TES SEBAGAI ALAT UKUR PRESTASI AKADEMIK

Suharman
Sekolah Tinggi Agama Islam Negeri Teungku Dirundeng Meulaboh
Email: suharmanalhamid@gmail.com

Abstrak

Proses belajar mengajar yang dilakukan harus selalu melalui proses akhir yaitu evaluasi agar hasil yang ingin dicapai menjadi lebih baik. Salah satu teknik evaluasi yang sering digunakan oleh lembaga pendidikan adalah tes prestasi. Tes prestasi adalah tes yang disusun secara terencana untuk mengungkap performansi maksimal subjek dalam menguasai bahan-bahan atau materi yang telah diajarkan. Dalam menyusun instrumen tes untuk tes prestasi, Intrumen tes harus melalui Validitas dan Reliabilitas Tes. Sehingga tes yang akan disusun bisa sepadan dengan kemampuan seseorang yang akan diberikan tes. Instrumen tes yang disusun tidak boleh terlalu jauh di bawah atau di atas kemampuan peserta tes, dan tingkat kesukaran item-item soal sebaiknya berada pada kategori sedang. Dalam kajian mengenai tes terdapat dua pendekatan yang dapat digunakan yaitu pendekatan secara klasik (Classical Test Theory/CTT) dan pendekatan secara modern yang berdasarkan pada Item Response Theory (IRT).

Kata Kunci: *Penilaian, Tes, CTT, IRT*

Abstract

Teaching and learning process carried out must always go through the final process of evaluation, so that the results to be achieved are better. One of evaluation technique that is often used by educational institutions is achievement tests. Achievement tests are tests that are arranged in a planned manner to reveal the subject's maximum performance in mastering the material or material that has been taught. In compiling test instruments for achievement tests, the test instrument must go through the Validity and Reliability Test. So the, test that will be prepared can be commensurate with the ability of someone who will be given a test. The test instruments that are prepared should not be too far below or above the ability of the test participants, and the level of difficulty of the items should be in the medium category. In the study of the test there are two approaches that can be used, namely the classical approach (CTT) and the modern approach based on Item Response Theory (IRT).

Keywords : Evaluation, Tests, CTT, IRT

A. Pendahuluan

Proses belajar mengajar yang dilakukan harus selalu diperbaiki agar hasil yang ingin dicapai menjadi lebih baik. Salah satu upaya dalam meningkatkan kualitas proses hasil dan prestasi akademik sebagai bagian dari peningkatan kualitas pendidikan adalah melalui sistem penilaian.

Salah satu teknik penilaian yang sering digunakan oleh lembaga pendidikan adalah tes. Sumardi Suryabrata dalam Chabib Toha (2003), menyatakan bahwa tes merupakan pertanyaan-pertanyaan yang harus dijawab dan atau perintah-perintah yang harus dijalankan, yang mendasarkan harus bagaimana testee menjawab pertanyaan atau melakukan perintah-perintah itu, penyelidik mengambil kesimpulan dengan cara membandingkan dengan standar atau testee lainnya.

Dalam melaksanakan penilaian, penilai harus memahami berbagai kegiatan teknis dalam menentukan metode dan format penilaian yang dapat digunakan untuk mendapatkan informasi yang dibutuhkan. Informasi tersebut diperlukan dalam menafsir dan menetapkan keputusan untuk kepentingan pendidikan. Penilai membutuhkan ketrampilan dalam mengidentifikasi dan memahami berbagai macam perspektif penilaian, baik penilaian kontekstual dan proses maupun penilaian hasil. Karena peniliaian merupakan pusat kontrol keberhasilan program pendidikan.

B. Pengertian Tes

Tes pada umumnya dimaksudkan untuk mengukur aspek-aspek perilaku manusia, seperti aspek pengetahuan (kognitif), sikap (afektif), maupun aspek keterampilan (psikomotorik). Bidang kognitif diukur melalui uji tes, bidang afektif diukur melalui kuesioner, wawancara, dan pengamatan, serta bidang psikomotor diukur melalui perbuatan dan pengamatan. (Naga, 1992).

Tes merupakan alat atau prosedur yang digunakan untuk mengetahui atau mengukur sesuatu dalam suasana, dengan cara dan aturan-aturan yng sudah ditentukan (Arikunto, 2008). Tidak jauh berbeda dengan Arikunto, Chaplin (2005) menyatakan bahwa tes adalah satu perangkat pertanyaan yang sudah dibakukan, yang dikenakan pada seseorang dengan tujuan untuk mengukur perolehan atau bakat pada suatu bidang tertentu.

Selanjutnya Depdiknas (2003) mendefinisikan bahwa tes adalah himpunan pertanyaan yang harus dijawab atau pernyataan-pernyataan yang harus dipilih dan ditanggapi, atau tugas-tugas yang harus dilakukan oleh orang yang dites dengan tujuan untuk mengukur suatu aspek (perilaku) tertentu dari orang yang di tes. Tes tersebut memenuhi empat aspek yaitu kegunaan, mungkin dikerjakan, legal atau sah, dan ketelitian. Tes itu merupakan hasil perakitan item-item soal yang telah dibakukan melalui proses analisis item, serta diadministrasikan, diskor, dan diinterpretasikan secara baku.

Beberapa pendapat para ahli lainnya tentang pengertian tes seperti yang disampaikan oleh Azwar (2007)

diantaranya, Anne Anastasi (2006) yang mengatakan bahwa tes pada dasarnya merupakan suatu pengukuran yang objektif dan standar terhadap sampel prilaku. Frederick G. Brown (1976) menyebutkan pengertian tes sebagai prosedur yang sistematik guna mengukur sampel prilaku seseorang. Sedangkan Lee J. Cronbach dalam buku *Essentials of psychological Testing* (1970) menyatakan pengertian tes yaitu, *"...a systematic procedure for observing a person"s behavior and describing it with the aid of a numerical scale or a catagory system"*.

Dari beberapa batasan mengenai tes tersebut di atas, Azwar (2007) menarik beberapa kesimpulan mengenai pengertian tes.

1. Tes adalah prosedur yang sistematik. Maksudnya (a) aitem-aitem dalam tes disusun menurut cara dan aturan tertentu, (b) prosedur administrasi tes dan pemberian angka (*scoring*) terhadap hasilnya harus jelas dan dspesifikasikan secara terperinci, dan (c) setiap orang yang mengambil tes harus mendapatkan aitem-aitem yang sama dalam kondisi yang sebanding.
2. Tes berisi sampel prilaku. Artinya (a) betapapun panjangnya suatu tes, aitem yang ada di dalamnya tidak akan dapat mencakup seluruh isi materi yang mungkin di tanyakan, dan (b) kelayakan suatu tes tergantung dari sejauhmana aitem-aitem dalam tes itu mewakili secara representatif kawasan (*domain*) perilaku yang diukur.
3. Tes mengukur prilaku. Artinya aitem-aitem dalam tes menghendaki agar subjek menunjukkan apa yang diketahui atau apa yang telah dipelajari subjek dengan cara menjawab pertanyaan-pertanyaan atau mengerjakan tugas-tugas yang dikehendaki oleh tes.

Suatu tes dikatakan baik manakala mampu memberikan hasil ukur yang cermat dan akurat. Oleh karena itu Suryabrata, (2006), mengemukakan syarat-syarat tes yang baik adalah : (1) tes harus valid, artinya tes tersebut hanya mengukur satu aspek saja atau satu domain saja sehingga tepat mengukur apa yang hendak di ukur, (2) tes harus reliabel, yaitu ajek atau konsisten, (3) tes harus standar, artinya setiap peserta tes (*testee*) harus mendapat perlakuan yang sama baik mengenai materi tes, penyelenggaraan, pemberian skor, dan interpretasi hasil tes sehingga seorang *testee* yang mendapat skor tertentu di suatu tempat akan mendapat skor yang sama di tempat lain, (4) tes harus objektif, yaitu penilaian yang dilakukan oleh pemberi tes (*tester*) yang satu dengan yang lain akan sama untuk satu *testee*, (5) tes harus bersifat diskriminatif, artinya tes harus dapat mengungkapkan perbedaan suatu gejala yang terdapat pada setiap individu.

Lebih lanjut Hayat & Setiadi (1998) menyatakan bahwa Tes yang baik dapat didefinisikan sebagai sekumpulan item-item yang berkualitas (valid) yang telah dikalibrasi dan dipilih untuk membentuk satu instrumen pengukuran.

Berdasarkan bentuknya, tes dapat diklasifikasikan ke dalam 2 (dua) bentuk, yaitu:

1. Objektif yang meliputi, (a) pilihan ganda, (b) bentuk item dua pilihan jawaban (benar-salah, ya-tidak), dan (c) tes menjodohkan.
2. Non-Objektif yang meliputi: (a) isian atau melengkapi, (b) jawaban singkat atau pendek, dan (c) item uraian, (Depdiknas, 2003).

Berbagai bentuk tes tersebut di atas mempunyai keunggulan dan kelemahannya masing-masing. Misalnya bentuk tes uraian, bentuk tes ini memiliki keunggulan berupa dapat mengukur kemampuan siswa (peserta didik) dalam hal menyajikan jawaban terurai secara bebas, mengorganisasikan pikirannya, mengemukakan pendapatnya, dan mengekspresikan gagasan-gagasan dengan menggunakan kata-kata atau kalimatnya sendiri. Disamping keunggulannya tersebut, bentuk tes uraian juga memiliki beberapa kelemahan diantaranya, jumlah materi atau pokok bahasan yang dapat ditanyakan relatif terbatas, waktu untuk memeriksa jawaban siswa cukup lama, penskorannya relatif subjektif, dan tingkat reliabilitasnya relatif lebih rendah dibanding dengan item bentuk pilihan ganda, karena reliabilitas pada item bentuk uraian sangat bergantung pada penskoran tes.

Berbeda halnya dengan tes uraian, bentuk tes pilihan ganda memiliki beberapa keunggulan lain, yaitu dapat mengukur berbagai jenjang kognitif (dari ingatan sampai dengan evaluasi), mudah dalam penskorannya, cepat, objektif dan dapat mencakup ruang lingkup materi yang luas dalam suatu tes untuk suatu jenjang pendidikan. Bentuk tes pilihan ganda ini juga sangat tepat digunakan untuk ujian dengan jumlah pesertanya sangat banyak atau yang sifatnya massal, sedangkan hasilnya harus segera diumumkan. Namun demikian tes pilihan ganda ini, memiliki beberapa kelemahan diantaranya, memerlukan waktu yang relatif lama dalam pembuatan soal, sulit membuat distraktor yang homogen dan berfungsi, dan terdapat peluang untuk menebak kunci jawaban (guessing).

Terlepas dari berbagai kelemahannya, tes bentuk pilihan ganda ini telah banyak digunakan di hampir seluruh pelosok Indonesia, bahkan untuk Ujian Akhir Nasional dan Ujian Seleksi Penerimaan mahasiswa baru di semua Perguruan Tinggi yang pelaksanaannya dikelola oleh Pemerintah.

C. Tes Prestasi

Tes prestasi adalah tes yang disusun secara terencana untuk mengungkap performansi maksimal subjek dalam menguasai bahan-bahan atau materi yang telah diajarkan. Dalam kegiatan pendidikan formal di kelas, tes prestasi belajar dapat berbentuk ulangan-ulangan harian, tes formatif, tes sumatif, bahkan ebtanas dan ujian-ujian masuk perguruan tinggi, (Azwar, 2007).

Tes prestasi belajar dapat dibagi menjadi dua jenis, yaitu tes kemampuan

(*power test*) dan tes kecepatan (*speed test*). Perbedaan tes kemampuan dan tes kecepatan adalah : (1) Prinsip dari *power test* adalah tidak adanya batasan waktu dalam mengerjakan tes. Jika waktu pengerjaan tes tidak dibatasi, maka hasil tes benar-benar mengungkapkan kemampuan seseorang secara maksimal atau menyeluruh. Pembatasan waktu dalam mengerjakan tes, kemungkinan akan menyebabkan orang menjadi tidak dapat menunjukkan kemampuan yang dimilikinya secara maksimal, sehingga skor yang dihasilkan tidak menunjukkan seluruh kemampuan yang sebenarnya dari peserta tes tersebut. (2) Pada *speed test* yang diukur ialah kecepatan di dalam memikirkan atau mengerjakan suatu tes atau tugas. Tes tersebut biasanya relatif mudah, sehingga yang diukur benar-benar kecepatan bekerja atau berpikir seseorang. (Puspendik, 2006).

Dalam menyusun instrumen tes untuk tes prestasi, hal pertama yang harus diperhatikan adalah bagaimana instrumen tes yang akan disusun tersebut bisa sepadan dengan kemampuan seseorang yang akan di berikan tes. Untuk hal ini, Hayat (2000), menyatakan bahwa pada tes prestasi belajar sebuah tes dengan jumlah item yang banyak dan seluruh itemnya bertaraf kesukaran sedang (*on-target*) bagi orang yang menempuh tes, akan mendapat informasi yang lebih teliti mengenai orang yang diukur jika dibandingkan dengan tes yang itemnya sedikit dan tingkat kesukarannya tidak *matching* dengan kemampuan peserta tes (*off-target*). Hal ini memberi gambaran bahwa instrumen tes yang disusun tidak boleh terlalu jauh di bawah atau di atas kemampuan peserta tes, dan tingkat kesukaran item-item soal sebaiknya berada pada kategori sedang. Sehingga dengan demikian instrumen tes yang disusun nantinya dapat berfungsi dengan baik.

Adapun fungsi tes prestasi belajar seperti yang disampaikan Ebel (1991), adalah sebagai berikut : (a) Fungsi utama tes prestasi adalah untuk mengukur keberhasilan siswa dalam belajar, (b) Tes juga dapat membantu guru dan instruktur dalam membuat nilai yang akurat dan bermakna, (c) Tes prestasi belajar juga berfungsi untuk memotivasi dan mengarahkan siswa dalam belajar. Siswa (peserta didik) akan cendrung belajar lebih giat bila mereka dihadapkan pada waktu-waktu dimana ujian akan berlangsung. Dengan kata lain, mereka akan belajar lebih serius pada materi-materi yang menurut pemikiran mereka akan diujikan pada saat berlangsungnya tes.

D. Validitas dan Reliabilitas Tes

1. Validitas Tes

Validitas sebuah tes memberitahukan kepada kita tentang apa yang bisa disimpulkan dari skor-skor tes. Dalam kaitan ini kita seharusnya waspada menerima tes sebagai indeks dari apa yang diukur. (Anastasi & Urbina, 2006). Validitas menurut Cronbach adalah sebagai proses dimana pembuat tes atau pengguna tes mengumpulkan bukti-bukti untuk

mendukung jenis kesimpulan yang akan diambil dari skor tes, (Crocker & Algina, 1986).

Suatu tes dikatakan memiliki validitas yang tinggi apabila tes tersebut dapat memberikan hasil ukur yang tepat dan akurat sesuai dengan maksud dikenakannya tes tersebut. Sebaliknya bila hasil ukur yang diperoleh dari tes tersebut tidak sesuai dengan tujuan dilaksanakannya, maka tes tersebut dikatakan tidak mempunyai validitas yang tinggi.

Hasil estimasi validitas suatu pengukuran dinyatakan secara empirik oleh suatu koefisien yang disebut dengan koefisien validitas. Koefisien validitas dapat dinyatakan oleh korelasi antara distribusi skor tes yang bersangkutan dengan distribusi skor suatu kriteria. Kriteria ini dapat berupa skor tes lain yang mempunyai fungsi ukur sama dan dapat pula berupa ukuran- ukuran lain yang relevan, (Azwar, 2007).

Bila skor tes diberi simbul X dan skor kriteria diberi simbul Y, maka koefisien validitasnya, yaitu r_{xy} . Dan salah satu tekhnik yang dapat digunakan untuk mengukur validitas tes adalah dengan menggunakan korelasi *product-moment* dengan simpangan yang dikemukan oleh Pearson sebagai berikut :

$$r_{xy} = \frac{\sum X}{\sqrt{\left(\sum x^2\right)\left(\sum y^2\right)}} \qquad (2.1)$$

Dimana :

r_{xy} Koefisien korelasi antara variabel x dan variabel y.

$\sum X$ jumlah perkalian antara x dan y

x^2 kuadrat dari x

y^2 kuadrat dari y

Pada tahun 1954, *The American Psychological Association* (APA) melalui *Technical Recommendation for Psychological Test and Diagnostic Techniques* merumuskan empat macam validitas, yaitu validitas isi (*content validity*), validitas konstruk (*construct validity*), validitas prediksi (*predictive validity*) dan validitas konkuren (*concurrent validity*).

Keempat macam validitas ini selanjutnya dijabarkan oleh Arikunto (2008) dengan menyebutnya sebagai empat *face validity*, yaitu validitas isi, validitas konstruk, validitas "ada sekarang", dan validitas prediksi:

a. Validitas Isi (*content validity*)

Sebuah tes dikatakan memiliki validitas isi apabila mengukur tujuan khusus tertentu yang sejajar dengan materi atau isi pelajaran yang diberikan. Oleh karena materi yang diajarkan tertera dalam kurikulum maka validitas isi ini sering juga disebut dengan validitas kurikuler.

Validitas isi dapat diusakan tercapai sejak saat penyusunan dengan cara merinci materi kurikulum atau materi buku pelajaran.

b. Validitas Konstruksi (*construct validity*)

Sebuah tes dikatakan memiliki validitas konstruksi apabila item-item yang membangun tes tersebut mengukur setiap aspek berpikir seperti yang disebutkan dalam Tujuan Instruksional Khusus (sekarang disebut indikator). Dengan kata

lain jika item-item soal mengukur aspek berpikir tersebut sudah sesuai dengan aspek berpikir yang menjadi tujuan instruksional (indikator).

Seperti halnya validitas isi, validitas konstruksi dapat diketahui dengan cara merinci dan memasangkan setiap item soal dengan setiap aspek dalam Tujuan Instruksional Khusus (indikator).

Apabila hal tersebut di atas tidak dapat dilakukan, maka cara yang paling sederhana adalah dengan melalukan analisis faktor konfirmatori. Analisis ini dilakukan untuk mengetahui validitas kontruk sebuah tes, sehingga tes yang dibagun benar-benar mengukur suatu aspek yang hendak diukur. Analisis faktor konfirmatori dapat dilakukan dengan bantuan program *LISREL*.

Dalam analisis dengan program LISREL, kriteria yang digunakan untuk dapat mengetahui nilai validitas konstruk adalah pada besar kecilnya *loading factors* yang diperoleh oleh setiap item pada saat dilakukan pengukuran.

Para ahli pengukuran mengungkapkan beberapa batasan nilai *loading factors* yang dapat digunakan untuk menentukan valididtas kostruk suatu tes diantaranya, Rigdon dan Ferguson, 1991, serta Doll, Xia Torkzadeh, 1994, yang menyatakan bahwa validitas konstruk yang baik adalah yang memiliki :

1) Nilai t muatan faktor (*loading factors*) lebih besar atau sama dengan nilai kritis (*t-values* □ 1.96 atau untuk praktisnya □ 2.00).
2) Muatan faktor standar (*standardized loading factors*) lebih besar atau sama dengan 0.70, (*standardized* □ 0.70), (Wijanto, 2008).

c. Validitas "ada sekarang" (*concurrent validity*)

Validitas ini lebih umum dikenal dengan validitas emperis. Sebuah tes dikatakan memiliki validitas emperis jika hasilnya sesuai dengan pengalaman. Jika ada kata "sesuai" tentu ada dua hal yang dipasangkan. Dalam hal ini hasil tes dipasangkan dengan hasil pengalaman. Pengalaman selalu mengenai hal yang telah lampau sehingga data pengalaman tersebut sekarang sudah ada, makanya validitas ini disebut validitas "ada sekarang" atau *concurrent*).

Dalam membandingkan hasil sebuah tes maka diperlukan suatu kriterium atau alat banding. Maka hasil tes merupakan sesuatu yang dibandingkan.

d. Validitas prediksi (*predictive validity*)

Memprediksi artinya meramal, dengan meramal selalu mengenai hal yang akan datang, jadi sekarang belum terjadi. Sebuah tes dikatakan memiliki validitas prediksi atau validitas ramalan apabila mempunyai kemampuan untuk meramalkan apa yang akan terjadi pada masa yang akan datang.

Validitas prediksi sangat penting artinya bila tes yang dimaksudkan berfungsi sebagai prediktor untuk memprediksi suatu keberhasilan di masa yang akan datang. Sebagai contoh berdasarkan hasil

tes seleksi penerimaan mahasiswa baru, peserta tes yaitu calon mahasiswa yang memiliki nilai tinggi pada tes seleksi diperkirakan akan berhasil dengan baik ketika mereka belajar di perguruan tinggi tersebut. Jika perkiraan ini tepat, maka tes seleksi tersebut dapat dikatakan memiliki validitas prediksi yang baik. Sebaliknya jika perkiraan tersebut tidak tepat, maka tes seleksi yang dilaksanakan sebelumnya tidak memiliki validitas prediksi yang baik.

Validitas prediksi dapat ditentukan dengan mengetahui hubungan antara skor tes sebagai prediktor dengan hasil prestasi belajar atau ukuran keberhasilan lainnya. Hasil prestasi belajar dan keberhasilan lain ini berfungsi sebagai skor kriteria. Koefisien korelasi antara skor tes dan skor kriteria merupakan petunjuk mengenai saling adanya hubungan antara skor keduanya dan dapat disebut dengan koefisien validitas prediksi. Apabila koefisien yang diperoleh ini adalah dari suatu penelitian dengan kelompok sampel yang representatif, maka tes yang telah teruji validitasnya akan mempunyai fungsi prediksi yang sangat berguna dalam memprediksi hasil prestasi belajar peserta tes pada masa akan datang.

Besarnya nilai koefisien validitas prediksi skor tes terhadap skor kriteria yang dianggap baik dalam memprediksi hingga saat ini masih terjadi perdebatan dikalangan pengembang tes. Sebagai contoh, *Scholastic Aptitude Test* (SAT) yang dikembangkan oleh *College Board* atau badan perguruan tinggi di Amerika Serikat memiliki koefisien validitas prediksi hanya sebesar 0.37, (Nairn dalam Weitzman, 1982). Sehingga keberadaan SAT sebagai tes seleksi sering dikritik. Demikian pula halnya dengan GRE (*Graduate Record of Examination*) dan tes-tes seleksi sejenisnya yang sering dikritik karena nilai prediksinya yang rendah terhadap indeks prestasi mahasiswa pada perguruan tinggi profesional (Nunnaly & Bernstein, 1994).

Namun demikian, terlepas dari tinggi rendahnya koefisien validitas prediksi, SAT atau Tes Bakat Skolastik (TBS) dan tes seleksi penerimaan lainnya tetap digunakan di hampir seluruh perguruan tinggi di Indonesia. Dan besarnya nilai koefisien validitas prediksi yang dianggap memuaskan adalah jika melebihi dari angka 0,30 (Azwar, 2000).

2. Reliabilitas Tes

Reliabilitas adalah kestabilan skor yang diperoleh orang yang sama ketika diuji ulang dengan tes yang sama pada situasi yang berbeda atau dari pengukuran ke pengukuran lainnya, (Puspendik, 2003). Selanjutnya Anastasi & Urbina (2006) mengemukakan bahwa realibilitas merujuk pada konsistensi skor yang dicapai oleh orang yang sama ketika mereka diuji-ulang dengan tes yang sama pada kesempatan berbeda, atau dengan seperangkat item-item ekuivalen (equivalent items) yang berbeda, atau dalam kondisi pengujian yang berbeda.

Reliabilitas alat ukur menurut Suryabrata (2006) menunjukkan sejauh mana hasil pengukuran dengan alat ukur tersebut dapat dipercaya. Hal ini ditunjukkan oleh taraf keajegan

(konsistensi) skor yang diperoleh oleh para subjek yang diukur dengan alat yang sama, atau diukur dengan alat yang setara pada kondisi yang berbeda.

Gambaran yang benar-benar ajeg pada sebuah instrumen tes memang sangat sulit untuk diperoleh karena unsur kejiwaan manusia itu sendiri. Manusia (sebagai peserta tes) mempunyai kemampuan, kecakapan, sikap dan lain sebagainya yang cenderung tidak tetap (berubah-rubah) dari waktu ke waktu. Di samping itu ada beberapa faktor ekternal lain yang dapat mempengaruhi keajegan (reliabilitas) tes seperti yang dikemukan oleh Gulford dalam Arvyaty (2005) adalah, (1) Jumlah item dalam suatu tes, semakin banyak item semakin reliabel tes, (2) Waktu untuk mengerjakan tes, semakin lama semakin reliabel tes, (3) ketergantungan suatu item dengan item yang lainnya dalam suatu tes akan mengurangi tingkat reliabilitas tes, (4) semakin objektif penskoran hasil tes, semakin reliabel suatu tes, (5) kemungkinan menebak dalam menjawab item-item pada tes, (6) semakin homogen materi tes semakin reliabel suatu tes.

Tinggi rendahnya reliabilitas, secara emperik ditunjukkan oleh suatu angka yang disebut koefisien reliabilitas. Pada awalnya, tinggi rendahnya reliabilitas dicerminkan oleh tinggi rendahnya korelasi antara dua distribusi skor dari dua alat ukur yang paralel yang dikenakan pada sekelompok individu yang sama (Azwar, 2007).

Selanjutnya untuk melakukan estimasi reliabilitas suatu tes dapat dilakukan dengan beberapa pendekatan umum, diantaranya metode tes-ulang (*test-retest method*), metode tes sejajar (*equivalent method*), metode konsistensi internal (*internal consistency method*), dan metode belah-dua (*split-half method*). Sebagian para ahli berpendapat bahwa pendekatan split-half merupakan bagian dari pendekatan internal concsstency.

a. Metode tes ulang (*test-retest method*)

Metode ini menunjukkan konsistensi pengukuran dari waktu ke waktu dan menghasilkan koefisien reliabilitas yang sering disebut koefisien stabilitas. Prinsip estimasinya adalah dengan menggunakan suatu instrument pengukur dua kali dengan tenggang waktu tertentu terhadap sekelompok subjek yang sama.

Kelemahan metode ini adalah kurang praktisnya pengenaan tes dua kali dan besarnya kemungkinan terjadi efek bawaan (*carry-over-effects*) dari pengenaan tes pertama ke kedua. (Azwar, 2007).

b. Metode bentuk paralel (*equivalent method*)

Pada metode tes ini digunakan dua buah tes yang mempunyai kesamaan tujuan, tingkat kesukaran dan susunan, tetapi item-item soalnya berbeda. Kelemahan dalam menggunakan metode ini adalah pengetes memiliki beban yang berat karena harus membuat dua instrumen tes yang setara.

c. Metode belah dua (*split-half method*)

Pada metode ini, ada dua cara untuk membelah item soal yaitu: (a) membelah

item menjadi item-item genap dan ganjil selanjutnya disebut belahan ganjil-genap, dan (b) membelah atas item-item awal dan item-item akhir yaitu separo jumlah pada nomor-nomr awal dan separo pada nomor-nomor akhir yang selanjutnya disebut belahan awal-akhir (Arikunto, 2008).

d. Konsistensi intenal (*internal consistency*)

Estimasi reliabilitas dengan pendekatan konsistensi internal didasarkan pada data dari sekali pengenaan satu bentuk alat ukur pada sekelompok subjek (*single trial administrartion*) Azwar (2007).

Tabel 2. 1 Metode Penentuan Reliabilitas

Bentuk reliabilitas	**Prosedur untuk memperoleh**
Test-retest method product moment dan korelasi intra kelas	Tes yang sama disajikan 2 kali kepada peserta tes yang sama dalam waktu yang berbeda dan kemudian hasilnya dikorelasikan.
equivalent method product moment dan korelasi intra kelas	Dua tes yang sama (paralel) disajikan pada peserta tes yang sama dalam waktu yang relatif tidak lama, kemudian hasilnya dikorelasikan.
Split-half method persamaan split-half dan persamaan Spearman-Brown	Tes yang dibelah dua disajikan kepada peserta tes kemudian hasilnya dikorelasikan antara dua belahan tersebut
Internal Consistency - Koefisien Alpja - Kuder-Richardson (KR-20) - Kuder-Richardson (KR-21)	Tes diberikan sekali, lalu digunakan persamaan. Tes diberikan sekali, lalu digunakan persamaan. Tes diberikan sekali, lalu digunakan persamaan.

Sumber: Surapranata (2006)

Secara statistik banyak perssamaan (rumus) yang dapat digunakan untuk menghitung besarnya koefisien atau indeks reliabilitas suatu tes. Dan rumus umum yang paling sering di pakai adalah rumus *Alpha Cronbach, Kuder-Richardson KR-20 dan KR-21*, yaitu:

1) Alpha-Cronbach

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum s_i^2}{s_t^2}\right] \qquad (2.2)$$

dimana:

α koefisien reliabilitas;
k jumlah item tes;
s_i^2 varians skor setiap item;
s_t^2 varians skor total.

2) Kuder-Richardson-20 (KR-20)

$$\rho = \frac{k}{k-1}\left[\frac{s_t^2 - \sum p_i q_i}{s_t^2}\right] \qquad (2.3)$$

dimana :

ρ koefisien reliabilitas;

k jumlah item tes;

p_i proporsi subyek menjawab benar item i; dan $q_i = 1 - p_i$;

s_t^2 varians skor total.

3) Kuder-Richardson-21 (KR-21)

$$\rho = \frac{k}{k-1}\left[1 - \frac{M(k-M)}{k\, s_t^2}\right] \qquad (2.4)$$

dimana :

ρ koefisien reliabilitas;

k jumlah item tes;

M mean skor total;

s_t^2 varians skor total.

E. Teori Tes Klasik dan Modern

Dalam kajian mengenai tes terdapat dua pendekatan yang dapat digunakan yaitu pendekatan secara klasik (*Classical Test Theory/CTT*) dan pendekatan secara modern yang berdasarkan pada *Item Response Theory* (*IRT*)

1. Teori Tes Klasik (*Classical Test Theory atau CTT*)

Sejak beberapa dekade lalu, teori tes klasik (*classical test theory*) telah mendominasi dan banyak berjasa dalam dunia pengukuran. Di antara konsep-konsep yang berdasarkan teori tes klasik yang sangat terkenal dan sangat berguna adalah formula-formula yang dikembangkan oleh *Kuder-Richardson*, formula *Spearman-Brown*, formula *error standar* dalam pengukuran dan lain-lain. Bahkan hampir keseluruhan formula reliabilitas dan validitas yang kita kenal sekarang ini dikembangkan atas konsep teori tes klasik (Azwar, 2007).

Sampai saat ini, teori tes klasik masih banyak di gunakan untuk menganalisis data-data penelitian yang sifatnya sederhana karena teori tes klasik memiliki beberapa kelebihan di antaranya: (1) murah atau tidak membutuhkan banyak biaya, (2) mudah dilaksanakan, (3) sederhana, (4) familier, dan (5) sampel yang dibutuhkan dalam jumlah kecil, Safari (2005).

Dalam perkembangannya teori tes klasik ini juga didasari pada beberapa asumsi. Menurut Suryabrata dalam Budiyono (2005), ada tujuh asumsi pada teori tes klasik, yaitu: (1) skor yang diperoleh peserta tes terdiri dari skor sebenarnya (*true score*) dan kesalahan pengukuran, (2) nilai harapan skor yang diperoleh sama dengan skor sebenarnya, (3) skor yang sebenarnya dan kesalahan pengukuran tidak berkorelasi, (4) kesalahan pengukuran pada dua tes yang mengukur kemampuan yang sama tidak berkorelasi, (5) pada dua tes yang mengukur kemampuan yang sama, kesalahan pengukuran pada tes pertama tidak berkorelasi dengan skor sebenarnya pada tes kedua, (6) dua tes yang menghasilkan skor yang memenuhi kelima asumsi pertama disebut *parallel tests* jika skor sebenarnya dan variasi kesalahan pengukuran yang diperoleh peserta tes

sama, dan (7) dua tes yang menghasilkan skor yang memenuhi kelima asumsi pertama disebut *essentially t-equivalent test* jika selisih skor sebenarnya yang diperoleh peserta tes pada tes pertama dan tes kedua merupakan bilangan konstan.

Sayangnya, dari beberapa kelebihan dan asumsi-asumsi yang telah di jelaskan di atas, ternyata dalam teori tes klasik terdapat beberapa keterbatasan yang kemudian menjadi permasalahan dalam pengembangan tes. Sebagai contoh, indeks kesukaran dan indeks daya beda (indeks diskriminasi) dalam teori tes klasik merupakan karakteristik item yang sangat bergantung pada kelompok sampel (*group-dependent*). Item akan nampak mudah bila kelompok yang dikenai tes rata-rata berkemampuan tinggi, sebaliknya bila kelompok yang dikenai tes berkemampuan rendah maka item tersebut akan kelihatan sulit serta memiliki tingkat kesukaran yang tinggi.

Secara rinci terdapat 4 keterbatasan teori tes klasik seperti yang dikemukakan oleh Hambleton, Swaminathan & Rogers (1991) sebagai berikut:

1) Indeks kesukaran item item dan indeks daya beda item (*discriminating power index*) bergantung pada kelompok sample (*group dependent*).
2) Koefisien reliabilitas dan validitas menjadi tinggi bila taraf kemampuan kelompok sampel heterogen (bervariasi tinggi). Sebaliknya koefisien reliabilitas dan validitas menjadi rendah bila kemampuan kelompok sampel cenderung seragam.
3) Asumsi kesetaraan (disamakan) terhadap error pengukuran bagi subyek-subyek yang dikenai tes. Sedangkan subyek ada yang konsisten dan ada yang tidak konsisten dalam menjawab item-item item.
4) Pada dasarnya pengujian tes melalui metode tes paralel, sulit untuk dilaksanakan bahkan dipastikan tidak ada tes paralel yang benar-benar setara.

Adapun pendekatan yang digunakan dalam teori tes klasik yaitu dengan cara menghitung tingkat kesukaran (*proportion correct*), indeks daya beda item (*point biserial*) , dan kehandalan atau keberfungsian distraktor (*proportion endorsing*).

a. Tingkat Kesukaran Item

Pada analisis item 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. Namun demikian cara yang paling mudah dan paling umum digunakan adalah skala rata-rata atau proporsi menjawab benar atau *proportion correct (p),* yaitu jumlah peserta tes yang menjawab benar pada item yang dianalisis dibandingkan dengan peserta tes seluruhnya, (Hayat, Surapranata, & Suprananto, 1999).

Persamaan yang digunakan untuk menentukan tingkat kesukaran dengan proporsi menjawab benar adalah :

$$p = \frac{\sum x}{S_m N} \qquad (2.5)$$

Dimana :

p proporsi menjawab benar / tingkat kesuakaran;

Σx banyaknya peserta tes yang menjawab benar;

S_m skor maksimum;

N jumlah peserta tes.

Sebenarnya dalam teori tes klasik, tingkat kesukaran item dapat dikatakan sebagai tingkat kemudahan karena semakin tinggi indeks tingkat kesukaran yang diperoleh oleh satu item/soal maka item tersebut semakin mudah, begitu juga sebaliknya item yang memiliki indeks tingkat kesukaran makin rendah maka item tersebut semakin sulit. Besarnya tingkat kesukaran berkisar antara 0 sampai dengan 1.

Tingkat kesukaran biasanya dibedakan menjadi tiga kategori. Item yang memiliki $p < 0.3$ biasanya disebut dengan item sukar. Item yang memiliki $p > 0.7$ biasanya disebut dengan item mudah. Dan item yang memiliki p antara 0.3 sampai dengan 0.7 biasanya disebut sebagai item sedang, seperti tampak pada tabel dibawah ini:

Tabel 2. 2 Kriteria Tingkat Kesukaran

Kriteria Tingkat Kesukaran (p)	Kategori Item
$p > 0.7$	Mudah
$0.30 \leq p \leq 0.70$	Sedang
$p > 0.30$	Sukar

Sumber : Hayat, et.al., 1999.

Item item yang memiliki indeks tingkat kesukaran mendekati 0 atau 1 maka item item tersebut dikategorikan ekstrim. Item item yang ekstrim mudah dan ekstrim sulit tidak memberikan informasi yang berguna bagi sebagian besar peserta tes. Oleh sebab itu, item item seperti ini kemungkinan distribusi jawaban pada alternatif jawaban ada yang tidak memenuhi syarat (Hayat *et.al*, 1999).

b. Daya Beda Item

Daya beda item atau daya pembeda item adalah kemampuan sesuatu item untuk membedakan antara siswa yang berkemampuan tinggi (pandai) dengan siswa yang berkemampuan rendah (Arikunto, 2008). Dan lebih rinci lagi Crocker & Algina dalam bukunya *Introduction To Classical & Modern Test Theory* (1986) menyebutkan bahwa secara umum daya pembeda item merupakan kemampuan suatu item dalam membedakan kelompok aspek yang diukur sesuai dengan perbedaan yang ada dalam kelompok itu. Parameter daya beda item disebut sebagai indeks daya beda yang hanya dapat diaplikasikan pada item yang bersifat dikotomus.

Angka yang menunjukkan besarnya daya beda disebut indeks diskriminasi disingkat dengan *D*. Dan besarnya angka yang menunjukkan daya beda item berkisar antara -1 sampai dengan +1. Tanda negatif menunjukkan bahwa peserta tes yang berkemampuan rendah dapat menjawab benar item tersebut sedangkan peserta tes yang berkemampuan tinggi menjawab salah. Dengan demikian dapat disebutkan bahwa daya beda item sama dengan

validitas item.

Adapaun Indeks daya beda item yang termasuk sudah dapat membedakan kelompok yang berkemampuan tinggi dengan kelompok yang berkemampuan rendah adalah di atas 0.30 (Surapranata, 2006). Item yang memiliki validitas di atas 0.30 merupakan item yang baik (Nunnally dalam Surapranata, 2006). Pendapat yang sama juga dikemukakan oleh Nitko (1983), item yang diterima adalah item yang memiliki indeks daya beda di atas 0.30, direvisi apabila memiliki indeks daya beda berada antara 0.10 sampai dengan 0.29, sedangkan item yang memiliki indeks daya beda di bawah 0.10 akan ditolak. Sementara Fernandes sedikit memberi kelonggaran dengan menyatakan bahwa item yang memiliki indeks daya beda di atas 0.20 sudah cukup baik untuk membedakan kelompok yang berkemampuan tinggi dengan kelompok yang berkemampuan rendah, (Kartowagiran, 2004).

Ada dua cara yang paling umum digunakan untuk menentukan besarnya daya beda item, antara lain dengan:

1) indeks diskriminasi;
2) indeks korelasi;

Untuk menghitung besarnya daya pembeda item dengan indeks diskriminasi dapat ditentukan dengan membagi kelompok responden menjadi dua kelompok, yaitu kelompok atas dan kelompok bawah. Variasi pembagian kelompok atas dan kelompok bawah berdasarkan perolehan skor total dapat dilakukan dengan tiga variasi, yakni 50%-50%, 33%-33%, atau 27%-27%. Dalam kebanyakan kasus jumlah peserta tes kelompok atas sama dengan jumlah peserta tes kelompok bawah, (Surapranata. 2006). Sehingga perhitungan daya pembeda dapat dinyatakan dengan persamaan:

$$D = \frac{\sum A}{n_A} - \frac{\sum B}{n_B} \qquad (2.6)$$

D indeks daya pembeda;
ΣA jumlah peserta tes yang menjawab benar pada kelompok atas;
ΣB jumlah peserta tes yang menjawab benar pada kelompok bawah;
n_A jumlah peserta tes kelompok atas
n_B jumlah peserta tes kelompok bawah

Sementara daya pembeda item dengan indeks korelasi dapat dijelaskan sebagai berikut:

1) Korelasi Point 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 item bersifat dikotomi sedangkan variabel skor total atau sub skor total bersifat kontinum. Variabel item dinamakan dikotomi karena skor-skor yang terdapat pada item hanya ada satu dan nol. Seperti halnya pada bentuk item pilihan ganda, item yang benar diberi angka 1 dan yang salah diberi angka 0. Variabel skor total atau sub skor total peserta tes bersifat kontinum atau non dikotomi yang diperoleh dari jumlah jawaban yang benar (Hayat at al, 1999).

Korelasi point biserial ditentukan

dengan menggunakan persamaan :

$$r_{pbis} = \frac{M_p - M_t}{S_t}\sqrt{\frac{p}{q}} \quad (2.8)$$

r_{pbis} koefisien korelasi *point biserial*;
M_p mean skor pada tes dari peserta tes yang memiliki jawaban benar pada *item* item;
M_t mean skor total;
S_t standar deviasi pada skor total;
p proporsi peserta tes yang jawabannya benar pada *item* item;
q proporsi jawaban salah peserta tes = 1 – *p*.

2) Korelasi Biserial

Korelasi *biserial* dapat ditentukan dengan persamaan:

$$r_{bis} = \frac{M_p - M_t}{S_t}\frac{p}{y} \quad (2.9)$$

r_{pbis} koefisien korelasi *point biserial*;
M_p mean skor pada tes dari peserta tes yang memiliki jawaban benar pada item item;
M_t mean skor total;
S_t standar deviasi pada skor total;
p proporsi peserta tes yang jawabannya benar pada *item* item;
y ordinat *p* atau nilai *y* dalam distribusi normal.

Nilai korelasi *point biserial* selalu lebih rendah jika dibandingkan dengan nilai korelasi *biserial*. Koefisien *point biserial* merupakan kombinasi hubungan antar item dengan kriteria dan taraf kesukaran.

c. Kehandalan Distraktor

Yang dimaksud dengan kehandalan (keberfungsian) distraktor disini adalah distribusi peserta tes dalam menentukan pilihan jawaban pada item bentuk pilihan ganda. Tes bentuk pilihan ganda terdiri dari item yang berisi permasalahan yang ditanyakan dan kemungkinan pilihan jawaban (penyebaran pilihan jawaban). Dan dari sekian banyak pilihan jawaban hanya terdapat satu jawaban yang paling benar yang disebut dengan kunci jawaban sedangkan selebihnya adalah pilihan jawaban yang tidak benar yang disebut distraktor (pengecoh).

Distraktor yang baik adalah distraktor yang memiliki homogenitas dengan kunci jawaban. Sebaliknya distraktor akan menjadi kurang baik apabila pilihan jawaban selain kunci atau distraktor tidak memiliki homogenitas dengan kunci jawaban. Kehandalan distraktor juga berfungsi sebagai pengidentifikasian peserta tes yang berkemampuan tinggi dan peserta tes yang berkemampuan rendah.

Distraktor akan berfungsi apabila dipilih secara merata oleh peserta tes. Dengan kata lain dapat disebutkan bahwa proporsi peserta tes yang menjawab pilihan jawaban tertentu, baik kunci jawaban maupun distraktor menyebar pada seluruh pilihan jawaban. Penyebaran pilihan jawaban berkisar antara 0 sampai dengan 1. Sehingga suatu pilihan jawaban selain kunci dikatakan berfungsi dengan baik apabila dipilih paling sedikit oleh 2.5 % (≥ 0.025) peserta tes. (Zulaiha, 2008).

Kehandalan distruktor atau penyebaran pilihan jawaban dapat diperoleh melalui perhitungan dengan menggunakan rumus sederhana, yaitu :

$$Ppj = \frac{Jpj}{n} \qquad (2.10)$$

Dimana :

Ppj penyebaran jawaban untuk pilihan jawaban tertentu;

Jpj banyak siswa yang memilih pilihan jawaban tertentu;

n banyaknya peserta tes.

Persamaan lain yang dapat digunakan untuk menunjukkan indeks daya beda item pada pilihan jawaban (alternatif) adalah dengan menggunakan persamaan koefisien *point-biserial* dan *biserial* korelasi *product moment Pearson*. Apabila indeks korelasi *point biserial* maupun *biserial* yang diperoleh pada pilihan jawaban selain kunci (distraktor) semakin negatif (< 0) , maka sudah dapat dikatakan distraktor sudah berfungsi dengan baik. Dan apabila tidak ada satu orang pun dari peserta tes yang menjawab pilihan jawaban item tersebut, maka nilai distraktornya adalah -9.000. Nilai -9.000 tersebut menunjukkan bahwa statistik item atas pilihan jawaban tidak dapat dihitung.

2. Teori Tes Modern (*Item Response Theory atau IRT*)

Item Respon Theory (*IRT*) bertujuan untuk mengatasi kelemahan-kelemahan yang terdapat pada teori tes klasik. Untuk itu para ahli pengukuran kemudian menyusun model alternatif yang mempunyai ciri-ciri dan sifat-sifat sebagai berikut : (1) karakteristik item tidak tergantung kepada kelompok peserta tes yang dikenai item tersebut, (2) skor yang menyatakan kemampuan peserta tidak tergantung kepada tes, (3) model dinyatakan dalam tingkatan (level) item, tidak dalam tingkatan tes, (4) model tidak memerlukan tes paralel untuk menghitung koofisien reliabilitas, dan (5) model menyediakan ukuran yang tepat untuk setiap skor kemampuan, (Hambleton, et.al., 1991).

Item Respon Theory (IRT) juga merupakan model pengukuran yang mempunyai dua postulat, yaitu: (1) performansi peserta tes pada suatu item dapat diprediksi oleh sekumpulan faktor yang disebut *traits, latent trait* atau *abilities* (kemampuan), dan (2) hubungan antara performansi peserta tes pada suatu item dan sekumpulan *traits* dapat digambarkan dalam sebuah fungsi monoton naik yang disebut fungsi karakteristik item (*item characteristic function*) atau kurva karakteristik item (*item characteristic curve*) (Hambleton, et.al. 1991). Fungsi karakteristik item ini adalah menggambarkan bahwa semakin meningkat level kemampuan seseorang, semakin meningkat pula peluangnya menjawab benar item tertentu.

Dalam teori tes modern juga terdapat beberapa asumsi seperti yang dikemukakan oleh Hambleton, et.al. (1991); dan Naga, (1992).

Asusmsi-asumsi tersebut antara lain :

a. Unidimensionalitas

IRT mengisyaratkan asumsi unidimensionalitas yang berarti bahwa setiap item hanya mengukur satu ciri *laten* peserta (kemampuan). Secara praktik asumsi ini sukar untuk dipenuhi sepenuhnya, sebab ada beberapa faktor lain dapat memengaruhi hasil suatu tes. Sebenarnya unidimensionalitas dalam teori tes modern ini adalah adanya faktor yang paling dominan memengaruhi hasil suatu tes. Dan faktor itu adalah kemampuan peserta tes.

b. Independensi lokal

Independensi dapat diartikan bahwa setiap item yang ada dalam suatu tes tidak saling berkorelasi satu sama lain akibat respon peserta tes. Dengan kata lain, kemampuan yang dinyatakan dalam model ini adalah satu-satunya faktor yang memengaruhi respons peserta tes pada item-item soal tertentu.

c. Fungsi karakteristik item.

Fungsi karakteristik item menyatakan hubungan sebenarnya antara variabel yang tak terobservasi (yaitu kemampuan) dengan variabel terobservasi (yaitu respons item).

Model teori tes modern atau *Item Respon Theory* (*IRT*) oleh safari (2005) dikategorikan dalam 4 model, yaitu, (1) model 1 parameter atau *rasch model* (tingkat kesukaran item/*threshold*), (2) model 2 parameter (*threshold* dan daya beda item/*slope*) dan , (3) model 3 parameter (*threshold, slope*, dan faktor menebak/*asymptote*), dan (4) model 4 parameter (*threshold, slope, asymptote*, dan faktor lain yang mempengaruhi). Akan tetapi model 4 parameter dalam analisis teori tes modern belum banyak digunakan.

Sementara Hambleton, et.al, (1991) dalam bukunya *Fundamentals of Item Response Theory* membagi model parameter logitik menjadi tiga model yaitu:

a. Model satu parameter (1PL atau *model rasch*),

yaitu model yang hanya menganalisis pada parameter tingkat kesukaran item (*threshold*/*b*). Model *IRT* ini murupakan model yang paling sering digunakan. Formula untuk model 1 PL ini adalah :

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad (2.11)$$

Dimana :

i	*1, 2..., n;*
$P_i(\theta)$	kemungkinan sampel menjawab secara tepat dengan kemampuan menjawab benar pada item tersebut
b_i	parameter tingkat kesukaran
n	jumlah item tes
e	bilangan transedental yang bernilai 2.718

Parameter b_i untuk sebuah item merupakan suatu titik pada skala kemampuan dimana probablitas menjawab benar peserta tes sebesar 0.5. Parametar b_i diperoleh melalui titik potong kurva probabilitas peserta tes yang menjawab benar dan kurva probabilitas peserta tes

yang menjawab salah, dimana titik potong tersebut berada pada level probabilitas sebesar 0.5 (50%).

Dalam kaitannya dengan skala kemampuan, pada Item karakteristik kurva (ICC) dapat digambarkan bahwa semakin tinggi nilai parameter b_i, semakin besar kemampuan yang diperlukan oleh peserta tes untuk mendapatkan 50% kesempatan menjawab benar item tersebut.

Rentang indeks tingkat kesukaran item (b_i) yang ideal adalah antara -2 sampai dengan 2, dimana pada rentang tersebut nilai tingkat kemampuan kelompok peserta tes telah ditransformasikan sehingga nilai rata-rata (*mean*) menjadi 0 dan standar deviasi 1. Nilai b_i yang mendekati -2 menggambarkan bahwa item tersebut sangat mudah, dan sebaliknya bila nilai b_i mendekati ke angka 2, maka tingkat kesukaran item tersebut dapat dikatakan sukar/sulit untuk dijawab oleh kelompok peserta tes.

b. Model dua parameter (2PL),

yaitu model yang digunakan untuk menganalisis data yang menitikberatkan pada parameter tingkat kesukaran (*threshold*/*b*) dan daya beda item (*slope*/*a*). Item karakteristik kurva (*ICC*) untuk model 2 parameter logistik ini dapat diperoleh dengan persamaan:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad (2.12)$$

dimana *:*

i , , P_i(θ), b_i, *n, dan e* telah didefinisikan di atas.

D=1,7 untuk logistik 1PL dan D=1 untuk Rasch

Pada model dua parameter (2PL) logistik ini elemen yang bertambah selain *D* adalah a_i, yaitu parameter diskriminasi/ daya beda item (*slope*). Parameter a_i merupakan proporsi terhadap garis singgung *slope pada titik 0 =* b_i. Item yang mempunyai daya beda tinggi adalah item yang lebih dapat membedakan kemampuan peserta tes dibandingkan dengan item yang mempunyai daya beda kecil. Landai atau curamnya Item Karakteristik Kurva (ICC) sangat bergantung pada daya beda item tersebut dalam membedakan kemampuan peserta tes. Item yang mempunyai daya beda kecil akan memiliki ICC yang landai, sebaliknya item dengan daya beda tinggi akan memiliki ICC yang curam.

Rentang indeks daya beda item pada model 2PL ini adalah -∞ sampai dengan +∞, akan tetapi item yang dikatakan mempunyai daya beda yang baik adalah item yang memiliki indeks daya beda (*slope*) antara 0 sampai dengan 2.

c. Model tiga parameter (3PL),

yaitu model yang digunakan untuk melakukan analisis data yang menitikberatkan pada parameter tingkat kesukaran item (*threshold*/*b*), daya beda item (*slope*/*a*), dan faktor menebak (*asymptote*/*c*) :

Rumus untuk model tiga parameter logistik adalah:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad (2.13)$$

Dimana *i , D,* P_i(θ), b_i, *n, e, dan* a_i telah

didefinisikan sebelumnya. Parameter yang bertambah pada model ini adalah c_i yang biasa disebut faktor *guessing* atau *pseudo-chance level*, dan dalam istilah lain dapat disebut dengan *asymptote*. Nilai *asymptote* akan mempengaruhi probabilitas peserta tes dalam menjawab benar satu item. *Asymptote* yang tinggi akan menyebabkan semakin tinggi pula probabilitas peserta tes untuk menjawab benar item tersebut.

F. Analisis Regresi Sederhana

Analisis regresi dikembangkan untuk mengkaji dan mengukur hubungan antar dua variabel atau lebih. Dalam analisis regresi dikembangkan persamaan estimasi untuk mendeskripsikan pola atau fungsi hubungan antar variabel. Sesuai dengan namanya, persamaan regresi itu digunakan untuk mengestimasi nilai dari suatu variabel berdasarkan nilai variabel lainnya. Variabel yang diestimasi itu disebut variabel dependen. Sedangkan variabel yang diperkirakan mempengaruhi variabel dependen itu disebut variabel independen, (Reksoatmodjo, 2007).

Analisis regresi terdiri dari analisis regresi sederhana dan analisis regresi berganda. Analisis regresi sederhana (*simple regression analysis*) adalah analisis regresi yang menggunakan hanya satu variabel independen (*independent variabel*) dan satu variabel dependen (*dependent variabel*). Sedangkan analisis regresi berganda (*multi regression analysis*) menggunakan lebih dari satu variabel independen. Namun demikian pada penelitian ini penulis menjelaskan tentang analisis regresi sederhana dikarenakan data penelitian ini hanya terdiri dari satu independen variabel.

Analisis regresi sederhana merupakan salah satu bagian dari tekhnik analisis regresi parametrik yang dapat memberikan dasar untuk memprediksi besarnya variasi serta menganalisis varian, (Triton, 2006).

Tujuan dilakukan analisis regresi adalah untuk, 1) menentukan persamaan garis regresi berdasarkan nilai konstanta dan koefisien regresi yang dihasilkan, 2) mencari korelasi bersama-sama antara variabel independen dengan variabel dependen, 3) mengitung besarnya variasi pada variabel dependen yang dapat dijelaskan oleh variabel independen, dan 4) menguji signifikansi pengaruh variabel independen terhadap variabel dependen melalui uji F atau uji t.

Bentuk hubungan antara variabel independen dengan variabel dependen dapat digambarkan dalam satu garis yang disebut garis regresi. Garis regresi dapat berbentuk garis lurus (linier) atau garis melengkung (non linier). Hubungan linier digambarkan oleh kesamaan perubahan variasi yang tetap baik penurunan atau peningkatan yang terjadi pada variabel dependen dan variabel independen. Sementara hubungan non liner kebalikan dari hubungan linier, yaitu perubahan peningkatan atau penurunan variasi yang terjadi tidak konsisten.

Sehubungan dengan kemungkinan bentuk garis hasil analisis regresi ini,

regresi linier maupun regresi non linier, maka sebelum melakukan analisis regresi sebaiknya dilakukan uji linearitas hubungan antara variabel independen dan variabel dependen sehingga analisis regresi yang dipilih akan sesuai.

Secara umum, persamaan garis regresi antara variabel independen (X) dan variabel dependen (Y) dapat digambarkan dengan formula :

$$Y = b_0 + b_1 X \tag{2.14}$$

dimana :

Y kriterium;

X prediktor;

b_0 konstanta regresi atau harga yang memotong sumbu Y; dan

b_1 koefisien regresi atau sering disebut slope.

Untuk dapat mengetahui seberapa besar variasi yang diberikan oleh X terhadap Y, terlebih dahulu harus diketahui besaran nilai b_0 dan b_1, dimana nilai keduanya dapat perolah dengan menggunakan persamaan sebagai berikut :

$$b_0 = \frac{\sum Y \sum X^2 - \sum X \sum XY}{N\sum X^2 - \left(\sum X\right)^2} \tag{2.15}$$

dan

$$b_1 = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - \left(\sum X\right)^2} \tag{2.16}$$

Pada persamaan regresi ini nilai b_1 disebut dengan *slope*. *Slope* menentukan seberapa besar variabel *Y* akan berubah ketika variabel *X* naik 1 poin, (Gravetter, F. J., & Wallnau L. B., 2007)

Untuk mengetahui bentuk korelasi antara variabel X dan variabel Y sesuai dengan tujuan regresi dapat dicari dengan teknik korelasi *product moment Pearson*, dengan rumus umumnya adalah :

$$R_{xy} = \frac{\sum xy}{\sqrt{\left(\sum x^2\right)\left(\sum y^2\right)}} \tag{2.17}$$

Dimana nilai xy, x^2, dan y^2 diperoleh melalui persamaan:

$$\sum xy = \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{N} \tag{2.18}$$

$$\sum x^2 = \sum X^2 - \frac{\left(\sum X\right)^2}{N} \tag{2.19}$$

dan

$$\sum y^2 = \sum Y^2 - \frac{\left(\sum Y\right)^2}{N} \tag{2.20}$$

Dalam kaitannya dengan regresi, nilai hasil korelasi antara variabel *Y* dan variabel *X* dapat dinamakan dengan validitas prediksi. Mengenai validitas prediksi telah penulis jelaskan sebelumnya pada bagian validitas dan reliabilitas.

Selanjutnya mencari koefisien determinasi (R^2) dengan menggunakan persamaan sebagai berikut :

$$R^2 = \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2} \tag{2.21}$$

Koefisien determinasi (R^2)

digunakan untuk menentukan besarnya variasi yang terjadi pada variabel dependen Y yang sering disebut kriterium atau kriteria berdasarkan data yang terdapat pada variabel independen X yang disebut prediktor.

Terakhir adalah menguji signifikansi pengaruh variabel independen terhadap variabel dependen melalui uji F, sehingga diperoleh persamaan sebagai berikut:

$$F = \frac{b^2 x \sum (X - \overline{X})}{S_e^2} \qquad (2.22)$$

Dimana nilai a, b dan S_e diperoleh dari persamaan-persaman di bawah ini :

$$a = \overline{Y} + b_1 \overline{X} \qquad (2.23)$$

dan

$$b = \frac{\sum XY - n.\overline{X}.\overline{Y})}{\sum X^2 - n.\overline{X^2}} \qquad (2.24)$$

serta

$$S_e = \sqrt{\frac{\sum Y^2 - a.\sum Y - b.\sum XY}{n-2}} \qquad (2.25)$$

G. Kesimpulan

Tes prestasi merupakan tes yang disusun secara terencana untuk mengungkap performansi maksimal subjek dalam menguasai bahan-bahan atau materi yang telah diajarkan. Dalam kegiatan pendidikan formal di kelas, tes prestasi belajar dapat berbentuk ulangan-ulangan harian, tes formatif, tes sumatif dan beberapa bentuk tes lainnya.

Dalam menyusun instrumen tes untuk tes prestasi, yang harus diperhatikan adalah bagaimana instrumen tes bisa sepadan dengan kemampuan seseorang yang akan di berikan tes. Tes prestasi belajar merupakan sebuah tes dengan jumlah item yang banyak dan seluruh itemnya bertaraf kesukaran sedang (on-target) bagi orang yang menempuh tes. Hal ini memberi gambaran bahwa instrumen tes yang disusun tidak boleh terlalu jauh di bawah atau di atas kemampuan peserta tes, dan tingkat kesukaran item-item soal sebaiknya berada pada kategori sedang. Sehingga dengan demikian instrumen tes yang disusun nantinya dapat berfungsi dengan baik.

Daftar Pustaka

Anastasi, A., & Urbina, S. (2006). *Tes Psikologi*. Edisi Ketujuh, (Imam, R. H. S. Penerjemah) Jakarta : Indeks

Arikunto, S. (2008). *Dasar-dasar Evaluasi Pendidikan*, Edisi Revisi, Jakarta : Bumi Aksara.

Arvyaty. (2005). *Komparasi Bentuk Tes ditinjau dari Tingkat Kesukaran Item, Daya Beda item, dan Reliabilitas Tes*, Tesis Pascasarjana, tidak diterbitkan, Universitas Negeri Jakarta.

Azwar, S. (2000). *Reliabilitas dan Validitas*, Yokyakarta : Pustaka Pelajar.

Azwar, S. 2007). *Tes Prestasi, Fungsi*

dan Pengembangan Pengukuran Prestasi Belajar, edisi II, Cetakan VI, Yogyakarta : Pustaka Pelajar.

Budiyono. (2005). *Perbandingan Metode Mantel-Haenszel, SIBTEST, Regresi Logistik, dan Perbedaan Peluang dalam Mendeteksi Keberbedaan Fungsi Item*, Yogyakarta : UNY (disertasi)

Chaplin, J.P. (2005). *Kamus Lengkap Psikologi*, Jakarta : PT. Raja Grafindo Persada

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*, New York : Holt, Rinehart and Winston, Inc.

Departemen Pendidikan Nasional, Direktorat Jenderal Pendidikan Dasar dan Menengah, Direktorat Tenaga Kependidikan. (2003). *Sistem Penilaian Kelas SD, SMP, SMA dan SMK*, Jakarta : Pengarang.

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of Educational Measurement*, New Jersey : Prentice Hall.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*, California. ; Sage Publication Inc.

Hayat, B. (2000). *Pengantar Model Rasch (Kalibrasi item)*, Jakarta : Pusat Penelitian Pendidikan Balitbang Depdiknas..

Hayat, B., & Setiadi, H. (1998). *Mendesain Instrumen tes Dengan Model Rasch*, Jakarta ; Pusat Penelitian dan Pengembangan Sistem Pengujian, Balitbang Dikbud.

Hayat, B., Surapranata, S,. & Suprananto. (1999). *Manual Item and Test analysis (ITEMAN) Pedoman Penggunaan "ITEMAN"*. Jakarta : Pusat Penelitian dan Pengembangan Pendidikan Nasional, Puspendik.

Kartowagiran, B. (2004). *Perbandingan Berbagai Metode Untuk Mendeteksi Bias Item*, Jogyakarta ; Fakultas Psikologi UGM.

Naga, S. D. (1992). *Pengantar Teori Sekor pada Pengukuran Pendidikan*, Jakarta: Gunadarma

Nunnaly, J.C., & Ira, H. B. (1994). *Psychometric Theory.3rd ed.* New York, McGraw-Hill, Inc

Pusat Penilaian Pendidikan Balitbang Depdiknas. (2006). *Urgensi Ujian Nasional*, Jakarta.

Reksoatmodjo, T. N. (2007). *Statistika untuk Psikologi dan Pendidikan*, Bandung; Refika Aditama

Safari. (2005). *Teknik Analisis Item Item Instrumen Tes dan Non-Tes*, Jakarta ; Depdiknas

Surapranata, S. (2006). *Analisis, Validitas, Reliabilitas dan Interpretasi Hasil Tes*, Bandung ; PT.Remaja Rosdakarya.

Suryabrata, S. (2006). *Pengembangan Alat Ukur Psikologis*, Yogyakarta : Andi.

Triton, (2006). SPSS 13.0 Terapan, *Riset Statistik Parametrik*, Yogyakarta : Andi.

Weitzman, R. A. (1982). *The Prediction of College Achievement by the Scholastic Aptitude Test and the High School Record*. Journal of Educational Measurement, Vol. 19, No. 3. (Autumn, 1982).

Wijanto, H.S. (2008). *Structural Equation Modeling dengan* Lisrel 8.8 Konsep & Tutorial. Yogyakarta : Graha Ilmu.

Zulaiha, R. (2008). *Analisis Item Secara Manual*, Jakarta : Pusat Penilaian Pendidikan.