PENGEMBANGAN *COMPUTERIZED ADAPTIVE TESTING* UNTUK MENGUKUR HASIL BELAJAR MAHASISWA UNIVERSITAS TERBUKA

Agus Santoso
Universitas Terbuka Jakarta
aguss@mail.ut.ac.id

Abstrak

Penelitian ini bertujuan untuk mengembangkan ujian berbasis komputer yang diaplikasikan Universitas Terbuka (CBT-UT) ke tes adaptif berbasis komputer (CAT) untuk Mata kuliah Bahasa Indonesia. Penelitian ini mempunyai tiga tahap, yaitu tahap pengembangan bank soal, simulasi algoritma, dan pengembangan program CAT. Pada tahap pertama, butir-butir soal yang memenuhi kriteria dipilih untuk masuk bank soal. Pada tahap kedua, studi simulasi untuk menguji apakah algoritma yang diterapkan berproses dengan baik ataukah tidak, yaitu algoritma desain CAT murni dan CAT yang dikendala modul (CCAT). Pada tahap ketiga, program aplikasi CAT pada sistem ujian UT dikembangkan. Hasil penelitian menunjukkan bahwa: (1) 127 butir yang memenuhi kriteria ideal sebagai bank soal CAT, (2) bank soal data empiris akurat untuk mengukur peserta tes dengan kemampuan sedang sampai tinggi, (3) untuk mengestimasi kemampuan peserta tes pada tingkat kemampuan sedang sampai tinggi, pada desain CAT murni diperlukan 8 sampai 15 butir soal, (4) Desain CAT yang dikendala modul (CCAT50) dapat diterapkan oleh UT sebagai bentuk ujian alternatif pada sistem ujian UT. Desain ini dapat juga digunakan oleh lembaga pendidikan lain yang dalam pengujiannya mensyaratkan keseimbangan isi.

Kata kunci: *tes adaptif berbasis komputer*

DEVELOPMENT OF COMPUTERIZED ADAPTIVE TESTING TO MEASURE LEARNING OUTCOMES OF THE UT STUDENTS

Agus Santoso
Dept. Open University Jakarta
aguss@mail.ut.ac.id

Abstract

This study aims to develop a computer-based testing applied by the Universitas Terbuka (CBT–UT) in the computer-based adaptive testing (CAT) for Indonesian Language course. This study has three stages, namely the development of item bank, the simulation of algorithms, and the development of the CAT program. First, the items that satisfied the criteria for a good test item were included in the item bank for CAT. Second, a simulation to test the capability of the algorithms of processing the data properly. The simulation involved two algorithm designs, a pure CAT and a module-constrained CAT (CCAT). While in the third stage, the CAT application program in the UT's testing system is developed. The study found: (1) 127 items which is ideal for CAT item bank. (2) The empirical data item bank accurately measures examinee with moderate to high level of abilities. (3) To estimate examinee with moderate to high level of abilities, the pure CAT designs need 8 to 15 items. (4) Module-Constrained CAT (CCAT50) design can be implemented by UT as an alternative test at the UT's testing system. This model can also be used in other educational institutions that requires content balancing in the assessment.

Key words: *computer based adaptive testing*

Pendahuluan

Sejak tahun 2006 Universitas Terbuka (UT) telah menerapkan sistem ujian berbasis komputer (*Computer Based Testing*, CBT), di samping menggunakan *paper and pencil test* (PPT) yang selama ini telah diselenggarakan. CBT dikembangkan berdasarkan pemanfaatan teknologi internet, dengan mempertimbangkan sarana komputer di Unit Program Belajar Jarak Jauh UT yang tersebar di 37 lokasi di Indonesia.

CBT yang dikembangkan UT didasarkan pada rancangan *nonaditif linear fixed-form test*, artinya tingkat kesukaran butir soal tes tidak disesuaikan dengan kemampuan peserta tes, setiap peserta mengerjakan sejumlah butir soal tertentu dengan jumlah butir soal adalah tetap. Penyelenggaraan tes yang memberikan sejumlah butir soal yang sama pada setiap peserta tes seperti pada sistem ujian akhir semester UT dengan CBT maupun PPT kurang efisien, khususnya untuk peserta tes dengan kemampuan rendah dan tinggi (Lord, 1980: 150; Hambleton, Swaminathan, & Rogers, 1991: 145). Hal ini karena banyak butir soal yang tidak mampu memberikan informasi berguna dalam membedakan peserta tes dalam rentang kemampuan tertentu. Peserta tes yang memiliki kemampuan tinggi mendapat beberapa butir soal yang mudah, di mana mereka memiliki peluang yang kecil menjawab salah. Dengan demikian, butir soal seperti itu tidak menyediakan informasi tentang kemampuan mereka. Sebaliknya, peserta tes dengan kemampuan rendah akan mendapatkan beberapa butir soal yang sukar, di mana mereka memiliki peluang yang kecil untuk menjawab butir soal dengan benar, akibatnya jawaban mereka yang salah hanya memberikan sedikit informasi mengenai kemampuan mereka.

Oleh karena itu, untuk meningkatkan efisiensi dan akurasi dalam mengukur kemampuan peserta tes, maka UT perlu menerapkan tes adaptif pada penyelenggaraan CBT. Adaptif memiliki pengertian bahwa butir soal (tes) yang diberikan disesuaikan dengan tingkat kemampuan setiap peserta tes atau *tailored testing* (Lord, 1980:151). Penyelenggaraan tes adaptif berbasiskan komputer ini populer disebut dengan *Computerized Adaptive Testing* (CAT).

CAT didasarkan pada item *response theory* (IRT). Pada CAT, komputer tidak hanya sekedar memindahkan butir soal ke dalam komputer, tetapi komputer diatur untuk menyeleksi dan memberikan butir soal, selanjutnya komputer menskor jawaban peserta. Kemudian komputer memilih butir soal baru untuk diberikan lagi kepada peserta. Butir soal yang diberikan adalah butir soal yang memberikan informasi tertinggi untuk peserta berdasarkan jawaban butir soal sebelumnya, proses ini berlanjut terus sampai aturan pemberhentian telah tercapai, yaitu tes dihentikan jika sejumlah butir tes tertentu telah diberikan atau presisi pengukuran yang telah ditentukan telah tercapai. Melalui proses ini umumnya peserta tes akan menerima butir soal yang sesuai dengan kemampuan mereka dan menghindari butir soal yang terlalu sulit atau terlalu mudah untuk mereka.

CAT sedikitnya mempunyai empat kelebihan dibandingkan dengan PPT atau CBT. Pertama, dengan memberikan butir-butir soal yang sesuai dengan kemampuan individu peserta tes maka CAT dapat lebih meningkatkan efisiensi, karena butir soal yang terlalu mudah atau terlalu sulit dapat dihindari sehingga panjang tes dapat berkurang tanpa mengurangi tingkat presisi pengukuran (Wainer, 1990:10; Hambleton, Swaminathan, & Rogers, 1991:146, Weiss & Schleisman, 1999:130). Kedua, karena CAT mengambil butir soal dari bank soal yang sudah terkalibrasi dan tersimpan secara elektronik maka keamanan tes lebih terjamin. Ketiga, skor peserta pada CAT dapat segera diketahui oleh peserta tes, karena komputer langsung menskor dan mengestimasi kemampuan peserta segera setelah peserta menjawab butir-butir soal yang diberikan. Keempat, dengan memanfaatkan kemajuan teknologi komputer, tampilan format butir soal yang tidak dapat dilakukan pada PPT seperti animasi dan suara dapat dilakukan pada CAT (Vispoel, 1999: Ackerman et al, 1999).

Namun demikian, ada beberapa hal yang perlu dipertimbangkan dalam mengembangkan CAT, seperti pertimbangan teknis dan teoretis. Masalah teknis yang berkaitan dengan komponen penting dari CAT, seperti metode untuk memilih item pertama dan metode untuk memilih item berikutnya, harus dievaluasi untuk memastikan bahwa skor yang dihasilkan dari desain CAT adalah valid dan dapat diandalkan. Selain itu, untuk memenuhi tes yang standar maka keseimbangan isi pada PPT mungkin

perlu dipertimbangkan ketika mengembangkan CAT. Menurut Green, et al. (1984) dan Kingsbury & Zara (1989) pengembangan CAT memerlukan evaluasi pada enam komponen yaitu: (1) model respons butir, (2) bank soal, (3) pemilihan butir soal awal, (4) metode pengestimasian tingkat kemampuan, (5) prosedur pemilihan butir soal, dan (6) aturan pemberhentian.

Penelitian ini bertujuan untuk mengembangkan ujian berbasis komputer, CBT ke ujian adaptif berbasis komputer (CAT). Dengan menerapkan CAT pada UAS UT beberapa keuntungan dapat diperoleh seperti: pengukuran tingkat kemampuan peserta tes lebih efisien dan reliabel, keamanan tes terkendali, proses perakitan soal tidak diperlukan lagi, skor dapat segera diketahui oleh peserta tes, dan memberikan layanan ujian alternatif pada ujian akhir semester UT.

Mata kuliah yang dipilih untuk di-CAT-kan adalah Bahasa Indonesia. Perangkat tes Bahasa Indonesia berformat pilihan ganda dengan empat pilihan jawaban, terdiri atas 50 butir soal tersebar pada enam modul/materi. Banyaknya butir soal yang menguji kompetensi pada modul 1 adalah 8 butir, banyaknya butir soal yang menguji kempetensi pada modul 2 adalah 6 butir, modul 3 sebanyak 10 butir, modul 4 sebanyak 12 butir, modul 5 sebanyak 4 butir, dan modul 6 sebanyak 10 butir.

Pada penelitian ini model respons butir atau model IRT yang digunakan untuk membangun CAT adalah model logistik 3 parameter. (Hambleton, Swaminathan, & Rogers, 1991: 17, Hambleton & Swaminathan, 1985 : 49).

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{D.a_i(\theta - b_i)}}{1 + e^{D.a_i(\theta - b_i)}} \qquad (1)$$

$P_i(\theta)$ adalah peluang peserta tes yang memiliki kemampuan θ dapat menjawab butir i dengan benar. θ adalah tingkat kemampuan peserta tes. D merupakan faktor skala = 1,702. a_i merupakan parameter daya beda dari butir ke-i. b_i merupakan kesukaran butir ke-i. c_i merupakan tebakan semu butir ke-i, dan e diaproksimasi sebesar 2,718.

Tiga konsep IRT yang digunakan dalam pengembangan CAT adalah: (1) fungsi informasi, (2) kesalahan baku pengukuran (*standard error of measurement*), dan (3) pendugaan tingkat kemampuan (*ability estimation*).

Nilai fungsi informasi butir menggambarkan seberapa akurat suatu butir soal dapat mengestimasi tingkatan kemampuan peserta tes. Dengan menggunakan fungsi informasi, ketepatan pengukuran pada pengestimasian kemampuan peserta dapat dihitung pada setiap tingkat kemampuan. Fungsi informasi butir dinyatakan oleh Birnbaum (Hambleton, Swaminathan , & Rogers, 1991: 91) dalam persamaan berikut.

$$I_i(\theta) = \frac{2{,}89a_i^2(1-c_i)}{\left[\ (c_i + \exp(1{,}7a_i(\theta - b_i))\ \right]\left[1 + \exp(-1{,}7a_i(\theta - b_i)\ )\ \right]^2} \quad (2)$$

Persamaan (2) menunjukkan bahwa nilai informasi hanya tergantung pada parameter butir (misalnya; *a*, *b*, dan *c* untuk model 3P) dan tingkat kemampuan, θ. Dengan demikian, untuk setiap tingkat kemampuan, θ, kontribusi informasi untuk setiap butir pada bank soal dapat dihitung.

Fungsi informasi tes merupakan jumlah dari fungsi informasi butir penyusun tes tersebut (Hambleton & Swaminathan, 1985: 94). Seperti fungsi informasi butir, fungsi informasi tes menggambarkan seberapa akurat perangkat tes mengestimasi tingkat kemampuan yang berbeda. Semakin besar informasi pada tingkat kemampuan yang diberikan semakin akurat kemampuan itu diestimasi dari perangkat tes itu.

Kesalahan baku pengukuran (*Standard Error of Measurement, SEM*) berkaitan erat dengan fungsi informasi. Fungsi informasi tes dengan *SEM* mempunyai hubungan yang berbanding terbalik kuadratik, semakin besar fungsi informasi tes maka *SEM* semakin kecil atau sebaliknya. Hubungan keduanya, menurut Hambleton, Swaminathan, & Rogers (1991: 94) dinyatakan dengan

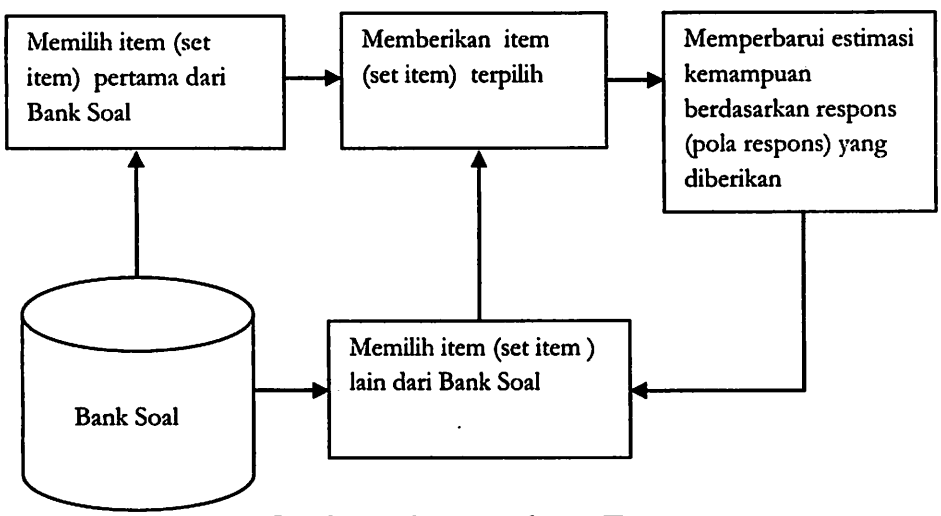
$$\mathrm{SEM}(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} \quad (3)$$

Metode yang umum untuk mengestimasi tingkat kemampuan peserta adalah metode *Maximum Likelihood Estimation* (MLE) (Baker, 1992). Tujuan MLE adalah menemukan nilai yang memaksimumkan fungsi kemungkinan.

Fungsi kemungkinan merupakan fungsi peluang dari pola respons peserta terhadap butir. Pada praktiknya, untuk mengestimasi tingkat kemampuan dengan MLE ini dilakukan dengan menggunakan prosedur iterasi Newton-Raphson (Hambleton & Swaminathan, 1985: 83).

Satu masalah dengan penerapan metode MLE pada tes adaptif adalah ketidakmampuan fungsi kemungkinan untuk menemukan solusi maksimum ketika peserta tes menjawab semua butir soal dengan benar atau salah. Untuk mengatasi masalah ketidakmampuan metode MLE dalam mengestimasi kemampuan peserta manakala respons peserta tes belum berpola pada penelitian ini digunakan metode *step size* (Dodd, 1990; Weiss, 2004).

Proses *adaptive testing* secara skematik disajikan pada Gambar 1.

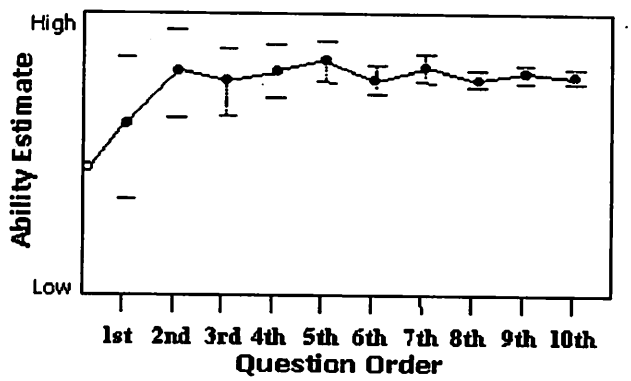


Gambar 1. Proses *Adaptive Testing*

Berdasarkan Gambar 1, proses *adaptive testing* dimulai dengan memilih butir soal atau kelompok butir soal pertama dari bank soal. Setelah butir soal atau kelompok butir soal dipilih, selanjutnya butir soal diberikan kepada peserta tes. Setelah peserta tes merespons (benar atau salah) butir soal atau kelompok butir soal pertama, kemudian tingkat kemampuan

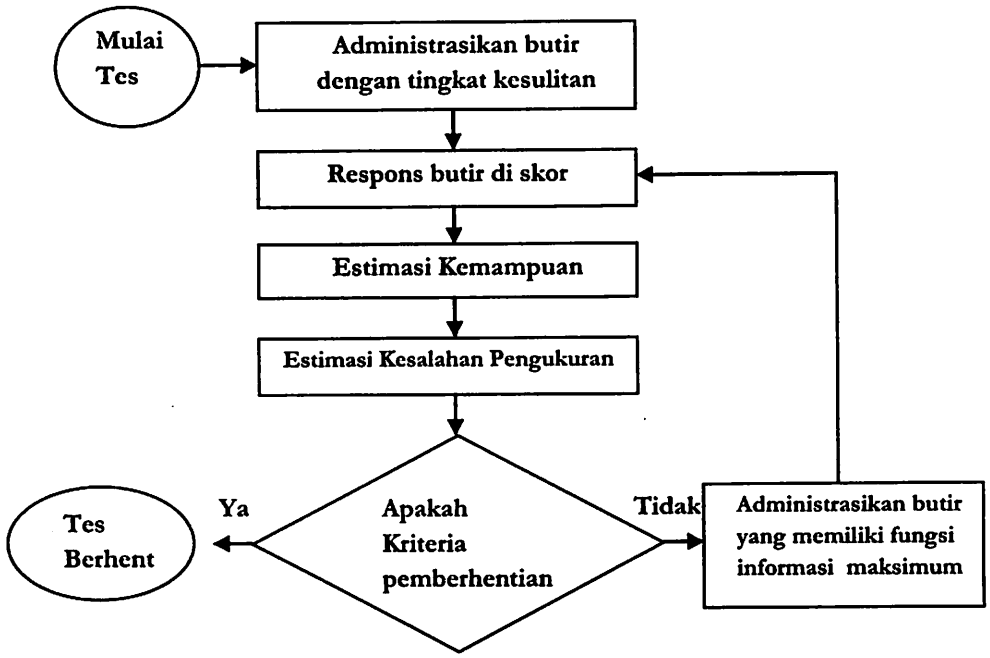
peserta diperbarui atau diestimasi kembali. Selanjutnya, berdasarkan estimasi tingkat kemampuan terbaru, butir soal atau kelompok butir soal yang lain dipilih kembali dari bank soal. Kemudian butir soal atau kelompok butir soal yang lain diberikan lagi kepada peserta tes, begitu seterusnya proses ini berlangsung dan diberhentikan setelah sebanyak butir soal yang ditentukan sudah diberikan atau setelah presisi estimasi tingkat kemampuan atau tingkat kesalahan baku pengukuran yang diinginkan telah dicapai.

Ilustrasi estimasi tingkat kemampuan peserta tes pada CAT disajikan pada Gambar 2 berikut.



Gambar 2. Estimasi Kemampuan pada CAT (Anonim, 2006)

Gambar 2 menunjukkan bagaimana tingkat kemampuan seorang peserta tes diestimasi lebih rendah setelah pertanyaan dijawab secara salah (pertanyaan 3, 6, 8, dan 10). Titik-titik vertikal mengindikasikan besarnya *error* dikaitkan dengan tingkat kemampuan yang diestimasi.



Gambar 3. Bagan Alur Pengujian Algorima CAT

Pengujian algoritma CAT disajikan pada Gambar 3. Berdasarkan gambar tersebut, tes dimulai dengan memilih butir soal awal dengan tingkat kesukaran sedang. Berikutnya respons terhadap butir diskor. Kemudian diestimasi (sementara) tingkat kemampuan peserta dengan menggunakan *Maximum Likelihood Estimation*. Selanjutnya, dicari nilai fungsi informasi butir pada tingkat kemampuan peserta yang telah diperoleh dan dihitung pula estimasi kesalahan baku pengukurannya. Kemudian dipilih lagi butir yang memiliki nilai fungsi informasi tertinggi atau yang mengurangi kesalahan pengukuran terbesar. Begitu seterusnya sampai tes dihentikan jika kriteria pemberhentian terpenuhi.

Terkait dengan aturan pemberhentian tes yang digunakan maka pada penelitian ini dikembangkan dua desain algoritma CAT, yaitu: algoritma CAT murni dan algoritma CAT yang dikendala modul (Modul-*Constrained* CAT, CCAT). Pada algoritma CAT murni tes dihentikan jika kesalahan

baku pengukuran telah mencapai 0,30 atau setara dengan tingkat reliabilitas sebesar 91% pada pengukuran menggunakan teori klasik (Thissen, 1990), sedangkan pada algoritma CCAT tes dihentikan jika telah mencapai sejumlah butir tertentu.

Metodologi Penelitian

Prosedur pengembangan model pengukuran pada penelitian ini terdiri atas tiga tahapan: tahap pertama membangun bank soal, tahap kedua simulasi algoritma CAT, dan tahap ketiga membangun program CAT.

Bank soal untuk keperluan CAT dibangun dari butir-butir soal yang telah diujikan dengan PPT. Data pola respons peserta tes dari tujuh masa ujian dipilih untuk dikalibrasi dengan program BILOG-MG (Mislevy & Bock, 1990). Selanjutnya, butir-butir soal yang memenuhi kriteria ideal yaitu yang memiliki tingkat daya beda pada rentang 0,4 sampai 2,0; tingkat kesukaran di antara -3,0 sampai +3,0; dan faktor *guessing* pada rentang 0,0 sampai 0,30 dipilih sebagai butir-butir soal pada bank soal untuk keperluan CAT.

Pada studi simulasi, dikembangkan dua desain algoritma CAT, yaitu desain algoritma CAT murni dan CAT yang dikendala modul (CCAT). Simulasi desain CAT murni dilakukan berdasarkan bank soal yang tersedia (empiris) dan dua bank soal ideal hasil data bangkitan. Bank soal data bangkitan Spesifikasi 1 (BS-S1) adalah bank soal bangkitan berukuran 250 butir dengan sebaran tingkat kesukaran merata pada setiap skala tingkat kemampuan peserta tes, sedangkan bank soal spesifik 2 (BS-S2) adalah bank soal bangkitan berukuran 250 butir dengan sebaran tingkat kesukaran mengikuti sebaran normal yaitu butir soal dengan tingkat kesukaran sedang banyak, sedangkan butir soal mudah dan sukar sedikit

Pada desain CCAT diskenariokan empat desain CCAT, yaitu: CCAT dengan panjang tes 25% (proporsional per materi) dari panjang PPT atau disingkat CCAT25, CCAT dengan panjang 50% dari panjang PPT (CCAT50), CCAT75, dan CCAT100. Berdasarkan rancangan algoritma CAT pada studi simulasi, selanjutnya dibuat program (*software*) CAT.

Rancangan algoritma dituliskan dalam bahasa pemrograman *Power Builder versi* 10.

Hasil Penelitian dan Pembahasan

1. Bank Soal CAT

Kalibrasi menghasilkan 127 butir soal sebagai bank soal untuk keperluan CAT. Ringkasan statistik parameter butir pada Bank Soal Empiris disajikan pada Tabel 1 berikut.

Tabel 1. Ringkasan Statistik Parameter Butir Soal pada Bank Soal Empiris

Parameter	Mean	Std-deviasi	Min.	Maks.
Daya beda (a)	1,207	0,479	0,457	2,501
Tingkat kesukaran (b)	0,530	1,266	-2,468	2,524
Guessing (c)	0,139	0,047	0,031	0,248

Selanjutnya, banyaknya butir soal per materi/modul dari BS-Empiris disajikan pada Tabel 2. Berdasarkan Tabel 2 tersebut, terlihat bahwa sebaran persentase banyaknya butir soal tiap modul antara bank soal untuk keperluan CAT mirip persentase banyaknya butir soal pada perangkat PPT, kecuali pada materi/modul 6 hanya tersedia 17,3 % sedangkan pada PPT sebanyak 24,0 %.

Tabel 2. Banyaknya dan Persentase Banyak Butir Soal per Materi/Modul pada BS-Empiris dan Perangkat PPT

Materi/Modul	BS-Empiris		Perangkat PPT	
	Banyak Butir Soal	Persentase	Banyak Butir Soal	Persentase
Modul 1	19	15,0	8	16,0
Modul 2	15	11,8	6	12,0
Modul 3	27	21,3	10	20,0
Modul 4	35	27,6	12	24,0
Modul 5	9	7,1	4	8,0
Modul 6	22	17,3	12	24,0
Jumlah	127	100	50	100

2. Hasil Simulasi

Berikut dipaparkan contoh hasil simulasi desain CAT murni berdasarkan bank soal empiris. Misalkan untuk theta = 0 yang diambil secara acak. Berdasarkan hasil simulasi, peserta ini dapat diestimasi dengan butir soal sebanyak 10 butir. Nomor induk soal (NIS), parameter butir, nomor modul, pola respons, estimasi theta, kesalahan baku pengukuran dan nilai fungsi informasi disajikan pada Tabel 3.

Tabel 3. Nomor Induk Soal, Modul yang Digunakan, dan Pola Respons serta Estimasi Theta, SEM, dan Nilai Fungsi Informasi

No.Urut	1	2	3	4	5	6	7	8	9	10
N.I.S	180	423	179	159	158	54	319	99	89	440
a	0,961	2.297	2.147	2.354	1.526	1.593	1.724	1.603	1.441	1.859
b	0,020	0.507	-0.105	-0.535	-0.602	-0.133	0.188	0.255	0.173	0.319
c	0,151	0.173	0.140	0.137	0.169	0.225	0.125	0.084	0.089	0.183
Modul	3	6	3	3	3	1	4	2	2	6
Respons	1	0	0	1	1	1	0	1	1	0
theta	**0.5**	**0.06319**	**-0.5238**	**-0.2596**	**-0.1798**	**-0.0702**	**-0.1648**	**-0.0221**	**0.07946**	**0.00744**
SEM		0.87389	0.64173	0.45163	0.41180	0.37575	0.35010	0.32504	0.30602	0.29157
info	0.495993	0.81343	1.11879	2.47442	0.99413	1.18566	1.07598	1.30636	1.21317	

Keterangan : Respons 1 = benar; 0 = salah

Berdasarkan Tabel 3 terlihat bahwa butir pertama yang terpilih adalah butir dengan NIS 180 dari modul 3, memiliki tingkat kesukaran, b = 0,02, artinya ini sesuai dengan kriteria yang diterapkan pada algoritma desain CAT murni bahwa butir soal awal yang dipilih adalah butir dengan tingkat kesukaran sedang, yang dipilih secara acak pada rentang tingkat kesukaran sedang (-0,5 sampai +0,5).

Dari Tabel 3, terlihat pula bahwa butir soal ini direspons 1, artinya dijawab benar, selanjutnya karena benar maka ditampilkan lagi butir dengan NIS 423 dari modul 2 dengan tingkat kesulitan, b sebesar 0,507. Butir dengan NIS 423 ini dipilih karena memiliki fungsi informasi maksimum

pada theta sebesar 0,5. Hal ini juga telah sesuai dengan kriteria pemilihan butir soal berikutnya yang diterapkan pada algoritma CAT murni yang menggunakan kriteria *step size* sebesar 0,5. Pada butir soal pertama ini, kesalahan baku estimasi atau kesalahan pengukuran belum bisa ditentukan karena belum ada pola respons.

Selanjutnya, ketika butir soal kedua direspons salah, maka pemilihan butir soal ketiga sudah didasarkan pada hasil pengestimasian theta. Hal ini karena metode MLE yang diterapkan pada algoritma desain CAT murni akan berproses setelah respons sudah berpola (minimal ada satu benar atau satu salah). Berdasarkan metode MLE setelah menjawab butir soal nomor urut 1 benar dan nomor urut 2 salah, maka berdasarkan metode MLE kemampuan peserta ini diestimasi sebesar 0,063, dan kesalahan baku pengukuran sebesar 0,8739, dan karena kesalahan baku pengukuran belum mencapai 0,30 maka tes masih berlanjut.

Berdasarkan nilai fungsi informasi, maka butir soal ketiga yang dipilih adalah butir dengan NIS 179. Butir ini terpilih karena memiliki nilai fungsi informasi maksimum di antara butir-butir lainnya di bank soal empiris untuk theta sebesar 0,063. Seperti terlihat pada Tabel 3, nilai fungsi informasi butir soal ini sebesar 0,8134. Selanjutnya butir soal ketiga ini direspons, kemampuan dan kesalahan baku pengukuran diestimasi kembali, kemudian butir soal keempat dipilih, direspons, kemampuan diestimasi ulang, begitu seterusnya sampai tes dihentikan pada butir soal ke-10 karena pada butir ke-10 kesalahan baku pengukurannya telah mencapai 0,30 dengan estimasi theta sebesar 0,0074.

Banyaknya butir soal yang diperlukan untuk setiap theta yang disimulasikan disajikan pada Tabel 4.

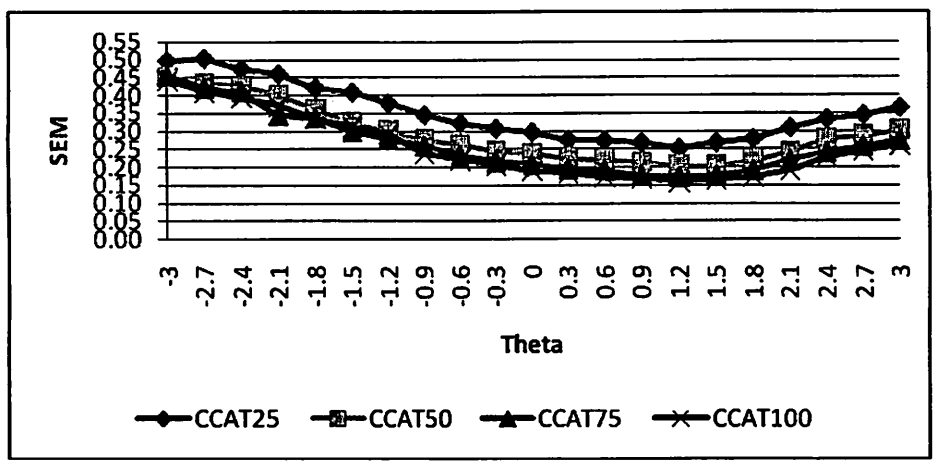
Tabel 4. Banyaknya Butir Soal yang Diperlukan pada 21 Tingkatan Theta yang Disimulasikan Berdasarkan Tiga Bank Soal

Theta	BS-Empiris	BS-S1	BS-S2
-3	112	12	202
-2.7	109	12	133
-2.4	117	11	80
-2.1	90	11	26
-1.8	70	11	12
-1.5	72	10	11
-1.2	34	10	10
-0.9	27	10	9
-0.6	15	10	9
-0.3	12	10	8
0	10	10	8
0.3	10	9	8
0.6	9	9	8
0.9	8	9	9
1.2	8	9	9
1.5	9	11	10
1.8	12	11	12
2.1	14	11	15
2.4	21	11	38
2.7	30	12	102
3	48	12	193

Berdasarkan Tabel 4 terlihat bahwa pada BS-Empiris untuk theta antara -3 sampai -1,5 diperlukan banyaknya butir cukup besar (masih lebih dari 70 butir), bahkan untuk theta -3 diperlukan banyak butir yang melebihi banyaknya butir pada bank soal. Hal ini menunjukkan bahwa proses CAT masih berlangsung karena presisi pengukuran belum tercapai walaupun semua butir soal pada bank soal sudah diberikan kepada peserta dengan

tingkat kemampuan -3. Kemudian untuk theta -1,2 dan -0,9 diperlukan banyaknya butir masing-masing sebanyak 30 dan 25 butir. Untuk theta antara -0,6 sampai +2,1 diperlukan banyaknya butir berkisar 8 sampai 15 butir, sedangkan untuk theta antara +2,4 sampai +3 diperlukan banyaknya butir sekitar 21 sampai 48 butir. Pada BS-S1 untuk setiap theta antara -3 sampai +3 diperlukan banyaknya butir yang hampir sama, yaitu 8 sampai 12 butir. Pada BS-S2 terlihat bahwa pada theta ekstrim (-3 sampai -2,7 dan +2,7 sampai +3) diperlukan lebih dari 100 butir, sedangkan pada theta antara -2,1 sampai 2,1 diperlukan sekitar 8 sampai 25 butir.

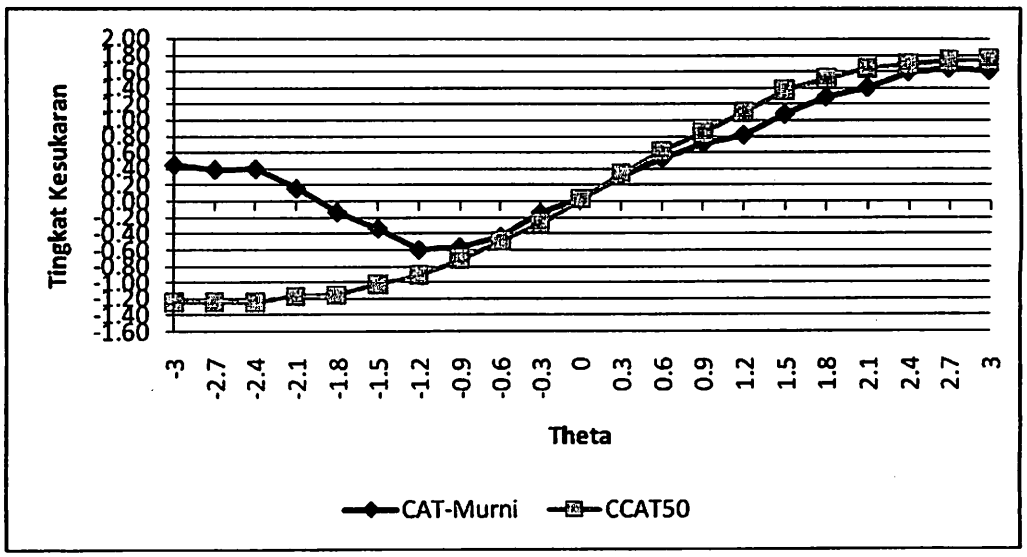
Kesalahan baku pengukuran (*Standard error of measurement* = *SEM*) pada empat desain CCAT disajikan pada Gambar 4.



Gambar 4. Kesalahan Baku Pengukuran Empat Desain CCAT

Pada Gambar 4 terlihat bahwa desain CCAT75 dan CCAT50 memiliki kesalahan baku pengukuran yang mirip dengan CCAT100 untuk setiap theta yang disimulasikan, sedangkan desain CCAT25 memiliki kesalahan baku pengukuran yang agak besar dibandingkan CCAT100. Dengan demikian dari sisi kesalahan baku pengukuran dapat ditemukan dua desain CCAT yang dapat dijadikan sebagai pilihan untuk diaplikasikan pada sistem ujian UT, yaitu desain CCAT50. Ditambahkan bahwa

kesesuaian tingkat kesukaran butir dengan tingkat kemampuan untuk desain ini juga cukup baik dibandingkan desain CAT murni. Gambar 5 berikut menyajikan tingkat kesesuaian tersebut.



Gambar 5. Kesesuaian Tingkat Kesukaran dan Theta desain CAT murni dan CCAT50

Kesesuaian antara tingkat kesukaran butir dan tingkat kemampuan karena banyak butir soal pada setiap modul cukup tersedia dengan persentase banyak butir soal yang proporsional dengan persentase banyak butir soal pada perangkat PPT (lihat Tabel 2).

3. Pembahasan

Kualitas bank soal CAT ditentukan oleh ukuran bank soal, parameter butir soal, dan struktur isi. Nilai parameter butir soal pada BS-Empiris dan sebaran banyaknya butir soal setiap modul cukup ideal, namun dari sisi ukuran dirasa masih belum cukup. Hasil kalibrasi butir dari 296 butir yang berasal dari tujuh masa ujian yang hanya menghasilkan 127 butir soal ini disebabkan karena (1) banyak butir soal yang kurang *fit* dengan model IRT

3P dan (2) banyak butir soal dengan nilai faktor *guessing* cukup besar atau di luar rentang nilai ideal. Peserta tes yang cukup besar (> 5000 peserta) untuk beberapa masa ujian tertentu sering menghasilkan kalibrasi parameter butir yang tidak *fit* dengan model yang diterapkan (Hambleton, Swaminathan, & Rogers (1991: 54), dan kalaupun *fit* nilai parameter butir yang dihasilkan berada di luar rentang kriteria butir ideal.

Berdasarkan hasil simulasi terlihat bahwa BS-Empiris akurat untuk mengestimasi theta (tingkat kemampuan) sedang sampai tinggi (-1,2 sampai +3), namun kurang akurat untuk mengestimasi pada tingkat kemampuan yang rendah (-3 sampai -1,5). Hal ini disebabkan karena pada BS-Empiris kurang cukup menyediakan butir-butir soal untuk peserta dengan kemampuan rendah. Dengan kata lain, dalam BS-Empiris kekurangan butir soal dengan tingkat kesukaran mudah atau sangat mudah.

Desain CAT murni merupakan desain yang mengukur tingkat kemampuan secara adil untuk semua peserta tes. Hal ini karena setiap peserta tes akan dihentikan tesnya ketika kesalahan pengukurannya telah mencapai 0,3. Namun demikian, pada desain ini tidak semua modul ditampilkan untuk setiap peserta tes, seperti terlihat pada tabel 3, butir soal dari modul 5 tidak muncul. Hal ini merupakan konsekuensi logis dari penerapan desain CAT murni yang menggunakan kriteria fungsi informasi maksimum untuk memilih butir soal berikutnya, yang mengabaikan keseimbangan isi.

Desain CCAT50 merupakan desain yang mempertimbangkan keseimbangan isi, dengan desain ini setiap modul/materi akan terwakili oleh butir soalnya untuk diberikan kepada peserta tes. Desain ini lebih direkomendasikan untuk diterapkan sebagai alternatif desain CBT pada aplikasi sistem ujian UT. Hal ini karena: (1) desain CCAT50 memiliki kesalahan baku estimasi yang cukup rendah, (2) desain CCAT memiliki tingkat kesesuaian yang cukup baik dibandingkan desain CAT murni. Ketika tingkat kemampuan dan tingkat kesukaran butir mirip atau sesuai maka prinsip dasar tes adaptif tercapai, dan (3) desain CCAT50 dengan panjang tes sebanyak 25 butir soal lebih rasional diterima dan sudah dapat menjamin keakuratan pengestimasian. Hasil penelitian yang dilakukan van der Linden & Pashley (2000: 22) memperkuat alasan bahwa bias dan

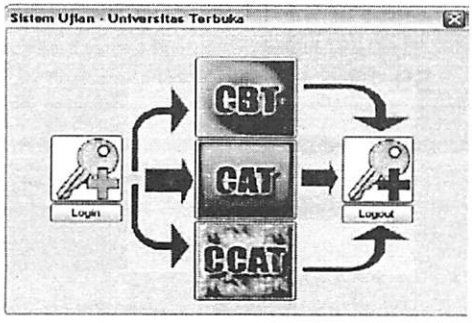
kesalahan pengukuran sangat kecil pada tes adaptif kètika panjang tes lebih dari 20 butir.

Aplikasi CAT pada sistem ujian lebih efisien dibandingkan dengan CBT (PPT), dalam arti bahwa dengan CAT peserta tes sudah dapat diestimasi kemampuannya hanya dengan 8 sampai 15 butir soal saja dibandingkan CBT maupun PPT yang harus menempuh sebanyak 50 butir soal. Namun, karena bank soal untuk keperluan CAT kurang menyediakan butir-butir soal bagi peserta dengan kemampuan yang rendah (-3 sampai -1,5), akibatnya aplikasi CAT yang dikembangkan hanya akurat untuk mengukur peserta dengan kemampuan sedang sampai tinggi, tetapi kurang akurat untuk mengukur peserta tes dengan kemampuan sangat rendah.

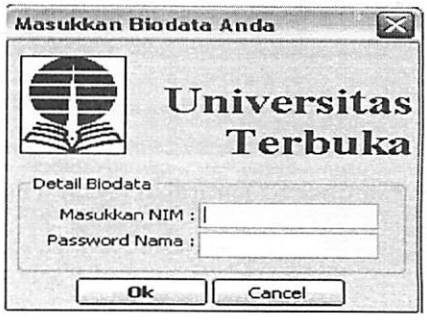
Aplikasi desain CCAT ini baik digunakan manakala Universitas Terbuka menginginkan untuk menerapkan aplikasi CAT yang mempertimbangkan keseimbangan isi. Namun, perlu disadari bahwa desain CCAT ini memiliki keterbatasan berupa ketidakadilan antar individu peserta tes karena kesalahan pengukuran setiap peserta tes dapat berbeda-beda.

Penelitian pengembangan ini hanya berdasarkan pada studi simulasi dan uji coba hanya dilakukan oleh peneliti dan beberapa sukarelawan, sedangkan uji coba lapangan belum dilakukan. Dengan demikian, faktor-faktor yang berkaitan dengan manusia seperti ketidaktahuan menggunakan komputer, kelelahan, kemampuan konsentrasi yang diduga mempengaruhi skor peserta tes diabaikan.

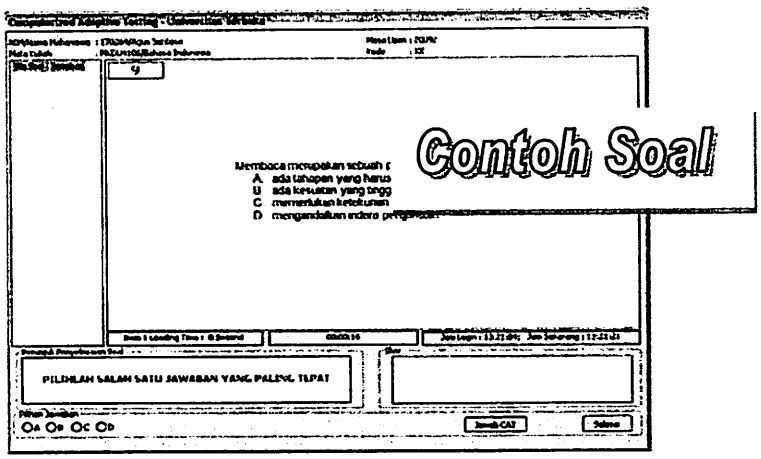
Berikut disajikan tampilan-tampilan program aplikasi CAT yang dikembangkan.



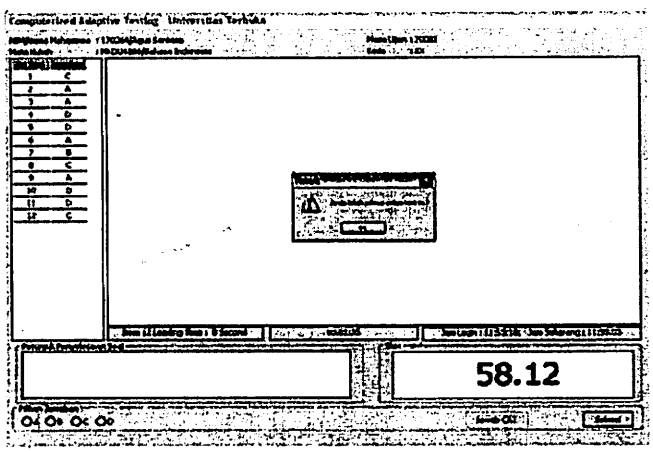
Gambar 5. Tampilan Awal Aplikasi



Gambar 6. Tampilan Login



Gambar 7. Tampilan Aplikasi CAT



Gambar 8. Tampilan Akhir Tes

Kesimpulan dan Rekomendasi

Banyaknya butir soal ideal untuk bank soal CAT adalah sebanyak 127 butir. Bank soal data empiris akurat untuk mengukur peserta tes dengan kemampuan sedang sampai tinggi (-1,2 sampai +3,0), namun kurang akurat dalam mengukur kemampuan yang rendah (-3 sampai -1,5). Banyaknya butir soal (panjang tes) yang diperlukan pada desain CAT murni berkisar antara 8 sampai 15 butir soal untuk mengestimasi kemampuan peserta tes pada tingkat kemampuan sedang sampai tinggi.

Kepada UT direkomendasikan untuk mengaplikasikan desain CAT yang dikendala oleh modul (CCAT50) karena desain ini telah mempertimbangkan keseimbangan isi, namun implementasi dari desain ini perlu dikaji lebih lanjut, sehingga dapat dimanfaatkan oleh UT sebagai bentuk ujian alternatif. Model desain ini juga dapat digunakan oleh lembaga pendidikan lain yang memerlukan keseimbangan isi dalam sistem penilaiannya.

Daftar Pustaka

Anonim (2006). *Adaptive testing.* Diambil pada tanggal 18 Juni 2006, dari http://www.windowsgalore.com/cert/adaptive_testing/.

Ackerman, T. A., Evans, J., Park, K., Tamassia, C., and Turner, R. (1999). Computer assessment using visual stimuli: A test of dermatological skin disorders. Dalam Drasgow, F., & Olson-Buchanan J. B. (Eds). *Innovations in Computerized Assessment* (pp. 137-150). Lawrence Erlbaum Associates, Publishers. Mahwah, New Jersey.

Baker, F.B. (1992). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker, Inc.

Dodd, B.G. (1990). The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement,* 4, 355 – 366.

Green, B.F., Bock, R.D., Humphyers, L.G., et al. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 4, 347 – 360.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications, Inc.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory*, Boston, MA: Kluwer Academic Publishers.

Kingsbury, G.G., & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 4, 359 – 375.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ : Lawrence Erlbaum Associates.

Mislevy, R. J., & Bock, R. D. (1990). *Item analysis and test scoring with binary logistic models: BILOG MG.* Chicago, IL: Scientific Software, Inc.

Thissen, D. (1990). Reliability and measurement precision. Dalam H. Wainer (Eds.), *Computerized Adaptive Testing: A Primer* (2nd ed., pp. 161–186). Hillsdale, NJ: Lawrence Erlbaum Associates.

van der Linden, W.J. & Pashley, P.J. (2000). Item selection and ability estimation in adaptive testing. Dalam van der Linden & Glas C.A.W. (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 1–25). The Netherlands: Kluwer Academic Publishers.

Vispoel, W.P. (1999). Creating computerized adaptive test of music aptitude : Problem, solusions, and future directions. Dalam F. Drasgow, & J. B. Olson-Buchanan (Eds.), *Innovations in Computerized Assessment* (pp. 151 –176). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Wainer, H. (1990). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Weiss, D.J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, 2, 70 - 84.

Weiss, D.J. & Schleisman, J.L. (1999). Adaptive testing. Dalam G. N. Masters & J. P. Keeves (Eds.), *Edvances in Measurement in Educational Research and Assessment* (pp. 129–137). Pergamon, NY: Elsevier Science Ltd.