

Item Response Theory Model Empat Parameter Logistik Pada Computerized Adaptive Test

Aslam Fatkhudin^{a*}, Bayu Surarso^b, Agus Subagio^c

^aPoliteknik Pratama Mulia, Tegal

^bJurusan Matematika, Fakultas Sains dan Matematika, Universitas Diponegoro

^cJurusan Fisika, Fakultas Sains dan Matematika, Universitas Diponegoro

Naskah Diterima : 10 Januari 2014; Diterima Publikasi : 30 Maret 2014

Abstract

One of the computer-based testing is the Computerized Adaptive Test (CAT), which is a computer-based testing system where the items were given to the participants adapted to test the ability of the participants. Assessment methods are usually applied in CAT is Item Response Theory (IRT). IRT models are most commonly used today is the model 3 Parameter Logistic (3PL), which is about the discrimination, difficulty and guessing. However 3PL IRT models have not provided information more objectively test the ability of participants. The opinion of the test participants were tested items were also to be considered. In this study using CAT in combination with IRT model of 4PL. In this research, the development of CAT which uses about 4 parameters, namely the discrimination, difficulty, guessing and questionnaires. The questions used were about UAS 1 English subjects. Samples were taken from 30 students answer with the best value of the total 172 students spread across 6 classes to measure the parameter estimation problem. Further testing using CAT application 4PL IRT models compared to CAT 3PL IRT models. From research done shows that the CAT application combined with IRT models 4PL can measure the ability of the test taker shorter or faster and also opportunities participants correctly answered the test items was done tend to be better than the 3PL IRT models.

Keywords: Ability; CAT; IRT; 3PL; 4PL; Probability; Test

Abstrak

Salah satu pengujian berbasis komputer adalah *Computerized Adaptive Test* (CAT), yaitu sistem pengujian berbasis komputer dimana butir soal yang diberikan kepada peserta tes disesuaikan dengan kemampuan peserta. Metode penilaian yang biasanya diterapkan dalam CAT adalah *Item Response Theory* (IRT). Model IRT yang paling umum digunakan saat ini adalah model 3 Parameter Logistik (3PL), yaitu daya beda soal, tingkat kesukaran soal dan pengecoh. Namun IRT model 3PL belum memberikan informasi kemampuan peserta tes lebih objektif. Pendapat peserta tes terhadap butir soal yang diujikan juga perlu dipertimbangkan. Pada penelitian ini menggunakan CAT yang dikombinasikan dengan IRT model 4PL. Pada penelitian ini dilakukan pengembangan CAT yang menggunakan 4 parameter soal, yaitu daya beda, tingkat kesulitan, pengecoh dan kuesioner. Soal-soal yang digunakan adalah soal UAS 1 mata pelajaran Bahasa Inggris. Sampel diambil dari jawaban 30 siswa dengan nilai terbaik dari total 172 siswa yang tersebar di 6 kelas untuk mengukur estimasi parameter soal. Selanjutnya dilakukan pengujian dengan menggunakan aplikasi CAT model IRT 4PL dibandingkan dengan CAT model IRT 3PL. Dari penelitian yang dilakukan diperoleh hasil bahwa aplikasi CAT yang dikombinasikan dengan IRT model 4PL dapat mengukur kemampuan peserta tes lebih singkat atau cepat dan juga peluang peserta tes menjawab benar butir soal yang dikerjakan cenderung lebih baik dibanding IRT model 3PL.

Kata Kunci : Tingkat kemampuan; CAT, IRT, 3PL, 4PL, Probabilitas, Tes

1. Pendahuluan

Tes adalah salah satu alat yang paling banyak digunakan dalam pendidikan untuk mengukur tingkat kemampuan siswa (Triantafillou *et al.*, 2008). Kemajuan terbaru dalam teknologi komputer dan teori pengujian telah mempercepat perubahan format uji dari konvensional (ujian menggunakan pensil dan kertas) ke pengujian berbasis komputer. Salah satu pengujian berbasis komputer adalah *Computerized Adaptive Test* (CAT). CAT adalah pengujian berbasis komputer dimana butir soal yang

diberikan kepada peserta tes disesuaikan dengan kemampuan peserta tes (Ozyurt *et al.*, 2013).

Metode penilaian yang biasanya diterapkan dalam CAT adalah *Item Response Theory* (IRT) atau Teori Respon Butir (Baker, 2001). IRT merupakan model matematika yang memperhitungkan kemungkinan orang memberikan jawaban dengan benar untuk setiap item (Huang, Lin dan Cheng, 2009).

Skor yang diperoleh pada akhir dari tes ini adalah bukan nilai ujian tetapi estimasi kemampuan yang dikenal sebagai *theta* (θ) yang

*) Penulis korespondensi: fatkhudin@gmail.com

memiliki nilai variasi antara +3 sampai -3. Pada skala θ , nilai 0 merupakan tingkat kemampuan rata-rata, nilai negatif merupakan tingkat kemampuan yang lebih rendah dari nilai rata-rata dan positif merupakan tingkat kemampuan lebih tinggi dari rata-rata (Huang, 2009).

Ada berbagai model yang digunakan dalam IRT. Di antara model IRT yang ada, Model Tiga Parameter Logistik (3PL) adalah salah satu model yang paling umum digunakan. Ketiga parameter tersebut adalah daya beda soal, tingkat kesulitan soal dan tebakan. (Huang, 2009). Namun IRT model 3PL belum memberikan informasi kemampuan peserta tes lebih objektif. Pendapat peserta tes terhadap butir soal yang diujikan juga perlu dipertimbangkan. Untuk itu diperlukan IRT model 4PL dengan kuisioner sebagai parameter keempatnya dari peserta tes dalam memberi penilaiannya terhadap butir soal yang diujikan.

Dari hal inilah yang melatarbelakangi penulis untuk melakukan penelitian dengan menerapkan *Item Response Theory* Model 4 Parameter Logistik pada *Computerized Adaptive Tests* (CAT).

2. Kerangka Teori

2.1. Pustaka Rujukan

Ada sejumlah studi dimana CAT dikombinasikan dengan IRT telah diterapkan. Penelitian pada implementasi modul penilaian individu dalam sistem *e-learning* yang berkaitan dengan proses desain dan pengembangan CAT yang diintegrasikan dengan UZWEBMAT (*Turkish abbreviation of Adaptive and INtelligent WEB based MAThematics teaching-learning system*). UZWEBMAT adalah sistem pakar pembelajaran elektronik (*e-learning*) yang mendukung kemampuan dan kecerdasan individu. Hal ini ditujukan untuk mengajar tingkat probabilitas subjek. Pertama, sebuah bank soal dibuat dalam penelitian ini. Bank soal ini diberikan kepada peserta didik dalam bentuk lembaran kerja. Tes ini dianalisis sesuai dengan IRT (*Item Response Theory*) atau Teori Respon Butir. Model-Data kemampuan beradaptasi diuji pada akhir analisis. Disimpulkan bahwa data dapat disesuaikan dengan IRT model 3PL. Pada akhir operasi ini, bank soal akhir yang berisi 752 pertanyaan secara total diperoleh. Peserta didik diajarkan dengan UZWEBMAT untuk menyelesaikan semua mata pelajaran dengan mengambil bagian terakhir dari tes yang berisi 30 pertanyaan/soal secara acak. Item/soal dalam tes ini secara dinamis dipilih dari bank soal sesuai dengan penilaian adaptif. Dari hasil tersebut diperoleh perkiraan tingkat kemampuan peserta ujian serta informasi tingkat kesulitan tiap pertanyaan untuk selanjutnya disampaikan kepada peserta ujian berikutnya. Tingkat kemampuan peserta ujian itu kembali diperkirakan dan diperbarui sesuai dengan jawaban dari setiap pertanyaan. Berkas modul ini, peserta didik secara individual diajarkan dengan UZWEBMAT disediakan dengan penilaian adaptif sesuai dengan kualifikasi dan kapasitas mereka sendiri. Berbeda dengan tes klasik, penilaian individu menjadi kenyataan melalui penilaian adaptif yang memberikan kesempatan untuk mengetahui

peserta didik sesuai dengan tingkat kemampuan mereka (Ozyurt *et al.*, 2012).

Penelitian lain yang mengkombinasikan antara CAT dengan IRT misalnya menganalisa multi faktor DIF (*Differential Item Functioning*) dalam memberikan ujian nasional mata pelajaran matematika sekolah dasar di Brasil. Hasil analisis data menunjukkan bahwa soal berdasarkan faktor jenis sekolah lebih memuaskan dan konsisten dalam mendistribusikan kemampuan peserta didik dibandingkan dengan soal yang berdasarkan faktor geografis (Goncalves *et al.*, 2013).

2.2. Dasar Teori

Computerized Adaptive Test adalah sistem pengujian berbasis komputer dimana butir soal yang diberikan kepada peserta tes disesuaikan dengan kemampuan peserta tes (Ozyurt *et al.*, 2013). Dalam CAT, peserta tes mengerjakan sejumlah butir soal sesuai dengan kemampuannya secara berulang kali sampai diperoleh kesalahan baku pengukuran $SE(\theta)$ yang semakin konstan serta harga mutlak selisih $SE(\theta)$ antar perulangan penyajian soal yang semakin kecil. Jika nilai $SE(\theta)$ sudah diperoleh, maka proses pengerjaan soal dapat dihentikan. Cara penghentian penyajian soal ini disebut dengan *stopping rule* (Samsul Hadi, 2013).

Untuk mendapatkan nilai kesalahan baku pengukuran $SE(\theta)$ dan nilai selisih $SE(\theta)$ adalah dengan menggunakan rumus persamaan sebagai berikut :

$$SE(\theta) = \frac{1}{\sqrt{\sum I_i(\theta)^2}}$$

Keterangan :

$SE(\theta)$ = nilai kesalahan baku pengukuran

$I_i(\theta)$ = nilai fungsi informasi butir

N = jumlah butir yang ada

Item Response Theory (IRT) atau yang biasa disebut Teori Tes Modern yaitu penelaahan butir soal dengan menggunakan teori jawaban butir soal. Teori ini merupakan suatu teori yang menggunakan fungsi matematika untuk menghubungkan antara peluang menjawab benar suatu skala dengan kemampuan siswa. Nama lain IRT adalah *Latent Trait Theory* (LTT), atau *Characteristics Curve Theory* (ICC). Asal mula IRT adalah kombinasi suatu versi hukum *phi-gamma* dengan suatu analisis faktor butir soal (*item factor analysis*) kemudian bernama Teori Trait Latent (*Latent Trait Theory*), sekarang secara umum dikenal menjadi teori jawaban butir soal (*Item Response Theory*) (McDonald, 1999).

IRT terdiri atas 4 model (Hambleton, 1993), yaitu :

- a. Model 1 Paramater Logistik (1PL)
→ *Difficulty* (tingkat kesulitan),
- b. Model 2 Paramater Logistik (2PL)
→ *Difficulty* (tingkat kesulitan) & *Discrimination* (daya beda),
- c. Model 3 Paramater Logistik (3PL)
→ *Difficulty* (tingkat kesulitan), *Discrimination* (daya beda) & *guessing* (pengecoh)
- d. Model 4 Paramater Logistik (4PL)
→ *Difficulty* (tingkat kesulitan), *Discrimination* (daya beda), *guessing* (pengecoh) & *carelessness*

Persamaan Model 4 Parameter Logistik dirumuskan dengan :

$$P(\theta) = \frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}} \quad (2)$$

Keterangan :

$P(\theta)$ = Probabilitas peserta ujian menjawab dengan benar

θ = estimasi kemampuan peserta ujian menjawab dengan benar

a = nilai parameter *discrimination* / daya beda

b = nilai parameter *difficulty* / tingkat kesulitan

c = nilai parameter *guessing* / pengecoh dari

d = nilai parameter *carelessness* dari soal

e = nilai eksponensial bernilai 2,718

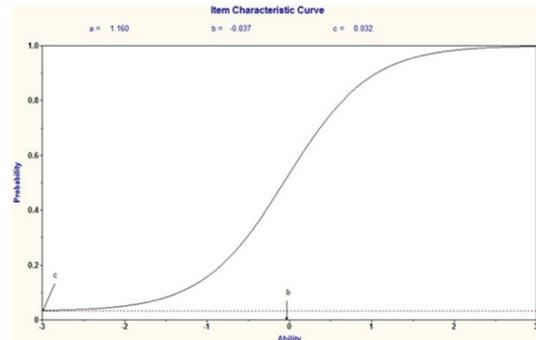
D = faktor penskalaan harganya 1,7

Untuk mendapatkan nilai d (parameter *carelessness*) pada penelitian ini kami menggunakan data kuisisioner untuk menilai masing-masing butir soal yang diujikan. Soal yang digunakan dalam penilaian hasil belajar haruslah *valid*. Soal-soal yang dibuat masih mungkin terjadi kekurangan atau kekeliruan yang menyangkut aspek kemampuan spesifik yang diukur, bahasa yang digunakan, kesalahan ketik dan sebagainya. Oleh karena itu soal terlebih dahulu harus ditelaah.

Dalam teori IRT terdapat kurva utama disebut Kurva Karakteristik Butir Soal (*Item Characteristic Curve*). Kurva ini mewakili karakteristik suatu soal yang menunjukkan kemungkinan seorang peserta tes dengan kemampuan tertentu (θ) dapat menjawab dengan benar. Kurva ini juga disebut *Item Response Function* (IRF) yang dinotasikan dengan $P(\theta)$.

Pada Gambar 1 menunjukkan contoh ICC dari item sesuai dengan model 3 Parameter Logistik (3PL). Ketiga paramater itu masing-masing yaitu : a adalah indeks diskriminasi butir dengan nilai pada gambar tersebut sebesar 1,16, b adalah tingkat kesulitan butir soal dengan nilai -0,037 dan c adalah parameter estimasi dengan nilai 0,032, θ adalah tingkat kemampuan seseorang; e adalah 2,718; $P_i(\theta)$ adalah probabilitas seseorang memberikan jawaban yang tepat untuk setiap i pertanyaan dengan tingkat kemampuan θ . Model 2PL dapat diturunkan dari 3PL dengan menetapkan $c = 0$ dan model 1PL dapat diperoleh dengan menetapkan $c = 0$ dan $a = 1$.

Kurva pada gambar 1. diperoleh dari perhitungan pada tabel 1. Tabel probabilitas model IRT 1PL. Pada tabel 1, angka 0,05; 0,12 dan seterusnya didapatkan dengan memasukkan persamaan 2.



Gambar 1. Contoh *Item Characteristic Curve* (Sumber : Ozyurt *et al.*, 2012)

Tabel 1. Tabel probabilitas model IRT 3PL

	Theta						
	-3	-2	-1	0	1	2	3
Soal	0,05	0,12	0,27	0,50	0,73	0,88	0,95

3. Metodologi

3.1. Bahan Penelitian

Sumber bahan yang dipergunakan dalam penelitian ini ada 3 (tiga), yaitu soal yang didapat dari guru pengampu mata pelajaran, jawaban dari siswa/peserta ujian dan kuisisioner penelaahan soal dari siswa/peserta ujian. Tujuan adanya kuisisioner penelaahan soal dari siswa/peserta ujian supaya penilaian lebih objektif karena bisa jadi soal yang diberikan oleh guru pengampu mata pelajaran, menyimpang atau tidak sesuai dengan materi soal yang telah diajarkannya.

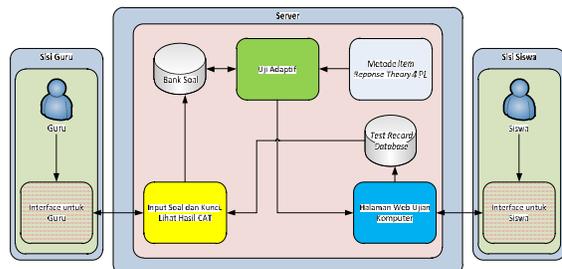
Soal yang didapat dari guru pengampu kemudian diinputkan ke dalam sistem untuk selanjutnya menjadi data bank soal. Penginput soal bisa dari *administrator* sistem atau guru yang bersangkutan, dalam hal ini berarti harus terdapat data guru yang didapat dari pihak sekolah yang sudah diinputkan ke dalam sistem oleh *admin*. Agar siswa/peserta ujian bisa mulai mengerjakan soal, siswa harus *login* terlebih dahulu ke dalam sistem. Karenanya perlu diinputkan juga data siswa oleh *admin*. Kuisisioner yang sudah didapat, kemudian diolah agar menghasilkan nilai d sebagai parameter keempat dari metode IRT untuk selanjutnya melekat pada masing-masing butir dari bank soal yang sudah diinputkan ke dalam sistem.

2.3. Alat Penelitian

Alat penelitian yang berkaitan dengan hardware yang dipergunakan adalah laptop sebagai alat utama penelitian sekaligus servernya dengan spesifikasi processor core-i3 2,4 GHz; RAM DDR3 dengan kapasitas 4 Gbyte; HardDisk 500 Gbyte; VGA onboard, Monitor dengan ukuran 14 inci; DVD+RW, LAN; WLAN serta Bluetooth. Sedangkan siswa/peserta ujian dalam mengerjakan soal dengan menggunakan komputer yang ada di laboratorium komputer sekolah mereka. Spesifikasi komputernya adalah processor dual core 1,75 GHz; RAM DDR2 dengan kapasitas 2 Gbyte, HardDisk 120 Gbyte, VGA onboard, LED Monitor dengan ukuran 15 inch serta DVD+RW. Karena sistem yang dibuat berbasis *localhost*, maka dibutuhkan alat tambahan berupa *switch* dan kabel UTP yang sudah diinstalasi konektor RJ-45 untuk menghubungkan semua komputer dalam laboratorium tersebut.

2.4. Prosedur Penelitian

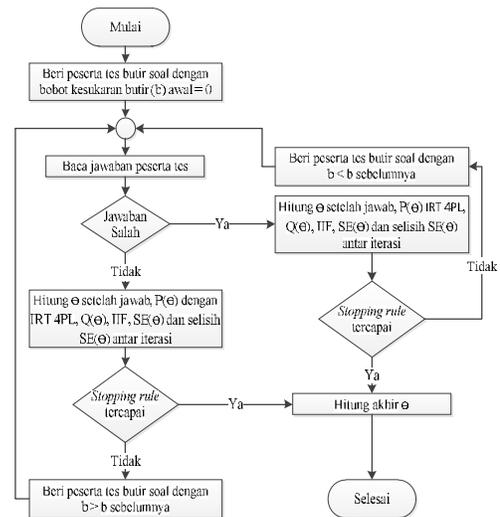
Pada tahap ini kami mencari dan mempelajari literatur pustaka dari sejumlah jurnal atau penelitian serta buku-buku referensi yang terkait dengan metode yang digunakan, yaitu *Item Response Theory* dan *Computerized Adaptive Test*. Selanjutnya dilakukan observasi untuk mengambil data dari tempat penelitian di SMA Muhammadiyah 1 Pekajangan. Data yang dibutuhkan adalah data soal beserta kunci jawaban dan kisi-kisinya, data guru dan data siswa serta kuisioner butir soal yang didapat dari siswa. Tahap Perancangan ini dibuat arsitektur sistem informasi dan *flowchart* programnya untuk kemudian dirancang DFD dan tabel. Arsitektur sistem terdiri dari dua bagian, yaitu sisi guru dan sisi siswa (Gambar 2). Disisi guru, dengan menggunakan desktop atau laptop, guru mengakses halaman untuk guru. Di halaman atau *interface* tersebut, guru bisa *inputkan* soal beserta kunci jawabannya. Setelah soal tersebut dikerjakan oleh peserta ujian/siswa, sistem akan melaksanakan sebuah penilaian ujian akhir semester untuk uji adaptif dari soal. Pada sisi siswa, dengan menggunakan desktop atau laptop seorang siswa mengerjakan ujian adaptif. Untuk mengembangkan sistem, diperlukan sebuah bank soal. Sistem akan menerapkan metode IRT model 4 parameter logistik untuk memperkirakan kemampuan siswa setelah ia menjawab soal tes.



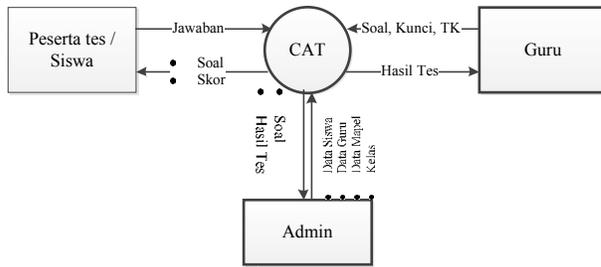
Gambar 2. Arsitektur CAT dikombinasikan dengan IRT Model 4PL

Gambaran *flowchart* programnya adalah peserta tes diberi soal dengan nilai parameter *difficulty* dari soal awal sedang (b awal = 0) karena dianggap semua peserta tes mempunyai tingkat kemampuan (θ) sedang (θ awal = 0). Jika soal awal tersebut dapat dijawab dengan benar, diberikan soal berikutnya yang lebih sulit, tetapi jika soal awal tersebut dijawab salah maka berikutnya diberi soal yang lebih mudah. Proses tersebut dilakukan berulang-ulang sampai *stopping rule* tercapai. Kemudian tingkat kemampuan akhir (θ) peserta tes dan probabilitas yang menggunakan IRT model 4PL dihitung. Berikutnya menghitung nilai total tes. Terakhir, sistem menampilkan informasi berupa diagram dan keterangan dari hasil perhitungan tersebut. Gambar 3 menunjukkan *flowchart* program *Computerized Adaptive Test* yang dikombinasikan dengan IRT model 4PL.

Sedangkan *Data Flow Diagram* (DFD)-nya adalah siswa dan guru memberikan datanya ke dalam sistem. Data diterima oleh administrator untuk dikonfirmasi kemudian dimasukkan ke dalam sistem sebagai data yang *valid*. Guru *inputkan* soal ke dalam sistem. Soal yang sudah dimasukkan ke sistem kemudian akan dikerjakan oleh siswa. Siswa dan guru otomatis akan menerima skor dan hasil tes dari sistem. Alur tersebut bisa digambarkan dengan Diagram Konteks sebagai ditunjukkan pada gambar 4.



Gambar 3. Flowchart CAT dikombinasikan dengan IRT Model 4PL



Gambar 4. Konteks Diagram

Dari DFD di atas, dihasilkan beberapa tabel sebagai berikut:

Tabel 2. Struktur Tabel Peserta

No.	Nama Field	Tipe dan Panjang Field	Keterangan
1	Kd_peserta	Int (10)	Kode siswa / peserta tes
2	Nama_peserta	Varchar (50)	nama lengkap siswa
3	Nomor_peserta	Varchar (40)	Nomor Induk Peserta Didik
4	Password	Varchar (10)	Password
5	Kelas	Varchar (10)	Kelas

Tabel 3. Struktur Tabel Kelas

No.	Nama Field	Tipe dan Panjang Field	Keterangan
1	Kd_kelas	Int (10)	Kode kelas
2	Kelas	Varchar (50)	Kelas

Tabel 4. Struktur Tabel Mapel

No.	Nama Field	Tipe dan Panjang Field	Keterangan
1	Kd_mapel	Int (10)	Kode mata pelajaran
2	Mapel	Varchar (100)	Nama mata pelajaran

Tabel 5. Struktur Tabel Guru

No.	Nama Field	Tipe dan Panjang Field	Keterangan
1	Kd_Guru	Int (5)	Nomor urut
2	Kode	Varchar (30)	Kode guru
3	Nama	Varchar (50)	Nama lengkap guru
4	Mapel	Varchar (50)	Mata pelajaran yang diampu
5	Password	Varchar (50)	Password

Tabel 6. Struktur Tabel Soal

No.	Nama Field	Tipe dan Panjang Field	Keterangan
1	Kd_Soal	Int (10)	Kode soal
2	Isi_Soal	Text	Isi soal
3	Kunci_Soal	char (1)	Kunci soal
4	Tingkat_kesulitan	float	Nilai tingkat kesulitan
5	Pengecoh	Char (1)	Pengecoh jawaban

Tabel 7. Struktur Tabel Daftar_Tes

No.	Nama Field	Tipe dan Panjang Field	Keterangan
1	Kd_judul_tes	Int (10)	Kode judul tes
2	Judul	Varchar (50)	Judul tes

Tabel 8. Struktur Tabel Tes

No.	Nama Field	Tipe dan Panjang Field	Keterangan
1	Kd_Tes	Int (10)	Kode test
2	Kd_judul_tes	Int (10)	Kode judul tes
3	Waktu_test	Timestamp	Waktu mengerjakan test
4	Kd_peserta	Int (10)	Nomor Induk Peserta Didik

Tabel 9. Struktur Tabel Detail_Tes

No.	Nama Field	Tipe dan Panjang Field	Keterangan
1	Kd_DetailTes	Int (20)	Kode detail tes
2	Kd_tes	Int (10)	Kode tes
3	Kd_Soal	Int (10)	Kode soal
4	b	float	Tingkat kesulitan
5	Jawaban	Char (1)	Jawaban butir soal
6	Skor_Jawaban	Int (4)	Skor jawaban
7	Teta_awal	float	Nilai teta awal
8	Teta_akhir	float	Nilai teta akhir
9	d	float	Kuisisioner
10	P	float	Probabilitas menjawab benar
11	Q	float	Probabilitas menjawab salah
12	I	float	Informasi butir
13	SE	float	Kesalahan baku
14	Selisih_SE	float	Selisih kesalahan baku

4. Hasil dan Pembahasan

4.1. Hasil

Penelitian ini mengambil subjek siswa/siswi peserta Ujian Akhir Semester (UAS) 1 mata

pelajaran Bahasa Inggris Tahun Pelajaran 2013/2014 di SMA Muhammadiyah 1 Pekajangan Pekalongan. Objek yang dipilih adalah butir soal objektif pilihan ganda dan respons butir peserta UAS 1 Bahasa Inggris. Data penelitian berupa butir soal pilihan ganda sejumlah 45 butir dan respons butir peserta UAS 1 Bahasa Inggris yang dikumpulkan dengan teknik dokumentasi dan kuisioner. Populasi penelitian ini adalah peserta UAS 1 Bahasa Inggris meliputi 6 kelas atau 172 peserta. Dari 172 peserta tersebut, sebagai sampel diambil 30 peserta yang memilih jawaban benar dengan rentang lebih dari 28 soal ($\pm 17,44\%$). Berdasarkan sampel yang diperoleh, selanjutnya dilakukan penskoran dan analisa butir soal dengan metode *Item Response Theory* (IRT) model 4 Parameter Logistik (PL) dan kemudian dibuatkan sistem *Computerized Adaptive Test* (CAT).

Hasil Analisis Butir Soal dengan metode IRT model 4 PL diketahui dengan langkah pertama adalah membuat tabel masukan yang berisi kunci jawaban dan pengecoh dari masing-masing butir soal serta respon butir atau jawaban dari masing-masing peserta. Dari tabel masukan kemudian dibuat tabel penskoran. Jika jawaban sesuai kunci jawaban (benar), maka akan bernilai 1, sebaliknya jika jawaban tidak sesuai kunci jawaban (salah), maka akan bernilai 0. Dari tabel penskoran akan didapatkan skor/nilai dengan menjumlah semua jawaban benar untuk masing-masing peserta ujian kemudian dikalikan 100 dan dibagi jumlah soal yaitu 45 butir soal. Dari proses pemasukan dan penskoran, didapatkan hasil sebaran 9 (sembilan) kelompok nilai seperti yang ditampilkan pada Tabel 10.

Selanjutnya tabel penskoran tersebut diurutkan secara *descending* atau dari nilai terbesar ke nilai terkecil. Dari tabel pengurutan bisa diketahui ranking, nilai terbesar yaitu 82 dan nilai terkecilnya yaitu 62. Nilai tersebut dipakai untuk menghitung *proportion correct* (proporsi jawaban benar peserta ujian) dan nilai estimasi kemampuan peserta ujian (*theta* (θ)). Nilai *theta* (θ) ini nantinya akan digunakan untuk menghitung *probabilitas* peserta ujian ($P(\theta)$).

Langkah berikutnya setelah tabel pengurutan terbentuk adalah membaginya dalam dua kelompok, yaitu 50% kelompok atas (kelompok nilai besar) dan 50% kelompok bawah (kelompok nilai kecil). Tujuannya untuk mendapatkan nilai daya beda (*a*), nilai tingkat kesulitan (*b*) dan nilai pengecoh (*c*). Tabel 11. menunjukkan keterangan kualitas soal hasil dari ketiga nilai parameter logistik tersebut.

Daya beda berfungsi untuk menghitung tingkat validitas dan menggambarkan tingkat kemampuan soal dalam membedakan antar peserta didik yang sudah memahami materi yang diujikan dengan peserta didik yang belum/tidak memahami materi yang diujikan. Adapun klasifikasinya adalah seperti berikut ini (Crocker dan Algina, 1986) :

1. 0,40 - 1,00 soal diterima baik (baik sekali)
2. 0,30 - 0,39 soal diterima tetapi perlu diperbaiki (baik)
3. 0,20 - 0,29 soal diperbaiki

4. 0,19 - 0,00 soal tidak dipakai/dibuang

Tingkat kesulitan soal berfungsi untuk mengetahui peluang menjawab benar suatu soal pada tingkat kemampuan tertentu yang biasanya dinyatakan dalam bentuk indeks. Indeks tingkat kesulitan ini pada umumnya dinyatakan dalam bentuk proporsi yang besarnya berkisar 0,00 - 1,00 (Aiken, 1994). Semakin besar indeks tingkat kesulitan yang diperoleh dari hasil hitungan, berarti semakin mudah soal itu. Suatu soal memiliki TK= 0,00 artinya bahwa tidak ada siswa yang menjawab benar dan bila memiliki TK=1,00 artinya bahwa siswa menjawab benar. Perhitungan indeks tingkat kesulitan ini dilakukan untuk setiap nomor soal.

Tabel 10. Tabel hasil sebaran kelompok nilai

No	Kelompok Nilai	Jumlah Siswa
1	82	2
2	78	1
3	76	2
4	71	4
5	67	1
6	69	3
7	67	3
8	64	5
9	62	9
Jumlah		30

Tabel 11. Tabel keterangan kualitas soal

No.	Parameter	Keterangan	Jumlah
1	Daya Beda	Soal baik sekali	2
		Soal baik	1
		Soal diperbaiki	13
		Soal dibuang	29
		Jumlah	45
2	Tingkat Kesulitan	Soal sukar	6
		Soal sedang	14
		Soal mudah	25
		Jumlah	45
3	Pengecoh	Fungsi	35
		Tidak Fungsi	10
		Jumlah	45

Adapun klasifikasinya adalah seperti berikut ini :

1. 0,00 - 0,30 soal tergolong sukar
2. 0,31 - 0,70 soal tergolong sedang
3. 0,71 - 1,00 soal tergolong mudah

Faktor pengecoh berfungsi untuk mengetahui berfungsi tidaknya jawaban yang tersedia. Butir soal yang baik, pengecohnya akan dipilih secara merata oleh siswa/siswi yang menjawab salah (5%).

Sedangkan parameter ke-4 (parameter *carelessness*) atau nilai *d* diambilkan dari kuisioner peserta ujian untuk menilai sendiri

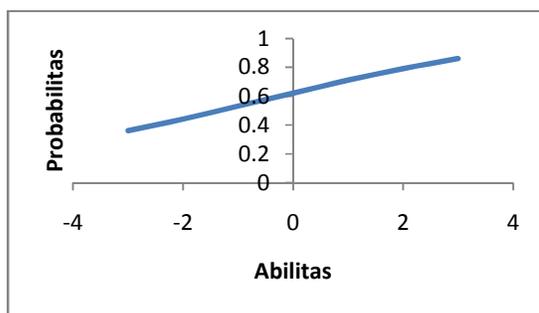
butir soal. Hal ini dimaksudkan agar penilaian terhadap butir soal lebih objektif. Setelah semua nilai parameter diketahui, berikutnya memasukkan nilai-nilai tersebut dalam program CAT.

Computerized Adaptive Test dilakukan dengan mengambil sebuah contoh seorang siswa yang mengerjakan *computerized adaptive test* didapatkan hasil seperti tertera pada tabel 12. Detail Hasil Tes.

Tabel 12 menggambarkan jumlah soal yang dijawab oleh peserta tes. Dari total jumlah soal 45 butir, dengan pola jawaban yang diinputkan, peserta tes hanya menjawab 10 butir soal secara acak seperti yang ditunjukkan pada kolom i. Kolom i merupakan kode soal. Kolom b adalah bobot atau nilai tingkat kesukaran butir soal. Pada kolom skor menampilkan nilai dari jawaban peserta tes.

Tabel 12. Tabel detail hasil tes

i	b	Skor	θ Awal	θ Stl Jwb	d	Pi(θ)	Q(θ)	Selisih SE
138	0	1	0	0	1,00	0,50	0,50	2,00
131	0,07	0	0	0	0,67	0,69	0,31	0,53
105	-0,02	1	0	0,02	0,33	0,17	0,83	0,18
144	0,18	1	-0,02	0,18	0,33	0,17	0,83	0,13
113	0,2	1	0,18	0,2	0,33	0,17	0,83	0,10
140	0,38	0	0,2	0,2	0,67	0,72	0,28	0,10
141	-0,11	0	0,2	0,2	1,00	0,37	0,63	0,09
130	-0,16	1	0,2	0,16	1,00	0,50	0,50	0,07
132	0,41	0	-0,16	0,16	0,67	0,82	0,18	0,04
119	-0,39	1	-0,16	0,39	0,33	0,17	0,83	0,03



Gambar 5. Grafik *Probabilitas* IRT model 4PL

Kolom skor ini berkorelasi dengan kolom θ setelah menjawab. Jika peserta tes menjawab dengan benar, maka akan bernilai 1 dan nilai pada kolom θ setelah menjawab akan berubah positif yang menunjukkan bahwa kemampuan peserta tes meningkat, sebaliknya jika jawaban peserta tes salah, maka akan bernilai 0 dan nilai pada kolom θ setelah menjawab akan berubah negatif, menunjukkan kemampuan peserta tes menurun. Nilai Omerupakan tingkat kemampuan rata-rata. Kolom d adalah nilai kuisioner yang diberikan oleh peserta tes di tiap butir soal yang dikerjakan. Kolom Pi(θ) menunjukkan nilai *probabilitas*/peluang peserta tes menjawab dengan benar.

Sedangkan Q(θ) menunjukkan nilai *probabilitas*/peluang peserta tes menjawab salah. Kolom selisih SE menunjukkan nilai mutlak selisih kesalahan baku pengukuran. Nilai pada kolom SE inilah yang menjadi acuan penghentian penyajian soal. Bila selisih SE sudah 0,01 maka tes akan berhenti. Di kolom selisih SE tabel 9 di atas nilainya yang terakhir belum 0,01 tetapi penyajian soal tetap berhenti di soal ke-10, itu terjadi karena kemungkinan bila penyajian soal dilanjutkan, selisih SE berikutnya akan bernilai dibawah 0,01 yaitu 0,00.

Dari data pada tabel 12 dapat diperoleh grafik hubungan *probabilitas* dengan tingkat kemampuan setiap kali peserta ujian menjawab soal yang diberikan oleh sistem. Dengan mengambil contoh soal terakhir yang dikerjakan oleh peserta tes yaitu soal ke-10 didapatkan grafik probabilitas yang naik seiring perubahan nilai kemampuan peserta ujian, seperti ditampilkan pada Gambar 5 di atas. Artinya peluang peserta tes untuk menjawab benar butir soal meningkat, *ekuivalen* dengan kemampuan peserta tes.

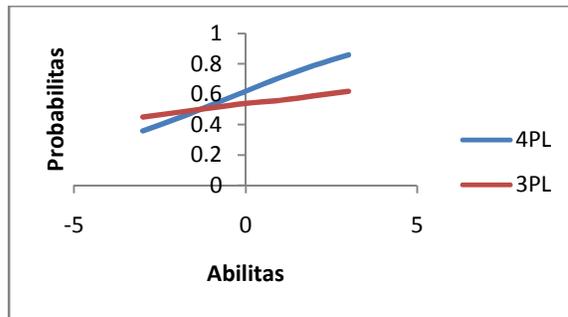
4.2. Pembahasan

Membandingkan CAT sebelumnya yang menggunakan metode IRT model 3PL tanpa menggunakan nilai d (nilai kuisioner dari peserta tes di tiap butir soal) sebagai parameter ke empatnya, dengan CAT pada penelitian ini yang menggunakan metode IRT model 4PL (menggunakan nilai d), menggunakan data dari hasil penelitian seperti pada tabel 12 dan gambar 5 grafik *probabilitas*, diperoleh hasil seperti ditampilkan pada tabel 13.

Tabel 13 menunjukkan peserta tes menjawab soal sebanyak 13 butir secara acak dari sejumlah butir soal yang ada yaitu 45 butir. Ditunjukkan pada nilai terakhir di kolom selisih SE (kesalahan baku pengukuran) yaitu 0,01 yang menjadi batas acuan untuk menghentikan penyajian soal.

Tabel 13. Tabel detail hasil tes CAT yang menggunakan metodel IRT model 3PL

i	b	Skor	θ Awal	θ Stl Jwb	Pi(θ)	Q(θ)	Selisih SE
138	0	1	0	0	0,50	0,50	2,00
131	0,07	0	0	0	0,53	0,47	0,58
105	-0,02	1	0	-0,02	0,50	0,50	0,26
144	0,18	1	-0,02	0,18	0,50	0,50	0,15
113	0,2	1	0,18	0,2	0,50	0,50	0,11
140	0,38	0	0,2	0,2	0,58	0,42	0,08
141	-0,11	0	0,2	0,2	0,37	0,63	0,06
130	-0,16	1	0,2	-0,16	0,50	0,50	0,05
132	0,41	0	-0,16	-0,16	0,72	0,28	0,03
119	-0,39	1	-0,16	-0,39	0,50	0,50	0,04
115	0,8	1	-0,39	0,8	0,50	0,50	0,03
137	1,4	0	0,8	0,8	0,73	0,27	0,02
110	-0,73	0	0,8	0,8	0,07	0,93	0,01



Gambar 6. Grafik perbandingan probabilitas IRT model 3PL dengan IRT model 4PL

Berbeda dengan CAT yang menggunakan metode IRT model 4PL seperti ditunjukkan pada tabel 9 di atas, dimana peserta tes hanya menjawab soal sebanyak 10 butir saja. Hal ini menjelaskan kelebihan dan kontribusi dari penelitian ini. Aplikasi CAT dengan menggunakan metode IRT model 4PL dapat mengukur kemampuan peserta tes lebih singkat atau cepat. Apabila data di tabel 13 dibuat grafik, dengan mengambil contoh soal terakhir yang dikerjakan oleh peserta tes yaitu soal ke-13 kemudian dibandingkan dengan grafik probabilitas IRT model 4PL yang terdapat pada gambar 5 di atas, akan diperoleh grafik perbandingan seperti gambar 6 di atas. Pada grafik ditunjukkan bahwa peluang (probabilitas) peserta tes menjawab benar pada IRT model 4PL lebih baik dibandingkan peluang (probabilitas) peserta tes menjawab benar pada IRT model 3PL. Grafik tersebut juga menjelaskan makin banyak parameter logistik yang digunakan, maka kurva probabilitasnya akan semakin meningkat. Sekali lagi ini menjelaskan kelebihan dari sistem pada penelitian, yaitu *probabilitas*/peluang peserta tes menjawab benar cenderung naik.

3. Kesimpulan

Setelah melakukan pengujian terhadap metode IRT model 4PL yang dikombinasikan dengan *Computerized Adaptive Test* (CAT) maka didapatkan kesimpulan bahwa estimasi terhadap parameter soal dapat dilakukan secara otomatis oleh sistem. Dengan menggunakan algoritma *Computerized Adaptive Test* (CAT) dengan *Item Response Theory* model 4PL, pembuat soal dapat mengestimasi parameter soal tanpa harus bergantung kepada para ahli.

Parameter soal yang diperoleh dari estimasi parameter secara otomatis tersebut dapat digunakan untuk *adaptive online test*. Dari data hasil penelitian yang terdapat pada pembahasan, menunjukkan bahwa aplikasi CAT yang menggunakan metode IRT model 4PL dapat mengukur tingkat kemampuan dan *probabilitas*/peluang peserta tes menjawab benar butir soal dengan dengan lebih baik. Kelebihan dari aplikasi CAT yang menggunakan metode IRT model 4PL adalah dapat mengukur kemampuan peserta tes lebih singkat atau cepat dan juga peluang peserta tes menjawab benar butir soal yang dikerjakan cenderung lebih baik dibandingkan dengan aplikasi CAT

yang menggunakan metode IRT model 4PL. Semakin banyak parameter logistik yang digunakan, maka kurva probabilitasnya akan semakin meningkat.

Ucapan Terima Kasih

Penulis mengucapkan terimakasih kepada Kepala Sekolah dan Guru Bahasa Inggris kelas X SMA Muhammadiyah 1 Pekalongan yang telah mengizinkan penulis melakukan penelitian pada institusinya.

Daftar Pustaka

- Aiken and Lewis R., 1994. *Psychological Testing and Assessment*, (Eight Edition), Boston : Allyn and Bacon.
- Baker, F., 2001. The basics of item response theory. ERIC clearinghouse on assessment and evaluation. College Park, MD: University of Maryland.
- Crocker, L. and Algina, J., 1986. Introduction to Classical and Modern Test Theory. New York : Holt, Rinehart and Winston, Inc.
- Goncalves, F.B., Gamerman, D., Soares T.M., 2013. Simultaneous multifactor DIF analysis and detection in Item Response Theory, *Computational Statistics and Data Analysis* 59, 144 – 160.
- Hambleton, R.K., 1993. Principles and Selected Applications of Item Response Theory. In Linn, Robert L. (Editor). *Educational Measurement*. Third Edition. Phoenix: American Council on Education, Series on Higher Education Oryx Press.
- Huang Y.M., Lin, Y.T. and Cheng S.C., 2009. An adaptive testing system for supporting versatile educational assessment, *Journal of Computers & Education* 52, 53–67.
- McDonald R.P., 1999. *Test Theory: A Unified Treatment*, Lawrence Erlbaum Associates Publishers, New Jersey.
- Ozyurt H., Ozyurt O., Baki A. and Guven B., 2012. Integrating computerized adaptive testing into UZWEBMAT : Implementation of individualized assessment module in an e-learning system, *Journal Expert System with Application*, 39, 9837 – 9847.
- Ozyurt H., Ozyurt O. and Baki A., 2013. Design and development of an innovative individualized adaptive and intelligent e-learning system for teaching-learning of probability unit: Details of UZWEBMAT, *Journal Expert System with Application*, 40, 2914 – 2940.

- Samsul, H., 2013. Pengembangan Computerized Adaptive Test Berbasis Web, Yogyakarta : Aswaja Pressindo.
- Sharkness J. and DeAngelo L., 2011. Measuring Student Involvement: A Comparison of Classical Test Theory and Item Response Theory in the Construction of Scales from Student Surveys, *Journal Springer Res High Educ* 52, 480–507.
- Triantafillou E., Georgiadou E. and Economides A.A., 2008. The design and evaluation of a computerized adaptive tes on mobile devices, *Journal of Computers & Education* 50, 1319–1330.