COMPUTERIZED ADAPTIVE TESTING: EFISIENSI DAN AKURASI PENYELENGGARAAN TES

Agus Santoso
(aguss@ecampus.ut.ac.id)

PENGANTAR

Seiring dengan perkembangan teknologi, tes dengan menggunakan komputer mulai dilakukan. Awalnya, komputer hanya digunakan untuk mengotomatisasikan aktivitas pengukuran yang biasa. Tes yang semula berada di kertas (*paper and pencil test*) dipindahkan ke dalam komputer. Penggunaan komputer seperti ini disebut *Computerized Testing* atau *Computerized-Based Testing* (CBT) dan merupakan generasi pertama penggunaan komputer untuk pengujian (Bunderson, Inouye, & Olsen, 1989). Kelebihan dari CBT, yaitu: meningkatkan standardisasi, meningkatkan keamanan tes, mengurangi penggunaan kertas (*paperless*), meningkatkan kemampuan tampilan tes, dan memperkecil kesalahan pengukuran, serta mempercepat pemberian skor dan interpretasi. Beberapa perguruan tinggi/pemerintah/lembaga pendidikan telah menggunakan CBT antara lain untuk: ujian akhir, penerimaan mahasiswa baru, penerimaan pegawai, maupun ujian akhir semester, bahkan untuk tes kejujuran.

Computerized Adaptive Testing (CAT) merupakan generasi kedua dari penggunaan komputer untuk pengujian (Bunderson, Inouye, & Olsen, 1989). Perkembangan di bidang teknologi komputer dan bidang pengukuran telah melahirkan penyelenggaraan tes dengan desain *adaptive test*. *Adaptive* berarti bahwa butir soal yang diberikan disesuaikan dengan tingkat kemampuan setiap peserta tes atau *tailored testing* (Lord, 1977).

Ide awal konsep *Adaptive test* dimulai oleh Alfred Binet tahun 1905 (Linacre, 2000). Menurut Binet tidak ada keadilan pada penyelenggaraan tes yang memberikan butir soal yang sama kepada semua peserta tes. Oleh karena itu, Ia melakukan tes secara individual. Secara singkat prosedur tes

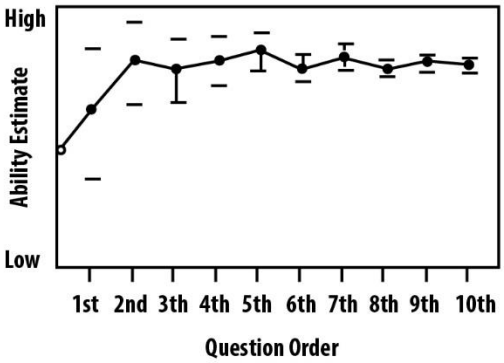
individual yang dilakukan Binet sebagai berikut: kemampuan intelegensi individu peserta tes ditebak, tingkat kesukaran butir diurutkan. Selanjutnya dengan cara manual sekumpulan butir soal yang ditargetkan diberikan, jika individu peserta tes menjawab benar banyak, maka akan diberikan sekumpulan butir soal yang lebih sukar. Sebaliknya jika menjawab salah banyak, maka akan diberikan sekumpulan butir soal yang lebih mudah, begitu seterusnya proses ini berulang sampai tes diberhentikan dan tingkat intelegensi peserta diestimasi. Ide *adaptive test* juga banyak diilhami dari kegiatan olah raga; mungkin seseorang hanya bisa sebatas jalan, yang lain hanya bisa joging, orang yang lain mungkin bisa sampai berlari, kecepatan lari antar individu juga berbeda sesuai dengan kemampuannya. Begitupun kemampuan seseorang dalam lompat tinggi, jika seseorang masih bisa berhasil melompat pada ketinggian tertentu, maka akan dinaikan terus ketinggiannya, sampai ia gagal atau tidak berhasil melompat dan kita dapat menilai kemampuan tinggi lompatannya. Kemampuan tinggi lompatan antar individu juga berbeda sesuai dengan kemampuannya.

Pada CAT yang berbasiskan *item response theory* (IRT), komputer tidak hanya sekedar memindahkan butir soal ke dalam komputer, tetapi komputer diatur untuk menyeleksi dan menyajikan 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 tes konvensional menggunakan *paper and pencil test* maupun CBT. Beberapa penelitian menyimpulkan bahwa jumlah butir atau panjang tes yang diperlukan pada penyelenggaraan CAT hanya memerlukan separoh bahkan kurang dibanding tes konvensional (Weiss, 2004; Eignor, Stocking, Way, & Steffen, 1993; McBride & Martin, 1983). Dalam hal ini, penggunaan CBT dan pengembangan CAT pada penyelenggaraan tes

merupakan salah satu upaya institusi pendidikan dalam memenuhi salah satu target SDGs yaitu peningkatan kualitas pendidikan.

PEMBAHASAN

Menurut Dunkel (1999), CAT adalah suatu metode penilaian secara teknologi di mana komputer menyeleksi dan menyajikan butir soal menurut perkiraan tingkat kemampuan peserta tes. Estimasi tingkat kemampuan peserta tes pada tes adaptif lebih ***akurat*** karena setiap peserta hanya diberi butir soal yang sesuai dengan kemampuannya, dengan kata lain kesalahan pengukuran (*measurement error*) akan lebih kecil. Ilustrasi estimasi tingkat kemampuan peserta tes pada CAT disajikan pada Gambar 1 berikut.



Sumber: Dunkel (1999)

Gambar 1. Estimasi Kemampuan pada CAT

Gambar 1 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 2 berikut menunjukkan proses *adaptive testing*.

Sumber: Modifikasi dari Wainer (1990)

Gambar 2. Proses Adaptive Testing

Berdasarkan Gambar 2, proses *adaptive testing* dimulai dengan memilih butir soal atau kelompok butir soal pertama dari bank soal. Biasanya butir soal pertama dipilih disesuaikan dengan tingkat kemampuan awal optimal dari populasi peserta tes atau dipilih dengan tingkat kemampuan setara dengan tingkat kemampuan awal peserta sedang. Setelah butir soal atau kelompok butir soal dipilih, selanjutnya butir soal diberikan kepada peserta tes. Setelah peserta tes merespon (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.

Komponen-Komponen CAT

Dalam mengaplikasikan sebuah tes ke dalam CAT perlu diperhatikan beberapa komponen. Menurut Wainer (1990) secara umum sistem CAT memiliki empat komponen, yaitu: bank soal, prosedur pemilihan butir soal, pendugaan kemampuan, dan aturan pemberhentian; sedangkan dua komponen CAT yang sering diperhatikan pada sistem CAT adalah

keseimbangan konten dan kontrol butir soal yang sering muncul (*item exposure*).

Menurut Green, Bock, & Humphyers. (1984) dan Kingsbury & Zara (1989) untuk mengembangkan CAT memerlukan evaluasi pada enam komponen berikut:

1) model Item Response Theory (IRT)
2) bank soal
3) pemilihan butir soal awal
4) metode pendugaan tingkat kemampuan
5) prosedur pemilihan butir soal
6) aturan pemberhentian.

1. Model IRT

Model IRT menggambarkan peluang menjawab butir soal secara benar berdasarkan tingkat kemampuan peserta tes dan butir soal yang diberikan. Dalam pendekatan IRT kemampuan individu atau *proficiency* (disimbolkan dengan θ) dan tingkat kesukaran butir atau *difficulty* (disimbolkan dengan *b*) berada pada satu dimensi yang sama (Lord, 1977; Hambleton, Swaminathan, & Rogers, 1991; Embretson & Reise, 2000). Model θ IRT mempunyai tiga asumsi yang mendasari; Pertama, unidimensi; asumsi ini menyatakan bahwa tes hanya mengukur satu kemampuan, θ. Kedua, independensi lokal; asumsi ini menyatakan bahwa jawaban peserta terhadap butir soal pada tingkat kemampuan θ tertentu bebas secara statistik. Ketiga, kecocokan spesifikasi model; asumsi ini menyatakan bahwa pemilihan model matematika cocok dengan data, artinya plot peluang menjawab benar pada suatu butir soal pada skala kemampuan, θ kurvanya mengikuti model IRT yang dipilih. Ketika spesifikasi model cocok dengan data tes maka dua sifat yang diinginkan dari IRT, yaitu sifat invariansi parameter butir atau independensi parameter butir soal terhadap kemampuan dan invariansi parameter kemampuan atau independensi kemampuan terhadap butir soal dapat diperoleh. Kedua sifat IRT ini sangat diinginkan untuk *adaptive test* yang akan memberikan butir soal yang berbeda untuk peserta tes yang berbeda pula.

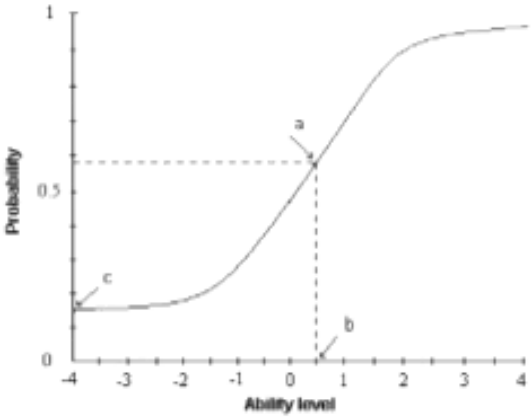
Tiga model IRT yang umum digunakan untuk butir-butir soal dengan format butir soal pilihan ganda adalah model-logistik 1 parameter (1P), 2P, dan 3P. Model 1P merupakan model IRT yang paling sederhana (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Embretson & Raise, 2000). Pada model ini butir-butir soal diasumsikan tidak dapat dijawab benar dengan cara menebak dan mempunyai daya beda yang sama tetapi setiap butir soal mempunyai tingkat kesukaran (*b*) bervariasi. Parameter *b* mengacu pada titik pada skala kemampuan (*ability*) dimana seorang peserta mempunyai peluang 50% menjawab butir soal dengan benar. Semakin besar *b* semakin sulit butir soal itu. Ketika butir-butir soal diasumsikan mempunyai parameter daya beda (dinyatakan dengan *a*) yang bervariasi maka model 2P lebih cocok. Nilai *a* tinggi menunjukkan bahwa butir soal lebih dapat membedakan peserta tes kedalam kelompok kemampuan yang berbeda dibandingkan dengan nilai *a* yang rendah. Model 3P digunakan ketika parameter tebakan semu (*pseudo-guessing*, dinyatakan dengan *c*) diasumsikan ada dalam model. Parameter ini merepresentasikan peluang seorang peserta tes dengan kemampuan sangat rendah menjawab butir soal dengan benar.

Model IRT- 3P sebagai berikut.

$$P(\theta) = c + \frac{1-c}{1+e^{-1,7.a(\theta-b)}}\ , \qquad (1)$$

Untuk setiap butir soal dapat ditampilkan kurva karakteristik (*item characteristic curve*)

Gambar 3 berikut adalah kurva karakteristik butir soal pada model IRT-3P



Gambar 3. Kurva Karakteristik Butir Soal Model IRT-3P

Gambar 3 menunjukkan kurva karakteristik soal, yang menyatakan besarnya peluang menjawab butir soal secara benar berdasarkan tingkat kemampuan peserta tes (*ability level*) dan butir soal yang memiliki tingkat kesukaran, *b;* daya beda, *a*; dan faktor *guessing*, *c*.

2. Bank Soal

Bank soal (*item bank*) adalah kumpulan butir-butir soal (*item pool*) yang sudah dikalibrasi dan sudah memiliki nilai parameter. CAT mengambil butir soal dari bank soal untuk diberikan kepada peserta tes. Ketersediaan butir-butir soal yang cukup dengan kualitas butir-butir soal yang baik pada bank soal sangat menentukan kualitas CAT. Menurut Wang & Vispoel (1998), ada tiga aspek yang memberikan kontribusi terhadap kualitas bank soal, yaitu: ukuran bank soal, parameter butir soal, dan struktur isi. Ukuran bank soal minimal dipengaruhi oleh panjang tes dan ukuran peserta tes. Way (1997) menyarankan rasio 1 berbanding 6 – 8 untuk panjang tes dan banyaknya butir soal minimal yang harus ada dalam bank soal, artinya jika panjang tes CAT dirancang sebanyak 20 butir soal maka banyaknya butir yang harus tersedia pada bank soal minimal sebanyak 120 sampai 160 butir.

Wang & Vispoel (1998) menyarankan bahwa bank soal untuk keperluan CAT sebaiknya memiliki butir-butir soal dengan tingkat daya beda tinggi dan berdistribusi seragam pada setiap tingkat kemampuan.

3. Pemilihan Butir Soal Awal (Starting Point)

Ketika CAT dimulai, belum ada butir soal yang diberikan pada peserta tes, belum ada respons yang diberikan oleh peserta tes sehingga tingkat kemampuan peserta belum dapat diestimasi. Walaupun belum ada informasi mengenai kemampuan peserta sebelumnya, penyelenggaraan CAT harus dimulai. Jika tidak ada informasi awal mengenai kemampuan peserta tes, maka CAT dapat dimulai dengan memilih butir soal awal yang sesuai dengan tingkat kemampuan peserta tes sedang (Green, et al. 1984; Vispoel, 1999; Mills, 1999).

Proses Melanjutkan (*Continue Process*)

Setelah memperoleh jawaban peserta tes terhadap butir soal yang diberikan, selanjutnya komputer menskor jawaban dengan benar atau salah, kemudian memutuskan apakah tes dilanjutkan ataukah tidak. Dua langkah untuk proses melanjutkan CAT, langkah pertama adalah mengestimasi tingkat kemampuan peserta tes; langkah kedua adalah bagaimana memilih butir soal berikutnya.

4. Metode Pendugaan Tingkat Kemampuan

Metode yang umum digunakan untuk mengestimasi kemampuan peserta tes adalah *Maximum Likelihood Estimation* (MLE) (Birnbaun, 1958; Baker, 1992), dan tiga metode Bayes: *Owen's Bayesian Procedure* (OWEN) (Owen, 1975), the *Expected a Posteriori Procedure* (EAP) (Bock & Aitken, 1981; Bock & Mislevy, 1982), and *Maximum a Posteriori Estimation* (MAP) (Samejima, 1969). Berikut dipaparkan secara singkat metode MLE.

Maximum Likelihood Estimation

Misalkan seorang peserta tes dengan tingkat kemampuan θ menjawab tes yang berisi *n* butir soal pilihan ganda dengan parameter butir soal diketahui.

Peluang bersama dari peserta tes dapat dituliskan sebagai $P(U_1, U_2, ..., U_n \mid \theta)$. Selanjutnya dengan asumsi independensi lokal maka fungsi kemungkinannya (*likelihood function*); $L(\theta)$, dituliskan sebagai berikut

$$L(\theta) = P(\mathbf{U} = \mathbf{u} \mid \theta) = P(U_1 = u_1, U_2 = u_2, ..., U_n = u_n \mid \theta) = \prod_{i=1}^{n} P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i}, \quad (2)$$

sedangkan $Q_i(\theta) = 1 - P_i(\theta)$ i = 1,2,..., n, $-\infty < \theta < \infty$.

Tujuan MLE adalah menemukan nilai yang memaksimumkan $L(\theta)$. Nilai parameter kemampuan yang memaksimumkan fungsi kemungkinan, L disebut dengan *the maximum likelihood estimate of ability* (Hambleton, 1993). Secara matematik, hal ini sama dengan untuk menemukan nilai yang memaksimumkan nilai logaritma natural, $\ell nL(\theta)$. Nilai ini dapat diperoleh dengan membuat turunan pertama dari $\ell nL(\theta)$ terhadap θ sama dengan nol.

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \sum_{i=1}^{n} \left[u_i - P_i(\theta) \right] \frac{P_i'(\theta)}{P_i(\theta) Q_i(\theta)} = 0 \ . \quad (3)$$

dan turunan kedua dari $\ell nL(\theta)$ terhadap θ , pada $\theta = \hat{\theta}$ kurang dari nol.

Pada prakteknya, untuk menyelesaikan sistem persamaan (3) dilakukan dengan menggunakan prosedur Newton-Raphson (Hambleton & Swaminathan, 1985; Embretson & Reise, 2000).

Kelemahan dari metode MLE adalah ketidakmampuan dalam mengestimasi kemampuan peserta manakala respons peserta tes belum berpola. Untuk mengatasi masalah ini, biasanya digunakan metode *step size* (Dodd, 1990; Weiss, 2004).

5. Prosedur Pemilihan Butir Soal Berikutnya

Setelah kemampuan peserta diestimasi, kemudian komputer harus memilih butir soal berikutnya. Metode yang digunakan untuk memilih butir butir soal berikutnya pada CAT antara lain: *Maximum Information*, *best matching b-value*, *Kullbak-Leibler Information atau Global Information* (Chang & Ying, 1996), *Likelihood Weight Information Criterian* (Veerkamp & Berger, 1997),

Gradual Maximum Information Ratio (Han, 2009), *Efficiency Balanced Information* (Han, 2012). Metode yang populer digunakan untuk memilih butir soal berikutnya pada CAT, yaitu *Maximum Information*. Berdasarkan prosedur ini, butir soal yang mempunyai nilai fungsi informasi terbesar pada kemampuan peserta tertentu dipilih untuk diberikan pada peserta tes. Hal ini menjamin bahwa nilai fungsi informasi tes untuk setiap peserta tes adalah maksimum, artinya kesalahan baku pengukuran minimum.

Secara matematis, fungsi informasi butir soal dituliskan sebagai berikut.

$$I_i(\theta) = \frac{\left[P_i'(\theta)\right]^2}{\left[P_i(\theta)\right]\left[Q_i(\theta)\right]}, \tag{4}$$

sedangkan *i* menyatakan nomor butir soal ke-*i*, dan $P'(\theta)$ adalah turunan pertama dari $P(\theta)$ pada θ (Lord, 1980; Hambleton et al., 1991). Persamaan (4) menunjukkan bahwa nilai informasi hanya tergantung pada parameter butir soal (misalnya; *a*, *b*, dan *c* untuk model IRT-3P), dan tingkat kemampuan, θ. Dengan demikian untuk setiap tingkat kemampuan, θ, kontribusi informasi untuk setiap butir soal pada bank soal dapat dihitung.

Nilai fungsi informasi butir soal menggambarkan seberapa akurat suatu butir dapat mengestimasi tingkat kemampuan peserta tes. Selanjutnya dengan mengasumsikan independensi lokal, jumlah fungsi informasi *n* butir soal dikenal sebagai fungsi informasi tes. Secara matematis fungsi informasi tes dinyatakan dengan rumus sebagai berikut.

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta), \tag{5}$$

Sedangkan kesalahan baku pengukuran (*standard error of measurement*) adalah

$$SEM = \frac{1}{\sqrt{I(\theta)}}, \quad (6)$$

Semakin besar informasi tes pada tingkat kemampuan yang diberikan semakin akurat kemampuan itu diestimasi dari perangkat tes itu. Dengan kata lain kesalahan baku pengukuran semakin kecil.

6. Aturan Pemberhentian (Stopping Rule)

Dua metode utama yang digunakan untuk memberhentikan CAT, yaitu *equal measurement precision* dan *fixed number of items*. Kedua metode ini menghasilkan variansi kesalahan pengukuran yang berbeda. Tujuan digunakannya metode *equal measurement precision* adalah menghasilkan skor tes dengan tingkat ketepatan estimasi yang sama. Namun, panjang tes diprediksikan bervariasi dari satu peserta dengan peserta tes lainnya. Sedangkan penerapan aturan *fixed number of items* akan berakibat pada ketepatan estimasi yang tidak sama dan mengakibatkan tes tidak adil, namun demikian, kriteria ini lebih mudah diterapkan.

Keunggulan CAT

1) Efisien; Jumlah butir soal yang diperlukan 67 persen bahkan hanya 50 persen dari panjang tes konvensional menggunakan *paper and pencil test* maupun CBT.
2) Akurat; butir soal yang diberikan sesuai dengan kemampuan individu peserta tes, sehingga memperkecil kesalahan pengukuran.
3) Adil; bisa mengontrol tingkat kesalahan pengukuran yang sama untuk setiap peserta tes. Dengan menetapkan tingkat kesalahan pengukuran tertentu, maka setiap individu peserta tes akan berhenti mengerjakan tes manakala telah mencapai tingkat kesalahan pengukuran tertentu.
4) Menambah keamanan tes; setiap peserta menerima butir soal yang berbeda, sehingga butir soal tidak mudah dikenali.

5) Memotivasi; peserta tes yang pandai tidak banyak membuang waktu untuk mengerjakan butir tes yang mudah, sebaliknya peserta tes yang kurang pandai tidak terbelenggu dengan butir soal yang sulit.
6) Penyelenggaraan tes bisa diulang lebih sering tanpa khawatir soal dihafal oleh peserta tes.
7) Mempercepat pemberian skor dan interpretasi.
8) Meningkatkan kemampuan tampilan tes (audio, video, hotspot, dll.).
9) Mudah mengumpulkan hasil tes.

Kelemahan CAT

1) Terkait dengan masyarakat; perlu menjelaskan dan meyakinkan kepada peserta tes dan atau orang tua tentang sistem CAT; adanya kemungkinan tes diberhentikan dengan cepat dan sudah bisa diketahui lulus dan tidak lulusnya peserta tes.
2) Peserta tes tidak bisa kembali ke soal sebelumnya.
3) *Item Exposure*; dimungkinkan adanya butir soal yang sering dimunculkan dibandingkan dengan butir soal yang lainnya, jika ini terjadi maka keamanan tes dapat terganggu.
4) Tidak fisibel/aplikatif untuk semua situasi; bank soal hasil kalibrasi dengan model IRT kurang akurat untuk ukuran sampel kecil. Tidak cocok untuk tes uraian.
5) Berdasarkan teori pengukuran modern, IRT; perlu rancangan software khusus.
6) Butuh ahli di bidang teori pengukuran modern, IRT.
7) Mahal; untuk mengembangkan satu sistem CAT memerlukan 350 juta sampai 1 milyar rupiah bahkan lebih.

Dengan memperhatikan keunggulan dan kelemahan CAT tersebut, upaya pengembangan CAT memiliki peluang besar untuk terus dilakukan di berbagai bidang. Dalam bidang pendidikan, penerapan CAT juga akan memungkinkan dikembangkannya pendidikan yang dapat diikuti oleh berbagai tingkat kemampuan individu peserta.

PENUTUP

Computerized Adaptive Testing (CAT) merupakan suatu metode penilaian menggunakan teknologi, dimana komputer menyeleksi dan menyajikan butir soal menurut perkiraan tingkat kemampuan peserta tes. CAT telah diaplikasikan untuk pengukuran di bidang pendidikan, kedokteran, musik, dan lain lain. Beberapa tes yang telah mengaplikasikan sistem CAT antara lain; GMAT (*Graduate Management Admission Test*), ASVAB (*The Armed Services Vocational Aptitude Battery*), ASCP (*American Society of Clinical Pathologies*), ETS (*Educational Testing Services*), PROMIS (*Prosecutor's Management Information System*), TOEFL (*Test of English as Foreign Language*). Di Indonesia: Pusat Penilaian Pendidikan Balitbang Kemendikbud sejak tahun 2010 telah mengkaji dan masih melakukan kajian terhadap sistem CAT untuk penyelenggaraan ujian nasional.

Masih banyak topik atau materi dari komponen-komponen dan penerapan CAT yang masih perlu dikaji dan diteliti. Beberapa pertanyaan penelitian, antara lain: Bagaimana membangun algoritma CAT, sehingga tes terstandarkan? Bagaimana kebermanfaatan bank soal? Bagaimana mengatasi butir soal yang sering dimunculkan (*item exposure*) pada CAT? Bagaimana waktu untuk menjawab butir soal perlu dibatasi ataukah tidak? dan masih banyak pertanyaan yang memerlukan jawaban dari penelitian.

Penggunaan CAT dalam bidang pendidikan sangat bermanfaat, penelitian terkait CAT dapat memberikan suatu inovasi pengukuran pembelajaran. Dengan demikian, pengembangan CAT dan penggunaannya oleh institusi pendidikan akan mendukung agenda SDGs yang memastikan pendidikan inklusif dan berkualitas setara serta mendorong kesempatan belajar seumur hidup bagi semua.

REFERENSI

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

Birnbaum, A. (1958). *On the estimation of mental ability*. Series Report 15, Project No. 7755–23. Randolph Air Force Base. Tx: USAF School of Aviation Medicine.

Bock, R.D. & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 4, 443–459.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 4, 431–444.

Bunderson, C.V., Inouye, D.K., & Olsen, J.B. (1989). *The four generations of computerized educational measurement*. Dalam R. L. Linn (Eds.), *Educational Measurement* (3nd ed., pp. 367–407). New York: American Council on Education & Macmillan Publishing Company.

Chang, H.-H., & Ying Z. (1996). A global information approach to computerized adaptive testing with beta blocking. *Applied Psychological Measurement,* 25, 333–341.

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.

Dunkel, P.A. (1999). Considerations in developing and using computer adaptive tests to asses second language proficiency. Diakses melalui http://www.cal.org/resources/digest/cat.html.

Eignor, D.R., Stocking, M.L., Way, W.D., & Steffen, M. (1993). *Case studies in computer adaptive test design through simulation (Research Report 93-56).* Princeton, NJ: Educational Testing Service.

Embretson, S.E. & Reise, S.P. (2000). *Item response theory for psychologist.* London: Lawrence Erlbaum Associates, Inc.

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

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

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

Hambleton, R.K. (1993). Principles and selected applications of item response theory. Dalam R. L. Linn (Eds.), *Educational Measurement (3rd ed.,* pp. 147-200). Phoenix, AZ: American Council on Education and the Oryx Press.

Han, K.T. (2009). *A gradual maximum information ratio approach to item selection in computerized adaptive testing.* Reseach Reports 09-07.McLean, VA: Graduate Management Admission Council.

---------- (2012). An Efficiency Balanced Information Criterion for Item Selection in Computerized Adaptive Testing. *Journal of Educational Measurement*, 3, 225-246

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

Linacre, J.M. (2000). *Development of computerized middle school achievement test*. Soul, South Korea: Komesa Press.

Lord, F.M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, 1 , 95–100.

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

Mc Bride, J.R., & Martin, J.T., (1983). Reliability and validity of adaptive ability tests in a military setting. Dalam D.J. Weiss (Eds). *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223–236). New York: Academic Press.

Mills, C.N. (1999). Development and introduction of a computere adaptive graduate record examinations general test. Dalam F. Drasgow & J. B. Olson-Buchanan (Eds), *Innovations in Computerized Assessment* (pp. 117–136). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Owen, R.J. (1975). A bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70, 351-356.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph,* No. 17.

Veerkamp, W.J.J., & Berger, M.P.F. (1997). Some new item-selection criteria for adaptive testing. *Journal of Behavioral Statistics*, 2, 203-226

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.

Wang, T., & Vispoel, W.P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, 2, 109–136.

Way, W.D. (1997). *Protecting the integrity of computerized testing item pools.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

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.