ADMINISTRASI TES PRESTASI DENGAN MODEL *COMPUTERIZED ADAPTIVE TESTING* (CAT) DAN DENGAN DIBATASI WAKTU RESPONSE BUTIR SOAL

(*Administration of Achievement Tests with Computerized Adaptive Testing (CAT) Model and Limited Response Time for Questions*)

Handaru Catu Bagus[1)], Burhanuddin Tola[2)], Awaluddin Tjalla[3)]

[1,2,3)]Pascasarjana UNJ, Jl. Pemuda No.28, RT.7/RW.14, Rawamangun, Kec. Pulo Gadung, Kota Jakarta Timur, Daerah Khusus Ibukota Jakarta 13220
e-mail: handcab@gmail.com, burhanuddin.tola@gmail.com, awaluddin.tjalla@gmail.com

Abstract. Assessment models that ignore the ability of individual variations to cause the information received will not be optimal. The computerized adaptive testing (CAT) model can overcome this weakness because the items that appear with the difficulty level will adjust to the test taker's ability. This study aimed to analyze the effectiveness, efficiency, and accuracy of the CAT model when used as an assessment model in school achievement tests. The research methodology is comparative quantitative. This study uses population data from the answers of students who took part in the UNBK in the province of the Special Capital Region (DKI) Jakarta in 2019 with mathematics as a subject. The results of this study show that the number of items selected by the CAT model is less than the CBT model, and the things adjust to the level of ability of the traveler, have a small measurement error value, and are almost the same as the CBT model. Therefore, the CAT model is more efficient in terms of time because the number of questions is less than the CBT model. It is more effective because it adapts to the participants' abilities and has the same accuracy as the CBT model.

Keywords: Adaptive model, Computer, Computerized adaptive testing, Computer-based test, Evaluation, Item response theory.

1. Pendahuluan

Ujian Nasional yang selanjutnya disebut UN adalah tes prestasi dengan standar nasional yang diukur kepada siswa kelas akhir di jenjang SMP sederajat dan SMA sederajat yang dilaksanakan diseluruh satuan Pendidikan di Indonesia. Dalam pelaksanaannya, UN menggunakan dua model administrasi tes yaitu administrasi tes secara tertulis selanjutnya disebut *paper and pencil test* (PPT) atau model konvensional dan administrasi tes berbasis komputer yang selanjutnya disebut *computer based tes* (CBT). Kedua administrasi tes tersebut menggunakna model penilaian dengan desain tes yang similar atau sama untuk setiap penempuh tesnya dengan tidak membedakan umur serta jenjang pendidikan. Asumsi ini mendasari bahwa penempuh tes dengan umur dan jenjang pendidikan yang sama memiliki kemampuan yang sama. Padahal secara prakteknya terdapat kemampuan yang sangat bervariasi untuk setiap penempuh tes tersebut.

Beragamnya kemampuan individu pada model penilaian kompetensi yang diabaikan

tersebut memiliki kelemahan, yaitu informasi tes yang didapatkan akan tidak optimal. Sebagai contoh, desain tes dalam kategori mudah namun diberikan kepada penempuh tes dengan kemampuan tinggi maka informasi yang diperoleh akan kurang berarti, karena dapat dipastikan bahwa penempuh tes tersebut akan menjawab benar pada desain tes tersebut. Dan sebaliknya, apabila desain tes dalam kategori sukar diberikan kepada penempuh tes dengan kemampuan rendah maka ada kemungkinan penempuh tes tersebut mendapatkan skor minimal atau bahkan nol. Oleh karena itu, masalah yang muncul yaitu keadilan dan informasi yang dihasilkan akan menjadi tidak akurat dan presisi.

Pasal 1 ayat 1 PERMEN DIKNAS No. 75 tahun 2009 menyatakan bahwa Ujian Nasional yang selanjutnya disebut UN adalah kegiatan pengukuran dan penilaian kompetensi peserta didik secara nasional pada jenjang pendidikan dasar dan menengah. UN resmi diselenggarakan pada tahun 2005 hingga tahun 2020 dan mulai tahun 2021 UN dihapus. Dan sejak tahun 2015 UN sudah tidak lagi sebagai penentu kelulusan di seluruh jenjang, sehingga UN hanya sebagai pemetaan terhadap ketercapaian proses belajar di kelas sesuai dengan kurikulum, namun penyelenggaraan ujiannnya dilakukan oleh pihak atau Lembaga mandiri lain yaitu Badan Standarisasi Nasional Pendidikan selanjutnya disebut BSNP dengan harapan agar lebih obyektif. Dan sejak tahun 2015 UN sudah menerapkan UN berbasis komputer selanjutnya disebut UNBK, sehingga sejak 2015 tersebut memiliki dua moda pelaksanaan UN yaitu moda PPT dan moda CBT [4].

Dalam teknis Prosedur Operasi Standar (POS) UN tahun pelajaran 2016 – 2017 pada bab IV tentang Bahan Ujian Nasional terlihat jelas bahwa paket tes yang diberikan kepada peserta didik mengabaikan variasi kemampuan individu, mengingat bahwa pada paket tes UN untuk satu rombongan belajar pada satu sekolah yang disiapkan, digandakan dan diberikan kepada peserta didik untuk kedua moda dengan menggunakan desain paket tes yang setara. Oleh karena itu sulit memperoleh informasi yang *on target* dikarenakan desain paket tes yang setara khususnya terhadap variansi kategori kemampuan penempuh tes [5]. Untuk mengatasi permasalahan tersebut, maka penulis mencoba untuk meneliti response penempuh tes UN yang menggukann model CBT dimodelkan dengan model tes adaptif.

Untuk mengatasi kelemahan model penilaian pendidikan yang berlangsung selama ini, model tes adaptif dapat menjadi alternatif administrasi ujian. Hal ini karena model ini memungkinkan desain tes memperoleh informasi yang on target, sebab tes dengan tingkat kesukaran butir soalnya diseuaikan dengan kemampuan penempuh tesnya. Oleh karena itu diharapkan dengan penggunaan desain tes tersebut akan menghasilkan informasi yang optimal sebab model tes akan terhenti setelah informasi kemampuan peserta didik dapat diestimasi. Teknik estimasi kemampuan penempuh tes dengan model tes adaptif menggunakan pendekatan teori tes modern atau *Item Response Theory* (IRT). Pendekatan IRT berorientasi pada butir soal yang berhubungan dengan kemampuan penempuh tesnya, dan tidak berorientasi pada instrument tes [8]. Oleh karena itu, dengan pendekatan IRT performa seseorang atau sekelompok orang dalam sebuah item dapat diprediksi.

Pada proses model tes adaptif dalam memilih dan menampilkan butir soal yang disesuaikan dengan informasi kemampuan peserta didik maka untuk mempermudahnya model ini dibantu oleh media komputer atau terkomputerisasi sehingga hasil yang diperoleh akan lebih cepat, efektif dan akurat dalam menghasilkan informasi yang optimal. Berdasarkan penjelasan diatas, dapat dipahami bahwa *Computerized Adaptive Testing* (CAT) adalah tes adaptif yang penyajian tesnya dibantu oleh media komputer termasuk dalam pemilihan butir soal hingga pengolahan hasil tes. Bunderson [3] mencatat beberapa kelebihan dari CAT, antara lain: meningkatkan kontrol dalam menampilkan item, meningkat keamanan tes, memperkaya kemampuan tampilan, diperoleh skor yang sama dengan waktu yang lebih singkat, mengurangi *error of measurement*, meningkatkan penyekoran dan pelaporan.

Penelitian model CAT ini sangat penting dilakukan sebagai alternatif pengganti model penilaian, khususnya tes prestasi baik di sekolah ataupun nasional atau UN, yang selama ini diterapkan di Indonesia. Selain itu, masalah kebocoran dan kecurangan yang selama ini terjadi dalam penyelenggaraan UN dapat diminimalisir.

Berdasarkan penjelasan di atas, penelitian ini bertujuan sebagai berikut pertama menganalisis efisiensi waktu pengerjaan dengan model CAT dibandingkan dengan model CBT apabila CAT dimodelkan dalam tes prestasi. Kedua menganalisis efektivitas jumlah butir soal yang disajikan pada model CAT dibandingkan dengan model CBT apabila CAT dimodelkan dalam tes prestasi. Ketiga menganalis uji hubungan atau korelasi antara hasil kemampuan penempuh tes yang menggunakan model CAT dibandingkan dengan model CBT. Hal ini untuk mengukur keakuratan model CAT.

Asesmen atau penilaian pendidikan merupakan bagian dari proses belajar mengajar. asesmen adalah serangkaian kegiatan untuk memperoleh, menganalisis, dan menafsirkan data tentang proses dan hasil belajar peserta didik yang dilakukan secara sistematis dan berkesinambungan, sehingga menjadi informasi yang bermakna dalam pengambilan keputusan oleh pihak sekolah atau pengambil keputusan [1].

Didalam proses belajar mengajar juga mengenal jenis penilaian diantaranya adalah *Asessment of Learning* atau lebih dikenal dengan tes prestasi belajar. Asesmen ini diberikan kepada siswa setelah siswa melakukan pembelajaran dan umumnya asesmen ini diberikan di periode akhir setiap semesternya. UN merupakan tes prestasi hasil belajar siswa secara sumatif yang dikerjakan oleh siswa pada tingkat akhir di masing-masing jenjang. Oleh karena itu penilaian sumatif digunakan untuk mengukur prestasi seorang siswa setelah adanya proses pengajaran dari guru, sehingga penilian sumatif bisa dikatakan sebagai tes prestasi akhir setiap proses pengajaran.

Tes Prestasi masuk dalam kelompok tes performansi maksimum. Tes performansi maksimum adalah kemampuan terbaik yang mampu diperlihatkan oleh penempuh tes sebagai jawaban terhadap butir soal. Pengonstruksian tes jenis ini harus memiliki banyak

stimulus (berupa pertanyaan) yang terstruktur secara jelas. Pertanyaan dan arah jawaban yang dikehendaki dalam alat ukur, harus benar-benar dapat dipahami oleh penempuh tes sebelum mereka memberikan respon. Karena respon penempuh tes berkaitan dengan kemampuan kognitifnya, maka respon yang dipilih oleh penempuh tes dapat dikatakan sebagai respon yang "benar" atau "salah" dan diberi skor yang sepadan.

Dalam pelaksanaan tes prestasi baik secara nasional atau untuk kepentingan penilaian kelas atau sekolah, asesmen kepada siswa dalam bentuk tes atau ujian kepada penempuh tes dapat dilakukan dengan berbagai cara, mulai dengan administrasi tes dengan cara konvensional, yaitu dengan menggunakan kertas (*paper-pencil test*), hingga pemanfaatan teknologi, seperti *Computer Based Test* (CBT) dan *Computerized Adaptive Test* (CAT). Berikut akan dijelaskan mengenai bentuk administrasi tes yang banyak dilakukan.

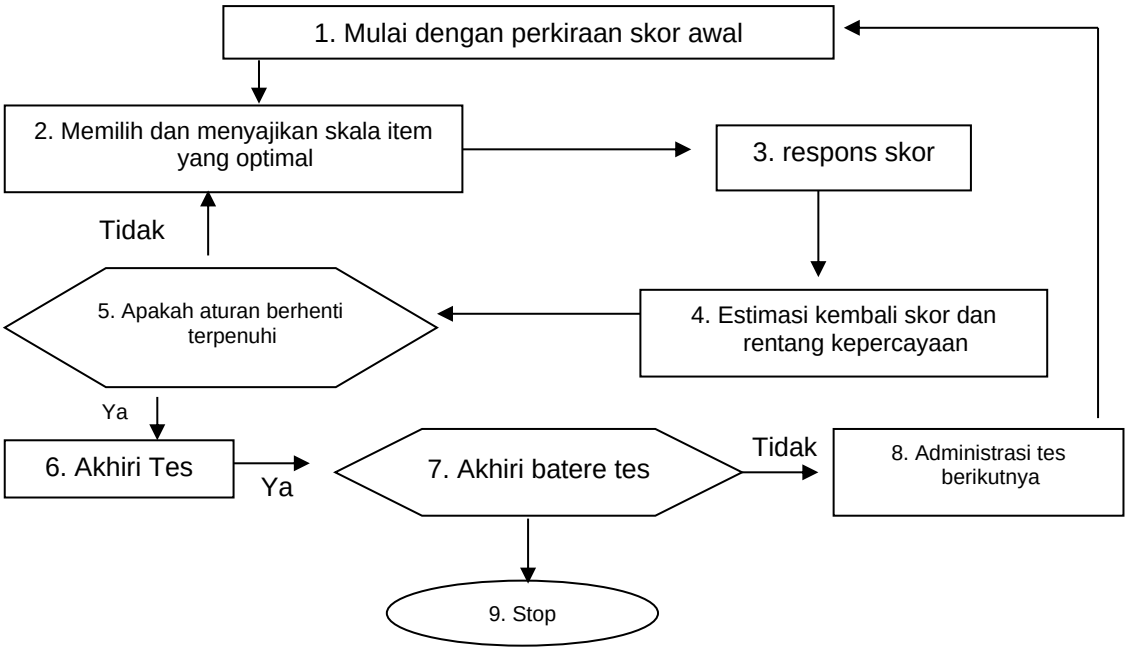
Computer Based Test (CBT) merupakan administrasi tes generasi pertama dalam penggunaan komputer dalam pengetesan. Dengan CBT penampilan butir soalnya dibantu oleh perangkat komputer. Dalam hal desain paket tesnya, CBT sama halnya dengan PPT, yaitu bentuk administrasinya juga masih konvensional karena semua penempuh tes menerima seperangkat butir soal yang sama, dan mengabaikan keberagaman kemampuan dari penempuh tesnya. Namun kelebihan CBT dibandingkan dengan PPT adalah seluruh aktifitas dari penempuh tes dapat terekam oleh sistem sehingga aktifitas pengerjaan siswa menjadi lebih mudah terkontrol dan seluruh informasi terekam secara cepat oleh sistem yang sudah terkomputerisasi. Oleh karena itu aktifitas penempuh tes dari awal tes hingga akhir tes tercatat dan terekam oleh sistem, termasuk respons siswa, waktu respon butir, pola respon Ketika singgah (visit) dari butir satu ke butir yang lain, dan beberapa aktifitas lainnnya. Dengan kedatangan pengujian berbasis komputer menjadikan pencatatan waktu respons sebagai bagian rutin dari moda tes. Hasil rekaman aktifitas ini yang akan digunakan untuk dapat mengetahui seberapa lamanya penempuh tes dalam meresponse jawaban setiap butir soalnya. Dalam pelaksanaan UN di tahun 2019, CBT sebagai administrasi tes yang digunakan dalam UNBK.

Computerized Adaptive Testing (CAT) merupakan generasi kedua dari penggunaan komputer untuk pengetesan [3]. Salah satu aplikasi dari pendekatan IRT adalah penggunaan CAT. Adaptif memiliki pengertian bahwa butir soal yang diberikan sesuai dengan kemampuan setiap penempuh tes, sehingga setiap individu akan mendapatkan seperangkat butir soal yang berbeda. Leung [12] mengatakan bahwa dalam CAT seorang penempuh tes diberikan butir soal yang dipilih berdasarkan kemampuannya yang diperkirakan (θ). Karena setiap individu mendapatkan seperangkat butir soal yang disesuaikan dengan kemampuannya maka CAT termasuk dalam tailored-testing. Dengan demikian, CAT berbasis IRT biasanya berisi lebih sedikit butir soal dibandingkan pengukuran PPT atau CBT yang biasa atau konvensional [6].

Dari penjelasan di atas, terlihat jelas bahwa CAT berbasis IRT biasanya berisi lebih sedikit item dibandingkan pengukuran PPT dan CBT yang konvensional [6]. Hasil

empirik juga dijelaskan oleh Olsen (dalam [3]) yang mencatat bahwa pada sebuah tes prestasi belajar hanya dibutuhkan 30% hingga 50% dari keseluruhan butir soal dalam paket tes untuk mencapai tingkat presisi yang sama dengan PPT atau CBT. Dengan berkurangnya jumlah butir soal yang diberikan kepada penempuh tes, maka secara langsung akan mengurangi jumlah waktu response yang dibutuhkan untuk mengadministrasikan tes [3].

Konsep algoritma yang dipakai oleh CAT adalah sebagai berikut: apabila penempuh tes tidak dapat menjawab benar pada butir soal yang diberikan maka komputer akan memberikan butir soal yang memiliki derajat kesukaran yang lebih rendah. Apabila penempuh tes tidak dapat menjawab benar, komputer akan memberikan butir soal yang memiliki derajat kesukaran lebih rendah. Sebaliknya, apabila penempuh tes dapat menjawab benar, butir soal yang diberikan selanjutnya adalah butir soal dengan kesukaran lebih tinggi. Gambar 1, memberikan bagan proses CAT yang dikemukakan oleh Wainer [16].



Gambar 1. Bagan proses CAT

Dalam penerapan sebuah tes ke dalam CAT yang berbasis IRT, ada beberapa hal yang harus diperhatikan. Embretson dan Reise [6] memaparkan lima faktor harus diperhatikan dalam CAT. Pertama, Item bank. Tujuan dari CAT adalah untuk mengadministrasikan serangkaian butir soal yang dapat memberikan informasi dan efisiensi yang maksimal untuk setiap penempuh tes. Untuk mewujudkan hal ini, penempuh tes yang berbeda akan menerima rangkaian item yang berbeda, dan skor mereka pada kemampuan laten diperkirakan berdasarkan respons mereka terhadap item-item yang berbeda tersebut. Dengan demikian, kapasitas untuk mewujudkan CAT bergantung pada sebuah item bank yang berisi sekumpulan besar item di mana parameter IRT dari setiap item telah diketahui.

Tujuan pengetesan adalah untuk mengukur dengan baik keseluruhan rentang kemampuan maka idealnya sebuah item bank berisi sejumlah item yang memiliki kemampuan daya beda tinggi dengan parameter kesukaran tersebar di antara rentang kemampuan. Ketika sebuah item bank memenuhi kriteria ini, seluruh penempuh tes dapat diadministrasikan pengujian yang tepat dan mereka dapat diukur secara akurat.

Kedua, Mengadministrasikan item pertama. Apabila diasumsikan kemampuan penempuh tes dalam populasi terdistribusi secara normal maka dapat dimulai dengan parameter kesukaran sebesar -0,5 hingga 0,5. Apabila diperoleh informasi mengenai kemampuan penempuh tes dalam kontinum kemampuan maka informasi tersebut dapat digunakan untuk memilih tingkat kesulitan pada butir soal di awal. Rata-rata θ dari populasi penempuh tes dapat digunakan sebagai perkiraan kemampuan sehingga dapat menjadi optimal [15].

Ketiga, Pemberian skor. Terdapat tiga metode utama untuk mengestimasi posisi penempuh tes dalam kontinum kemampuan, yaitu ML (*maximum likelihood*), MAP (*maximum a posterori*), dan EAP *(expected a posteriori*). Beberapa peneliti tidak menganjurkan penggunaan informasi sebelumnya karena dapat berpotensi untuk mempengaruhi skor. Misalnya, apabila hanya sedikit item yang diadministrasikan maka tingkat kemampuan yang diestimasikan akan tertarik ke arah nilai rata-rata dari distribusi awal. Untuk itu digunakan prosedur step-size untuk memberikan skor di tahapan awal CAT.

Keempat, Pemilihan item berikutnya. Pemilihan item berikutnya terkait dengan pemberian skor. Strategi yang dapat digunakan untuk memberikan item berikutnya adalah *maximum information* dan *minimum expected posterior standard deviation*, yang disebut juga *Bayesian estimation* [15]. Pada *maximum information* dilakukan dengan memilih item pada setiap tahap yang memiliki nilai *b* mendekati perkiraan θ saat itu.

Kelima, Menghentikan Tes. Dalam CAT, setiap kali kemampuan penempuh tes diperkirakan kemampuannya berdasarkan respons terhadap item dan *standard error* diperkirakan kembali, komputer kemudian memilih item selanjutnya untuk diberikan. Ada dua kriteria untuk menghentikan administrasi CAT, yaitu *variable length dan fixed length*. Pada *variable length*, administrasi CAT berhenti ketika *standard error measurement* sudah mencapai batasan yang telah ditetapkan. Thissen dan Mislevy (1990) menyebut kriteria ini sebagai *target precision*. Penentuan *standard error*, menurut Hornke [11], dengan standard error lebih kecil atau sama dengan 0,38, akan sepadan dengan koefisien reliabilitas sebesar 0,85. Di lain pihak, Blais dan Raiche [2] menemukan apabila standard error of measurement lebih kecil atau sama dengan 0,40 maka *standard error* dari tingkat kemampuan individu hanya berbeda sebesar 0,03. Prosedur *fixed length* marupakan pemberhentian pengetesan apabila sejumlah item tertentu telah diadministrasikan. Thissen dan Mislevy [15] menyebut kriteria ini sebagai *maximum number of items*. Kelebihannya adalah mudah untuk dilakukan dan penggunaan item

dapat diperkirakan dengan tepat.

Penelitian ini menggunakan kerangka teori seperti disebutkan di atas sebagai dasar dalam metodologi proses penelitian model CAT. Oleh karena itu, lima faktor di atas menjadi sangat penting sebagai dasar model CAT dapat diaplikasikan.

Waktu Response Butir Soal

Waktu respon butir soal adalah jumlah waktu yang diperlukan untuk menyelesaikan setiap butir soal. Setiap butir soal memiki tingkat kesukaran, semakin sulit butir soal yang diberikan maka secara *time intensity* waktu yang dibutuhkan oleh penempuh tes semakin lama, dan sebaliknya semakin cepat waktu mengerjakan maka butir soal semakin mudah. Waktu respon butir soal dapat dimanfaatkan untuk berbagai macam peruntukan. Salah satu yang umum dipakai adalah durasi rerata yang diperlukan untuk menempuh setiap butir soal dapat digunakan untuk untuk menjustifikasi jumlah soal yang tepat diujikan kepada siswa untuk satu sesi ujian, sehingga tidak terjadi kekurangan waktu (*speed test*) atau terlalu banyak waktu.

Manfaat yang lain dari waktu respons butir soal adalah untuk meningkatkan akurasi estimasi kemampuan siswa, karena siswa yang dapat menyelesaikan lebih cepat diasumsikan kemampuan lebih tinggi dibandingkan yang lambat dalam pengerjaan. Dari segi butir soal, terdapat sejumlah soal yang pertanyaannya sangat kompleks, sehingga metode pemecahannya menjadi sangat sulit dan membutuhkan waktu yang relatif tidak terbatas dan batasnya adalah kesanggupan respons dari penempuh tesnya. Umumnya tes prestasi hasil belajar adalah *power test*, artinya tidak ada batasan waktu bagi penempuh tes untuk menjawab seluruh pertanyaan. Namun tetap saja dengan tingkat kesukaran item semakin lama semakin meningkat sehingga dengan waktu yang lebih luangpun sebagian orang tidak akan mampu menjawab dengan benar.

Ketika penerapan waktu pengerjaan tes dibatasi, dan khususnya ketika tes memiliki batas waktu yang ketat, penempatan butir soal dengan kategori mudah ditempatkan dan ditampilkan pada nomor awal ujian akan menghasilkan skor yang lebih tinggi dibandingkan ketika butir soal dengan kategori sulit ditampilkan dinomor awal dan butir soal kategori mudah ditempatkan di nomor terakhir. Hal ini menguatkan konsep bahwa sifat *power test* tidak membatasi waktu dalam pengerjaan tes. Namun setiap berlangsungnya tes prestasi di sekolah waktu pengerjaan tes tetap harus dibatasi. Perkiraan waktu respon butir untuk model soal pilihan ganda yaitu antara 40 sampai dengan 60 detik, sementara model tes isian singkat setiap butir memiliki perkiraan waktu respon butir yaitu antara 15 sampai dengan 20 menit. Untuk waktu respon butir pada UNBK merupakan waktu agregat respon butir yaitu 120 menit. Sehingga waktu agregat tersebut tidak menyesuaikan kondisi dari tingkat kesukaran setiap butir tersebut.

Apabila UN menggunakan moda pelaksanaan CAT maka pengaturan waktu respon butir

dapat langsung dimodifikasi sesuai dengan kebutuhan tes. Jadi dengan moda pelaksanaan CAT selain pembangkitan butir yang diberikan kepada penempuh tes sesuai dengan kemampuannya, dengan CAT juga dimungkinkan setiap butir yang tampil hasil pembangkitan tersebut dapat diatur waktu respon butir. Selain pembangkit butir yang disesuaikan dengan kemampuan penempuh, waktu respon butir, dimungkinkan juga pengukuran estimasi kemampuan penempuh tes setiap kali merespon butir tersebut. Setiap penempuh tes dengan kemampuan yang berbeda membutuhkan jumlah waktu yang sama untuk menyelesaikan tes. Dengan demikian, penempuh tes yang lebih lambat mungkin tidak dapat menyelesaikan semua butir soal pada tes dengan waktu pengerjaan yang dibatasi.

True Theta (Kemampuan Sebenarnya)

Setiap Penilaian, baik dalam tes diagnostik maupun tes prestasi selalu mengandung kesalahan pengukuran. Setiap kemampuan (*theta*) yang diperoleh dari penempuh tes terdiri atas tiga hal, pertama nilai kemampuan amatan yang sering pula disebut sebagai estimasi kemampuan penempuh (*estimated theta*), kedua nilai kemampuan yang sebenarnya (*true score*) yaitu nilai acuan yang sesuai dengan kemampuan penempuh tes yang sebenarnya, dan ketiga kesalahan pengukuran, yaitu faktor yang mempengaruhi ketidakajegan suatu pengukuran sehingga mempengaruhi perolehan skor. Oleh karena true theta dipengaruhi oleh nilai estimasi kemampuan penempuh dan kesalahan pengukuran maka dapat dirumuskan pada persamaan satu di bawah dalam persamaan matematis sebagai berikut [14]:

$$T = X + e \qquad 1)$$

T : Kemampuan sebenarnya (*true theta*)
X : Estimasi kemampuan (*estimated theta*)
e : Kesalahan pengukuran (e*rror of measurement*)

Dalam kenyataan pengukuran bahwa true theta memang sulit diketahui, bisa dikatakan hanya Tuhan yang mengetahui kemampuan sebenarnya dari setiap penempuh tes, dan manusia hanya bisa mengukur estimasi kemampuan penempuh tesnya. Namun untuk kebutuhan penelitian khususnya penelitan dengan metode komparasi atau ingin membandingkan dua atau lebih model pengukuran, true theta dapat dikaitkan dengan nilai perolehan kemampuan dari setiap penempuh tes dari beberapa kali observasi pengukuran. Atau penentuan true theta biasanya menggunakan nilai perolehan kemampuan dari hasil pengukuran yang penempuhnya mengerjakan dengan serius, misalnya tes prestasi di sekolah atau tes seleksi.

2. Metodologi

Metodologi penelitian ini menggunakan pendekatan kuantitatif dan bersifat komparatif. Data penelitian adalah jawaban atau respon siswa SMP yang mengikuti UNBK atau UN dengan administrasi tes dengan CBT tahun 2019 di Provinsi DKI jakarta. Hal demikian karena data respon UN yang dimiliki Provinsi DKI jakarta bervariasi sehingga mudah untuk diamati dan dianalisis. Sementara itu, peneliti memfokuskan mata pelajaran matematika, karena merupakan *core competence* peneliti, selain juga mata pelajaran matematika adalah matapelajaran bersifat ilmu pasti dan sejalan dengan perkembangan teknologi. Sampel dipilih secara sistematik dari sejumlah populasi SMP yang mengikuti UNBK di Provinsi DKI.

Cara pemilihannya adalah dengan mengurutkan data dari skor terendah hingga skor tertinggi, selanjutnya dipilih secara acak hingga total sampel menjadi 341 data yaitu terdiri dari 19 data dari persentil kurang dari 25, 217 data dipilih dari persentil antara 25 hingga 75 dan 105 data sisanya dari persentil lebih dari 75. Terhadap 341 data penempuh tes terpilih tersebut akan diperoleh informasi response jawaban butir soal dan waktu respons butir soal. Setiap data response butir tersebut diolah ulang dengan pemodelan menggunakan administrasi model CAT sehingga akan diperoleh estimasi kemampuan penempuh tes, dan jumlah soal yang muncul setiap penempuh tesnya. Sementara itu informasi waktu response butir soal digunakan untuk mengukur lamanya tes dikaitkan dengan banyaknya jumlah butir yang muncul setiap penempuh tes dengan pemodelan administrasi CAT.

Sementara itu karena metodologi penelitian ini bersifat komparatif, maka perlu ditentukan nilai true theta sebagai nilai acuan yang diperoleh dari nilai hasil obeservasi ketika penempuh tes mengikuti tes matapelajaran matematika pra-UNBK dan ketika UNBK. Hasil kedua tes tersebut dihitung reratanya. Dan hasil rerata tersebut yang digunakan oleh peneliti sebagai nilai *true theta* atau nilai acuan untuk memperoleh kesalahan pengukuran dari hasil estimasi kemampuan dengan administrsi tes dengan CBT dan CAT.

Dengan data-data tersebut di atas dan dengan dilakukan pengolahan ulang dengan menerapkan pemodelan administrasi tes dengan model CAT maka akan dilakukan analisis dan pembahasan tentang efisiensi terhadap banyaknya jumlah butir soal yang muncul setiap penempuh tes apabila administrasi tesnya dengan model CBT dan CAT, serta akan dilakukan analisis dan pembahasan tentang efektifitas dan akurasi terhadap hasil estimasi kemampuan penempuh tes antara model CAT dan CBT.

Sebelum menjelaskan metodologi penelitian lebih jauh, penulis akan menjelaskan definisi dari efektifitas, efisiensi, dan akurasi sebagai batasan penelitian ini. Menurut Hidayat [10], efektifitas adalah suatu ukuran yang menyatakan seberapa jauh target (kuantitas, kualitas dan waktu) telah tercapai. Semakin besar presentase target yang dicapai, semakin

tinggi efektifitasnya. Menurut Mahmudi [13] efektivitas adalah sejauh mana unit yang dikeluarkan mampu mencapai tujuan yang ditetapkan.

Sementara itu efisiensi menurut Hasibuan [9] yang mengutip pernyataan [7], efisiensi adalah perbandingan yang terbaik antara *input* (masukan) dan *output*. Dengan kata lain, efisiensi adalah sesuatu yang kita kerjakan berkaitan dengan hasil yang optimal dengan tidak membuang banyak waktu dalam proses pengerjaannya. Sementara itu arti dari akurasi adalah seberapa dekat nilai hasil pengukuran (*estimated theta*) dengan nilai sebenarnya (*true theta*) atau nilai acuan. Semakin dekat dengan nilai acuan tersebut maka error pengukuran akan semakin kecil, hal ini menyatakan bahwa pengukurannya lebih akurat.

Jika model CAT dikaitkan dengan efisiensi, efektivitas dan akurasi seperti dijelaskan di atas maka dapat dihipotesiskan awal bahwa untuk efektifitas dipahami ketika pemilihan butir soal yang ditentukan dan dipilih sistem kepada penempuh tes sesuai dengan kemampuan penempuhnya, sehingga butir soal yang tidak sesuai dengan penempuh tesnya tidak akan dipilih oleh sistem administrasi model CAT. Oleh karena itu efektifitas ini akan menyebabkan banyaknya jumlah butir soal yang dimunculkan oleh sistem. Dari data, sudah diinformasikan bahwa administrasi tes dengan CBT setiap penempuh tes akan diberikan butir soal sebanyak 40 butir, namun dengan administrasi tes dengan model CAT, butir soal yang tidak sesuai dengan penempuh tesnya tidak akan dipilih oleh sistem maka butir soal yang dipilih dan dimunculkan oleh sistem setiap penempuh tesnya akan lebih sedikit dibandingkan dengan administrasi model CBT.

Selanjutnya kaitannya dengan efisiensi sejalan dengan hasil efektifitas di atas, bahwa banyaknya jumlah butir soal yang dipilih dan dimunculkan oleh sistem setiap penempuh tes dengan model CAT lebih sedikit maka akan menyebabkan akumulasi terhadap waktu response butir soal ketika administrasi tesnya dengan model CAT akan lebih menghemat waktu dibandingkan dengan model CBT. Sementara itu akurasi dipahami bahwa informasi *error* yang diperoleh antara *true theta* dengan estimasi kemampuan (estimated theta) penempuh tes. dengan kata lain administrasi tes dengan model CBT akan menghasilkan estimasi kemampuan penempuh tes begitu juga administrasi tes dengan model CAT. Selisih antara *true theta* dan estimasi kemampuan akan diperoleh *error* pengukuran. Keakuratan ini dapat teruji dengan menghitung MSE (*mean square error*) setiap model administrasinya. Dan administrasi tes dengan model CAT akan menghasilkan MSE lebih kecil dibandingkan dengan model CBT, oleh karena itu model CAT lebih akurat dibandingkan dengan model CBT.

Prosedur dengan administrasi tes model CAT sudah dijelaskan pada kerangka teoritis di atas. Oleh karena itu dalam penelitian ini menggunakan prosedur sebagai berikut: 1) Item bank, Bank soal yang dipakai adalah informasi statistik soal pada UNBK tahun 2019 dengan menggunakan mata pelajaran matematika dengan jumlah butir soal yaitu 40 butir; 2) Pemilihan soal pertama, dipilih oleh sistem secara random, dan dipilih soal yang

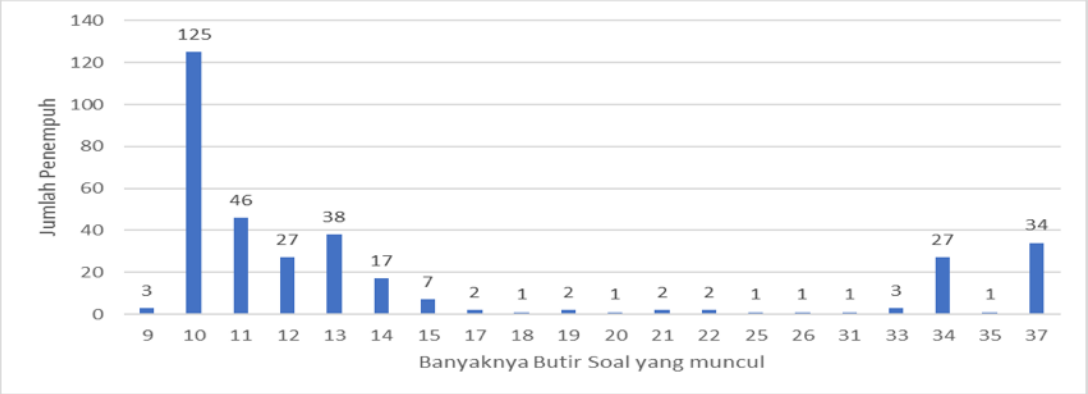
memiliki tingkat kesukaran sedang atau dengan nilai tingkat kesukarannya antara -0,5 hingga 0,5 dalam skala logit; 3) Perhitungan estimasi kemampuan, model CAT dalam estimasi kemampuan menggunakan pendekatan *maximum likelihood* dan dengan prosedur *step-sizing* dalam setiap penentuan kemampuan di tahapan awal; 4) Pemilihan butir soal berikut, model CAT dalam pemilihan butir soal berikut menggunakan maximum information dilakukan dengan memilih butir soal pada setiap tahap yang memiliki nilai b mendekati perkiraan θ saat itu; 5) menghentikan tes pada administrasi tes dengan model CAT menggunakan *variable length*, dan oleh karena itu administrasi CAT berhenti ketika *standard error measurement* sudah mencapai batasan kurang atau sama dengan 0,4.

3. Hasil dan Pembahasan

Dari hasil pengolahan penelitian, terdapat tiga hal yang dianalisis dan dibahas oleh peneliti, semuanya memiliki kaitan dengan tujuan dalam penelitian ini. Hal yang dibahas adalah 1) Analisis efisiensi model CAT dibandingkan dengan model CBT; 2) Analisis efektivitas antara model CAT dibandingkan dengan model CBT; 3) Analisis akurasi antara hasil yang menggunakan model CAT dibandingkan dengan model CBT.

Analisis dan bahasan efisiensi model CAT dibandingkan dengan CBT

Seperti telah dijelaskan di atas bahwa efisiensi dengan penerapan model CAT, yaitu ada kaitannya dengan banyaknya jumlah soal yang dimunculkan oleh sistem setiap penempuh. Dari hasil pengolahannya dapat ditunjukkan pada Grafik 1.



Sumber : Data Primer,diolah

Gambar 1. Perbandingan antara banyaknya jumlah soal yang muncul dengan Jumlah Penempuh tesnya dalam penerapan model CAT.

Pada Gambar 1, terlihat bahwa terdapat 125 penempuh tes yang yang hanya memperoleh 10 butir soal. Dan jumlah penempuh tes yang memperoleh 10 butir soal paling banyak

jumlah penempuhnya dibandingkan dengan jumlah butir soal yang lainnya. Dari Gambar 1, juga terlihat banyaknya butir soal yang dimunculkan oleh sistem antara 9 hingga 15 butir soal. Oleh karena itu dapat dikatakan bahwa sistem model CAT sudah dapat mengestimasi kemampuan penempuh tes hanya dengan 15 butir soal saja.

Dari Gambar 1, juga terlihat bahwa ada banyak penempuh tes yang memperoleh jumlah butir soal yang maksimal atau jumlahnya sama dengan ketika model CBT diterapkan. Dari grafik tesbut terdapat 34 penempuh tes memperoleh 37 butir soal dan 27 penempuh tes memperoleh 34 butir soal, hal ini karena penempuh tersebut pada awal butir soal dapat meresponse dengan jawaban selalu benar atau selalu salah, sehingga metode estimasi kemampuan penempuh tes dengan maksimum likelihood tidak mampu mengestimasi kemampuan penempuh tesnya, sehingga penentuan estimasi kemampuannya dengan metode step sizing, dimana apabila penempuh tes masih memiliki pola response jawaban dari awalnya selalu benar, maka estimasi kemampuannya dihitung dengan menambahkan nilai 0.5 dari nilai estimasi kemampuan penempuh tes pada butir soal sebelumnya. Dan apabila penempuh tes masih memiliki pola response jawaban dari awalnya selalu salah, maka estimasi kemampuannya dihitung dengan mengurangi nilai 0.5 dari nilai estimasi kemampuan penempuh tes pada butir soal sebelumnya.

Sumber: Data Primer, diolah

Gambar 2. Variasi jumlah soal dengan administrasi tes model CAT

Gambar 2 di atas dengan administrasi tes dengan CAT memperlihatkan bahwa dari 341 data sampel yang terpilih terdapat 77,13% data penempuh tes yang sudah cukup hanya menempuh 15 butir soal akan tetapi sistem sudah mampu mengestimasi kemampuan penempuh tes tersebut. Dari gambar 2 di atas juga diperlihatkan hanya 19,35% penempuh tes dengan jumlah soal yang ditempuh lebih dari 31 butir. Apabila dibandingkan dengan tes dengan model CBT dimana setiap penempuh tes akan memperoleh jumlah butir yang sama yaitu 40 butir soal, oleh karena itu implementasi administrasi tes dengan model CAT lebih efisien dibandingkan dengan model CBT.

Pembahasan berikutnya yaitu efisiensi dari sisi lamanya penempuh dalam menjawab seluruh butir soal yang ditempuh, maka model CAT akan lebih efisien dibandingkan model CBT. Hal ini dapat dibuktikan dengan banyaknya jumlah butir soal yang ditempuh

dikalikan dengan rerata waktu response butir soal. Dari 341 sample data dapat dihitung rerata waktu respons butir soalnya, yaitu 160,141 detik, sehingga dapat dikatakan bahwa penempuh tes dalam mengerjakan satu butir soal membutuhkan waktu 160,141 detik. Dengan waktu response setiap butir soal tersebut maka dapat dikatakan bahwa sebagian besar atau sebanyak 77,13% penempuh tes memiliki waktu dalam menyelesaikan 15 butir soal yang ditempuh yaitu 40 menit. Hal ini jika dibandingkan dengan model CBT yang setiap penempuh tes harus mengerjakan jumlah butir soal yang sama yaitu 40 butir soal dan dengan waktu 120 menit. Sehingga dapat disimpulkan bahwa model CAT lebih efisien waktu pengerjaannya 1/3 kali dibandingkan dengan model CBT.

Analisis dan bahasan efektivitas model CAT dibandingkan dengan CBT

Dalam pembahasan efektivitas seperti yang sudah dijelaskan pada metodologi di atas, yaitu pada pemilihan butir soal yang dimunculkan disesuaikan dengan kemampuan penempuh tesnya. Oleh karena itu, dalam penerapan administrasi dengan model CAT, maka sistem tidak akan memilih butir soal yang tidak sesuai dengan kemampuan penempuh tes. Hal ini jelas bahwa model tes dikatakan efektif apabila butir soal yang diberikan kepada penempuh tesnya disesuaikan dengan kemampauan penempuh tes tersebut. Jadi, karena butir soal disesuikan kemunculannya maka tidak semua butir soal diberikan kepada penempuh tes, sehingga setiap penempuh tes akan mendapat soal yang berbeda disesuaikan dengan kemampuan penempuhnya dan jumlah soal yang dikerjakan setiap penempuh tes tidak sama.

Gambar 3, adalah contoh hasil pengolahan dari tiga penempuh dengan administrasi tes model CAT dengan menggambarkan soal yang dipilih oleh sistem model CAT disesuaikan dengan kemampuan penempuhnya.

Gambar 4a **Gambar 4b** **Gambar 4c**

Sumber: Data Primer, diolah

Gambar 3. Hasil efektivitas model CAT

Gambar 4a **Gambar 4b** **Gambar 4c**

Sumber: Data Primer, diolah

Gambar 4 Hasil efektivitas model CBT

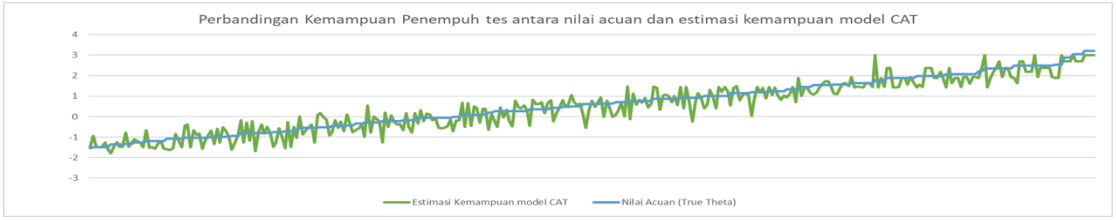
Gambar 3, menunjukkan bahwa butir soal yang dipilih oleh sistem disesuaikan dengna kemampuan penempuh tesnya. Gambar 3a, diperlihatkan contoh penempuh tes dengan persentil di atas 75, dari gambar tersebut kemampuan cenderung meningkat karena setiap diberikan butir soal penempuh tes merespon dengan jawaban benar. Namun diperlihatkan bahwa butir soal yang harusnya muncul tidak tersedia untuk penempuh tersebut, sehingga dari nomor 5 sampai nomor 13, tingkat kesukaran butir soal tidak tersedia dengan meningkatnya estimasi kemampuan penempuh tes.

Gambar 3b, diperlihatkan contoh penempuh tes dengan persentil antara 25 – 75, dari gambar tersebut diperlihatkan bahwa butir soal yang dipilih oleh sistem sangat sesuai dengan kemampuan penempuh tesnya. Sehingga terlihat jelas keefektifan dari pemilihan butir soal oleh sistem model CAT. Gambar 3c, diperlihatkan contoh penempuh tes dengan persentil di bawah 25, dari gambar tersebut kemampuan penempuh tesnya menurun dan mulai butir soal di nomor 8 dan seterusnya kemampuannya meningkat, atau response jawaban penempuh tesnya benar. Dan tingkat kesukaran butir soal yang diberikan dari nomor 8 – 13 juga meningkat sesuai dengan meningkatnya kemampuan penempuh tesnya.

Sementara itu apabila kita bandingkan dengan administrasi tes dengan model CBT pada gambar 4 diatas terlihat bahwa ketiga gambar (4a, 4b, dan 4c) butir soal yang dipilih oleh sistem tidak menyesuaikan dengan kemampuan penempuh tesnya. Oleh karena itu berdasarkan analisis di atas jelas bahwa model CAT lebih efektif dibandingkan dengan model CBT, dan akan lebih efektif manakala item bank yang dimiliki oleh model CAT tersedia menyebar untuk setiap skala kemampuan penempuhnya. Jika *item bank* yang dimiliki tidak menyebar untuk setiap kemampuan penempuh maka penempuh tes dengan model CAT akan dipilihkan butir soal oleh sistem jauh dari kemampuan penempuh tesnya, hal ini jelas terlihat dari gambar 3a.

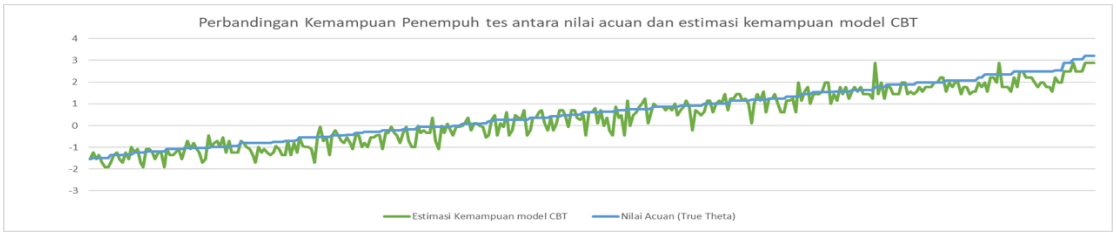
Analisis dan pembahasan akurasi model CAT dibandingkan dengan CBT

Dalam pembahasan akurasi seperti yang sudah dijelaskan di atas bahwa model dikatakan akurat apabila memiliki error pengukuran yang kecil. Oleh karena itu dalam analisis kali ini, akan dibandingkan nilai error pengukuran dengan uji MSE (Mean Square Error) antara model CBT dan model CAT. Untuk memperoleh nilai error pengukuran seperti persamaan 2 di atas maka nilai acuan (true theta) dalam penelitian ini menggunkaan nilai acuan berupa nilai rerata setiap penempuh tes dari pelaksanaan pra-UN dan UN. Sehingga error pengukuran diperoleh dengan menghitung selisih antara nilai acuan dan hasil estimasi kemampuan penempuh tesnya.



Sumber: Data Primer, diolah

Gambar 5. Perbandingan antara nilai acuan dengan estimasi kemampuan penempuh dengan model CAT



Sumber: Data Primer, diolah

Gambar 6. Perbandingan antara nilai acuan dengan estimasi kemampuan penempuh dengan model CBT

Dari gambar 5 dan gambar 6 di atas terlihat bahwa estimasi kemampuan penempuh tes hasilnya mendekati nilai acuan, dan jarak perbedaan dengan nilai acuan tersebut itu yang di sebut error. Apabila dilihat memang tidak ada perbedaan yang cukup signifikan antara nilai acuan dan hasil estimasi kemampuan penempuh tes baik yang menggunakan model CBT maupun model CAT.

Dari hasil uji MSE dapat dibuktikan bahwa dengan menggunakan model CAT nilai MSE yaitu sebesar 0.158411619 sedangkan dengan menggunakan model CBT nilai MSE lebih besar sedikit yaitu 0.160009318. sehingga dapat disimpulkan walaupun perbedaan nilai MSE antara pengukuran estimasi kemampuan penempuh tes dengan administrasi model CBT dan CAT sangat kecil, namun dapat disampaikan bahwa model CAT lebih akurat dibandingkan dengan model CBT.

4. Kesimpulan

Dari Penelitian ini menghasilkan tiga simpulan. Pertama, Model CAT lebih efisien dibandingkan dengan model CBT. Hal ini dapat dipahami dari hasil kajian bahwa banyaknya jumlah butir soal yang dikerjakan oleh penempuh tes dengan model CAT lebih sedikit dibandingkan dengan model CBT, sehingga waktu yang dibutuhkan lebih hemat 1/3 kali dibandingkan dengan model CBT. Model CAT dapat memiliki jumlah butir soal yang sama dengan model CBT apabila bank soal yang dimiliki tidak tersebar dalam variasi skala kemampuan penempuh. Selain itu, konsistensi jawaban peserta akan mempengaruhi efisiensi model CAT. Kedua, Model CAT lebih efektif dibandingkan dengan model CBT. Hal ini dapat dipahami dari hasil kajian bahwa butir soal yang dipilih oleh sistem dengan model CAT dan ditempuh oleh penempuh tes disesuaikan dengan kemampuan penempuhnya, sehingga informasi yang diperoleh dari estimasi kemampuan penempuh tesnya lebih optimal dibandingkan dengan model CBT. Ketiga, Hasil uji MSE antara model CAT dan CBT menghasilkan nilai yang sama, walaupun model CAT lebih kecil dibandingkan dengan model CBT. Hal ini dapat dipahami dari hasil kajian bahwa model CAT memiliki keakuratan yang lebih dibandingkan dengan CBT.

Daftar Pustaka

[1] *Apa yang Harus Dilakukan Guru dalam Mengembangkan Silabus,* (2007). (http://rbaryans.wordpress.com/2007/07/27/, diakses 28 Juli 2010).

[2] Blais, J. & Raiche, G., 2002, Some Features of the sampling distribution of the ability estimate in computerized adaptive testing according to two stopping rules. *Makalah, 11th International Objective Measurement Workshop,* New Orleans, April 2002.

[3] Bunderson, C.V., Inouye, D.K., Olsen, J.B., (1989), *The Four Generations of Computerized Educational Measurement*. Dalam Robert L. Linn. Educational Measurement. 3rd ed. New York: American Council on Education & Macmillan Publishing Company.

[4] Departemen Pendidikan Nasional, (2009), *Peraturan Menteri Pendidikan Nasional No. 75 Tahun 2009 Tentang Ujian Nasional SMP/MTs, SMPLB, SMA/MA, SMALB dan SMK Tahun Pelajaran 2009/2010.*

[5] Departemen Pendidikan Nasional, (2009), *Prosedur Operasi Standar (POS) UN SMP, MTs, SMPLB, SMA, MA, SMALB dan SMK Tahun Pelajaran 2009/2010.*

[6] Embretson, S.E., & Reise, S.P., (2000), *Item Response Theory for Psychologist*, Lawrence Erlbaum Associates, Inc., New Jersey.

[7] Emerson, H., (1986), *Sistem Birokrasi Pemerintah*. CV. Mas Gunung Agung, Jakarta.

[8] Hambleton, R.K., Swaminathan, H., Rogers, H.J., (1991), Fundamental of Item Response Theory, *Sage Publications, Inc*., *California*, **2**.

[9] Hasibuan, S.P., (1984), *Manajemen Dasar dan Suatu Pengantar*, Haji Masagung, Jakarta.

[10] Hidayat, (1986), *Teori Efektifitas Dalam Kinerja Karyawan*, Gajah Mada University Press, Yogyakarta.

[11] Hornke, L.F., (2000), Item Response Times in Computerized Adaptive Testing. *Psicológica*. **21**, 175-178.

[12] Leung, C., Chang, H., Hau, K., (2005), Computerized Adaptive Testing: A Mixture Item Selection Approach for Constrained Situations, *British Journal of Mathematical & Statistical Psychology*, *Proquest Psychology Journals*, **58**, 239.

[13] Mahmudi, (2010), *Manajemen Kinerja Sektor Publik*, Penerbit UUP STIM YKPN, Yogakarta.

[14] Sumarna Surapranata, (2009), *Analisis Validitas, Reliabilitas, dan Interpretasi Hasil Tes*, PT. Remaja Rosda Karya, Bandung.

[15] Thissen, D., & Mislevy, R.J., (1990), *Testing Algorithms*, Dalam H. Wainer, N.J. Dorans, R. Flugher, & B.F. Green, Computerized Adaptive Testing: a Primer, Lawrance Erlbaum Associates, Publishers, New Jersey.

[16] Wainer, H., 1990, *Introduction and History*. Dalam H. Wainer, N.J. Dorans, R. Flugher, & B.F. Green. Computerized Adaptive Testing: a Primer, Lawrance Erlbaum Associates, Publishers, New Jersey.