ANALISIS BUTIR SOAL DENGAN PENDEKATAN *CLASSICAL TEST THEORY* DAN *ITEM RESPONSE THEORY*

Wiwin Mistiani
Teacher Training and Tarbiyah Faculty, IAIN Palu

Abstract

The test is one of the easiest and most inexpensive ways that can be done to take pictures of students' progress in the cognitive domain. Therefore, the existence of a quality test device is a necessity so that the students' cognitive abilities may be disclosed. The quality of a test device can be seen to perform qualitative and quantitative analysis. Qualitative analysis is an analysis that is performed before the tests given to participants to test their compatibility with aspects of material, construction and language, while the quantitative analysis can be done by using classical test theory and item response theory.
Key Word: classical test theory, item response theory

Pendahuluan

Mutu pendidikan merupakan masalah yang senantiasa diupayakan peningkatannya oleh pemerintah. Peningkatan pendidikan menjadi salah satu prioritas pembangunan pendidikan nasional. Pembangunan pendidikan juga merupakan upaya peningkatan sumber daya manusia. Untuk mencapai sumber daya manusia yang berkualitas, pendidikan harus bermutu. Namun untuk mencapai pendidikan yang bermutu, masih banyak kendala yang dihadapi dunia pendidikan di Indonesia antara lain mutu pendidik dan tenaga kependidikan, siswa, sarana prasarana, proses pembelajaran termasuk di dalamnya proses evaluasi dan penilaian.

Berdasarkan peraturan pemerintah nomor 19 tahun 2005 tentang standar Nasional Pendidikan Pasal 66 ayat 1, menyatakan bahwa penilaian hasil belajar oleh pemerintah bertujuan untuk menilai pencapaian kompetensi lulusan secara nasional pada mata pelajaran tertentu dalam kelompok mata pelajaran ilmu

pengetahuan teknologi dan dilakukan dalam bentuk ujian nasional[1]. Penilaian merupakan komponen penting dalam sistem pendidikan untuk mengetahui perkembangan dan tingkat pencapaian hasil pembelajaran. Penilaian memerlukan data yang baik. Salah satu sumber data adalah hasil pengukuran. Kegiatan pengukuran ini biasanya dilakukan melalui tes.[2] Tes yang bermutu baik sangat menentukan baik buruknya mutu rumusan hasil penilaian, dan selanjutnya akan menentukan mutu berbagai keputusan atau kebijakan pendidikan.

Tes prestasi belajar bertujuan untuk mengukur kemampuan atau prestasi seseorang setelah menjalani proses pembelajaran.[3] Tes seperti ini penting sekali dilakukan oleh guru, sekolah maupun lembaga kependidikan untuk mengetahui seberapa jauh siswa telah mencapai tujuan pembelajaran yang diharapkan.[4] Hasil tes dapat digunakan oleh guru, sekolah, atau institusi kependidikan lain untuk mengambil keputusan atau umpan balik bagi perbaikan proses belajar mengajar

Namun, terkadang pemberian tes yang terlalu susah atau terlalu mudah menyebabkan pendidik sulit membedakan kemampuan peserta didik. Oleh karena itu, diperlukan analisis terhadap butir soal, dengan harapan hasil tes dapat merepresentasikan kemampuan peserta didik, karena sebuah tes yang baik akan bisa mengungkapkan keadaan sebenarnya dari siswa, dan tes yang tidak baik tidak akan bisa mengungkap apa kemampuan sebenarnya siswa.

Selanjutnya sebuah tes yang baik itu, harus valid dan reliabel. Validitas merupakan penilaian menyeluruh dimana bukti

---

[1]Depdiknas, *Peraturan Pemerintah RI Nomor 19, Tentang Standar Nasional Pendidikan,* (Jakarta: Depdiknas, 2005) p. 66.

[2]Suharsimi Arikunto, *Dasar-Dasar Evaluasi Pendidikan,* ,(Jakarta: Bumi Aksara, 2009) p. 6

[3]Saifuddin Azwar, *Reliabilitas dan Validitas,* (Yogyakarta: Pustaka Pelajar, 2010) p.13

[4]Sumarna Surapranata, *Analisis, Validitas, Reliabilitas, dan Interpretasi Hasil Tes Implementasi Kurukulum 2004,* (Bandung: PT. Remaja Rosdakarya, 2004) p 19.

empiris dan logika teori mendukung pengambilan keputusan serta tindakan berdasarkan skor tes atau model-model penilaian yang lain. Validitas sebuah tes dapat dilakukan dalam berbagai bentuk seperti *content validity*, *criterion validity* dan *construct-related validity*. Meskipun idealnya validasi dapat dilakukan dengan memakai semua bentuk validitas tes tersebut, tetapi pengembang tes dapat memilih bentuk validasi dengan melihat tujuan pengembangan tes. Selain valid, alat ukur yang baik juga harus reliabel. Sebuah tes dikatakan reliabel jika skor yang diperoleh oleh peserta relatif sama meskipun dilakukan pengukuran berulang-ulang. Untuk memperoleh skor yang sama, maka tidak boleh ada kesalahan pengukuran. Dengan demikian, keandalan sebuah alat ukur dapat dilihat dari dua petunjuk yaitu kesalahan baku pengukuran dan koefisien reliabilitas. Kedua statistik tersebut masing-masing memiliki kelebihan dan keterbatasan

Selain valid dan reliable, tes yang baik juga tergantung dari banyaknya butir-butir soal berkategori baik yang terdapat dalam tes. Semakin banyak butir soal yang baik, semakin baiklah perangkat tes tersebut. Sebaliknya, semakin sedikit jumlah butir soal yang baik, semakin buruklah kualitas tes itu. Untuk melihat kualitas sebuah tes dapat dilakukan dengan menggunakan analisis kualitatif (teoretik) dan kuantitatif (empiris). Secara kualitatif tes dikatakan baik jika telah memenuhi persyaratan penyusunan dari sisi materi, konstruksi dan bahasa. Adapun secara kuantiatif dapat dilakukan dengan dua teknik yaitu teori tes klasik (*classical true-score theory)* dan teori respon butir *(Item Response Theory)*. Dalam tulisan berikut ini, penulis hanya akan memberikan pengantar tentang analisis butir soal menggunakan teori tes klasik *(classical test theory)* dan modern dengan item *response theory (IRT).*

Hakekat Tes

a. Pengertian Tes

Allen & Yen menyebut tes sebagai "*a test is device for obtaining Secara a sample of an individual's behavior",*[5] yang mendefinisikan tes sebagai instrument atau prosedur sistematis untuk mengukur perilaku sampel. Pengartian tes sebagai prosedur yang sistematis untuk mengamati dan menggambarkan satu atau lebih karakteristik seseorang dengan bantuan skala numerik atau sistem kategorik. Menurut Zaenal Arifin pada hakekatnya tes adalah suatu alat yang berisi serangkaian tugas yang harus dikerjakan atau soal-soal yang harus dijawab oleh peserta didik untuk mengukur suatu aspek perilaku tertentu.[6] Dengan demikian tes berfungsi sebagai alat ukur. Dalam tes prestasi belajar, aspek perilaku yang hendak diukur adalah tingkat kemampuan peserta didik dalam menguasai materi pelajaran yang telah disampaikan. Prosedur yang dilakukan dalam mendapat informasi yang bersifat kualitatif dan kuanttatif melalui tes disebut pengukuran. Tes diartikan juga sebagai sejumlah pertanyaan yang membutuhkan jawaban, atau sejumlah pernyataan yang harus diberi tanggapan dengan tujuan mengukur tingkat kemampuan seseorang atau mengungkap aspek tertentu dari orang yang dikenai tes.

Secara umum tes diartikan sebagai alat pengukur yang mempunyai standar objektif, sehingga dapat digunakan secara meluas, serta betul-betul dapat digunakan untuk mengukur dan membandingkan keadaan psikis atau tingkah laku individu dan prosedur yang sistematis untuk mengamati atau mendeskripsikan satu atau lebih karakteristik seseorang dengan menggunakan standar numerik atau sistem kategori. Selanjutnya Cronbach dalam Djemari Mardapi, menjelaskan bahwa semua tes pada dasarnya

[5]Allen et al., *Introduction to Measurement Theory*, (Monterey: McGraw-Hill, 1979) p.1

[6] Zaenal Arifin, *Evaluasi Pembelajaran,* (Bandung: PT. Remaja Rosdakarya, 2009) p. 119

adalah untuk mengukur unjuk kerja dalam suatu segi[7]. Namun tes unjuk kerja biasanya digunakan terhadap suatu tugas yang membutuhkan respon nonverbal. Tes unjuk kerja mengacu pada suatu standar yang ingin dicapai atau yang ditetapkan sebagai batas minimum yang harus bisa dilakukan siswa, misalnya operasi hitung, melakukan komunikasi, membaca, menyimak, dan sebagainya. Oleh karena itu standar yang ingin dicapai harus ditetapkan terlebih dahulu. Pengukuran pada prinsipnya bertujuan untuk mengetahui karakteristik suatu objek yang berkaitan dengan aspek kognitif, aspek afektif dan aspek psikomotor.

Tes pilihan ganda biasanya terdiri dari sejumlah item soal. Tes yang baik harus terdiri atas item-item soal yang baik. Pada tes pilihan ganda, item yang baik harus mempunyai tingkat kesulitan yang memadai, daya beda yang baik, dan berfungsinya pengecoh. Tingkat kesulitan menunjuk kepada perbandingan antara banyaknya peserta tes yang menjawab benar dengan banyaknya seluruh peserta tes. Daya pembeda menunjuk kepada selisih proporsi yang menjawab benar pada kelompok atas dan proporsi yang menjawab benar pada kelompok bawah. Pada perkembangannya, daya pembeda suatu item didefinisikan sebagai korelasi antara skor item tersebut dengan skor total. Berfungsinya pengecoh menunjuk kepada seberapa banyak peserta yang memilih pengecoh tersebut, dan pengecoh yang dipilih paling sedikit 5% dari seluruh peserta tes. Item soal pilihan ganda dikatakan memenuhi persyaratan apabila besarnya tingkat kesukaran berkisar antara 0,30 hingga 0,80, besarnya daya pembeda 0,30 atau lebih, dan pengecoh dipilih oleh paling sedikit 5% dari seluruh peserta tes. Butir yang baik dalam suatu kerangka uji tes adalah butir yang tidak terlalu sukar dan tidak juga terlalu mudah yakni butir dengan taraf kesukaran 0,5, selain memang terletak ditengah-tengah juga memberikan variasi yang maksimum.[8]

---

[7]Djemari Mardapi, *Teknik Penyusunan Instrumen Tes dan Nontes*, (Yogyakarta: Mitra Cendekia Press, 2008) p.76

[8] Dali S. Naga, *Pengantar Teori Skor*, (Jakarta: Gunadarma, 1992). p. 57

b. Fungsi Tes

Secara umum ada tiga macam fungsi tes dalam dunia pendidikan, yakni (1) tes dapat berfungsi sebagai alat untuk mengukur prestasi belajar siswa; (2) sebagai motivator dalam pembelajaran; (3) sebagai upaya perbaikan kualitas pembelajaran[9]. Sejalan dengan pendapat di atas Djemari Mardapi (2008: 68) mengemukakan bahwa fungsi tes yang penting adalah untuk; 1) mengetahui tingkat kemampuan peserta didik, 2) mengukur pertumbuhan dan perkembangan peserta didik, 3) mendiagnosis kesulitan belajar peserta didik, 4) mengetahui hasil pengajaran, 5) mengetahui hasil belajar, 6) mengetahui pencapaian kurikulum, 7) mendorong peserta didik belajar, dan 8) mendorong pendidik mengajar yang lebih baik dan peserta didik belajar lebih baik.

` Jika ditinjau dari fungsinya ada empat macam tes yang banyak digunakan dalam lembaga pendidikan, yaitu (1) tes penempatan; (2) tes diagnostik; (3) tes formatif; dan (4) tes sumatif. Selanjutnya dijelaskan bahwa tes penempatan dilaksanakan pada awal pelajaran, tes diagnostik berguna untuk mengetahui kesulitan belajar yang dihadapi peserta didik termasuk kesalahan pemahaman konsep. Tes formatif bertujuan untuk memperoleh masukan tentang tingkat keberhasilan pelaksanaan proses pembelajaran, sedangkan tes sumatif dilaksanakan pada akhir semester. Sebagai alat ukur, maka tes harus dapat memberikan informasi mengenai pengetahuan dan kemampuan obyek yang diukur, sebagaimana yang diungkapkan Miller:

> Test are formal assessment instrument used to judge student's cognitive ability in an academic discipline as well as to gather quantitative information about student's psychomotor performance (physical skills) and affective characteristic."[10]

c. Penggolongan Tes

---

[9]Djaali dkk., *Pengukuran dalam Bidang Pendidikan,* (Jakarta: PT Grasindo, 2008) p. 7

[10]Patrick W. Miller, *Measurement and Teaching,* (USA: Patrick W. Miller and Associates, 2008) p. 1

Berdasarkan bentuknya tes dapat diklasifikasikan ke dalam (1) tes pilihan ganda, (2) tes benar-salah, (3) tes isian/jawaban singkat, (4) tes menjodohkan, dan (5) tes uraian. Sementara itu ada dua kategori butir tes yang biasa digunakan dalam dunia pendidikan. Pertama adalah pernyataan berbentuk objektif (seperti: pilihan ganda, benar-salah, dan penjodohan), kedua adalah pernyataan subjektif (seperti: esai, studi kasus, dan jawaban pendek). Bentuk tes yang sering digunakan di lembaga pendidikan adalah bentuk tes objektif dan bentuk tes non objektif. Tes objektif memiliki sistem penskoran yang berlaku sama untuk korektor, sedangkan untuk tes non objektif berlaku sebaliknya.

Tes objektif yang sering digunakan adalah tes bentuk pilihan ganda, benar salah, menjodohkan, dan uraian objektif. Sedangkan untuk tes non objektif digunakan pada ilmu-ilmu sosial yang jawabannya luas dan tidak hanya satu jawaban yang benar, tergantung argumentasi peserta tes. Selanjutnya Djemari menyatakan bahwa tes uraian objektif sering digunakan pada mata pelajaran yang batasnya jelas, seperti mata peljaran Matematika, Fisika, Kimia, dan Biologi yang mana pada mata pelajaran tersebut biasanya memerlukan jawabannya hanya satu, mulai dari memilih rumus yang tepat, memasukan angka dalam rumus, menghitung hasil, dan menafsirkan hasilnya.

d. Tes Prestasi Belajar

Di sekolah kegiatan tes prestasi belajar dapat berupa ulangan harian, ulangan semester, ujian sekolah dan ujian akhir sekolah. Khusus untuk kelas XII pada tingkat SMA biasanya masih ditambahkan lagi dengan ujian try out sebagai tahap persiapan untuk menghadapi ujian nasional. Tes prestasi belajar adalah himpunan pertanyaan yang harus dijawab atau pernyataan-pernyataan yang harus dipilih/ditanggapi, atau tugas yang harus dilakukan oleh orang yang dites dengan tujuan untuk mengukur suatu aspek. Idealnya penguasaan kemampuan matematika harus dikuasai anak sedini mungkin. Anak mendasarkan diri pada pengetahuan yang telah mereka miliki untuk menyempurnakan kompetensi matematikanya dan memperluas pemahamannya

tentang pengetahuannya itu.Kaitannya dengan itu tes prestasi belajar merupakan tes yang disusun secara terencana untuk mengungkap kemampuan maksimal siswa dalam menguasai bahan atau materi tertentu yang diajarkan kepadanya. Oleh karena itu agar memperoleh informasi yang akurat maka dibutuhkan tes yang handal.

Menurut Miller mengatakan bahwa tes yang baik minimalnya harus memenuhi syarat validitas, reliabilitas, objektivitas dan feasibilitas.[11] Sementara Bott menyatakan bahwa tes yang baik harus memenuhi empat karakteristik yaitu: valid, reliabel, obyektif dan praktis.[12] Selanjutnya Saifuddin Azwar menyatakan bahwa suatu tes dapat berfungsi secara efektif maka haruslah memiliki minimal tiga kualitas yaitu valid, reliabel, dan unbiased.[13]

Validitas artinya ketepatan interpretasi hasil prosedur pengukuran, dan reliabel dapat artinya konsistensi hasil pengukuran, dan usabilitas artinya praktis prosedurnya, atau validitas dapat diartikan sebagai sejauh mana tes mampu mengukur atribut yang seharusnya diukur." Suatu alat ukur yang tinggi validitasnya akan menghasilkan eror pengukuran yang kecil, artinya skor setiap subjek yang diperoleh melalui alat ukur tersebut tidak jauh berbeda dari skor yang sesungguhnya. Terdapat lima sumber validitas untuk menginterpretasi skor tes untuk berbagai tujuan. Sumber tersebut dapat berasal dari a) isi/kontent tes, b) proses respon, c) struktur internal, d) hubungan dengan variabel lain, e) konsekuensi tes. Sehubungan dengan sumber validitas di atas, terdapat empat jenis validity yaitu *content validity*, *construct validity*, *concurrent validity* dan *predictive validity*. Sementara Messick menyatakan bahwa terdapat tiga pendekatan untuk

[11]Miller. *Measurement*, p. 83

[12]Bott, P. A. *Testing and Assessment in Occupational and Technical Education*, (Boston: Allyn & Bacon, 1995) p.56

[13]Azwar Saifuddin, *Tes*, p. 35

menentukan validitas tes, yaitu validitas yang berhubungan dengan isi, kriteria (prediktif dan konkuren), serta konstruk.[14]

Di samping validitas, informasi tentang reliabilitas tes sangat diperlukan. Karakteristik suatu tes yang baik hendaknya memiliki keterandalan yang baik yang ditunjukkan oleh indeks koefisien reliabilitas. Reliabilitas menurut Brennen menunjukkan adanya konsistensi skor yang diperoleh dari hasil pengukuran.[15] Dari penjelasan diatas yang dimaksud dengan reliabilitas alat ukur menunjukkan "sejauh mana hasil pengukuran dengan alat tersebut dapat dipercaya". Reliabilitas ditunjukkan oleh taraf keajegan (konsistensi) skor yang diperoleh subjek yang diukur dengan alat sama, atau diukur dengan alat yang setara pada kondisi yang berbeda.

Sebuah tes dikatakan reliabel apabila skor amatannya berkorelasi tinggi dengan skor sebenarnya. Besarnya indeks reliabilitas berada pada rentang nilai 0-1. Feld & Brennan mengatakan bahwa suatu instrument sudah dianggap reliabel jika memiliki koefisien reliabilitas minimal 0,7.[16] Pendapat yang sama diungkapkan oleh Nitko dan Brookhart yang menyatakan bahwa koefisien reliabilitas sebesar 0,7 merupakan batas minimal yang dapat ditolerir.[17] Terdapat tiga pendekatan yang dapat digunakan dalam mengestimasi reliabilitas tes, yaitu metode estimasi *test-retest*, *parallel-forms* dan *internal consistency* sementara itu, menurut Saifuddin Azwar metode estimasi tes ulang (*test-retest*) akan menghasilkan koefisien stabilitas, metode estimasi parallel (*parallel-forms*) akan menghasilkan koefisien ekuivalensi, dan metode estimasi penyajian tunggal (*single trial administration*) bertujuan untuk menghasilkan koefisien konsistensi internal. [18]

---

[14]Messick, S., *Validity,* (London: Macmillan Publishers. 1989) p. 16

[15]Feld et al., *Educational Measurement,* (Westport: Greenwood Publishing Group, 2006) p. 102

[16]Feld et al., *Educational,* p.106

[17]Nitko et al., *Educational Assessment of Students,* (New Jersey: Prentice Hall, 2011) p 80.

[18]Azwar Saifuddin, *Tes*, p. 43

Koefisien reliabilitas yang diperoleh melalui metode estimasi tes-ulang sangat sensitif terhadap perubahan keadaan subjek yang terjadi selama tenggang waktu diantara tes pertama dan penyajian ulangnya. Efek bawaan dari tes pertama terhadap tes kedua seringkali tidak dapat diprediksi dan akhirnya mempengaruhi koefisien yang diperoleh. Pada sisi lain estimasi terhadap reliabilitas dengan metode paralel bermasalah terhadap sukarnya untuk memenuhi kondisi paralel antara dua bentuk tes yang bersangkutan. Sehingga metode konsistensi internal lebih memiliki nilai praktis yang tinggi dalam mengestimasi reliabilitas. Selanjutnya teknik estimasi reliabilitas yang berkembang banyak mengacu kepada metode konsistensi internal, diantaranya teknik split-half, rumus Rulon, rumus Flanagan, teknik KR-20, teknik KR-21, teknik analisis varians dan koefisien alpha.Validitas dan reliabilitas tes sangat tergantung pada ciri-ciri butir soal. Menurut Anastasi dan Urbina "reliabilitas dan validitas yang tinggi dapat dibangun terlebih dahulu dalam tes melalui analisis butir soal". Analisis butir soal merupakan salah satu cara untuk mengetahui karakteristik dan kualitas butir soal.[19]

Analisis Butir Soal

a. Analisis butir soal secara kualitatif

Analisis soal secara kualitatif bertujuan untuk mengetahui kualitas butir soal dari saspek materi, konstruksi, dan bahasa yang digunakan. Aspek materi berkaitan dengan substansi keilmuan yang ditanyakan serta tingkat berfikir yang terlibat, aspek konstruksi berkaitan dengan tehnik penulisan soal, dan aspek bahasa berkaitan dengan kejelasan hal yang ditanyakan. Telaah butir secara kualitatif dapat menggunakan format penelaahan butir soal yang dikembangkan oleh kementerian pendidikan nasional. Menurut Depdiknas pedoman analisis butir soal untuk setiap aspek adalah sebagai berikut:

1) Aspek Materi
   a) Soal sesuai dengan indikator

[19]Anastasi et al., *Psychological Testing,* (New Jersey: Prentice Hall Inc., 2008) p. 190

b) Materi yang diukur sesuai dengan kompetensi
c) Pilihan jawaban homogen dan logis
d) Hanya ada satu kunci jawaban yang tepat

2) Aspek Konstruksi
a) Pokok soal dirumuskan dengan singkat, jelas, dan tegas
b) Rumusan pokok soal dan pilihan jawaban merupakan pertanyaan yang diperlukan
c) Pokok soal tidak memberi petunjuk ke kunci jawaban
d) Gambar, grafik, tabel, diagram, wacana, dan sejenisnya yang terdapat pada soal jelas dan berfungsi
e) Panjang rumusan soal relatif sama
f) Pilihan jawaban tidak menggunakan pernyataan "semua jawaban di atas salah" atau "semua pilihan di atas benar" dan sejenisnya
g) Pilihan jawaban yang berbentuk angka atau waktu harus disusun berdasarkan urutan besar kecilnya angka tersebut atau kronologisnya
h) Butir-butir soal tidak bergantung pada jawaban soal sebelumnya

3) Aspek Bahasa
a) Menggunakan bahasa yang sesuai dengan kaidah bahasa Indonesia
b) Menggunakan bahasa yang komunikatif
c) Pilihan jawaban tidak mengulang kata/kelompok kata yang sama, kecuali merupakan satu kesatuan pengertian
d) Tidak menggunakan bahasa yang berlaku setempat (bias budaya)[20]

Telaah butir soal secara kualitatif dapat melibatkan ahli atau teman sejawat yang menguasai bidang ilmu, tata bahasa, dan evaluasi.

b. Analisis Butir Soal Secara Empiris

1) Analisis Item Menurut Teori Klasik *Classical Test Theory*

[20]Depdiknas, *Kurikulum 2004 SM, Pedoman Khusus Pengembangan Silabus dan Penilaian,* (Jakarta: Depdiknas, 2004) p.35

Teori klasik merupakan teori pengukuran yang telah dikembangkan dan diaplikasikan sejak lama diberbagai bidang kehidupan. Keunggulan teori tes klasik terletak pada konsepnya yang mudah dipahami serta penggunaannya dapat dilakukan pada tes berskala kecil. Inti teori tes klasik adalah berupa asumsi-asumsi yang dirumuskan secara matematis yang modelnya disebut skor murni (*true score model).* Teori tes klasik mengasumsikan bahwa setiap orang memilki nilai yang benar (True= T), yang akan diperoleh jika tidak ada kesalahan dalam pengukuran. Terdapat lima asumsi yang digunakan dalam teori tes klasik. Asumsi-asumsi teori tes klasik tersebut dikembangkan dalam berbagai formula yang berguna dalam melakukan pengukuran psikologis. Formula tersebut diantaranya estimasi indeks kesukaran, indeks daya beda, efektifitas distraktor, reliabilitas tes dan validitas.

a) Tingkat kesukaran

Taraf kesukaran suatu item dinyatakan oleh suatu indeks yang dinamakan indeks kesukaran item dan disimbolkan oleh huruf p. Menurut Saifuddin Azwar menyatakan bahwa indeks kesukaran item merupakan rasio antara penjawab item dengan benar dan banyaknya penjawab item[21]. Secara teoritik dapat dikatakan bahwa p sebenarnya merupakan probabilitas empirik untuk lulus item tertentu bagi kelompok siswa tertentu. Pada pengukuran item tes, tingkat kesulitan butir berhubungan dengan persentase orang-orang yang dapat menjawab soal dengan benar. Tingkat kesukaran soal merupakan proporsi jumlah peserta tes yang menjawab benar dari suatu butir soal, yaitu perbandingan antara jumlah peserta tes yang menjawab benar dengan jumlah peserta tes seluruhnya.

Tingkat kesukaran item sebagaimana dinyatakan oleh Blerkom adalah *"proportion of students who answered the item correctly."* [22] Besarnya harga tingkat kesukaran item terletak antara 0 hingga 1. Semakin besar tingkat kesukaran yang diperoleh semakin mudah item tersebut, sebaliknya semakin kecil tingkat

[21]Azwar Saifuddin, *Tes*, p. 13

[22]Blerkom et al., *Measurement and Statistic for Teachers,* (New York Routledge, 2009) p. 127

kesukaran yang diperoleh maka item tersebut semakin sukar. Tingkat kesukaran soal pada tes tergantung pada tujuan tes yang disusun. Indeks kesukaran soal yang paling sederhana dan paling sering digunakann adalah taraf kesukaran p, yaitu banyaknya proporsi jawaban yang benar terhadap semua jawaban. Ukuran mudahnya soal dilihat dari nilai p jika nilai p rendah maka soal tersebut semakin sukar, sebaliknya jika nilai p tinggi maka menunjukkan soal tersebut mudah. Cara mudah dan umum digunakan untuk menentukan nilai p adalah skala rata-rata atau proporsi menjawab benar (p) yaitu jumlah peserta tes yang menjawab benar pada soal yang dianalisis dibandingkan dengan peserta tes seluruhnya Harga tingkat kesukaran item pada perangkat tes dapat ditentukan dengan rumus sebagai berikut:

$\text{Pi} = \frac{B}{T}$

Keterangan:

Pi : tingkat kesukaran butir

B : banyaknya peserta tes yang menjawab soal dengan benar

T : jumlah peserta tes yang mengerjakan soal[23]

Miller menyatakan tentang indeks kesukaran item butir soal yang dapat diterima adalah 0,30- 0,70, karena pada interval ini informasi tentang kemampuan siswa dapat diperoleh dengan maksimal. Butir soal yang memiliki indeks kesukaran di atas 0,7 termasuk soal yang mudah, sementara yang memiliki indeks kesukaran di bawah 0,3 termasuk butir soal yang sukar[24]. Senada dengan pendapat di atas Allen & Yen juga mensyaratkan bahwa tingkat kesukaran butir soal yang baik adalah pada rentang nilai 0,3 -0,7.[25] Besarnya tingkat kesukaran berkisar antara 0 sampai 1. Tingkat kesukaran dikategorikan sebagai berikut.

[23]Azwar Saifuddin, *Tes*, p. 56

[24]Patrick W. Miller, *Measurement*, p. 131

[25]Allen et al., *Introduction*, p.121

Tabel 1: Kategori Tingkat Kesukaran Butir Soal

Tingkat Kesukaran (p)	Kategori
$p > 0,70$	Mudah
$0,30 \leq p \leq 0,70$	Sedang
$p < 0,30$	Sukar

Berdasarkan tabel di atas dapat dijelaskan bahwa jika indeks kesukaran di atas 0,70 maka butir soal tersebut masuk dalam kategori mudah, antara 0,30-0,70 masuk dalam kategori sedang dan jika indeks kesukarannya di bawah 0,30 maka butir soal tersebut masuk dalam kategori sukar. Secara umum dapat disimpulkan bahwa semakin tinggi indeks kesukaran semakin mudah butir soalnya, sebaliknya semakin rendah indeks kesukaran semakin sukar butir soalnya.

b) Daya pembeda

Daya pembeda atau daya beda suatu soal berfungsi untuk menentukan dapat tidaknya suatu soal membedakan kelompok dalam aspek yang diukur sesuai dengan perbedaan yang ada pada kelompok tersebut. Beberapa cara yang digunakan untuk menentukan daya pembeda antara lain dengan indeks diskriminasi, indeks korelasi dan indeks keselarasan. Daya pembeda juga berhubungan dengan seberapa baik sebuah item dapat membedakan kemampuan peserta tes. Daya beda tes menunjukkan kemampuan tes tersebut memisahkan siswa yang pandai dan kurang pandai. Indeks daya pembeda butir soal pada dasarnya merupakan suatu indeks pada butir soal yang dapat membedakan siswa yang memperoleh skor tinggi dengan siswa yang memperoleh skor rendah.[26]

Indeks daya pembeda ditetapkan dari selisih proporsi yang menjawab dari masing-masing kelompok. Indeks menunjukkan kesesuaian antara fungsi soal dengan fungsi tes secara keseluruhan. Indeks daya beda dapat dihitung dengan korelasi point biserial dan

[26]Feld et al., *Educational*, p. 338

korelasi biserial. Indeks daya beda secara klasik terletak antara -1,00 sampai +1,00. Indeks daya beda bernilai positif berarti item tersebut mampu membedakan subjek kelompok atas menjawab dengan benar dan kelompok rendah menjawab salah, dan negatif jika subjek yang terdiri dari kelompok pandai menjawab salah sedangkan subjek kelompok rendah menjawab benar, menyarankan bahwa butir soal yang dapat diterima adalah butir soal yang memiliki indeks daya pembeda lebih dari 0,30. Butir soal yang memiliki indeks daya pembeda di bawah 0,30 dapat diperbaiki atau dibuang. Daya beda dikatakan baik untuk butir soal acuan norma jika minimum besar indeks pembedanya 0,3.[27] Kriteria besarnya koefisien daya beda diklasifikasikan sebagai berikut.

Tabel 2: Kategori Penerimaan Daya Beda Butir Soal

Indeks daya beda (r_{bis})	Kategori
0,40 – 1,00	Bagus sekali
0,30 – 0,39	Lumayan bagus, perlu peningkatan
0,20 – 0,29	Belum memuaskan, perlu diperbaiki
-1,00 – 0,19	Jelek dan harus dibuang

Berdasarkan tabel di atas dapat dijelaskan bahwa koefisien korelasi dari 0,40 sampai 1,00 soal dinyatakan mempunyai kategori baik, 0,30 sampai 0,39diterima tanpa direvisi, sedangkan antara 0,20 – 0,29 diterima dengan revisi, dan -1,00 sampai dengan 0,19 termasuk soal dengan daya pembeda yang jelek sehingga harus dibuang. Selanjutnya Anas Sudijono menafsirkan baik tidaknya daya beda dengan kriteria sebagai berikut:

[27]Djemari Mardapi, *Teknik*, p. 143

Tabel 3: Kategori Daya Beda Butir Soal

Indeks Daya Beda (r_{bis})	Kategori
> 0,3	Baik
$0,2 \leq r_{bis} \leq 0,3$	Cukup
< 0,2	Tidak Baik

Dari tabel dapat dijelaskan bahwa butir soal yang mempunyai indeks daya beda di atas 0,3 termasuk dalam kategori baik, sedangkan butir dengan indeks daya beda antara 0,2-0,3 termasuk dalam kategori cukup, dan butir dengan indeks daya beda kurang dari 0,2 termasuk dalam kategori tidak baik.

c) Keberfungsian pengecoh

Analisis pengecoh (distraktor) memberikan informasi seberapa banyak siswa pada kelompok atas dan kelompok bawah memilih masing-masing option pada butir soal berbentuk pilihan berganda. Satu tujuan analisis soal adalah untuk mngetahui tentang distribusi jawaban subyek pada alternatif jawaban yang tersedia. Melalui distribusi jawaban dapat diketahui; (1) banyaknya peserta tes yang menjawab benar, (2) pengecoh yang tidak dipilih peserta tes berkemampuan rendah, (3) pengecoh yang menyesatkan, dan (4) pengecoh yang menjadi daya tarik bagi peserta yang berkemampuan rendah. Sebuah pengecoh dikatakan baik apabila memiliki nilai korelasi *point biserial* negatif. Nilai negatif menunjukkan bahwa siswa yang berkemampuan rendah cendrung memilih pengecoh tersebut sebagai jawaban, sebaliknya siswa yang berkemampuan tinggi akan memilih kunci jawaban sebagai jawabannya. Selain itu setiap pilihan jawaban harus ada yang memilih setidaknya oleh 5% siswa. Sementara Miller menyatakan bahwa pengecoh dikatakan efektif jika dipilih oleh paling sedikit 2% siswa.[28]

d) Reliabilitas

Reliabilitas adalah ketetapan atau keajegan suatu tes apabila diteskan kepada subyek yang sama. Tes dikatakan

[28]Patrick W. Miller, *Measurement*, p. 29

memiliki reliabilitas yang tinggi atau terdapat korelasi yang tinggi antara hasil tes pertama dengan hasil tes kedua apabila hasil skor tesnya sama. Kalau antara hasil tes pertama dengan hasil tes kedua tidak terdapat hubungan atau hubungannya rendah, maka tes tersebut dikatakan tidak reliabel.

Hal ini ditunjukkan oleh taraf keajegan (konsistensi) skor yang diperoleh para subjek yang diukur dengan alat yang sama atau diukur dengan alat yang setara pada kondisi yang berbeda. Secara sederhana reliabilitas tes adalah tingkat sejauh mana skor deviasi seseorang tetap konsisten dengan pengulangan tes yang sama atau tes yang dianggap sama.Menurut Linn koefisien reliabilitas yang baik adalah di atas 0,70.[29] Saifuddin Azwar menyatakan bahwa secara teoritik besarnya koefisien reliabilitas berkisar dari 0,00 sampai 1,00, namun pada kenyataannya koefisien sebesar 1,00 dan sekecil 0,00 tidak pernah dijumpai.

e) Kesalahan Pengukuran

Kesalahan pengukuran (*Standard error of measurement*) membantu pemakai tes dalam memahami kesalahan yang bersifat random (acak) yang mempengaruhi skor seseorang peserta tes dalam pelaksanaan tes. Kesalahan pengukuran dapat dihitung dengan rumus sebagai berikut.

$$\sigma_E = \sigma_X \sqrt{1-\rho_{xx'}}$$

Keterangan:

σ_E = SEM

σ_X = deviasi standar dari skor total

$\rho_{XX'}$ = koefisien reliabilitas.[30]

Jika menggunakan analisis dengan program Iteman maka nilai SEM dapat langsung dilihat pada skala statistik.Karakteristik

---

[29]Linn, R. L., *Education Measurement* (3^{th} ed.), (New York: Macmilan Publishing Company, 1989) p. 23

[30] Depdikbud, *Pengelola Pengujian bagi Guru Mata Pelajaran*, (Jakarta: Proyek Peningkatan Mutu Sekolah Menengah Umum, 1997) p. 23

butir soal dengan pendekatan klasik di atas memilki keterbatasan yang mendasar antara lain pertama, hasil estimasi parameter targantung pada karakteristik peserta ujian (*group dependen)*. Hal ini berimplikasi pada tingkat kesukaran soal akan menjadi rendah jika tes diujikan pada kelompok peserta tes berkemampuan tinggi, sebaliknya jika tes diujikan pada peserta dengan kemampuan rendah maka tingkat kesukaran tes tersebut menjadi tinggi. Kedua, hasil estimasi kemampuan peserta tergantung pada karakteristik butir soal *(item dependen)*. Keterbatasan ini menyebabkan estimasi kemampuan peserta akan rendah jika soal yang diberikan berada di atas kemampuannya. Mengatasi berbagai kelemahan pada teori tes klasik para ahli pengukuran mengembangkan model teori tes modern atau Teori Respon Butir.

2) Analisis Item menurut *Item Response Theory* (IRT)

Berdasarkan kelemahan-kelemahan yang dimiliki teori tes klasik maka dikembangkanlah teori tes baru, yaitu Teori tes modern atau *Item Response Theory* (IRT) pertama kali dimunculkan oleh Lord dalam disertasinya tahun 1952. Analisis butir soal secara modern yaitu penelaahan butir soal dengan menggunakan IRT. Tujuan utama IRT adalah memberikan kesamaan antara statistik soal dan estimasi kemampuan.

Menurut Hambleton, Swaminathan & Rogers teori respon butir dikembangkan berdasarkan dua postulat, yaitu: 1) prestasi peserta uji pada suatu tes dapat diprediksikan dengan seperangkat faktor yang disebut kemampuan laten (*latens traits*), *trait* adalah dimensi kemampuan seseoraang seperti kemampuan verbal, kemampuan psikometer, kemampuan kognitif, dan sebagainya, dan 2) hubungan antara prestasi uji pada suatu butir tes dan perangkat kemampuan yang mendasarinya sesuai dengan grafik fungsi naik monoton tertentu yang disebut kurva karakteristik butir (*item characteristic curve*)[31]. Kurva ini menggambarkan bahwa semakin tinggi kemampuan peserta uji semakin meningkat pula peluang menjawab benar suatu butir tes.

[31]Hambleton et al., *Fundamentals of Item Response Theory*. (Newbury Park, CA: Sage Publication Inc. 1991) p. 7

Beberapa asumsi yang melandasi teori respons butir yaitu 1) satu dimensi (*unidimensional*), artinya dimensi karakter peserta yang diukur oleh suatu tes tunggal. Asumsi ini sangat sulit untuk dipenuhi karena banyaknya faktor-faktor yang mempengaruhi tes seperti kognitif, kepribadian, dan bahasa. Akan tetapi dapat diasumsikan jika dalam hasil analisis terdapat satu faktor yang lebih dominan dari faktor yang lain, maka dapat dikatakan bahwa tes tersebut unidimensi; dan 2) kebebasan lokal (*local independence*), artinya respon peserta tes terhadap suatu butir tidak berhubungan dengan butir lainnya dalam tes tersebut.

Pada umumnya dalam teori respons butir digunakan model distribusi logistik. Beberapa model logistik dalam teori respon butir, diantaranya: 1) model logistik satu parameter, ditentukan oleh satu karakteristik butir yaitu tingkat kesukaran, 2) model logistik dua parameter ditentukan oleh dua karakteristik butir yaitu tingkat kesukaran dan daya pembeda, dan 3) model logistik tiga parameter ditentukan oleh tiga karakteristik butir yakni tingkat kesukaran, daya pembeda, dan faktor tebakan. Model logistik dua parameter yang digunakan pada analisis penelitian ini menggunakan rumus:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad i = 1, 2, 3, \dots, n$$

Keterangan:

$P_i(\theta)$ = probabilitas menjawab benar butir i oleh peserta berkemampuan θ

θ = kemampuan peserta

e = bilangan transenden yang besarnya mendekati 2,718

n = banyaknya butir dalam tes

a_i = daya pembeda butir i

b_i = tingkat kesukaran butir i

D = faktor skala, 1,7[32]

Model logistik digunakan pada data dengan jawaban butir tes dikotomi. Hal ini senada dengan IRT bentuk tes pilihan ganda

[32] *Ibid., h. 34*

sering digunakan pada penelitian pendidikan dan psikologis dan cocok jika dianalisis menggunakan IRT. Beberapa aspek atau karakteristik butir yang perlu diperhatikan dalam teori tes medren, yaitu:

1) Tingkat Kesukaran

Pada teori respon butir, indeks tingkat kesukaran, b, adalah matriks yang sama dengan keahlian atau sifat. Butir yang memiliki tingkat kesukaran tinggi maka memiliki indeks tingkat kesukaran sulit. Menurut Hambleton & Swaminathan indeks tingkat kesukaran yang baik berada di antara -2 -+2.[33]

2) Daya Beda

Daya beda yang tinggi dapat diartikan bahwa butir tersebut dapat membedakan (diskriminasi) antara peserta ujian dengan berbagai tingkat. Pada teori respon butir, a sebagai simbol dari indeks daya beda yaitu sebuah pengukuran dari daya beda butir. Indeks tersebut terkadang disebut *slope*, karena menunjukkan seberapa tajam kemungkinan perubahan respon yang menjawab benar seperti kemampuan atau kenaikan trait (ciri). Pada teori tes klasik dan teori respon butir yang baik, nilai daya beda yang lebih tinggi menunjukkan indeks diskriminasi yang lebih besar (baik). Menurut Hambleton & Swaminathan, indeks daya beda berada di antara 0-2.[34]

3) Kecocokan Butir dengan Model Logistik (*goodness of fit ststistic*)

IRT dapat direalisasikan hanya bila terdapat kesesuaian antara model yang digunakan dengan data tes. Kecocokan suatu item dengan model dapat dilihat dari nilai chi kuadrat item dibandingkan dengan harga kritik distribusi chi kuadrat sesuai dengan dk item yang bersangkutan pada taraf signifikansi α = 0,01 atau α = 0,05. Butir dikatakan cocok model jika nilai chi kuadrat butir lebih kecil dari harga distribusi chi kuadrat pada nilai kritisnya.

---

[33] *Ibid., h. 36*

[34] *Ibid., 36*

4) Fungsi Informasi

Pada analisis teori tes modern dapat diketahui fungsi informasi tes dan kemampuan peserta didik. Fungsi informasi tes merupakan penjumlahan dari fungsi informasi seluruh butir pada tes tersebut pada tingkat kemampuan θ. Menurut Dali S. Naga fungsi informasi tes akan berubah-ubah menurut nilai θ. Pada nilai θ tertentu, nilai fungsi informasi mencapai maksimum. Titik maksimum berarti bahwa jika butir itu dikerjakan oleh peserta dengan θ tersebut, maka akan diperoleh informasi yang paling tinggi.

Pada model logistik dua parameter, fungsi informasi butir memenuhi persamaan berikut

$$I_i(\theta) = \frac{D^2 a^2 e^{Da(\theta-b)}}{[1+e^{Da(\theta-b)}]^2}$$

Keterangan:

D = konstanta, 1,7

θ = skala kemampuan

e = 2,71828

a = daya beda butir soal

b = tingkat kesukaran butir soal[35]

Setiap butir soal memiliki fungsi informasi dan jumlahnya merupakan fungsi informasi tes tersebut tes sehingga fungsi informasi paket tes akan tinggi jika butir penyusunnya mempunyai fungsi informasi yang tinggi pula. Secara matematis fungsi informasi tes dapat ditulis:

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta)$$[36]

5) Standar Error Pengukuran (SEM)

Kesalahan error pengukuran menurut teori respons butir dinyatakan dengan SEM yang besarnya tergantung pada tingkat kemampuan seseorang dan fungsi informasi tes. Adanya kesalahan yang melekat pada data hasil pengukuran ini disebabkan oleh banyak faktor diantaranya adalah alat ukur itu sendiri, pelaksanan pengukuran, objek pengukuran, dan teknik analisis yang

[35]Dali S. Naga, *Pengantar*, p. 324

[36]Hambleton et al., *Fundamentals*, p. 94

digunakan. Fungsi informasi dengan kesalahan baku pengukuran (*Standar Error of Measurement*/SEM) mempunyai hubungan yang berbanding terbalik kuadratik Makin besar nilai fungsi informasi berarti SEM semakin kecil dan sebaliknya. Jika, fungsi informasi dinyatakan dengan $I(\theta)$ dan kesalahan baku pengukuran dinyatakan dengan $SEM(\theta)$, bentuk hubungan keduanya dirumuskan sebagai berikut:

$$SEM(\theta) = \frac{1}{\sqrt{I(\theta)}}$$ [37]

Kesimpulan

1. Teori tes klasik adalah berupa asumsi-asumsi yang dirumuskan secara matematis yang modelnya disebut skor murni (*true score model).* Teori tes klasik mengasumsikan bahwa setiap orang memilki nilai yang benar (True =T), yang akan diperoleh jika tidak ada kesalahan dalam pengukuran. Terdapat lima asumsi yang digunakan dalam teori tes klasik. Asumsi-asumsi teori tes klasik tersebut dikembangkan dalam berbagai formula yang berguna dalam melakukan pengukuran psikologis. Formula tersebut diantaranya estimasi indeks kesukaran, indeks daya beda, efektifitas distraktor, reliabilitas tes dan validitas.
2. Teori respon butir dikembangkan berdasarkan dua postulat, yaitu: 1) prestasi peserta uji pada suatu tes dapat diprediksikan dengan seperangkat faktor yang disebut kemampuan laten (*latens traits*), *trait* adalah dimensi kemampuan seseoraang seperti kemampuan verbal, kemampuan psikometer, kemampuan kognitif, dan sebagainya, dan 2) hubungan antara prestasi uji pada suatu butir tes dan perangkat kemampuan yang mendasarinya sesuai dengan grafik fungsi naik monoton tertentu yang disebut kurva karakteristik butir (*item characteristic curve*). Kurva ini menggambarkan bahwa semakin tinggi kemampuan peserta uji semakin meningkat pula peluang menjawab benar suatu butir tes.

[37] *Ibid,*, p. 96

Daftar Pustaka

Allen & Yen. 1979. *Introduction to Measurement Theory*. Monterey: McGraw- Hill.

Blerkom, Malcolm L. Van. 2009. *Measurement and Statistic for Teachers*. New York: Routledge, Taylor & Francis Group.

Bott, P. A. 1995. *Testing and Assessment in Occupational and Technical Education*. Boston: Allyn & Bacon.

Dali S. Naga. 1992. *Pengantar Teori Skor*. Jakarta: Gunadarma.

Depdikbud. 1997. *Pengelola Pengujian bagi Guru Mata Pelajaran*. Jakarta: Proyek Peningkatan Mutu Sekolah Menengah Umum.

Depdiknas. 2005. *Peraturan Pemerintah RI Nomor 19, tentang Standar Nasional Pendidikan.* Jakarta: Depdiknas.

---------*Kurikulum 2004 SMA. Pedoman Khusus Pengembangan Silabus dan Penilaian.* Jakarta: Depdiknas.

Djaali & Pudji Muljono. 2008. *Pengukuran dalam Bidang Pendidikan.* Jakarta: PT Grasindo.

Djemari Mardapi. 2006. *Teknik Penyusunan Instrument Tes dan Nontes*. Yogyakarta: Mitra Cendikia Press.

Feld & Brennan, R.L. 2006. *Educational Measurement*. Westport: Greenwood Publishing Group.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. 1991. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publication Inc.

Linn, R. L. 1989. *Education Measurement* 3th ed. New York: Macmilan Publishing Company.

Messick, S. 1989. *Validity.* London: Macmillan Publishers.

Miller, Patrick W. 2008. *Measurement and Teaching*. USA: Patrick W. Miller and Associates.

Saifuddin, Azwar. 2010. *Tes Prestasi. Fungsi Pengembangan Pengukuran Prestasi Belajar*. Yogyakarta: Pustaka Pelajar.

-------. *Reliabilitas dan Validitas*. 2010. Yogyakarta: Pustaka Pelajar.

Suharsimi Arikunto. 2009. *Dasar-dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara.

Sumarna Surapranata. 2004. *Analisis, Validitas, Reliabilitas, dan Interpretasi Hasil Tes Implementasi Kurukulum 2004*. Bandung: PT. Remaja Rosdakarya.

Zaenal Arifin. 2009. *Evaluasi Pembelajaran*. Bandung: PT. Remaja Rosdakarya.