EFISIENSI DAN AKURASI COMPUTERIZED ADAPTIVE TESTING PADA SISTEM UJIAN AKHIR SEMESTER UNIVERSITAS TERBUKA

Agus Santoso (aguss@ut.ac.id)
Jurusan Statistika FMIPA Universitas Terbuka

ABSTRACT

Universitas Terbuka (UT) applied online examination system (sistem ujian online – SUO) for end of semester examination (ujian akhir semester-UAS), beside the paper and pencil test (P & P test). In order to improve efficiency, adaptive test application should be analyzed, as an alternative to present UAS system. The aim of the research was to compare the efficiency and accuracy level of the computerized adaptive testing (CAT) design and conventional test using both P & P test and SUO. The research was conducted by simulation procedure. The item bank for the simulation used calibrated 404 test items using item response theory model. In the research, CAT and P & P test algorithm was developed. To measure efficiency, the required number of the CAT design was analyzed, while to measure accuracy of the estimation, the bias and standard error of measurement of both design were compared. The simulation result showed that (1) CAT design was more efficient, since it required only half of the number of item which was used in P & P test, to estimate the ability of examinee, (2) CAT design was more accurate in estimating ability of examinee, compared to P & P test design, since it resulted lower bias and standard error of measurement compared to conventional test design. Therefore, CAT design could be applied in UT's UAS system, while considering the balance of content for each modules.

Key words: computerized adaptive testing, item response theory, paper and pencil test.

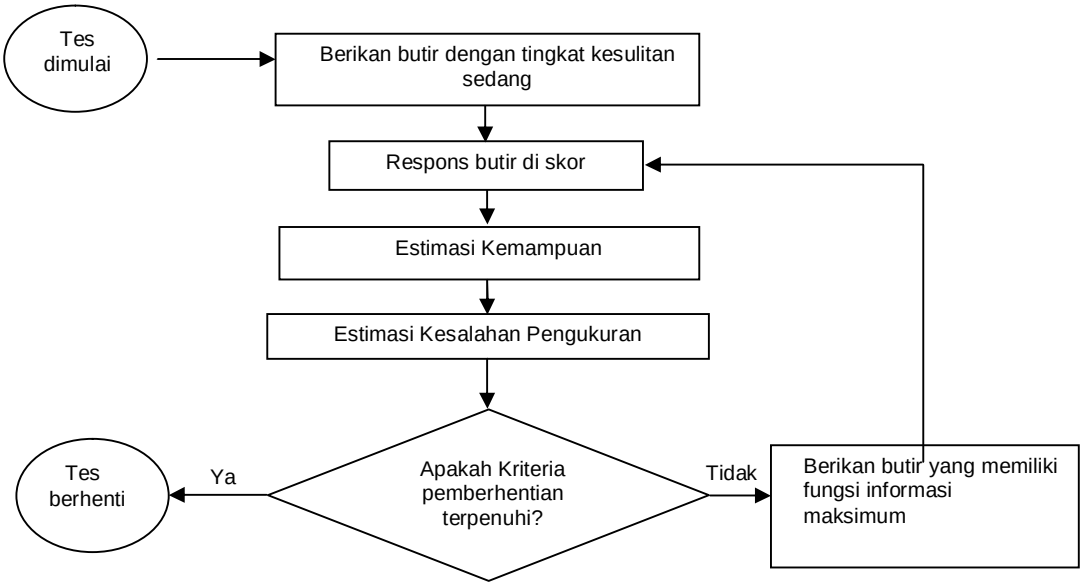
Sejak tahun 2006, Universitas Terbuka (UT) telah menerapkan sistem ujian berbasis komputer (UBK) pada sistem Ujian Akhir Semester (UAS), di samping menggunakan *paper and pencil test* (P&P *test*) yang selama ini telah diselenggarakan. UBK dikembangkan berdasarkan pemanfaatan teknologi internet, dengan mempertimbangkan sarana komputer di (UPBJJ-UT) Unit Program Belajar Jarak Jauh Universitas Terbuka. Pada sistem UBK, butir-butir soal yang diberikan kepada peserta tes tidak secara langsung disajikan/diberikan kepada mahasiswa namun didahului dengan proses *download* terlebih dahulu di UPBJJ UT, selanjutnya butir-butir soal itu diberikan kepada peserta tes melalui seorang administrator. Seiring dengan bertambahnya kapasitas *banchwidt* pada jaringan komputer di UT dan ketersediaan butir-butir soal pada bank soal di pusat pengujian UT, maka mulai tahun 2008 sistem UBK disempurnakan menjadi Sistem Ujian Online (SUO), dimana pemberian butir soal kepada peserta tes dapat dilakukan secara langsung dari bank soal yang ada di Pusat Pengujian UT. Pemberian butir soal pada SUO didasarkan pada kisi-kisi yang telah dibuat oleh pengampu matakuliah.

Pemberian butir soal dengan jumlah butir soal dan desain tes yang sama kepada seluruh peserta tes akan mengabaikan variasi kemampuan peserta tes dan menyebabkan adanya pemberian

butir-butir soal yang terlalu mudah untuk beberapa peserta tes yang memiliki kemampuan sangat tinggi dan pemberian butir-butir soal yang terlalu sukar untuk beberapa peserta tes yang memiliki kemampuan sangat rendah. Jawaban benar dari individu peserta tes yang memiliki kemampuan tinggi yang mengerjakan soal mudah dan jawaban salah dari individu peserta tes yang berkemampuan rendah yang mengerjakan soal sukar sesungguhnya kurang memberikan informasi mengenai kemampuan mereka. Hal tersebut mengakibatkan tes kurang efisien, kurang adil, dan kurang akurat (Hambleton, Swaminathan & Rogers, 1991).

Untuk menyempurnakan sistem ujian, pada UAS UT sebaiknya UT mengembangkan sistem ujian yang bersifat adaptif atau *Computerized Adaptive Testing* (CAT). *Adaptive* berarti bahwa butir soal yang diberikan disesuaikan dengan tingkat kemampuan setiap peserta tes (Lord, 1980). Pada CAT yang berbasiskan *item response theory* (IRT), komputer diatur untuk menyeleksi dan menyajikan/memberikan butir soal menurut perkiraan tingkat kemampuan peserta tes. Hal ini mengakibatkan individu peserta tes yang memiliki tingkat kemampuan tinggi akan mendapatkan butir soal yang lebih sulit dibandingkan dengan individu yang memiliki tingkat kemampuan rendah. Sebaliknya individu peserta tes yang memiliki tingkat kemampuan rendah akan mendapatkan butir soal yang lebih mudah dibandingkan dengan individu peserta tes yang memiliki tingkat kemampuan tinggi. Dengan demikian CAT lebih efisien karena dapat mengestimasi kemampuan peserta tes dengan jumlah butir soal yang lebih sedikit dibandingkan P&P *test* maupun SUO tanpa mengurangi ketepatan pengukuran (Wainer *et al*.,1990; Hambleton, Swaminathan, & Rogers, 1991; Weiss & Schleisman, 1999). Namun demikian, sebelum mengaplikasikan sistem CAT di UT perlu dikaji apakah model CAT yang akan dikembangkan lebih efisien dan akurat dibandingkan dengan P&P test.

Tujuan penelitian ini adalah untuk membandingkan tingkat efisiensi dan akurasi desain CAT dibandingkan *paper and pencil test* (P&P test) di Universitas Terbuka.



Gambar 1. Bagan Alur Pengujian Algorima CAT (Vispoel,1999)

Gambar 1 adalah bagan algoritma CAT yang dikembangkan pada penelitian ini yang mengacu pada algoritma CAT menurut Vispoel (1999). Berdasarkan Gambar 1, tes dimulai dengan memilih butir soal awal dari bank soal dengan tingkat kesukaran sedang, kemudian respons terhadap butir diskor dan tingkat kemampuan peserta serta kesalahan pengukurannya diestimasi, butir soal berikutnya yang diberikan kepada peserta adalah butir soal yang memiliki nilai fungsi informasi tertinggi atau yang mengurangi kesalahan pengukuran terbesar. Proses ini berlanjut, butir per butir soal diberikan kepada peserta tes sampai tes dihentikan jika kriteria pemberhentian terpenuhi.

METODOLOGI

Penelitian ini dilakukan dengan studi simulasi. Bank soal untuk keperluan simulasi CAT diambil dari butir-butir soal dari matakuliah yang memiliki jumlah peserta tes cukup besar (lebih dari 1.000) dan butir-butir soal yang tersedia lebih dari 400 butir soal. Berdasarkan kisi-kisi, jumlah butir soal untuk satu set soal tes matakuliah terpilih umumnya adalah 35 butir soal. Dari sebanyak 464 butir soal yang berasal dari 13 masa ujian dikalibrasi menggunakan model IRT 1 parameter atau model Rasch (Bond & Fox, 2007). Berdasarkan model ini, peluang seseorang yang berkemampuan (theta) tertentu menjawab butir soal dengan benar bergantung hanya pada satu parameter butir soal, yaitu tingkat kesukaran (tercantum pada Lampiran). Berdasarkan hasil kalibrasi sebanyak 404 butir soal *fit* atau cocok dengan model IRT 1 parameter dan dipilih sebagai butir-butir soal pada bank soal untuk keperluan simulasi. Selanjutnya, desain CAT dan UBK atau P&P test dikembangkan.

Prosedur simulasi untuk desain CAT berdasarkan pada 2100 simuli peserta tes yang disimulasikan, yang mewakili 100 simuli peserta tes untuk setiap 21 titik skala tingkat kemampuan, theta (θ) dari -3,0 sampai +3,0 dengan kenaikan 0,3.

Langkah simulasi untuk desain CAT sebagai berikut:

1. Untuk tingkat kemampuan peserta tes, θ tertentu, tes adaptif diberikan. Berdasarkan metode pemilihan butir awal, satu butir soal dipilih dan diberikan. Peluang peserta tes menjawab benar pada butir soal ke-*i*, $P_i(\theta)$ dihitung. Untuk membangkitkan jawaban atau respons dari peserta tes, nilai $P_i(\theta)$ dibandingkan dengan peubah acak *x* yang diambil dari sebaran uniform [0,1]. Jika $P_i(\theta)$ lebih besar dari *x* maka respons diskor 1, sebaliknya jika $P_i(\theta)$ kurang dari atau sama dengan *x* maka respons diskor 0. Berdasarkan respons dan parameter butir soal selanjutnya kemampuan peserta tes, θ diestimasi. Estimasi theta dan butir soal yang diberikan dicatat untuk dianalisis lebih lanjut.
2. Berdasarkan metode pemilihan butir soal, diberikan butir soal berikutnya untuk peserta tes, θ tersebut sampai mencapai tingkat kesalahan baku pengukuran (*standard error of measurement, SEM*) sebesar 0,30.
3. Langkah 1 dan 2 diulang untuk seluruh 2100 simuli peserta tes.
4. Banyaknya butir soal dan estimasi tingkat kemampuan dicatat untuk dianalisis.

Metode pemilihan butir soal awal menggunakan tingkat kesukaran sedang yaitu dimulai dengan rentang antara -0,50 sampai 0,50 yang dipilih secara acak. Metode pendugaan tingkat kemampuan menggunakan *maximum likelihood estimation = MLE* (Baker, 1992), namun ketika pola respons belum berpola pendugaan tingkat kemampuan menggunakan metode *step size* berukuran 0,5 (Dodd, 1990). Metode pemilihan butir soal berikutnya menggunakan kriteria fungsi informasi

maksimum yaitu butir soal yang mempunyai nilai fungsi informasi terbesar pada kemampuan tertentu dipilih untuk diberikan pada peserta tes. Pada penelitian ini, kriteria pemberhentian tes yang digunakan adalah tes dihentikan jika nilai estimasi kesalahan baku pengukuran (*standard error of measurement,* SEM) sudah mencapai 0,30. Nilai SEM sebesar 0,30 ini setara dengan reliabilitas sebesar 0,91 pada tes konvensional dengan P&P test (Thissen,1990).

Langkah simulasi untuk desain BK atau P&P test prinsipnya sama, namun pengestimasian kemampuan peserta tes dilakukan setelah peserta menjawab 35 butir soal yang diberikan dan bank soal yang digunakan hanya berasal dari satu perangkat tes P&P test pada masa ujian tertentu.

HASIL DAN PEMBAHASAN

Ringkasan statistik tingkat kesukaran butir dari 404 butir soal yang digunakan sebagai bank soal untuk keperluan simulasi CAT disajikan pada Tabel 1.

Tabel 1.Ringkasan Statistik Tingkat Kesukaran Butir Soal pada Bank Soal

Statistik	Tingkat Kesukaran Butir
Rata-rata	0,015
Standard deviasi	0,818
Minimum	-3,14
Maksimum	2,08

Berikut dipaparkan contoh hasil simulasi desain CAT. Misalkan untuk theta = 0 yang diambil secara acak, berdasarkan hasil simulasi, peserta ini sudah dapat diestimasi dengan butir soal sebanyak 17 butir. Nomor induk soal (NIS), tingkat kesukaran butir soal, pola respons untuk setiap urutan butir soal yang ditampilkan serta estimasi theta, kesalahan baku pengukuran dan nilai fungsi informasi disajikan pada Tabel 2.

Berdasarkan Tabel 2 terlihat bahwa butir pertama yang terpilih adalah butir soal dengan NIS 388, memiliki tingkat kesukaran, *b* = -0,37, 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).

Tabel 2. Nomor Induk Soal, Pola Respons, Estimasi Theta, SEM, dan Nilai Fungsi Informasi

No.Urut	1	2	3	4	5	6	…	15	16	17
N.I.S	388	359	26	210	322	5	…	16	288	178
Tk. Kesukaran	-0.37	0,49	0,04	0,48	0,15	-0,12	…	0,05	-0,04	0,04
Respons	1	0	1	0	0	1	…	0	1	0
Theta	0,5	0,0535	0,4888	0,1570	-0,104	0,1040	…	-0,024	0,0504	-0,023
SEM		0,8983	0,7338	0,6313	0,5606	0,5085	…	0,3136	0,3032	0,2938
Info	0,6270	0,6123	0,6176	0,6526	0,6725	0,6853	…	0,7045	0,7050	0,7054

Keterangan : Respons 1 = benar; 0 = salah

Berdasarkan Tabel 2, terlihat pula bahwa butir soal ini direspons 1, artinya dijawab benar, selanjutnya karena benar maka ditampilkan lagi butir soal dengan NIS 359. Butir soal dengan NIS 359 ini dipilih karena memiliki fungsi informasi terbesar pada theta sebesar 0,5, yaitu sebesar 0,6270. Hal ini juga telah sesuai dengan kriteria pemilihan butir soal berikutnya yang diterapkan pada algoritma CAT yang menggunakan kriteria *step size* sebesar 0,5. Berdasarkan kriteria *step size* ini

maka ketika butir soal pertama dijawab benar maka butir soal kedua dipilih adalah butir soal yang mampu memberikan informasi maksimum bagi peserta dengan kemampuan pada tingkat 0,5, sebaliknya jika butir pertama dijawab salah maka butir soal kedua dipilih adalah butir soal yang memberikan informasi maksimum bagi peserta dengan kemampuan pada tingkatan -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 tes adaptif 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 0,0535, dan kesalahan pengukuran sudah dapat dihitung, yaitu sebesar 0,8983, dan karena kesalahan baku pengukuran belum mencapai 0,30 maka tes masih berlanjut.

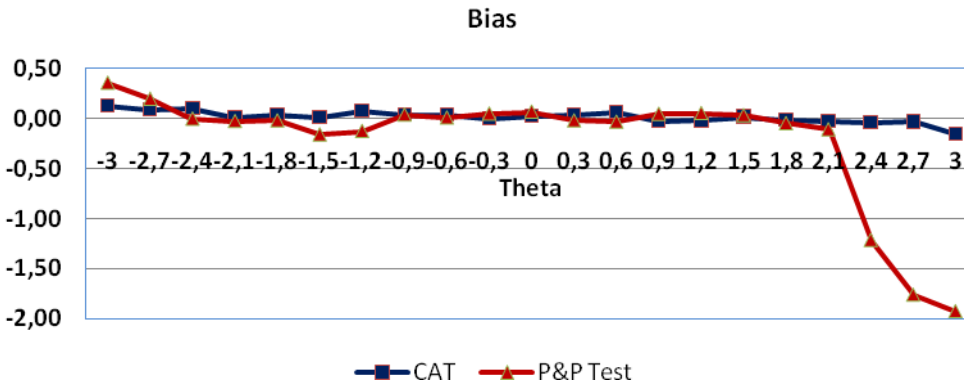
Berdasarkan nilai fungsi informasi maksimum, maka butir soal ketiga yang dipilih adalah butir soal dengan NIS 26. Butir soal ini terpilih karena memiliki nilai fungsi informasi terbesar diantara butir-butir soal lainnya di bank soal untuk theta sebesar 0,0535. Seperti terlihat pada Tabel 2, nilai fungsi informasi butir soal ini sebesar 0,6123. 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-17 karena pada butir ke-17 kesalahan baku pengukurannya telah mencapai 0,30 dengan estimasi theta sebesar -0,023.

Tabel 3. Banyaknya Butir yang Diperlukan pada 21 Tingkatan Theta Yang Disimulasikan

Theta	Banyaknya Butir yang Diperlukan
-3	49
-2,7	36
-2,4	25
-2,1	21
-1,8	19
-1,5	18
-1,2	17
-0,9	17
-0,6	17
-0,3	17
0	17
0,3	17
0,6	17
0,9	17
1,2	17
1,5	18
1,8	20
2,1	24
2,4	39
2,7	65
3	105

Berdasarkan hasil simulasi diperoleh bahwa panjang tes (banyaknya butir soal yang diperlukan) untuk desain CAT untuk setiap tingkat kemampuan yang disimulasikan disajikan pada Tabel 3. Dari Tabel 3 terlihat bahwa banyaknya butir yang diperlukan untuk mengestimasi tingkat kemampuan peserta tes untuk rancangan yang tidak dirandom dan yang dirandom berkisar antara 17 sampai 25 butir soal untuk rentang tingkat kemampuan (theta) antara -2,4 sampai 2,1. Hal ini menunjukkan bahwa dengan rancangan tes adaptif tingkat peserta tes sudah dapat diestimasi kemampuannya hanya dengan 17 sampai 25 butir soal saja untuk rentang theta antara -2,4 sampai 2,1. Namun untuk tingkat kemampuan ekstrim (rendah maupun tinggi) dibutuhkan jumlah butir (panjang tes) yang lebih banyak, yaitu antara 35 sampai 60 butir, bahkan untuk tingkat kemampuan yang paling tinggi (theta=+3) membutuhkan jumlah butir sebanyak 105 butir. Hal ini dikarenakan bank soal yang digunakan kurang menyediakan butir soal dengan tingkat kesukaran tinggi untuk peserta dengan kemampuan tinggi, seperti terlihat pada Tabel 1 bahwa tingkat kesukaran tertinggi dari bank soal yang digunakan pada penelitian ini maksimum adalah sebesar 2,08.

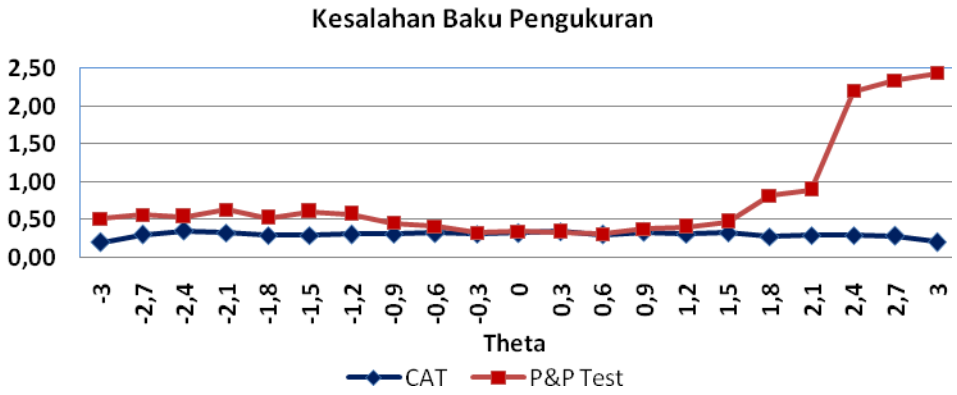
Bias adalah selisih antara estimasi kemampuan (θ) dengan sesungguhnya. Bias untuk setiap kemampuan (theta) yang disimulasikan berdasarkan desain CAT dan P&P test disajikan pada Gambar 2.



Gambar 2. B i a s

Berdasarkan Gambar 2 terlihat bahwa pola bias untuk desain CAT adalah acak pada rentang theta antara -3 sampai +3 dibandingkan desain P&P test. Bias desain CAT umumnya cukup kecil atau hampir mendekati nol untuk setiap theta, hal ini berarti hasil estimasi theta untuk desain CAT sangat akurat, sedangkan bias desain P&P test bervariasi cukup besar untuk setiap theta yang disimulasikan khususnya untuk tingkatan theta yang ekstrim besar.

Gambar 3 menunjukkan kesalahan baku pengukuran (*standard error of measurement*) desain CAT dan P&P test. Kesalahan ini mengindikasikan kesalahan acak dalam pendugaan. pada tingkat kemampuan (theta) tertentu. Kesalahan ini juga menggambarkan presisi dari pendugaan kemampuan.



Gambar 3. Kesalahan Baku Pengukuran

Dari Gambar 3 terlihat bahwa kesalahan baku pengukuran pada setiap tingkat theta untuk desain CAT umumnya lebih kecil dibandingkan desain P&P test, kecuali untuk theta antara -0,3 sampai 0,6. Kesalahan baku pengukuran desain P&P test sangat besar untuk theta ekstrim tinggi.

Berdasarkan banyaknya butir soal yang diperlukan, dari simulasi diperoleh hasil bahwa dengan desain CAT peserta tes sudah dapat diestimasi kemampuannya hanya dengan 17 sampai 25 butir soal saja atau 50% dari jumlah butir soal yang diperlukan pada P&P test atau UBK. Begitupun hasil analisis perbandingan kedua desain menggunakan kriteria bias dan kesalahan baku pengukuran yang menunjukkan bahwa desain CAT memiliki bias dan kesalahan baku pengukuran yang lebih kecil dibandingkan desain P&P test. Dengan demikian, desain CAT tidak hanya efisien tetapi juga lebih akurat dalam mengestimasi tingkat kemampuan peserta tes dibandingkan desain P&P test atau UBK. Desain CAT lebih efisien dan akurat, namun perlu disadari bahwa pada desain CAT murni ini ada kemungkinan butir soal yang diberikan kepada peserta tes tidak mewakili semua materi/modul, dengan kata lain ada butir soal dari materi/modul tertentu yang tidak terwakili sebagai butir soal CAT yang diberikan kepada peserta tes.

Penelitian ini hanya berdasarkan pada penelitian simulasi, pengembangan program aplikasi CAT pada sistem ujian UT belum dilakukan, sehingga kendala-kendala pada proses pengembangan program aplikasi CAT pada sistem ujian UAS di UT belum diketahui. Keterwakilan dari setiap materi/modul atau keseimbangan konten berdasarkan kisi-kisi sebagai syarat tes yang standar dan adanya butir soal yang sering dimunculkan (*item exposure*) pada penelitian ini belum diperhatikan.

KESIMPULAN DAN SARAN

Berdasarkan hasil simulasi maka dapat disimpulkan bahwa: (1) desain CAT lebih efisien karena hanya memerlukan sebanyak 17 sampai 25 butir soal untuk mengestimasi kemampuan peserta tes, (2) desain CAT lebih akurat dalam mengestimasi kemampuan peserta tes dibandingkan desain P&P test.

Penelitian tentang CAT di Indonesia masih sangat jarang, oleh karena itu perlu terus dilakukan penelitian tentang CAT. Perlu dilakukan penelitian lanjutan mengenai pengembangan CAT yang memperhatikan keseimbangan konten dan mengontrol butir soal yang sering dimunculkan sebelum model ini diaplikasikan.

DAFTAR PUSTAKA

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

Bond, T.G. & Fox, C.M. (2007). *Applying the rasch model: Fundamental measurement in the human sciences (2nd ed)*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

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.

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

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

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.

Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., Mislery, R.J., Steinberg, L. et al. (1990). *Computerized adaptive testing*: *A primer* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

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.

Lampiran

A. Konsep IRT yang digunakan untuk mengembangkan CAT

1. Model IRT 1 Parameter (Model Rasch):

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$, dengan *i* : 1,2,3, …,*n*

$P_i(\theta)$: probabilitas peserta tes yang memiliki kemampuan θ yang dipilih secara acak dapat menjawab butir i dengan benar

θ : tingkat kemampuan peserta tes (sebagai variabel bebas)

b_i : indeks kesukaran butir ke-i

e : bilangan natural yang nilainya mendekati 2,718

n : banyaknya butir dalam tes.

2. Fungsi Informasi Item (butir soal):

$$I_i(\theta) = E\left\{\left[\frac{\partial \ln f(U_i;\theta)}{\partial\theta}\right]^2\right\} = \frac{[P_i'(\theta)]^2}{[P_i(\theta)][Q_i(\theta)]}$$

3. Fungsi Informasi Tes

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

4. Kesalahan Baku Pengukuran (*Standard Error of Measurement* = SEM)

$$\text{SEM}(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$$

B. Kriteria Akurasi yang Digunakan Pada Studi Simulasi

5. $Bias(\theta) = \sum_{r=1}^{R}\left(\hat{\theta}_r - \theta\right)/R$ dan $SEM(\theta) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_r - \bar{\theta}\right)^2}$

R = banyaknya replikasi