BAB II

LANDASAN TEORI

2.1 Computerized Adaptive Testing

Menurut Meijer dan Nering (1999), tujuan *computer adaptive testing* atau tes adaptif terkomputerisasi adalah penyediaan tes yang optimal berdasarkan tingkat kemampuan setiap peserta yang diestimasi saat tes berlangsung, sehingga butir soal dipilih akan disesuaikan dengan tingkat kemampuan. Jika jawaban peserta benar, tingkat kesukaran akan dinaikkan. Begitu juga sebaliknya, jika jawaban peserta salah, tingkat kesukaran akan diturunkan. Berikut adalah prosedur tes adaptif secara umum (Cisar dkk., 2010):

1. Dari seluruh butir soal yang disimpan di bank soal, akan dipilih satu butir yang belum ditampilkan dan paling sesuai dengan estimasi tingkat kemampuan peserta tes saat ini.
2. Butir soal yang dipilih akan ditampilkan sebagai butir soal yang akan dijawab oleh peserta.
3. Berdasarkan respon atau hasil jawaban peserta, estimasi kemampuan peserta akan dihitung kembali berdasarkan seluruh respon pada butir soal yang telah diberikan.
4. Jika belum ditemui adanya kriteria dan aturan pemberhentian tes, sistem akan melakukan tahap satu sampai tahap tiga secara berulang.
5. Ketika kriteria aturan pemberhentian terpenuhi, sistem akan menghitung hasil akhir peserta.

Menurut beberapa jurnal, titik awal pengujian dalam algoritma tes adaptif terkomputerisasi dapat dimodifikasi. Saat pertama kali peserta mengikuti tes, dilakukan pengujian sementara berupa soal-soal dasar dari mata kuliah secara acak dimana hasil dari pengujian sementara tersebut akan dijadikan nilai estimasi kemampuan awal untuk tes yang sesungguhnya (Krishna, 2001). Alternatif lain adalah butir soal yang pertama kali dipilih jika belum ada informasi mengenai kemampuan peserta adalah butir soal dengan tingkat kesukaran sedang, dalam jangkauan nilai dari -0,5 sampai 0,5 (Agus Santoso dkk., 2010).

Tahap pemeriksaan aturan dan kriteria pemberhentian merupakan komponen penting dalam tes adaptif terkomputerisasi. Tes adaptif terkomputerisasi akan berhenti pada umumnya ketika memenuhi salah satu kriteria berikut (Jian-quan dkk., 2007):

1. Ketika seluruh butir soal pada bank soal sudah diambil untuk ditampilkan pada saat satu sesi tes tersebut. Hal ini disebabkan karena bank soal memiliki butir soal yang tidak banyak.
2. Ketika satu sesi tes telah mencapai batas jumlah maksimal soal yang harus diberikan kepada peserta. Aturan pemberhentian ini dikategorikan sebagai *fixed length test* (Thissen dan Mislevy, 2000). Aturan *fixed length test* mudah diimplementasikan dan setiap peserta akan diberikan butir soal dengan jumlah yang sama. Namun, pemberhentian ini menyebabkan pengukuran setiap peserta tidak dalam skala yang sama.
3. Ketika pengukuran kemampuan yang diestimasi sudah memenuhi presisi yang cukup atau di bawah dari standar *error* yang ditentukan. Aturan pemberhentian ini dikategorikan sebagai *variable length test* (Thissen dan Mislevy, 2000).

Dengan aturan pemberhentian ini, seluruh peserta akan mendapatkan pengukuran estimasi kemampuan dalam skala yang sama. Namun, ada kemungkinan peserta dengan estimasi tingkat kemampuan tinggi dan rendah akan mendapat jumlah butir soal yang lebih banyak jika tidak banyak butir soal yang memadai untuk tingkat kemampuan.

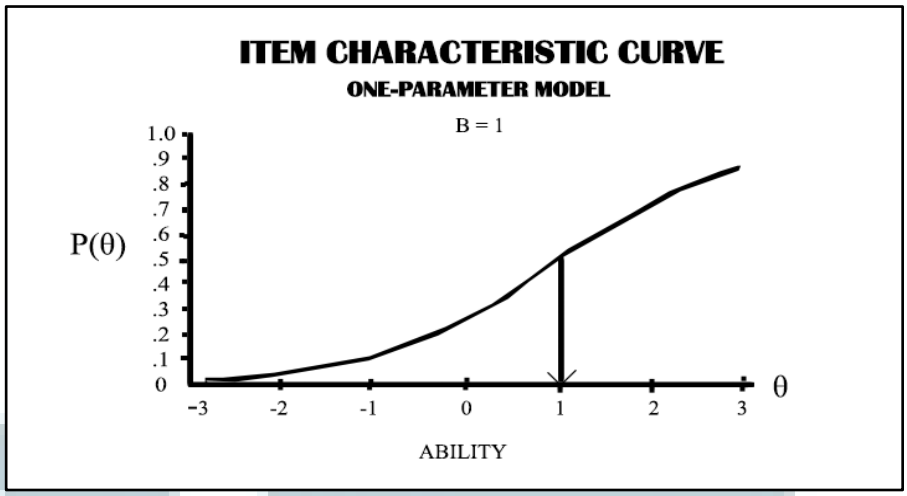
2.2 Item Response Theory

Item response theory merupakan suatu studi mengenai nilai suatu butir berdasarkan asumsi perhitungan matematis antara kemampuan dan respon butir soal (Rudner, 2001). Pemodelan *item characteristic curve* yang dijabarkan oleh Baker (2001) dalam bukunya yang berjudul The *Basic of Item Response Theory* terdiri dari tiga bentuk:

1. Model Rasch atau Model Logistik Satu-Parameter adalah pemodelan yang dipublikasikan pada 1960 oleh matematikawan Georg Rasch. Pendekatan model Rasch hanya menggunakan satu parameter saja dalam mengestimasi tingkat kemampuan siswa, yaitu tingkat kesulitan dari butir soal. Berikut adalah persamaan pada model Rasch :

$$P(\theta) = \frac{1}{1+e^{-(\theta-b)}} \quad \ldots \text{ Rumus 2.1}$$

dimana θ adalah tingkat kemampuan, b adalah parameter tingkat kesulitan dari butir soal, $P(\theta)$ adalah probabilitas dengan tingkat kemampuan seseorang menjawab benar dan e adalah nilai eksponensial. Gambar 2.1 adalah salah satu contoh kurva karakteristik butir untuk paramater b = 1 .



Gambar 2.1 Kurva Karakteristik Butir Soal 1 Parameter Logistik (Baker, 2001)

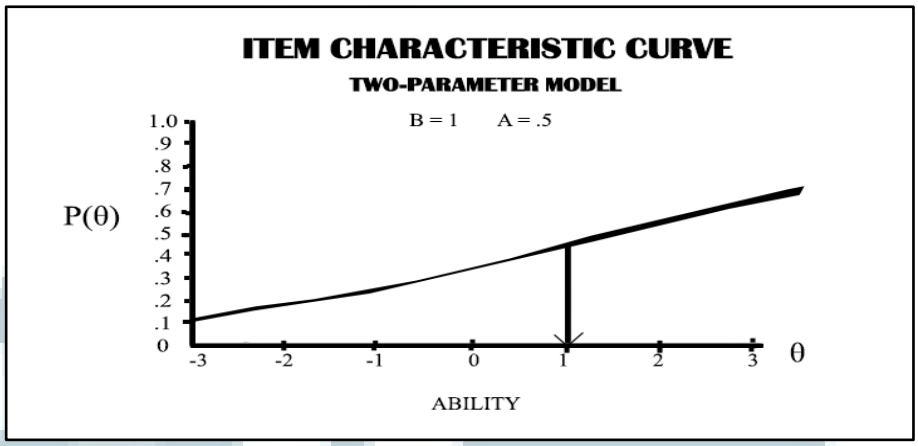
Jika θ semakin bernilai negatif, probabilitas untuk menjawab dengan benar semakin kecil. Sebaliknya, jika θ semakin bernilai positif, probabilitas untuk menjawab benar semakin besar.

2. Fungsi Logistik atau Model Logistik Dua-Parameter adalah model matematika strandar untuk kurva karakteristik butir soal dalam bentuk kumulatif. Pada model ini, terdapat dua parameter yang digunakan, yaitu tingkat kesulitan butir soal dan daya pembeda atau *discrimination*. Parameter daya pembeda ini merupakan suatu parameter dimana butir soal dapat membedakan peserta dengan kemampuan yang tinggi dan rendah. Berikut adalah persamaan model logistik:

$$P(\theta) = \frac{1}{1+e^{-a(\theta-b)}} \quad \ldots \text{ Rumus 2.2}$$

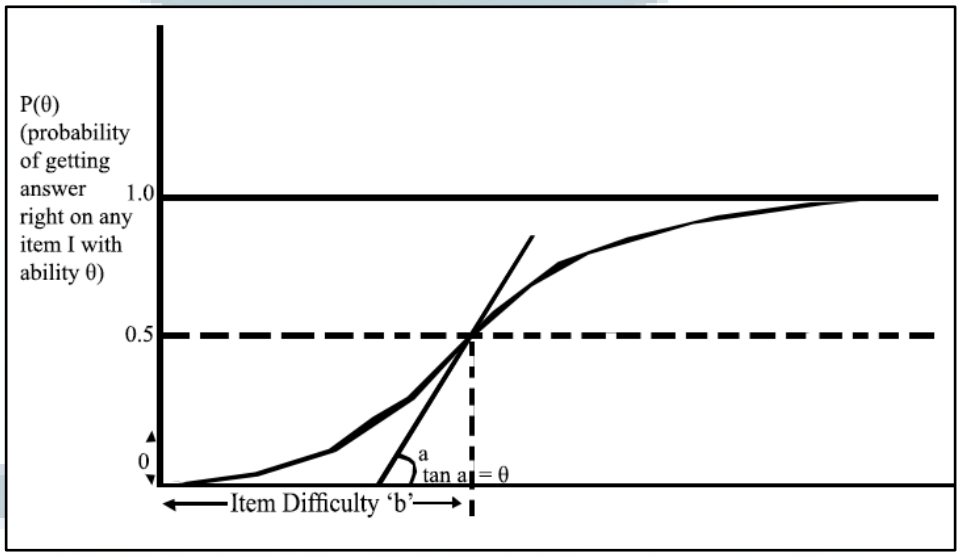
dimana θ adalah tingkat kemampuan, a adalah parameter daya pembeda, b adalah parameter tingkat kesulitan dari butir soal, $P(\theta)$ adalah adalah probabilitas dengan tingkat kemampuan seseorang menjawab benar dan e

adalah nilai eksponensial. Gambar 2.2 adalah salah satu bentuk kurva karakteristik butir soal dengan parameter *b* = 1 dan *a* = 0.5.



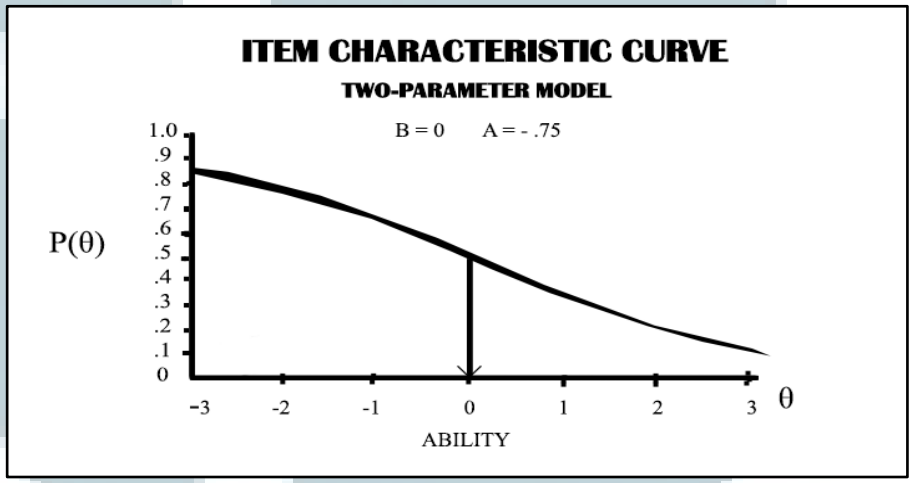
Gambar 2.2 Kurva Karaketristik Butir Soal 2 Parameter Logistik (Baker, 2001)

Berbeda dengan model satu parameter, model dua parameter menghitung *slope* atau kemiringan yang disebut dengan daya beda butir soal. Pada Gambar 2.3 dijelaskan bagaimana melihat daya beda suatu butir dari kurva karakteristik butir soal.



Gambar 2.3 *Slope* Kurva Karakteristik Butir Soal (Natarajan, 2009)

Jika nilai *a* semakin mendekati nol, bentuk kurva karakteristik butir soal akan menjadi semakin datar. Hal ini menandakan probabilitas setiap tingkat kemampuan dalam menjawab butir soal tidak akan berbeda jauh. Dalam kasus tertentu, beberapa butir soal dapat memiliki daya beda yang bernilai negatif. Gambar 2.4 adalah salah satu contoh kurva karakteristik butir soal untuk daya beda yang bernilai negatif.



Gambar 2.4 Kurva Karakteristik Daya Beda Negatif (Baker, 2001)

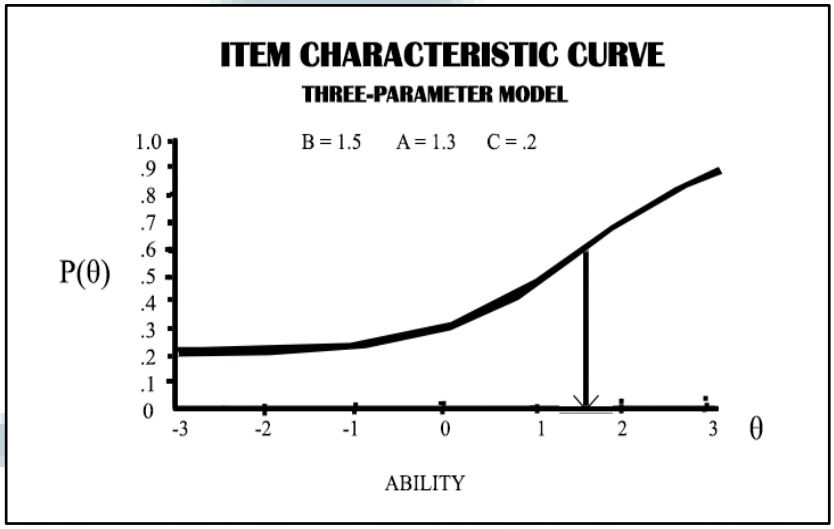
Butir soal yang memiliki nilai daya beda negatif akan memberikan probabilitas yang besar bagi θ yang mendekati nilai negatif tak terhingga untuk menjawab dengan benar. Sebaliknya, nilai θ yang semakin mendakati nilai positif tak terhingga akan mendapat probabilitas yang kecil untuk menjawab benar. Butir soal dengan nilai negatif dapat terjadi karena kunci jawaban dari butir soal tersebut memiliki indeks daya beda yang bernilai negatif (Baker, 2001). Dengan kata lain, pilihan yang salah seharusnya bernilai negatif. Butir soal dengan daya beda negatif dapat diindikasikan sebagai butir soal yang dirancang

kurang baik atau terjadinya kesalahan informasi yang ditangkap oleh siswa yang memiliki kemampuan tinggi (Baker, 2001).

3. Model Tiga-Parameter adalah model logistik dua-parameter yang ditambahkan satu parameter oleh Birnbaum pada 1968, yaitu probabilitas peserta menerka jawaban dari butir soal yang diberikan dengan benar. Hal ini disebabkan karena fakta dalam suatu tes terutama butir soal yang bersifat objektif, peserta dapat menjawab benar dengan cara menerka. Berikut adalah persamaan dari model tiga-parameter:

$$P(\theta) = c + (1 - c)\frac{1}{1+e^{-a(\theta - b)}} \quad \ldots \text{ Rumus 2.3}$$

dimana θ adalah tingkat kemampuan, a adalah parameter daya pembeda, b adalah parameter tingkat kesulitan dari butir soal, c adalah parameter *guessing*, $P(\theta)$adalah probabilitas dengan tingkat kemampuan seseorang menjawab benar dan e adalah nilai eksponensial. Gambar 2.4 adalah salah satu contoh kurva karakteristik butir soal untuk b = 1.5 a = 1.3 dan c = 0.2.



Gambar 2.5 Kurva Karakteristik Butir Soal 3 Parameter Logistik (Baker, 2001)

Berbeda dengan kedua model sebelumnya, dimana b dapat dilihat dari titik pada skala tingkat kemampuan yang memiliki $P(\theta)$ = 0.5. Pada model tiga parameter, batas terendah dari kurva karakteristik adalah nilai *c*. Hal ini menyebabkan parameter tingkat kesukaran butir soal tidak berada saat probabilitas menjawab benar sama dengan 0.5.

Item characteristic model ini digunakan dalam prosedur perhitungan estimasi kemampuan. Salah satu metode prosedur perhitungan yang sesuai untuk fungsi logistik adalah *maximum likelihood*. Perhitungan ini dimulai dari nilai priori untuk kemampuan peserta dan nilai parameter yang telah diketahui. Penyesuaian nilai estimasi kemampuan didapat dari setiap respon butir soal peserta. Proses akan terus diulang hingga perubahan pada estimasi kemampuan dapat diabaikan karena pertambahan nilai estimasi yang kecil pada iterasi berikutnya. Hasilnya berupa estimasi nilai untuk parameter kemampuan peserta. Berikut adalah persamaan estimasi (Baker, 2001):

$$\theta_{s+1} = \theta_s + \frac{\sum_{i=1}^{N} -a_i\,[\,u_i - P_i(\theta_s)]}{\sum_{i=1}^{N} a_i^2\,P_i(\theta_s)\,Q_i(\theta_s)} \quad \ldots \text{ Rumus 2.4}$$

dimana θ_s adalah estimasi kemampuan peserta pada iterasi ke-s, a_i adalah parameter daya pembeda pada butir soal ke-*i*, *i* adalah butir soal kesatu sampai ke-N jumlah butir soal yang disajikan, u_i adalah respon yang diberikan peserta pada butir soal ke-*i*, jika respon peserta dalam menjawab butir soal dengan benar, u_i = 1 dan begitu juga sebaliknya, $P_i(\theta_s)$ adalah probabilitas dari respon yang benar oleh peserta pada butir soal ke-*i* melalui model kurva *item characteristic* pada θ_s , dan $Q_i(\theta_s)$ adalah probabilitas respon yang salah oleh peserta pada butir soal ke-*i* dengan perhitungan $Q_i(\theta_s)$ = 1 - $P_i(\theta_s)$. Metode perhitungan pada prosedur ini dimulai dengan

mengambil nilai sembarang pada θ_s , kemudian mengitung probabilitas tingkat kemampuan seseorang merespon dengan benar dengan menggunakan pemodelan kurva *item characteristic*. Kemudian, hasil tersebut dimasukkan ke suatu variabel yang dinotasikan $\Delta\theta$ dimana nilai θ_{s+1} akan ditambah dengan θ_s . Hal ini menyebabkan θ_{s+1} akan menjadi θ_s pada iterasi berikutnya.

Prosedur tersebut hanya merupakan nilai estimasi kemampuan peserta, bukan nilai kemampuan peserta yang sebenarnya. Untuk itu, ditentukan *standard error* sebagai indikasi keakuratan estimasi. Berikut adalah persamaan *standard error* yang digunakan (Barker, 2001):

$$SE(\theta) = \frac{1}{\sqrt{\sum_{i=1}^{N} a_i^2 P(\theta) Q(\theta)}} \quad \ldots \text{ Rumus 2.5}$$

Nilai standar *error* ini akan mengukur keragaman nilai dari tingkat kemampuan yang diestimasi berada disekitar nilai parameter tingkat kemampuan peserta yang sesungguhnya tidak dapat diketahui.

Namun, menurut Baker (2001), prosedur estimasi dengan menggunakan *maximum likelihood* ini memiliki keterbatasan, yaitu jika respon peserta dalam menjawab butir soal tidak berpola. Respon peserta tidak berpola, seperti seluruh butir soal yang dijawab benar semua atau salah semua, akan menyebabkan perhitungan *maximum likelihood* tidak akan dapat berhenti dari proses iterasi. Pola seluruh respon benar akan menghasilkan nilai positif tak terhingga, sedangkan pola seluruh respon salah akan menghasilkan nilai negatif tak terhingga. Pada implementasi yang sebenarnya, *maximum likelihood* juga tidak dapat digunakan untuk menghitung pola respon pada awal tes. Untuk pertimbangan praktis, sampai respon peserta ujian sudah memiliki pola yang berbeda, perhitungan estimasi tingkat kemampuan sementara

dilakukan menggunakan prosedur *fixed step-size* (McKinley dan Reckase, 1981). Jika jawaban peserta terhadap butir soal yang dimunculkan adalah benar, estimasi kemampuan akan ditambahkan dengan suatu nilai konstan, sedangkan jika jawaban peserta adalah salah, estimasi kemampuan akan dikurang dengan suatu nilai konstan. Nilai konstan ini dapat ditentukan dengan bebas, misalnya 0.25, 0.4, atau 0.5.

2.3 Metode Pemilihan Butir Soal

Metode pemilihan butir soal pada tes adaptif terkomputerisasi umumnya melakukan pencarian nilai maksimum pada fungsi informasi butir soal atau yang disebut dengan *maximum information*. Berikut adalah persamaan fungsi informasi pada butir soal untuk model logistik dua-parameter (Baker, 2001) :

$$I_i(\theta) = a_i^2 P_i(\theta) Q_i(\theta) \quad \ldots \text{ Rumus 2.6}$$

dimana a_i merupakan parameter daya pembeda dari suatu butir soal ke-i yang ditampilkan kepada peserta, $P_i(\theta)$ adalah probabilitas seseorang merespon jawaban yang benar pada tingkat kemampuan tertentu, dan $Q_i(\theta)$ adalah probabilitas seseorang merespon jawaban yang salah. Nilai θ pada persamaan tersebut didapatkan dari hasil perhitungan *maximum likelihood*.

Prosedur pemilihan dengan *maximum information* yang murni akan mengambil butir soal dengan informasi tertinggi tanpa melihat apakah butir soal tersebut sudah diadministrasikan dan tidak akan melihat indikator materi dari butir soal tersebut. Hal ini dapat menyebabkan kemungkinan peserta akan mendapat soal dari indikator materi yang sama beberapa kali dan tidak terkendali sampai peserta menyelesaikan tes. Untuk menjaga tes adaptif yang berjalan sesuai dengan spesifikasi materi yang dituju, diperlukan prosedur sederhana untuk menjaga agar butir soal dari suatu

indikator materi tidak melebihi batas penyajiannya. Kingsbury dan Zara (1989) mengajukan prosedur yang dinamakan dengan dengan *Constrained-Computer Adaptive Testing* (C-CAT). Dalam C-CAT, suatu tes dirancang dengan menetapkan proporsi butir soal dari beberapa indikator materi atau area konten berbeda yang akan diberikan. Misalnya, dalam tes adaptif untuk pelajaran matematika, terdapat 20% soal yang mencakup materi penjumlahan, 20% mengenai pengurangan, 30% mengenai perkalian, dan 30% mengenai pembagian.

Terdapat beberapa macam metode butir soal berdasarkan konten area. Kingsbury dan Zara (1989) membuat persentase yang sudah dispesifikasikan diawal untuk setiap konten, kemudian butir soal pada setiap konten yang sudah diadministrasikan dalam tes akan dihitung persentasenya dan dibandingkan dengan persentase dari spesifikasi awal. Butir soal yang dipilih adalah butir soal dari konten area yang paling besar perbedaannya. Boyd (2003) menggunakan metode pemilihan pertama pada suatu spesifikasi konten secara acak kepada setiap peserta dan kemudian selanjutnya menggunakan metode Kingsbury dan Zara. Bergstrom dan Lunz (1999) mengajukan variasi metode yang dapat digunakan seperti memilih butir soal dari konten area yang sudah disusun berdasarkan urutan yang pasti atau mendistribusikan butir soal dari konten area secara acak. Agus dkk. (2010) memilih berdasarkan konten yang diinginkan, misalnya untuk butir pertama pada konten area pertama, butir kedua pada konten area kedua, dan seterusnya sampai setelah mengadministrasikan butir soal dari konten area terakhir, butir soal berikutnya kembali diambil dari konten area pertama.

Permasalahan lain dari *maximum information* adalah butir soal yang sering digunakan dalam suatu tes sehingga menyebabkan peserta dapat menghapal soal dan

jawaban tersebut yang dapat mempengaruhi pengukuran tingkat kemampuan yang kurang akurat. Permasalahan tersebut disebut dengan *item exposure*. Agar dapat mengontrol frekuensi kemunculan butir soal tersebut, Kingsbury dan Zara (1989) menggunakan prosedur pemilihan acak atau disebut dengan pemilihan *Randomesque*. Metode ini memilih dua atau sampai sepuluh butir soal yang memberikan informasi maksimum. Kemudian, dari daftar kandidat butir soal tersebut, satu butir soal dipilih secara acak. Pemilihan acak ini dilanjutkan sepanjang pengujian untuk mengurangi kemunculan butir soal. Adapun cara efisien dalam mengatur *item exposure* dengan menggunakan *counter*, dimana setiap butir soal tersebut diadministrasikan kepada peserta, *counter* ditambah dengan satu (Parathyras, 2011).

2.4 Tes Pilihan Ganda

Penilaian adalah terminologi umum yang meliputi serangkaian penuh prosedur untuk mendapatkan informasi mengenai pembelajaran siswa dan pembentukan pertimbangan penilaian mengenai kemajuan pembelajaran (Linn dan Miller, 2005). Tes merupakan salah satu bentuk penilaian dan pengukuran yang umumnya terdiri dari sekumpulan pertanyaan yang diberikan. Pengukuran tersebut memiliki tujuan untuk menetapkan kuantitatif dari hasil tes seperti menghitung jumlah jawaban benar atau memberikan poin untuk tes uraian.

Dalam membangun suatu tes, terdapat dua bentuk butir soal, yaitu penilaian objektif dan penilaian performa. Penilaian objektif hanya memiliki satu jawaban yang benar dan tepat, sedangkan penilaian performa memiliki keberagaman jawaban yang dapat dianggap benar dan tepat. Contoh bentuk penilaian objektif, sepeti pilihan ganda, *true-false*, *short answer*, *fill-in-blank*, dan *matching*, sedangkan bentuk

penilaian perfoma seperti soal uraian atau *essay* (Linn dan Miller, 2005). Butir soal dengan format pilihan ganda adalah salah satu yang paling banyak digunakan dalam strategi pengujian terkomputerisasi karena mudah diimplementasikan dan mudah diberikan skor (Roy dan Armarego, 2003). Bahkan menurut Linn dan Miller (2005), butir soal pilihan ganda dapat mengukur hasil pembelajaran yang sederhana seperti butir soal *true-false* dan *short answer*, dan mengukur hasil pembelajaran yang kompleks seperti pengetahuan, pemahaman, dan penerapan. Menurut Roy dan Armarego (2003), terdapat beberapa komponen yang dalam pilihan ganda:

1. *Stem* adalah komponen pertanyaan yang sebaiknya ringkas dan jelas, menghindari petunjuk secara tata bahasa untuk kunci jawaban, dan berisi kata-kata yang sebagian besar mengurangi beban membaca.
2. *Key* adalah kunci jawaban yang benar. Kunci jawaban sebaiknya memiliki panjang yang sama dengan distraktor, dan diacak posisinya dalam daftar alternatif pilihan.
3. Distraktor adalah alternatif jawaban, atau jawaban yang salah. Distraktor sebaiknya dibuat sebagai alternatif yang realistik, menutupi berbagai pilihan dan mudah mengidentifikasi kesalahpahaman atau sebagai jebakan bagi siswa yang ragu dalam memilih jawaban yang benar.

Pertanyaan dapat dibuat berdasarkan tingkat kognitif dan jenis pertanyaan.

Berikut adalah beberapa contoh konstruksi jenis pertanyaan dalam mengukur kemampuan pemrograman (Clark, 2004):

1. Pengetahuan. Contoh jenis pertanyaan ini umumnya mengenai pengetahuan dasar, seperti bahasa sintaks dan lingkungan pemrograman.

2. Pemahaman. Contohnya seperti menelusuri suatu kode yang diberikan dan mencari nilai akhirnya atau keluarannya.
3. Aplikasi. Contohnya seperti mengukur kemampuan dalam menuliskan dan menerapkan kode, melengkapi suatu kode atau membuat *function*.
4. *Problem Solving*. Jenis pertanyaan ini bertujuan untuk mengukur tingkat kemampuan analisis siswa. Contoh jenis pertanyaan dalam programming adalah apa yang dilakukan oleh suatu *function*, atau memberikan suatu kode yang harus dianalisis dan dijelaskan apa yang dilakukan oleh kode tersebut.

2.5 Analisis Butir Soal

Terminologi analisis butir soal digunakan untuk menetapkan komputasi dan pengujian statistik berdasarkan respon peserta ujian terhadap individu butir soal (Crocker dan Algina, 1986). Analisis butir soal atau kalibrasi butir soal digunakan untuk mengevaluasi butir soal dan mengidentifikasikan himpunan butir soal yang memiliki kontribusi besar dalam reliabilitas dan validitas.

Beberapa parameter yang dapat diukur untuk individu butir soal adalah tingkat kesukaran butir soal dan daya beda butir soal. Tingkat kesukaran atau *p-value* adalah proporsi peserta ujian yang menjawab butir soal tersebut dengan benar (Crocker dan Algina, 1986). Dari definisi tersebut, butir soal yang dijawab benar oleh 85% peserta ujian akan mendapat *p-value* sebesar 0.85. Jika nilai *p-value* mendekati nilai satu, butir soal tersebut dikategorikan mudah bagi peserta ujian. Berikut adalah rumus untuk menghitung *p-value*.

$$p_i = \frac{\text{Jumlah orang yang menjawab benar pada butir soal } i}{N} \quad \ldots \text{ Rumus 2.7}$$

dimana N adalah total keseluruhan peserta ujian. Rentang nilai *p-value* berkisar dari 0.00 sampai 1.00. Kemudian, parameter daya beda atau diskriminan untuk melihat seberapa efektif suatu butir soal membedakan peserta dengan kemampuan yang relatif tinggi dan peserta dengan kemampuan yang relatif rendah, berdasarkan dari kriteria internal atau nilai total skor tes (Crocker dan Algina, 1986). Tujuan dari perhitungan daya beda adalah mengidentifikasi butir soal dimana peserta dengan total skor yang tinggi akan memperoleh probabilitas yang tinggi untuk menjawab dengan benar, dan peserta dengan total skor rendah akan memperoleh probabilitas yang rendah. Dalam kasus tertentu perlu dicurigai apabila butir soal dimana baik peserta dengan total skor rendah dan tinggi memiliki kesuksesan yang sama dalam menjawab dengan benar atau bahkan memiliki nilai diskriminan yang negatif. Beberapa prosedur perhitungan yang dapat digunakan adalah indeks diskriminan, *point biserial correlation*, dan *biserial correlation* (Crocker dan Algina, 1986).

Kedua parameter yang dijabarkan di atas digunakan untuk menganalisis butir soal pada teori tes klasik. Walaupun terdapat persamaan dalam terminologi parameter tingkat kesulitan dan daya beda, perhitungan parameter yang digunakan untuk mengestimasi nilai parameter dalam teori respon butir berbeda dengan teori tes klasik. Prosedur untuk mengestimasi nilai parameter dalam teori respon butir dapat menggunakan *joint maximum likelihood*, *conditional maximum likelihood*, atau *marginal maximum likelihood* (Naga, 1992).

Beberapa aplikasi dapat digunakan untuk melakukan perhitungan analisis butir soal, seperti ITEMAN dan BILOG-MG. ITEMAN adalah salah satu program yang menyediakan perhitungan statistik analisis butir menggunakan teori tes klasik (Assessment System Corporation, 1989), sedangkan BILOG-MG (Zimowski dkk,

2003) adalah aplikasi yang menyediakan hasil perhitungan analisis butir soal dikotomus teori tes klasik dan teori respon butir satu parameter, dua parameter dan tiga parameter. Analisis butir soal dengan menggunakan teori tes klasik digunakan sebagai hasil evaluasi butir soal dan prediksi tingkat kesulitan serta daya beda, sedangkan hasil perhitungan estimasi dari BILOG-MG digunakan sebagai nilai parameter karakteristik butir soal dalam tes adaptif.