linn dan Gronlund (1995) menjelaskan validitas mengacu pada kecukupan dan kelayakan interpretasi yang dibuat dari penilaian, berkenaan dengan penggunaan khusus. Sedangkan Azwar (1996) menjelaskan suatu tes dapat dikatakan mempunyai validitas yang tinggi apabila tes tersebut menjalankan fungsi ukurnya, atau memberikan hasil ukur yang tepat dan akurat sesuai dengan maksud dikenakannya tes tersebut. Sisi lain yang sangat penting dalam konsep validitas adalah kecermatan pengukuran, yakni kemampuan untuk mendeteksi perbedaan-perbedaan kecil sekalipun yang ada pada atribut yang diukurnya.

Dalam pengukuran terhadap atribut psikologis, validitas sangat sulit dicapai. Hal ini dapat difahami karena pengukuran terhadap variabel psikologis dan sosial mengandung

kesalahan yang lebih banyak daripada pengukuran variabel yang bersifat fisik. Oleh karena sulitnya menentukan validitas yang sebenarnya, maka yang dapat dilakukan adalah mengestimasi validitas instrumen dengan perhitungan tertentu.

Pengukuran psikologi itu mempunyai fungsi : (1) penegakan suatu hubungan statistik dengan variabel khusus, (2) representasi isi dari sesuatu, dan (3) pengukuran sifat-sifat psikologis. Oleh karenanya, validitas itu dapat dikelompokkan menjadi tiga tipe, yaitu: (1) validitas kriteria, (2) validitas isi, dan (3) validitas konstruk (Nunnally, 1978, Allen & Yen, 1979, Fernandes, 1984, Woolfolk & McCane, 1984, dan Lawrence, 1994).

Validitas berdasarkan kriteria dibedakan menjadi dua, yaitu validitas prediktif dan validitas konkuren. Fernandes (1984) mengatakan validitas berdasarkan kriteria dimaksudkan untuk menjawab pertanyaan: *"How well test performance predicts future performance (predictive validity) or estimate current performance on some valued measure other than the test itself (concurrent validity)?"*. Hal senada juga disampaikan oleh Lawrence (1994) yang mengatakan bahwa tes dikatakan memiliki validitas prediktif bila tes itu mampu memprediksikan kemampuan yang akan datang. Dalam analisis validitas prediktif, performansi yang hendak diprediksikan disebut dengan kriteria. Besar kecilnya harga estimasi validitas prediktif suatu instrumen digambarkan dengan koefisien korelasi antara prediktor dengan kriteria tersebut.

Validitas isi suatu instrumen adalah sejauhmana butir-butir dalam instrumen itu mewakili komponen-komponen dalam keseluruhan kawasan isi objek yang hendak diukur dan sejauh mana butir-butir itu mencerminkan ciri perilaku yang hendak diukur (Fernandes, 1984; Nunnally, 1978). Sementara itu Lawrence (1994) menjelaskan bahwa validitas isi itu representativitas pertanyaan terhadap kemampuan khusus yang harus diukur.

Validitas konstruk adalah validitas yang menunjukkan sejauhmana instrumen mengungkap suatu trait atau konstruk teoretis yang hendak diukurnya (Fernandes, 1984; Nunnally, 1978). Prosedur validasi konstruk diawali dari suatu identifikasi dan batasan mengenai variabel yang hendak diukur dan dinyatakan dalam bentuk konstruk logis berdasarkan teori mengenai variabel tersebut. Dari teori ini ditarik suatu konstruksi praktis mengenai hasil pengukuran pada kondisi tertentu, dan konstruksi inilah yang akan dibuktikan

secara empiris. Apabila hasilnya sesuai dengan harapan maka instrumen itu dianggap memiliki validitas konstruk yang baik.

Untuk tes hasil belajar, yang utama adalah validitas isi, yakni butir-butir soal yang ditanyakan kepada peserta didik sesuai dan mewakili kompetensi yang harus dicapai oleh peserta didik. Hal ini dapat dilihat dari sejauh mana butir-butir soal itu sesuai dengan indikator yang merupakan jbaran dari kompetensi dasar.

b. Reliabilitas

Reliabilitas dapat diartikan sebagai keajegan atau kestabilan hasil pengukuran. Alat ukur yang reliabel adalah alat ukur yang mampu membuahkan hasil pengukuran yang stabil (Lawrence, 1994). Artinya suatu alat ukur dikatakan memiliki reliabilitas tinggi manakala digunakan untuk mengukur hal yang sama pada waktu berbeda hasilnya sama atau mendekati sama.

Reliabilitas alat ukur yang juga menunjukkan derajat kesalahan pengukuran tidak dapat ditentukan dengan pasti, melainkan hanya dapat diestimasi. Menurut Nunnally (1978) ada tiga cara mengestimasi reliabilitas, yaitu: (1) konsistensi internal, (2) tes paralel, dan (3) belah dua. Dalam cara konsistensi internal tes dilakukan hanya sekali pada sekelompok subjek kemudian dilakukan analisis atau diestimasi besarnya reliabilitas. Secara umum rumus untuk mengestimasi reliabilitas ini dapat digunakan rumus Koefisien Alpha. Namun apabila pilihan jawaban butir-butir pertanyaan/ pernyataan yang ada dalam instrumen/tes itu dikotomi maka dapat digunakan persamaan KR 20.

Tipe tes lainnya yang sering digunakan untuk mengestimasi reliabilitas adalah tipe tes paralel. Dalam tipe ini, tes dilakukan dua kali pada subjek yang sama namun tesnya berbeda meskipun paralel. Seperti yang telah dijelaskan di muka jarak antara ke dua tes ini sekitar dua minggu. Hasil kedua tes ini dikorelasikan, apabila koefisien korelasi ini kecil berarti tes itu kurang reliabel.

Selain konsistensi internal dan tes bentuk paralel, ada cara lain untuk mengestimasi reliabilitas, yaitu belah dua. Cara ini hanya menuntut satu kali tes untuk subjek yang sama kemudian hasilnya dibelah dua. Idealnya pembelahan ini harus dilakukan secara random, namun adakalanya yang menggunakan cara skor dari butir-butir pertanyaan/pernyataan

bernomor ganjil dipisahkan dengan skor dari butir-butir pertanyaan/ Pernyataan yang ber-nomor genap. Skor dari kelompok ini kemudian dikorelasikan dan selanjutnya digunakan rumus Spearman - Brown.

Salah satu cara untuk meningkatkan besarnya koefisien reliabilitas adalah memperpanjang tes, asalkan butir-butir yang ditambahkan harus homogen atau mengukur hal yang sama. Apabila butir yang ditambahkan tidak homogen maka reliabilitas tes tidak meningkat tetapi sebaliknya, malah menurun.

a. Analisis Soal Tes

Untuk mencapai butir-butir soal yang valid dan reliabel maka butir soal perlu dianalisis, yakni analisis secara teoritik atau telaah butir dan analisis kuantitatif untuk melihat tingkat kesulitan butir, daya beda butir, dan keberfungsian distraktor. Penjelasan analisis butir, baik kualitatif maupun kuantitatif adalah sebagai berikut.

1) Analisis Kualitatif (Telaah Butir)

Telaah kualitatif atau analisis teoritik dilakukan sebelum butir-butir soal diuji-cobakan dan di analisis secara empirik. Aspek-aspek yang diperhatikan dalam telaah kualitatif adalah aspek materi, konstruksi, dan bahasa/budaya ditelaah berdasarkan kaidah-kaidah yang telah ditentukan. Menurut Tim Pusbangsisjian, (1997/ 1998) kaidah-kaidah yang harus diperhatikan dalam menelaah butir soal yang berbentuk objektif pilihan ganda dapat dilihat pada Tabel 3 berikut.

Tabel 3. Lembar Telaah Butir Soal Pilihan Ganda

a)	Aspek materi
	(1) Soal sesuai dengan indikator; (2) Distraktor berfungsi; (3) Hanya ada satu kunci jawaban yang paling tepat
b)	Aspek konstruksi
	(1) Pokok soal dirumuskan dengan singkat, jelas dan tegas; (2) Rumusan pokok soal dan pilihan jawaban merupakan pertanyaan yang diperlukan

	<p>(3) Pokok soal tidak memberi petunjuk ke kunci jawaban;</p> <p>(4) Pokok soal bebas dari pernyataan yang bersifat negatif ganda;</p> <p>(5) Gambar, grafik, tabel, diagram, wacana, dan sejenisnya yang terdapat pada soal jelas dan berfungsi;</p> <p>(6) Panjang pilihan jawaban relatif sama;</p> <p>(7) Pilihan jawaban tidak menggunakan pernyataan "Semua jawaban di atas salah" atau "Semua pilihan jawaban di atas benar" dan sejenisnya;</p> <p>(8) Pilihan jawaban yang berbentuk angka atau waktu harus disusun berdasarkan urutan besar kecilnya angka tersebut atau kronologis;</p> <p>(9) Butir-butir soal tidak bergantung pada jawaban soal sebelumnya;</p>
c)	Aspek bahasa/budaya
	<p>(1) Menggunakan bahasa yang sesuai dengan kaidah bahasa Indonesia;</p> <p>(2) Menggunakan bahasa yang komunikatif;</p> <p>(3) Tidak menggunakan bahasa yang berlaku setempat (bias budaya);</p> <p>(4) Pilihan jawaban tidak mengulang kata/kelompok kata yang sama.</p>

Dalam analisis soal tes secara teoritik yang dikaji adalah kesesuaian antara butir-butir soal dengan tujuan atau indikator dan apakah soal tes sudah memenuhi validitas isinya. Soal tes juga dicermati penggunaan bahasa, kejelasan dan kesingkatannya, juga dilihat kejelasan dan kefungsiian tabel dan atau gambar. Pilihan jawaban juga dicermati homogenitas dan kejelasannya.

Selain kaidah untuk telaah butir secara teoritik, pedoman penyekoran juga harus jelas agar objektivitas pemberian skor oleh guru dapat dipertanggung-jawabkan. Pedoman pemberian skor untuk setiap butir soal uraian harus disusun sesegera mungkin setelah kalimat-kalimat butir soal tersebut selesai dirumuskan. Pedoman pemberian skor tidak boleh disusun saat koreksi akan dimulai.

Ada perbedaan pedoman penyekoran antara soal bentuk pilihan ganda dan soal bentuk uraian. Hal ini dikarenakan adanya perbedaan karakteristik di antara ke duanya yang secara rinci dapat dilihat pada Tabel 4 berikut.

Tabel 4. Perbandingan Antara Soal Bentuk Pilihan Ganda dan Uraian

Karakteristik	Uraian	Pilihan Ganda
Penulisan soal	Relatif mudah	Relatif sukar
Jumlah pokok bahasan yang ditanyakan	Terbatas	Lebih banyak
Aspek yang diukur	Dapat lebih dari satu	Hanya satu
Persiapan siswa	Penekanannya pada kedalaman materi	Lebih menekankan pada keluasan materi
Jawaban siswa	Mengorganisasikan jawaban	Memilih jawaban
Kecenderungan menebak	Tidak ada	Ada
Penyekoran	Sukar, lama, kurang konsisten (reliabel) dan subjektif	Mudah, cepat, sangat konsisten dan objektif

Pemilihan bentuk soal mana yang akan dipakai harus memperhatikan karakteristik soal seperti yang telah diuraikan di atas, tujuan penilaian dan efisiensi. Untuk ujian yang jumlah pesertanya sangat banyak maka soal pilihan ganda lebih efisien, baik dilihat dari segi waktu maupun dari segi biaya yang dikeluarkan.

a. Analisis Kuantitatif

Analisis kuantitatif dilakukan dengan menggunakan dua pendekatan, yaitu pendekatan teori klasik dan pendekatan teori modern atau teori respon butir (Item Respon Theory =IRT). Dalam penelitian ini, hanya dijelaskan cara analisis butir kuantitatif dengan pendekatan teori klasik. Penjelasan analisis butir menurut teori klasik adalah sebagai berikut.

1) Pendekatan Teori Tes Klasik

Skor sebenarnya (*true score* = T) dan skor kesalahan (*error score* = E) adalah konstruk teoritik yang tidak dapat diamati. Hanya skor amatan (*observed score* = X) yang dapat diperoleh, dan skor amatan = skor sebenarnya + kesalahan ($X = T + E$). Jika kita berbicara skor sebenarnya, penting diingat bahwa skor sebenarnya yaitu skor rata-rata yang diperoleh dari pengulangan tes secara independen dengan menggunakan tes yang sama, adalah teoritis belaka. Skor ini tidak menunjukkan dengan lengkap karakteristik sebenarnya dari peserta tes kecuali kalau tes tersebut memiliki validitas sempurna, yaitu bahwa tes tersebut mengukur dengan tepat apa pokok isi yang diukur.

Menurut para ahli, ada beberapa kelemahan yang ada pada pendekatan teori klasik. Beberapa di antaranya adalah Hambleton, dkk (1991) dan Lord (1980). Mereka menjelaskan bahwa kelemahan-kelemahan tes teori klasik adalah: (1) statistik butir tes sangat tergantung pada karakteristik subjek yang dites; (2) taksiran kemampuan peserta tes sangat tergantung pada butir tes yang diujikan; (3) kesalahan baku penaksir skor berlaku untuk semua peserta tes, sehingga kesalahan baku pengukuran tiap peserta dan butir soal tidak ada; (4) informasi yang disajikan terbatas pada menjawab benar atau salah saja tidak memperhatikan pola jawaban peserta tes; dan (5) asumsi tes paralel susah dipenuhi.

Walaupun teoriklasik ini memiliki beberapa kelemahan namun masih banyak yang menggunakan karena tidak menuntut responden besar (lebih 100) dan mudah mengaplikasikannya (melakukan analisis butir dengan pendekatan klasik ini). Oleh karenanya, untuk pengukuran yang melibatkan responden kecil misal pada pengukuran melalui tes harian pada bidang pendidikan, atau pengukuran pada bidang psikologi pada umumnya masih menggunakan pendekatan teori tes klasik.

Analisis kuantitatif menurut pendekatan teori tes klasik menghasilkan karakteristik butir yang meliputi tingkat kesukaran (p), daya pembeda (d), dan efektivitas distraktor. Selain itu, dengan analisis kuantitatif pendekatan teori klasik juga dapat diketahui reliabilitas soal tes, dan kesalahan baku pengukuran. Untuk melihat tingkat kesukaran, daya pembeda, dan efektivitas

distraktor dilakukan analisis setiap butir tes, sedangkan reliabilitas dan kesalahan pengukuran baku dapat dilihat dengan cara menganalisis soal tes secara keseluruhan.

Kesesuaian karakteristik butir dengan jenis dan tujuan tes sangat menentukan kualitas butir tes. Pada analisis butir secara klasik, tingkat kesukaran (p) dapat diperoleh dengan beberapa cara, antara lain: (1) skala kesukaran linier; (2) skala bivariat; (3) indeks Davis; dan (4) proporsi menjawab benar. Cara yang paling mudah dan paling banyak digunakan adalah skala rata-rata atau proporsi menjawab benar atau *proportion correct* (p), yaitu jumlah peserta tes yang menjawab benar pada butir yang dianalisis dibandingkan dengan peserta tes seluruhnya.

Tingkat kesukaran (p) mengandung banyak kelemahan, antara lain tingkat kesukaran sebenarnya merupakan ukuran kemudahan butir karena semakin tinggi indeks p , semakin mudah butir tersebut. Sebaliknya semakin rendah p semakin sulit. Oleh karenanya ada beberapa ahli pengukuran yang menyebut tingkat kesukaran ini dengan tingkat kemudahan. Tingkat kesukaran merupakan salah satu parameter butir soal, yang disimbolkan (P_i), yakni rasio antara jawaban benar dan banyaknya penjawab butir soal. Formulasi tingkat kesukaran butir soal adalah:

$$P_i = \frac{n}{N}$$

P_i = Tingkat kesukaran butir soal ke i

i = nomor butir soal

n = banyaknya siswa yang menjawab butir soal dengan benar

N = banyaknya siswa yang menjawab butir soal

Besarnya tingkat kesukaran berkisar antara nol dan satu. Suatu butir kadang-kadang dikategorikan ke dalam ekstrim sukar yaitu apabila nilai p mendekati nol dan ekstrim mudah apabila nilai p mendekati satu. Menurut Fernandes (1984), butir soal yang menghasilkan rerata skor sekitar 50 % dari skor maksimum dapat dikatakan bahwa butir soal itu mempunyai tingkat kesukaran yang tepat. Sementara itu, Thomas dan Dawson (1972) menjelaskan bahwa butir soal yang memiliki tingkat kesukaran 0,25 - 0,75 sudah dikatakan baik.

Daya pembeda atau daya beda suatu butir tes berfungsi untuk menentukan dapat tidaknya suatu butir tes membedakan kelompok dalam aspek yang diukur sesuai dengan perbedaan yang ada pada kelompok itu. Tujuan dari penelaahan daya pembeda adalah untuk melihat kemampuan butir tes tertentu dalam membedakan antara pengambil tes yang berkemampuan tinggi dan pengambil tes yang berkemampuan rendah.

Ada beberapa cara yang digunakan untuk menghitung daya pembeda, yaitu: (1) indeks diskriminasi, (2) indeks korelasi, dan (3) indeks keselarasan. Pada penelitian ini hanya dibahas dua cara untuk menghitung daya pembeda dengan metode korelasi yaitu korelasi *point biserial* dan korelasi *biserial*. Korelasi *point biserial* maupun korelasi *biserial* adalah korelasi *product moment* yang diterapkan pada data, variabel-variabel yang dikorelasikan sifatnya masing-masing berbeda satu sama lain. Variabel butir tes bersifat dikotomi yaitu bernilai 1 untuk jawaban benar dan 0 jika jawaban salah. Di sisi lain, variabel skor total atau sub skor total bersifat kontinum yang diperoleh dari jumlah jawaban yang benar. Nilai koefisien korelasi *point biserial* selalu lebih jika dibandingkan dengan nilai koefisien korelasi *biserial*. Koefisien *point biserial* merupakan kombinasi hubungan antara butir tes, kriteria atau skor total, dan tingkat kesukaran. Korelasi *point biserial* cenderung lebih mengutamakan butir tes yang memiliki tingkat kesukaran rata-rata dan akan maksimum apabila tingkat kesukarannya $p = 0.5$ (Bahrul Hayat, 1996 dan Sumadi Suryabrata, 1987). Korelasi biserial merupakan korelasi antara butir tes dan kriteria, bebas dari pengaruh tingkat kesukaran butir tes. Menurut Crocker & Algina (1986) koefisien *point biserial* ditentukan dengan rumus:

$$\rho_{pbis} = \frac{\mu_+ - \mu_\tau}{\sigma_\tau} \sqrt{\frac{p}{q}}$$

ρ_{pbis} = Korelasi point biserial

μ_+ = Rerata-rata skor peserta tes yang menjawab benar butir soal

μ_τ = Rerata skor total

σ_τ = Simpangan baku skor total

p = Proporsi banyaknya peserta yang menjawab benar

q = 1- p

Terkait dengan rumus daya beda ini, Sumadi Suryabrata (2000) menuliskan dalam bentuk lain, yaitu:

$$\rho_{pbis} = \frac{Mb - Ms}{\sigma_{\tau}} \sqrt{\frac{pq}{1}}$$

ρ_{pbis} = Korelasi point biserial

Mb = Rerata skor peserta tes yang menjawab benar

Ms = Rerata skor peserta tes yang menjawab salah

σ_{τ} = Simpangan baku skor total

p = Proporsi banyaknya peserta yang menjawab benar

q = 1 - p

Sementara untuk menghitung indeks daya pembeda dengan korelasi *biserial* digunakan rumus:

$$\rho_{bis} = \frac{\mu_{+} - \mu_{\tau}}{\sigma_{\tau}} \left(\frac{p}{Y} \right)$$

ρ_{bis} = Korelasi biserial

μ_{+} = Rerata-rata skor peserta yang menjawab benar

μ_{τ} = Rerata skor total

σ_{τ} = Simpangan baku skor total

p = Proporsi banyaknya peserta yang menjawab benar

Y = Ordinat p dalam distribusi normal

Bentuk lain untuk menghitung daya beda dengan korelasi biserial juga ditampilkan oleh Sumadi Suryabrata (2000), yakni:

$$\rho_{bis} = \frac{Mb - Ms}{\sigma_{\tau}} \left(\frac{p}{Y} \right)$$

ρ_{bis} = Korelasi biserial

Mb = Rerata-rata skor peserta yang menjawab benar

Ms = Rerata skor total

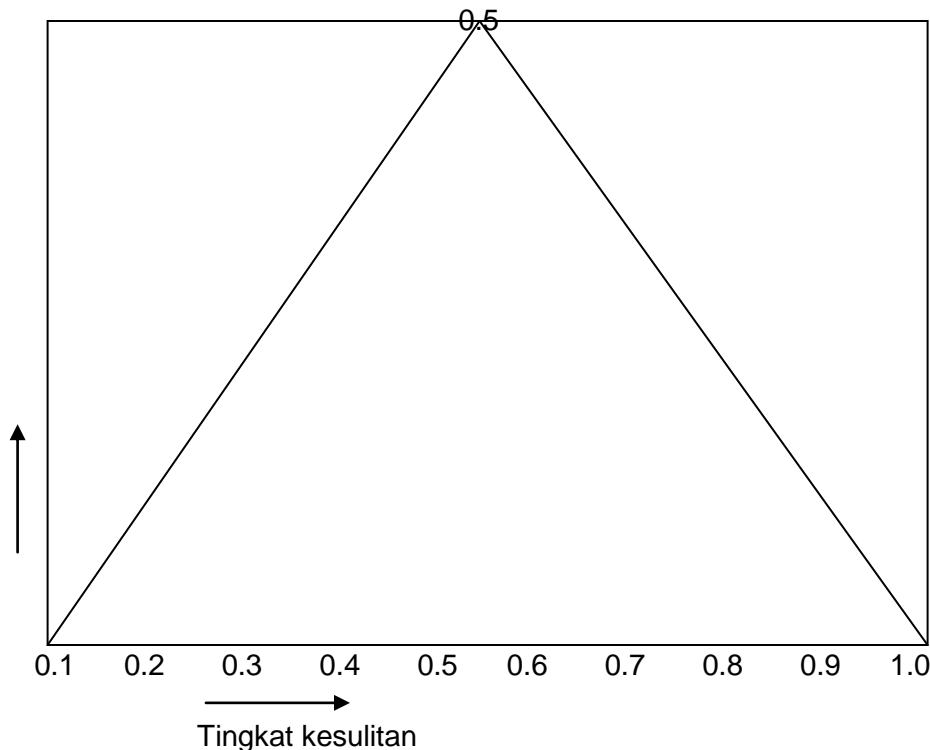
σ_{τ} = Simpangan baku skor total

p = Proporsi banyaknya peserta yang menjawab benar

Y = Ordinat p dalam distribusi normal

Indeks daya beda butir soal dapat digunakan sebagai bahan pertimbangan sebuah butir baik atau tidak baik. Butir soal yang baik adalah butir soal yang mempunyai indeks daya beda lebih dari 0,2. seperti yang dinyatakan Fernandes(1984). Sementara Ebel (1972) menjelaskan suatu butir soal dikatakan berkualitas apabila indeks diskriminasi atau daya pembedanya paling sedikit 0,41. Selanjutnya, Fernandes (1984) menggambarkan hubungan antara tingkat kesukaran dan indeks daya pembeda seperti Gambar 1.

Hal penting yang juga harus diperhatikan dalam menganalisis empirik butir soal adalah kemampuan distraktor atau alternatif jawaban yang disediakan menarik peserta tes untuk memilihnya. Jangan sampai tidak seorang peserta tes-pun memilih alternatif jawaban yang disediakan. Fernandes (1984) yang mengutip pendapat Brawn menjelaskan distraktor dikatakan baik apabila paling tidak dipilih oleh 2 % dari seluruh peserta. Sementara itu, Nitko (1996) mengatakan distraktor dikatakan berfungsi manakala paling tidak dipilih oleh seorang peserta tes dari kelompok rendah. Pemilih dari kelompok rendah harus lebih banyak daripada kelompok atas. Distraktor juga dapat dikatakan berfungsi manakala peserta tes (siswa) dari kelompok atas dapat membedakan antara distraktor dan kunci jawaban sehingga yang memilih kunci jawaban lebih banyak daripada yang memilih distraktor.



Gambar 1. Hubungan Antara Daya Pembeda dan Tingkat Kesulitan

Dalam menganalisis distribusi jawaban juga perlu memperhatikan kemungkinan salah kunci, yaitu manakala siswa dari kelompok atas yang memilih pengecoh lebih banyak daripada yang memilih kunci jawaban. Selain itu, juga perlu dideteksi ada tidaknya unsur tebakan dalam memilih alternatif jawaban. Hal ini dapat dilihat apabila jawaban peserta tes (siswa) merata, baik jawaban dari siswa kelompok atas maupun kelompok bawah.

Dengan demikian jelaslah bahwa untuk menilai kualitas butir tes tidak cukup hanya memperhatikan tingkat kesukaran dan daya pembeda butir tes yang bersangkutan. Penilaian kualitas butir tes juga harus melihat fungsi pilihan jawaban, terutama distraktor-distraktornya, yaitu harus tampak sebagai jawaban yang benar bagi subjek dari kelompok yang berkemampuan rendah. Sebaliknya harus tampak sebagai jawaban yang salah bagi subjek dari kelompok yang berkemampuan tinggi. Sekalipun suatu butir tes terlalu sukar atau terlalu mudah, namun apabila (1) daya pembeda butir tes, dan (2) distribusi jawaban, memenuhi kriteria, maka butir tes tersebut masih dapat diterima sebagai butir tes yang baik. Kriteria yang dimaksud adalah indeks daya pembeda butir tes $r_{bis} > 0,3$, dan indeks daya pembeda pilihan jawaban negatif kecuali kunci.

Hasil tes hendaknya juga membentuk distribusi normal. Hal ini dapat dicapai manakala butir-butir soal yang dipilih itu tepat, baik dilihat dari tingkat kesulitan maupun daya beda. Butir-butir soal yang tingkat kesulitannya tinggi cenderung menghasilkan skor yang memiliki distribusi juling positif atau hanya sebagian kecil peserta tes yang mendapat skor tinggi. Sebaliknya, bila butir-butir soal itu terlalu mudah maka skor yang diperoleh (hasil tes) cenderung membentuk juling negatif atau banyak sekali siswa yang mendapat skor tinggi.

4. Contoh Analisis Kuantitatif

Dalam contoh analisis kuantitatif ini akan diberikan contoh cara menghitung tingkat kesulitan, daya pembeda, dan reliabilitas. Perhatikan hasil uji coba tes seperti yang tertera pada Tabel 5. Apabila distribusi pilihan jawaban butir 5 seperti tampak pada Tabel 5, hitung:

- a. Tingkat kesulitan butir 5
- b. Daya beda butir 5
- c. Berapa reliabilitas soal tes yang hasil uji cobanya seperti tampak pada Tabel 5.

Tabel 5. Hasil uji coba soal tes Matematika yang berbentuk Pilihan ganda dengan 4 pilihan terhadap 30 orang siswa.

NO	RESP	BUTIR										
		1	2	3	4	5	6	7	8	9	10	
1	AA	0	0	1	1	1	0	0	1	1	1	6
2	AB	0	0	0	0	0	1	1	1	1	1	5
3	AC	1	1	1	1	1	1	1	1	1	0	9
4	AD	1	1	1	0	0	0	0	1	1	1	6
5	AE	1	0	0	1	1	0	1	0	1	0	5
6	AF	1	1	1	1	0	0	1	1	0	0	6
7	AG	1	0	1	1	1	1	0	0	1	1	7
8	AH	0	1	1	1	1	1	1	1	1	1	9
9	AI	1	0	0	0	0	1	1	1	0	0	4
10	AJ	0	1	1	1	1	1	1	1	0	0	7
11	AK	1	1	0	0	0	1	1	1	1	1	7
12	AL	1	0	0	0	1	1	1	0	0	1	5
13	AM	1	1	1	1	0	0	0	1	1	0	6
14	AN	1	1	1	1	1	1	1	1	1	1	10
15	AO	0	0	0	0	0	0	0	1	1	1	3
16	AP	0	0	0	0	1	1	1	1	0	0	4
17	AQ	1	1	1	1	1	1	1	1	1	1	10
18	AR	0	0	0	1	1	1	1	1	0	0	5
19	AS	1	0	0	0	0	0	1	1	1	1	5
20	AT	1	1	1	1	1	1	1	1	1	1	10
21	AU	1	0	0	0	0	1	1	1	1	1	6
22	AV	1	0	0	0	0	1	1	1	0	0	4
23	AW	1	1	1	1	0	0	1	1	1	1	8
24	AX	1	1	0	1	0	1	0	1	0	1	6
25	AY	0	1	1	0	0	1	1	1	1	1	7
26	AZ	0	0	0	0	0	0	0	1	1	1	3
27	BA	1	1	1	1	1	1	0	1	1	1	9
28	BB	0	1	1	1	1	1	1	0	0	1	7
29	BC	1	1	1	1	1	1	0	0	1	1	8
30	BD	1	0	0	1	1	1	1	1	1	1	8
KUNCI		A	A	B	C	B	A	C	B	D	D	
n		20	16	16	18	16	21	21	25	21	21	
p		0,67	0,53	0,53	0,60	0,53	0,70	0,70	0,83	0,70	0,70	
(1-p)		0,33	0,47	0,47	0,40	0,47	0,30	0,30	0,17	0,30	0,30	
p(1-p)		0,22	0,25	0,25	0,24	0,25	0,21	0,21	0,14	0,21	0,21	$\sum p(1-p)$ 2,19

Tabel 5 menunjukkan bahwa siswa yang mendapat skor 10, 9, 8, atau 4 masing-masing ada tiga orang, siswa yang mendapat skor 7 atau 5 masing-masing ada lima (5) orang, siswa yang mendapat skor 6 ada enam (6) orang, dan yang mendapat skor 3 ada dua (2) orang. Dari tiga orang yang mendapat skor 10, ternyata tiga-tiganya memilih pilihan B (atau betul) pada butir 5; demikian pula halnya siswa yang mendapat skor 9. Sementara itu, dari tiga orang yang mendapat skor 8, hanya ada dua orang yang betul pada butir 5, yakni yang memilih B, sedangkan yang satu (1) orang memilih C, boleh juga ditaruh pada pilihan A, atau D yang penting bukan B.

Dari lima (5) orang siswa yang mendapat skor 7, ada 3 orang yang pada butir 5 menjawab B (betul), sedangkan yang dua orang siswa salah, dia memilih A dan D. Demikian cara penjelasannya untuk siswa yang mendapat skor 6, 5, atau 4; yang penting kolom paling kanan dan jumlah siswa yang menjawab B pada butir 5 harus sama dengan yang ditunjukkan pada Tabel 5. Data ini selanjutnya disalin dan ditampilkan pada Tabel 6.

Tabel 6. Distribusi pilihan jawaban pada butir 5

Kriteria/skor total	PILIHAN JAWABAN				Distribusi Total
	A	B	C	D	
10		3			3
9		3			3
8		2	1		3
7	1	3		1	5
6	2	1	1	2	6
5		3	1	1	5
4	1	1	1		3
3	1	-		1	2
2					-
1					-

Tabel 6 menunjukkan bahwa tidak ada seorangpun yang mendapat skor 2 dan 1, selanjutnya dari data yang tertayang pada Tabel 6 dapat digunakan untuk menghitung tingkat kesulitan butir dan rerata skor tiap-tiap pilihan jawaban. Tabel 6 dapat dimanfaatkan untuk menjawab pertanyaan (a) yakni tingkat kesulitan butir yang selengkapnya dapat dilihat pada Tabel 7 berikut.

Jawab:

- a. Menghitung tingkat kesulitan butir 5

Tabel 7 menunjukkan bahwa tingkat kesulitan butir 5 adalah 16 (jumlah siswa yang menjawab butir 5) dibagi 30 (jumlah peserta) sama dengan 0,53.

Tabel 7. Distribusi pilihan jawaban pada butir 5 adalah sebagai berikut.

Kriteria/skor total	PILIHAN JAWABAN				Distribusi Total
	A	B	C	D	
10		3			3
9		3			3
8		2	1		3
7	1	3		1	5
6	2	1	1	2	6
5		3	1	1	5
4	1	1	1		3
3	1	-		1	2
2					-
1					-
N	5	16	4	5	
P	0,1667	0,53	0,133	0,1667	
Mean	5,2	7,4375	5,75	5,4	

b. Menghitung Daya beda butir 5

Perhatikan Tabel 7

Kolom paling kiri adalah skor total, dalam hal ini skor total maksimum = jumlah butir soal, yang selanjutnya dianggap = X. Kolom paling kanan adalah distribusi total yang besarnya dapat dilihat kolom paling kanan dari Tabel 6, yang selanjutnya dianggap = F. Dari sini selanjutnya dapat dihitung FX, FX^2 , dan seterusnya dapat dilihat pada Tabel 8.

Tabel 8. Tabel persiapan menghitung standar deviasi dan varian.

X	F	FX	X ²	FX ²
10	3	30	100	300
9	3	27	81	243
8	3	24	64	192
7	5	35	49	245
6	6	36	36	216
5	5	25	25	125
4	3	12	16	48
3	2	6	9	18
2				
1				
	N = 30	$\sum FX = 195$ M _x = 6,5		$\sum FX^2 = 1387$

$MB = (10.3 + 9.3 + 8.2 + 7.3 + 6.1 + 5.3 + 4.1)/16 = 7,4375 \approx 7,44$ $MA = (7.1 + 6.2 + 4.1 + 3.1)/5 = 5,20$ $MC = (8.1 + 6.1 + 5.1 + 4.1)/4 = 5,75$ $MD = (7.1 + 6.2 + 5.1 + 3.1)/5 = 5,40$ $Ms = (5,20 + 5,75 + 5,40)/3 = 5,45$	$\sigma_{\tau}^2 = \{(\sum FX^2)/(N) - (M^2)\}$ $\sigma_{\tau}^2 = \{(1387/30) - (6,5)^2\}$ $= (46,23 - 42,25)$ $\sigma_{\tau}^2 = 3,98$ jadi $\sigma_{\tau} = 1,99$
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

$$rpbis = \{(Mb - Ms) / \sigma_{\tau}\}(\sqrt{p \cdot q})$$

$$rpbis = \{(7,44 - 5,45) / 1,99\}(\sqrt{0,53 \cdot 0,47})$$

$$rpbis = 1 \times 0,499 = 0,499.$$

c. Menghitung reliabilitas butir

Karena hasil jawaban soal itu dikotomi (data pisah, misal 0 dan 1) maka reliabilitasnya dihitung dengan menggunakan rumus KR 20.

$$KR20 = \left(\frac{k}{k-1}\right)\left(\frac{S^2 - \sum pq}{S^2}\right) \quad \text{Dalam hal ini } S = \sigma_{\tau}$$

$$KR20 = \left(\frac{10}{9}\right)\left(\frac{3,98 - 2,19}{3,98}\right)$$

$$KR20 = 0,495$$

Harga ini lebih kecil daripada harga yang ditetapkan Feldt dan Brehmman (1989), sehingga dapat dikatakan bahwa soal tes yang digunakan ini kurang reliabel.

Setelah tingkat kesulitan dan daya beda tiap-tiap butir dihitung, reliabilitas juga sudah dihitung, langkah selanjutnya adalah mencocokkan hasil itu dengan rambu-rambu penerimaan butir seperti Tabel 9 berikut.

Tabel 9. Rambu –rambu penerimaan butir menurut Teori Klasik

1. Ebel (1972) yang mengatakan bahwa alat ukur yang memiliki koefisien reliabilitas 0,8 sudah baik Nunnally (1978) berpendapat bahwa soal uraian yang memiliki koefisien reliabilitas 0,6 – 0,7 dan untuk soal pilihan ganda yang memiliki koefisien reliabilitas 0,75 – 0,90 sudah dapat dikatakan baik. Feldt dan Brehmman (1989) mengatakan bahwa suatu instrumen yang memiliki koefisien reliabilitas $r_{xy} \geq 0,7$ sudah dikatakan reliabel.
2. Menurut Dawson (1972) butir soal yang memiliki tingkat kesukaran 0,25 – 0,75 dikatakan baik.
3. Menurut Ebel (1972) butir yang memiliki daya pembeda 0,41 atau lebih besar dikatakan baik, sedangkan Fernandes mengatakan butir yang memiliki daya pembeda lebih besar dari 0,2 dikatakan baik.
4. Fernandes (1984) yang menjelaskan distraktor dikatakan baik apabila paling tidak dipilih oleh 2 % dari seluruh peserta. Nitko (1996) mengatakan distraktor dikatakan berfungsi manakala paling tidak dipilih **oleh seorang peserta** tes dari kelompok rendah.

5. Analisis butir dengan ITEMAN

Untuk menghemat waktu, analisis butir biasanya dilakukan dengan menggunakan jasa program komputer. Untuk pendekatan teori klasik digunakan program ITEMAN.

Dalam kesempatan ini dipilih soal objektif bentuk pilihan ganda dengan 4 alternatif jawaban. Data dituliskan dalam bentuk alfabetik (A, B, C, D). Langkah pertama yang dilakukan adalah memasukkan data ke file dalam bentuk ASCII atau DOS Text yang dapat diketik dalam program Notepad.

Contoh:

030 o N 10
DCABBCADAACBACDACBDC
444444444444444444444444
YYYYYYYYYYYYYYYYYYYY
0042001 BADCCDABBDABBCDABDCD
0042002 BCADDACBCDABCADBADCC
dst
0042040 ABDCABDACCDAACBC

Keterangan :

Baris pertama

Kolom 1 – 3 : jumlah butir soal, maksimum 250 butir (dalam hal ini 020)
4 : kosong/spasi
5 : untuk jawaban kosong (omit) (dalam contoh o)
6 : kosong/spasi
7 : untuk butir soal yang belum dikerjakan (dalam contoh N)
8 : kosong/spasi
9 – 10 : jumlah identitas data siswa (maks 80) (dalam contoh 10)

Baris kedua : Kunci jawaban

Baris ketiga : Jumlah pilihan jawaban

Baris ke empat : kode Y = Yes, butir soal dianalisis, N = butir soal tidak dianalisis

Baris ke lima dst : jawaban responden. Bila data sudah masuk semua kemudian disimpan dalam sub direktori tersendiri, yaitu satu sub direktori dengan program ITEMAN. Dalam kesempatan ini data ditulis dengan "notepad" kemudian disimpan dengan menggunakan tipe *All files* dalam folder, yaitu jadi satu folder dengan program ITEMAN. Misalkan data yang baru selesai dimasukkan itu diberi nama CONTOH-1.DAT

Menjalankan program

Apabila program ITEMAN dan data sudah dalam 1 folder maka cara menjalankannya adalah sebagai berikut.

1. Klik dua kali program ITEMAN, maka di layar akan tampak

```
-----  
ITEM & TEST ANALYSIS PROGRAM  
>>>***** <<<<  
Enter the name of the input file:  
-----
```

2. Ketik nama file yang akan dianalisis, misalnya CONTOH-1.DAT Enter

```
-----  
Enter the name of the outout file:  
-----
```

Komputer meminta diisikan nama output file (file hasil analisis)

5. Ketik nama file output (hasil) yang dikehendaki, misal CONTOH-1.OUT lalu tekan ENTER, maka muncul:

```
-----  
Do you want the score written to a file?:  
-----
```

Y = bila dikehendaki hasil analisis direkam

N = bila hasil analisis tidak direkam.

Bila diketik Y maka akan muncul:

```
-----  
Enter the name of the score file:  
-----
```

Komputer minta diisikan nama file untuk skor peserta tes.

6. Ketikkan nama file untuk hasil skor, misal CONTOH-1.SCR lalu tekan ENTER

Dalam waktu beberapa detik, akan muncul tampilan:

ITEM ANALYSIS IS COMPLETE

Ini menunjukkan bahwa proses analisis telah selesai. Hasil dapat diprint, namun sebelumnya ditata dulu agar hasilnya tidak terpotong-potong. Contoh hasil analisis dapat dilihat pada halaman berikut.

Item and Test Analysis Program -- ITEMAN (tm) Version 3.00

Item analysis for data from file LAT1.DAT

Page 1

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics								
		Prop. Correct	Biser. Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser. Biser.	Point Biser.	Key				
1	0-1	0.350	0.179	0.139	A	0.200	0.218	0.152	?				
					B	0.400	-0.056	-0.044					
					CHECK THE KEY				C	0.050	-1.000	-0.485	
					D was specified, A works better				D	0.350	0.179	0.139	*
					Other	0.000	-9.000	-9.000					
2	0-2	0.600	-0.223	-0.176	A	0.050	0.365	0.173					
					B	0.300	0.237	0.180	?				
					CHECK THE KEY				C	0.600	-0.223	-0.176	*
					C was specified, B works better				D	0.050	-0.330	-0.156	
					Other	0.000	-9.000	-9.000					
3	0-3	0.400	0.594	0.468	A	0.400	0.594	0.468	*				
					B	0.150	0.177	0.115					
					C	0.200	-0.358	-0.251					
					D	0.250	-0.536	-0.393					
					Other	0.000	-9.000	-9.000					

dst

There were 20 examinees in the data file.

Scale Statistics

Scale:	0
N of Items	20
N of Examinees	20
Mean	8.900
Variance	7.790
Std. Dev.	2.791
Skew	-0.325
Kurtosis	-0.543
Minimum	3.000
Maximum	14.000
Median	9.000
Alpha	0.250
SEM	2.418
Mean P	0.297
Mean Item-Tot.	0.214
Mean Biserial	0.294

Keterangan Statistik Butir Soal:

1. Seq. No. adalah nomor urut butir soal
2. Scala-Item adalah nomor urut butir soal dalam tes/instrumen
3. Prop-Correct adalah proporsi peserta tes yang menjawab benar butir soal
4. Biser adalah indeks daya beda butir soal dengan menggunakan koefisien korelasi biserial. Nilai positif menunjukkan bahwa peserta tes yang menjawab benar butir soal, mempunyai skor yang relatif tinggi dalam tes tersebut. Untuk statistik pilihan jawaban (alternatif) korelasi biserial negatif sangat tidak dikehendaki untuk kunci jawaban.
5. Point biserial indeks daya beda butir soal dengan menggunakan koefisien korelasi point-biserial. Keterangan selanjutnya sama dengan yang ada pada Biser.

Keterangan Statistik Tes:

1. N of Item adalah jumlah butir soal
2. N of Examinees adalah jumlah peserta tes
3. Mean adalah skor rerata peserta tes
4. Variance adalah varian dari distribusi skor peserta tes yang memberikan gambaran tentang sebaran skor peserta tes.
5. Std.dev. adalah standar deviasi dari distribusi skor peserta tes.
6. Skew adalah kemiringan distribusi skor peserta tes. Juling negatif menunjukkan bahwa sebagian besar skor berada di bagian atas (skor tinggi) dari distribusi skor, dan sebaliknya
7. Kurtosis adalah puncak distribusi skor yang menggambarkan kelandaian distribusi skor peserta tes dibanding dengan distribusi normal. Nilai positif menunjukkan distribusi lebih lancip, dan nilai negatif menunjukkan distribusi yang lebih landai (merata). Kurtosis untuk distribusi normal adalah nol.
8. Alpha adalah koefisien reliabilitas alpha untuk tes tersebut.
9. SEM (*standard error of measurement*) adalah kesalahan baku pengukuran untuk setiap tes.
10. Mean P adalah rata-rata tingkat kesukaran semua butir soal dalam tes secara klasikal dihitung dengan cara mencari rata-rata proporsi peserta tes yang menjawab benar untuk semua butir dalam soal tes tersebut.
11. Mean Item-Tot adalah nilai rata-rata indeks daya beda dari semua butir dalam tes yang diperoleh dengan menghitung nilai rata-rata point biserial dari semua butir dalam tes/skala.
12. Mean Biserial adalah nilai rata-rata indeks daya beda dari semua butir dalam tes yang diperoleh dengan menghitung nilai rata-rata biserial dari semua butir dalam tes/skala.

PENUTUP

Dalam melakukan analisis empirik suatu instrumen menggunakan program ITEMAN, hal penting yang harus diperhatikan adalah jenis instrumen. Apakah instrumen itu memiliki butir

dengan jawaban dikotomi (*dichotomously scored items*) atau instrumen yang memiliki butir dengan jawaban ganda (*multipoint items*).

Untuk instrumen yang memiliki butir dengan jawaban dikotomi tidak perlu menentukan tipe *missing data*, sedangkan pada instrumen yang memiliki butir dengan jawaban ganda (*multipoint items*) harus menentukan tipe *missing data*. Oleh karenanya dalam melakukan analisis empirik pada instrumen yang memiliki butir dengan jawaban ganda (*multipoint items*) harus lebih hati-hati.

Untuk mendapatkan instrumen berkualitas tinggi memang tidak mudah. Meskipun demikian harus selalu diupayakan, lebih-lebih bila instrumen itu akan digunakan secara meluas atau secara nasional. Dengan instrumen yang berkualitas tinggi akan diperoleh data yang akurat dan pada gilirannya akan dapat diambil keputusan yang tepat.

DAFTAR PUSTAKA

- Allen, M.J. & Yen, W.M. 1979. *Introduction to measurement theory*. Monterey, CA: Brooks/Cole Publishing Company.
- Anderson, L.W., dan Krathwoh, D.R , 2001. *A Taxonomy for learning, teaching, and assessing: A Revision of Bloom's taxonomy of educational objectives*. A Bridged edition. New York: David Mc KeyCompany, Inc.
- Bahrul Hayat. 2002. Sertifikasi, ujian akhir, dan otonomi daerah. *Makalah*. Disampaikan pada Seminar Sistem ujian akhir di era otonomi, di Jakarta, 19 Mei 2002.
- Bahrul Hayat. 1996. Interpretasi hasil analisis iteman. *Buletin Pengujian dan penilaian pendidikan*, 2 (5) 7 – 9.
- Crocker, L. & Algina, J. 1986. *Introduction to classical and modern test theory*. New York : Holt, Rinehard and Winston Inc.
- Dawson, J.B. & Thomas, G.H. 1972. *Item analysis and examination statics*. Birmingham: The Union of Educational Institutions.
- Ebel, R.L. 1972. *Essentials of educational measurement*. (3rd. ed.) Englewood Cliffts,NJ: Prentice Hall Inc.
- Feldt, L.S. and Brennan, R.L. 1989. "Reliability", *Educational measurement*, edited by Robert L Linn. New York: Macmillan Publishing Company.
- Fernandes, H.J. X. 1984. *Evaluation of educational program*. Jakarta: National Education Planning , Evaluating and Curriculum Development.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. 1991. *Fundamentals of item response theory*. London: Sage Publications, Inc.

- Lawrence M.R. 1994. Question to ask when evaluating test. *Eric digest*. Artikel: ED385607. Sumber: <http://www.ericfacility.net/ericdigest/ed.385607.html> tanggal 10 Februari 2003.
- Lord, F.M. 1980. *Application of item response theory to practical testing problems*. Hillsdale, NJ.: Lawrence Erlbaum Associates, Publisher.
- Moore, B., Stanly, T. 2010. *Critical thinking and formative assessments*. Larchmount, NY: Eye On Education, Inc
- Nitko, A.J. 1996. *Penilaian berkelanjutan berdasarkan kurikulum (PB2K): Kerangka, konsep, prosedur, dan kebijakan* (terj. AM. Ahmad) Jakarta: Pusat Pengembangan Agribisnis.
- N. N. 1986. *User' manual for ITEMAN, RASCAL and ASCAL*. N.C. ASCAR
- Nunnally, J.C. 1978. *Psychometric theory*. New York: McGraw Hill Book Company. Inc
- Peraturan Menteri Pendidikan Nasional Republik Indonesia Nomor 20 Tahun 2007 Tentang Standar Penilaian.
- Sumadi Suryabrata. 2000. *Pengembangan Alat Ukur Psikologis*. Penerbit Andi Yogyakarta.
- Thomas, A. dan Thorne, G. (2007). *Higher Order Thinking*. Center for Development and learning. Diambil dari CDL pada tanggal 6 Agustus 2011.
- Tim Pusiasjian (1997/1998). *Bahan penataran: Pengujian Pendidikan*. Jakarta: Balitbang Dikbud.
- Tim. 1999. *Manual ITEMAN*. Jakarta: Pusiasjian Balitbang Depdiknas.
- Undang-Undang Republik Indonesia Nomor 20 Tahun 2003 Tentang Sistem Pendidikan Nasional
- Woolfolk, A.E. & McCune, L.N. 1984. *Educational Psychology for Teachers*. Englewood Cliffs, NJ: Prentice Hall, Inc.

LAMPIRAN-LAMPIRAN

