Jurnal Psikologi *Vol. 17 No. 1 April 2018, 1-17*

ANALISIS RESPON BUTIR PADA TES BAKAT SKOLASTIK

Farida Agus Setiawati, Rita Eka Izzaty, Veny Hidayat

Jurusan Psikologi, Fakultas Ilmu Pendidikan, Universitas Negeri Yogyakarta
Jl. Colombo No.1, Caturtunggal, Depok, Caturtunggal, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55281

farida_as@uny.ac.id

Abstract

This study aims to analyze the characteristics of the Scholastic Aptitude Test (SAT), consisting of both verbal and numerical subtests. We used a descriptive quantitative approach by describing the characteristics of SAT based on the degree of item difficulty, item discrimination index, pseudoguessing index, test information function and standard error measurement. The data are responses of the SAT instrument, collected from 1,047 subjects in Yogyakarta using the documentation technique. Data were then analyzed by Item Response Theory (IRT) approach with the help of the BILOG program on all logistic parameter models, preceded by identifying item suitability with the model. Analysis concludes that: verbal subtest tends to compliment the 2-PL and 3-PL model, meanwhile, numerical subtest only fit the 2-PL model. Majority items of SAT have a good characteristic on index of item difficulty, item discrimination, and pseudoguessing, and based of test information function, SAT is accurate to be used in the 1-PL, 2-PL, and 3-PL IRT models for all level of ability.

Keywords: item analysis; item response theory; scholastic aptitude test

Abstrak

Penelitian ini bertujuan untuk menganalisis karakteristik instrumen tes bakat skolastik yang terdiri dari subtes verbal dan subtes numerikal. Penelitian ini menggunakan pendekatan kuantitatif deskriptif dengan cara mendeskripsikan karakteristik instrumen *Scholastic Aptitude Test* (SAT) ditinjau dari tingkat kesukaran butir, indeks daya beda butir, indeks tebakan semu butir, fungsi informasi tes, dan kesalahan pengukuran tes. Data penelitian yang dikumpulkan melalui teknik dokumentasi berupa data respon dari instrumen SAT yang dikerjakan oleh 1.047 subjek di Daerah Istimewa Yogyakarta (DIY). Data yang dikumpulkan selanjutnya dianalisis dengan pendekatan *Item Response Theory* (IRT) dengan bantuan program BILOG pada semua model parameter logistik yang didahului dengan identifikasi kecocokan butir terhadap modelnya. Berdasarkan hasil analisis dapat disimpulkan bahwa subtes verbal cenderung cocok dengan model 2-PL dan 3-PL, sedangkan subtes numerikal hanya cocok dengan model 2-PL. Sebagian besar butir-butir SAT memiliki indeks kesukaran butir, indeks daya beda butir, dan indeks tebakan semu yang baik, serta berdasarkan fungsi informasi tes, instrumen SAT akurat digunakan pada model IRT 1-PL, 2-PL, dan 3-PL untuk semua level kemampuan.

Kata kunci: analisis butir; teori respon butir; tes bakat skolastik

PENDAHULUAN

Setiap orang memiliki kemampuan yang berbeda dengan orang lain. Salah satu bentuk perbedaan individu tersebut dapat diketahui dari bakatnya yang berbeda-beda. Bakat merupakan salah satu jenis perbedaan individu yang unik dan khas pada masing-masing individu (Nazimuddin, 2015; Jung, Ryman, Vakhtin, Carraso, Wertz, & Flores, 2014; Robinson, 2012; Ehrman, Leaver, & Oxford, 2003). Perbedaan individu ini secara umum disebabkan oleh dua faktor, yakni faktor bawaan dan faktor lingkungan. Perbedaan individu atas dasar faktor bawaan atau hereditas disebabkan adanya variasi yang berasal dari kromosom dan gen, sedangkan perbedaan individu yang terjadi akibat faktor lingkungan di antaranya disebabkan oleh lingkungan sosial psikologis di mana anak lahir, pola asuh orang tua, status sosial ekonomi keluarga, serta interaksi lingkungan antara anggota keluarga dengan lingkungan sekitar maupun

lingkungan sekolah (Sauce & Matzel, 2013; Van der Aa, Bartels, te Velde, Boomsma, de Geus, & Brug, 2012).

Sebagian besar orang seringkali menggunakan istilah bakat secara bersamaan dengan kecerdasan dan prestasi, padahal ketiga istilah tersebut sebenarnya memiliki makna yang berbeda. Bakat didefinisikan sebagai karakteristik yang berhubungan dengan kapasitas seseorang untuk mendapatkan pengetahuan atau keterampilan dari pelatihan atau pengalaman (Mankar & Chavan, 2013; Salkind & Rasmussen, 2007; Kubiszyn & Borich, 2003). Bakat berbeda dengan inteligensi, karena inteligensi lebih menekankan pada kemampuan seseorang yang sifatnya umum. Sedangkan bakat juga berbeda dengan prestasi yang diartikan sebagai kemampuan yang didapat dari hasil belajar.

Dalam dunia pendidikan, bakat terbukti sebagai prediktor yang baik dalam meramalkan keberhasilan individu terutama dalam bidang akademik (Curabay, 2016; Pyari, Mishra, & Dua, 2016; Ahnaldi, 2015; Mankar & Chavan, 2013; Salkind & Rasmussen, 2007; Oyetunde, 2007). Oleh sebab itu, kegiatan penelusuran bakat di berbagai jenjang pendidikan banyak dilakukan untuk pengembangan karir siswa.

Tes yang digunakan untuk mengukur bakat seseorang disebut dengan tes bakat atau *aptitude test*. Beberapa ahli sepakat bahwa tes bakat memiliki ciri sebagai berikut: (1) tes bakat mencakup area yang lebih luas bila dibandingkan tes prestasi, (2) tes bakat kurang berkaitan dengan pelajaran sekolah dan kurang terikat dengan budaya, (3) tes bakat memiliki indeks heritabilitas yang lebih tinggi bila dibandingkan tes prestasi, (4) tes bakat mengukur kemampuan khusus yang sudah terakumulasi, sedangkan tes prestasi cenderung mengukur hasil belajar, serta (5) tes bakat lebih valid untuk mengukur kinerja seseorang di masa depan (Hashmi, Zeeshan, Saleem, & Akbar, 2012; Sattler, 1998; Mehrens & Lehkarn, 1987).

Tes bakat biasa digunakan untuk mengukur berbagai jenis bakat (*multiple aptitude test*), maka terdapat beberapa contoh tes bakat yang sangat populer digunakan, yakni *Differential Aptitude Test* (DAT), *Flanagan Aptitude Classification Test* (FACT), *General Aptitude Test Battery* (GATB) (Awasthy & Kaur, 2009; Ereme, 2005). Dari ketiga contoh tes bakat multipel tersebut, DAT adalah tes bakat yang paling banyak diaplikasikan dalam dunia pendidikan (Mahakud, 2013). DAT dibuat oleh Bennett, Seashore, & Wesman pada tahun 1948. Instrumen DAT terdapat tujuh subtes, yakni *verbal reasoning (vr), numerical ability (na), abstract reasoning (ar), clerical speed and accuracy (csa), mechanical reasoning (mr), space relations (sr),* serta *language usage* yang terdiri atas *spelling and sentences*. Instrumen dapat digunakan secara bersamaan maupun terpisah tiap subtesnya (Mankar & Chavan, 2013; Vosloo, Coetzee, & Claassen, 2000).

Dari ketujuh subtes DAT, kombinasi subtes *verbal reasoning* dan subtes *numerical ability* merupakan subtes bakat yang dikenal dengan sebutan tes bakat skolastik (*Scholastic Aptitude Test* atau SAT). *Verbal reasoning* adalah kemampuan kognitif secara umum untuk memperoleh informasi sedangkan *numerical ability* adalah kemampuan untuk memahami hubungan angka. Kedua kemampuan tersebut merupakan kemampuan yang sangat dibutuhkan pada proses belajar. Tes ini terbukti signifikan sebagai prediktor prestasi belajar, karena merupakan kemampuan dasar siswa yang berperan besar untuk memperoleh pelajaran di sekolah (DAT for Selection General Abilites Battery, 2013; Oyetunde, 2007; Agronow & Studley, 2007; Grove,Wasserman, & Grodner., 2006; Kobrin & Michel, 2006; Geiser & Studley, 2002; Stumpf & Stanley, 2002).

Suatu alat tes dikatakan baik apabila memiliki karakteristik psikometrik yang baik. Karakteristik psikometrik ini merupakan suatu atribut yang terkait dengan tes

psikologi (Furr & Bacharach, 2008). Terdapat dua pendekatan tes yang diacu untuk mengetahui karakteristik psikometrik suatu instrumen, yaitu teori tes klasik atau *classical test theory* (CTT) dan teori respon butir atau *item response theory* (IRT) (Awopeju & Afolabi, 2016; Zoghi & Valipour, 2014; Guler, Uyanik, & Teker, 2013; Abedalaziz & Leng, 2013; Thorpe & Favia, 2012; DeMars, 2010; Adedoyin, Nenty, & Chilisa, 2008; Crocker & Algina, 2008; Hambleton, Swaminathan, & Rogers, 1991; Hambleton & Swaminathan, 1985). Teori respon butir merupakan teori tes yang berisi seperangkat variabel laten yang dirancang khusus untuk memodelkan interaksi antara *trait* atau kemampuan dari subjek uji terhadap karakteristik butir maupun pola respons butirnya (Brzezinska, 2016; Fox, 2007). Teori ini dibangun berdasarkan dua postulat, yaitu: (1) prestasi subjek pada suatu tes dapat diprediksikan dengan seperangkat faktor yang disebut kemampuan laten (*latent traits*), serta (2) hubungan antara prestasi uji pada suatu butir tes dan perangkat kemampuan yang mendasarinya sesuai dengan grafik fungsi naik monoton tertentu yang disebut *Item Characteristic Curve* (Zoghi & Valipour, 2014; Olufemi, 2013; Abedalaziz & Leng, 2013; Adedoyin, dkk., 2008; Hambleton, 1990; Hambleton, Swaminathan & Rogers, 1991). Oleh sebab itu, analisis dalam teori respon butir menggunakan konsep matematika yang menyatakan bahwa probabilitas subjek menjawab suatu butir dengan benar bergantung pada kemampuan subjek dan karakteristik butir (Lalor, Wu, & Yu, 2016; Erguven, 2014).

Pada awalnya, kemunculan teori respon butir disebabkan karena banyaknya kelemahan yang ada pada teori tes klasik, yaitu: (1) tingkat kesukaran dan daya beda butir soal tergantung pada kelompok peserta yang mengerjakannya (*sample dependent*); (2) penggunaan metode dan teknik untuk desain dan analisis tes dengan memperbandingkan kemampuan siswa pada pembagian kelompok atas, tengah, dan bawah; (3) konsep reliabilitas skor didefinisikan dari istilah tes paralel; (4) tidak ada dasar teori untuk menentukan bagaimana peserta memperoleh tes yang sesuai dengan kemampuan peserta yang bersangkutan; dan (5) *Standard Error Measurement* (*SEM*) berlaku pada seluruh peserta tes (Magno, 2009; Adedoyin, dkk., 2008; Fan, 1998; Hambleton, 1989; Hambleton & Swaminathan, 1985). Pada analisis menggunakan teori respon butir, kelemahan-kelemahan tersebut dapat diatasi. Hal ini disebabkan karena dalam teori respons butir: (1) sampel bersifat invarians yang artinya karakteristik butir maupun tes tidak tergantung pada tingkat kemampuan sampel; (2) estimasi-estimasi yang tidak bias bisa diperoleh meskipun sampelnya tidak representatif*;* (3); skor tes memiliki arti manakala dibandingkan dengan karakteristik item-item; (4) skala yang bersifat interval dicapai dengan menggunakan model pengukuran yang lebih logis; dan (5) *Standar Error Measurement* tidak memiliki nilai yang berbeda-beda antar skor (atau pola-pola respon) tetapi bersifat umum antar populasi (Yang & Kao, 2014; Olufemi, 2013; Anderson & Morgan, 2008; Embretson & Reise, 2000). Selain itu, Fan (1998) juga menambahkan bahwa analisis yang didasarkan pada teori respon butir lebih menekankan pada level informasi butir tes, sedangkan pada teori tes klasik analisis lebih menekankan pada level informasi perangkat tes. Jadi berdasarkan paparan tersebut, dapat disimpulkan bahwa analisis menggunakan teori respon butir akan memberikan hasil yang lebih cermat atau *powerfull* (Pollard, Dixon, Dieppe, & Johnston, 2009).

Teori respon butir menyediakan model analisis butir yang beragam tergantung jenis data yang akan dianalisis, misalnya: (1) untuk data dikotomi dapat dianalisis menggunakan model *latent linear*, model *perfect scale*, model *latent distance*, model *parameter normal ogive*, maupun model parameter *logistic*; (2) untuk data *multicategory,* dapat dilakukan dengan menggunakan model *nominal response*,

model *graded response*, maupun model *partial credit*; serta (3) untuk data *continuous*, dapat dianalisis menggunakan *continuous response* (de Ayala, 2009; van der Linden & Hambleton, 1996). Meskipun demikian, untuk data dikotomi, model analisis yang sering digunakan adalah model parameter logistik. Hal ini disebabkan karena hitungan matematis yang lebih sederhana pada penggunaan model distribusi logistik dibandingkan distribusi normal (Chung, 2005).

Analisis interpretasi model logistik memiliki model yang sama dengan model normal, namun ditambahkan dengan faktor penskalaan D sebesar 1,702 pada persamaannya (Haley, 1952). Berdasarkan distribusi logistik ini, model IRT diklasifikasikan berdasarkan jumlah parameter butirnya dibagi menjadi empat, yaitu *one parameter logistic model* (1-PL), *two parameter logistic model* (2-PL), *three parameter logistic model* (3-PL), dan *four parameter logistic model* (4-PL) (Magis, 2013; Hambleton, 1989; Hambleton & Swaminathan, 1985). Namun, dari keempat model tersebut model 1-PL, 2-PL, dan 3-PL merupakan model yang paling umum digunakan (Hambleton & Swaminathan, 1985). Model 4-PL jarang digunakan karena kurangnya konsensus terkait kegunaannya dan adanya kesulitan teknis dalam memperkirakan asimtot secara akurat (Loken & Rulison, 2010). Pada model 1-PL, hanya memuat satu parameter butir yaitu tingkat kesukaran butir; model 2-PL memuat dua parameter yaitu tingkat kesukaran dan indeks daya beda butir; sedangkan model 3-PL di samping memuat tingkat kesukaran dan indeks daya beda butir, juga memuat parameter *pseudo-guessing* atau tebakan semu (Philip & Ojo, 2017; Thorpe & Favia, 2012; DeMars, 2010; Crocker & Algina, 2008; Courville, 2004; Baker, 2001; Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991).

Analsis respon butir pada tes bakat skolastik telah dilakukan di beberapa penelitian. Abed, Al-Absi, & Shindi (2016) melakukan analisis IRT pada subtes numerikal yang melibatkan 504 mahasiswa dari 8 universitas di Jordan. Hasil penelitiannya menunjukkan karakteristik psikometrik butir dilihat dari tingkat kesulitannnya tergolong baik karena berada dalam rentang -2 sampai +2; selain itu indeks daya beda butir juga berfungsi baik karena berada dalam rentang 0 sampai +2. Lord (1968) melakukan penelitian menggunakan data subtes verbal untuk menentukan besarnya jumlah sampel dan subjek uji untuk menghasilkan karakteristik parameter butir 3-PL yang memadai. Penelitian tersebut akhirnya memberikan kesimpulan bahwa setidaknya diperlukan 50 butir dan 1000 subjek agar menghasilkan karakteristik tingkat kesukaran butir, indeks daya beda, dan indeks tebakan semu yang baik.

SAT banyak digunakan di bidang pendidikan terutama untuk penelusuran atau pemilihan program di SMA dan seleksi masuk perguruan tinggi, namun upaya untuk melakukan analisis IRT pada alat ukur skolastik belum penulis temukan. Parameter butir yang dihasilkan dari analisis IRT ini akan bemanfaat untuk pengembangan alat ukur ini lebih lanjut, misalnya dalam bentuk bank soal, pembuatan soal paralel, *test equiting* dan *computer adaptive testing*. Dengan demikian, urgensi penelitian untuk mengetahui parameter butir SAT perlu dilakukan mengingat tes ini banyak digunakan untuk pengukuran kemampuan atau potensi seseorang di sekolah menengah maupun perguruan tinggi. Artikel ini bertujuan menggambarkan hasil analisis fit model dan parameter butir pada butir-butir tes skolastik dengan menggunakan satu, dua dan tiga parameter butir. Parameter butir merupakan atribut psikologis yang melekat pada butir, yaitu indeks kesukaran soal, indeks daya beda, dan tebakan semu dan informasi butir atau tes.

METODE

Data Penelitian

Penelitian ini menggunakan data sekunder, dimana data dalam penelitian didapatkan melalui teknik dokumentasi. Data yang terkumpul merupakan hasil dari kumpulan data hasil tes skolastik yang terdapat di biro *testing* yang ada di Universitas Negeri Yogyakarta yang dikerjakan oleh 1.047 siswa remaja di Daerah Istimewa Yogyakarta (DIY). Semua data yang terkumpul selanjutnya dilakukan analisis parameter butir.

Instrumen Pengukuran

Data yang didapatkan ini merupakan data respon dari instrumen skolastik yang terdiri atas subtes kemampuan verbal dan numerikal. Kedua sub tes ini merupakan bagian dari tes DAT. Tes ini berbentuk *multiple choice* dengan lima option alternatif pilihan jawaban. Subjek diminta untuk memilih salah satu option yang merupakan jawaban dari soal yang disajikan. Adapun spesifikasi instrumen tes bakat skolastik disajikan dalam Tabel 1.

Tabel 1.
Spesifikasi Instrumen Tes Bakat Skolastik

Subtes	Jumlah Butir	Alokasi Waktu (menit)
Verbal	50	30
Numerikal	40	30

Analisis Data

Data pada tes bakat skolastik dilakukan secara terpisah pada subtes verbal dan numerikal. Hal ini disebabkan kedua data tersebut memiliki dimensi instrumen yang berbeda. Kedua data sebelumnya dianalisis dimensinya melalui analisis faktor eksploratori. Hasilnya menunjukkan baik subtes verbal maupun numerikal memiliki kecenderungan pada satu dimensi, terlihat dari *scree plot* kedua subtes. Hasil tersebut juga dapat digunakan untuk memenuhi prasyarat IRT terkait dengan unidimensi dan independendensi lokal. Asumsi terpenuhinya unidimensionalitas otomatis membuktikan asumsi independensi lokal (Retnawati, 2014).

Data selanjutnya dianalisis menggunakan IRT untuk melihat karakteristik parameter butir dari instrumen SAT, yang meliputi tingkat kesukaran (*b*), indeks daya beda (*a*), dan indeks tebakan semu (*c*). Akan tetapi, sebelum analisis karakteristik parameter butir dilakukan, terlebih dahulu diidentifikasi kecocokan parameter butir berdasarkan statistik *chi-square*. Kecocokan butir-butir dalam instrumen SAT ini dianalisis menggunakan statistik Chi-kuadrat dengan cara taraf signifikansi (α) = 0,01 yang merupakan nilai *default* dari program BILOG (Mislevy & Bock, 1990). Analisis model (1-PL, 2-PL, dan 3-PL) juga dilakukan dengan bantuan program BILOG. Estimasi parameter kemampuan dilakukan menggunakan formula yang disajikan oleh Hambleton & Swaminathan (1985) sebagai berikut:

Model 1-PL $$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1+e^{D(\theta - b_i)}} \quad (1)$$

Model 2-PL $$P_i(\theta) = \frac{e^{Dai(\theta - b_i)}}{1+e^{Dai(\theta - b_i)}} \quad (2)$$

Model 3-PL
$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Dai(\theta - b_i)}}{1+e^{Dai(\theta - b_i)}} \quad (3)$$

Keterangan :
θ : Tingkat kemampuan (ability) peserta tes
$P_i(\theta)$: Probabilitas peserta tes yang memiliki kemampuan θ dapat menjawab butir dengan benar
a_i : Indeks daya beda butir ke-*i*
b_i : Tingkat kesukaran butir ke-*i*
c_i : Indeks tebakan semu butir ke-*i*
e : bilangan natural yang nilainya mendekati 2,718
D : faktor penskalaan yang harganya 1,702 (sering dibulatkan menjadi 1,7)

Selain melihat karakteristik parameter butir, akan dihitung pula *Test Information Function* (TIF) maupun *Standard Error* (SE) yang

merupakan representasi koefisien reliabilitas dan kesalahan pengukuran dalam teori tes modern menggunakan formula berikut (Hambleton, Swaminathan, & Rogers, 1991; Hambleton & Swaminathan, 1985):

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \qquad (4)$$

$$SE(\theta) = \frac{1}{\sqrt{I_i(\theta)}} \qquad (5)$$

Keterangan :
$I(\theta)$: fungsi informasi tes
$I_i(\theta)$: fungsi informasi butir
$SE(\theta)$: kesalahan pengukuran

HASIL DAN PEMBAHASAN

Instrumen SAT dianalisis secara terpisah pada setiap subtesnya. Adapun karakteristik yang dimaksud meliputi kecocokan model, tingkat kesukaran butir (b), indeks daya beda butir (a), indeks tebakan semu (c), fungsi informasi tes, maupun kesalahan pengukuran.

Kecocokan Model

Kecocokan butir-butir dalam instrumen SAT dianalisis menggunakan statistik chi-kuadrat. Tabel 2 menyajikan butir-butir SAT yang teridentifikasi cocok dengan model (*fit*) maupun butir-butir yang teridentifikasi tidak cocok dengan model (*misfit*). Butir-butir yang telah *fit* dengan modelnya menunjukkan bahwa model IRT tersebut memberikan tanggapan (respon) yang paling baik terhadap butir-butir yang diuji (Embretson & Reise, 2000). Butir dikatakan memiliki *fit model* jika probabilitas $\chi^2 \geq 0{,}01$ atau taraf signifikansi (α) lebih dari 0,01. Berdasarkan Tabel 2 terlihat bahwa butir-butir dalam SAT yang memiliki banyak butir yang *misfit* adalah model 1-PL. Untuk model 2-PL dan 3-PL memiliki jumlah butir yang *fit* lebih banyak dibanding model 1-PL. Pada subtes verbal memberikan jumlah butir *fit* pada 2 dan 3 PL yang sama banyak; sedangkan untuk subtes numerikal, butir *fit* paling banyak terdapat pada model 2-PL. Berdasarkan hasil analisis sebagaimana yang disajikan dalam Tabel 2 dapat disimpulkan bahwa butir-butir dalam subtes verbal cenderung cocok dengan model 2-PL dan 3-PL, sedangkan butir-butir dalam subtes numerikal hanya cocok dengan model 2-PL.

Tabel 2 menunjukkan banyaknya butir yang fit pada model 2 PL lebih banyak dari 1 PL, dan pada model 3 PL tidak demikian. Hal ini menunjukkan bahwa model yang dipilih akan mempengaruhi fit tidaknya butir tersebut. Model 1-PL mendapatkan hasil butir-butir fit yang lebih sedikit dibanging model 2-PL dan 3-PL, hal ini terjadi karena pada model 1 PL, indeks daya beda pada butir-butir dibatasi atau tidak diakomadasi, dengan demikian variasinya tidak dijadikan pertimbangan sebagai parameter butir sebagaimana ada pada model 2-PL dan 3-PL. Dengan keterbatasan ini butir-butir akan menjadi tidak cukup kuat untuk menjadi butir yang fit (Crocker & Algina, 2008).

Ringkasan Karakteristik Parameter Butir

Tabel 3 menyajikan ringkasan statistik deskriptif karakteristik parameter butir instrumen SAT pada semua model IRT. Kriteria tingkat kesukaran butir yang baik berada pada -2 sampai +2. Kriteria indeks daya beda butir yang baik yakni berada antara 0 hingga +2. Sedangkan kriteria indeks tebakan semu pada tes pilihan ganda terletak di sekitar satu berbanding banyaknya pilihan jawaban. Jumlah distraktor pada subtes skolastik adalah 5, maka nilai maksimum c_i yang baik adalah $^1/_5$ atau 0.2. Berdasarkan tabel 3 terlihat bahwa untuk kedua subtes yang diteliti, rata-rata tingkat kesukaran butir, indeks daya beda, dan indeks tebakan semu tegolong baik, meskipun apabila dilihat dari nilai maksimum dan nilai minimum masih ada butir yang miliki parameter butir melibihi kriteria. Jika dibandingkan ketiga model, parameter terbaik didapatkan dari hasil analisis model IRT 3-PL. Sementara itu, apabila ditinjau dari parameter indeks tebakan semu, butir-butir subtes numerikal memiliki indeks tebakan semu yang lebih tinggi dibandingkan dengan subtes verbal.

Tabel 2.
Ringkasan Kecocokan Butir

Subtes	Model	Keterangan Butir	Nomor Butir	Jumlah	Persentase
Verbal	1-PL	*Fit*	1, 2, 3, 6, 7, 8, 9, 12, 22, 25, 30, 31, 32, 33, 34, 36, 37, 39, 43, 45, 47, 48, 49	23	46,00 %
		Misfit	4, 5, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 23, 24, 26, 27, 28, 29, 35, 38, 40, 41, 42, 44, 46, 50	27	54,00 %
	2-PL	*Fit*	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 44, 45, 46, 48, 49, 50	47	94,00 %
		Misfit	23, 41, 47	3	6,00 %
	3-PL*	*Fit*	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 41, 42, 43, 44, 45, 47, 48, 49, 50	47	95,92 %
		Misfit	22, 40, 46	2	4,08 %
Numerikal	1-PL	*Fit*	1, 2, 3, 4, 10, 11, 12, 17, 18, 20, 24, 25, 26, 29, 31, 32, 39	17	42,50 %
		Misfit	5, 6, 7, 8, 9, 13, 14, 15, 16, 19, 21, 22, 23, 27, 28, 30, 33, 34, 35, 36, 37, 38, 40	23	57,50 %
	2-PL	*Fit*	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 22, 24, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40	36	90,00 %
		Misfit	19, 21, 23, 28	4	10,00 %
	3-PL	*Fit*	1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 21, 22, 23, 24, 25, 26, 27, 30, 31, 32, 33, 34, 35, 36, 37, 39	32	80,00 %
		Misfit	6, 12, 19, 20, 28, 29, 38, 40	8	20,00 %

Tabel 3.
Ringkasan Statistik Deskriptif Karakteristik Parameter Butir

Subtes	Statistik	Model					
		1-PL	2-PL		3-PL		
		b	b	a	b	a	c
Verbal	Minimum	-2,878	-4,133	0,117	-1,324	0,272	0,025
	Maksimum	3,709	8,617	1,446	5,969	2,204	0,303
	Rerata	0,000	0,352	0,777	0,551	1,108	0,160
Numerikal	Minimum	-2,310	-5,456	0,272	-4,708	0,127	0,017
	Maksimum	2,516	7,307	1,655	2,580	2,200	0,264
	Rerata	0,000	-0,211	0,883	0,255	1,157	0,164

Tingkat Kesukaran Butir

Tabel 4 menyajikan ringkasan tingkat kesukaran butir (b) hasil estimasi pada subtes verbal maupun numerikal. Berdasarkan tabel tersebut tingkat kesukaran butir, baik mudah, sedang dan sulit pada ketiga model. Berdasarkan kriteria yang sudah dijelaskan sebelumnya, menunjukkan sebagian besar butir memiliki tingkat kesukaran yang sama pada ketiga model, meskipun ada beberapa butir yang memiliki tingkat kesukaran yang tidak sama pada ketiga model.

Pada Tabel 3 menunjukkan parameter tingkat kesukaran butir baik tes verbal maupun tes numerikal memiliki rerata yang *moderate*, sehingga dapat disimpulkan bahwa butir-butir instrumen SAT memiliki tingkat kesukaran sedang.

Tabel 4.
Ringkasan Tingkat Kesukaran Butir

Subtes	Model	b	Nomor Butir	Jumlah	Persentase
Verbal	1-PL	Mudah	1, 3, 10, 43	4	8,00 %
		Sedang	2, 4, 5, 6, 7, 8, 9, 11, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 28, 29, 30, 31, 32, 33, 34, 35, 37, 38, 39, 40, 41, 42, 44, 45, 46, 47, 48, 49, 50	43	86,00 %
		Sukar	13, 26, 36	3	6,00 %
	2-PL	Mudah	1, 3, 8, 10, 24, 44	6	12,00 %
		Sedang	4, 5, 6, 7, 11, 12, 14, 15, 17, 18, 19, 20, 21, 22, 23, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 38, 40, 41, 42, 43, 45, 48, 49,	33	68,00 %
		Sukar	2, 9, 13, 16, 27, 37, 39, 46, 47, 50	10	20,00 %
	3-PL	Mudah	1, 3, 8, 10, 43	5	10,20 %
		Sedang	4, 5, 6, 7, 12, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 28, 29, 30, 31, 32, 33, 34, 35, 37, 39, 40, 41, 42, 44, 47, 48	33	67,35 %
		Sukar	2, 9, 11, 13, 16, 26, 36, 38, 45, 49, 50	11	22,45 %
Numerikal	1-PL	Mudah	2, 3, 19	3	7,50 %
		Sedang	1, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 40	35	87,50 %
		Sukar	38, 39	2	5,00 %
	2-PL	Mudah	1, 2, 3, 5	4	10,00 %
		Sedang	4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 35, 36, 37, 40	33	82,50 %
		Sukar	34, 38, 39	3	7,50 %
	3-PL	Mudah	2, 3	2	5,00 %
		Sedang	1, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 35, 36, 37, 40	34	85,00 %
		Sukar	23, 34, 38, 39	4	10,00 %

Sedangkan kriteria tingkat kesukaran tiap butir dapat dilihat pada Tabel 4. Meskipun sebagian besar butir memiliki tingkat kesukaran butir yang sedang, masing-masing subtes dalam instrumen SAT ternyata memiliki rentang tingkat kesukaran yang bervariasi, yakni -2,878 - 3,709 (verbal 1-PL); -4,133 - 8,617 (verbal 2-PL); -1,324 - 5,969 (verbal 3-PL); -2,310 - 2,516 (numerikal 1-PL); -5,456 - 7,307 (numerikal 2-PL); serta -4,708 - 2,580 (numerikal 3-PL). Berdasarkan rentang tersebut terlihat adanya butir yang memiliki tingkat kesukaran ekstrim, yakni ekstrim mudah (b = -5,456) maupun yang ekstrim sukar (b = 7,307). Butir dengan tingkat kesukaran ektrim mudah mengindikasikan bahwa butir ini mampu dijawab oleh sebagian besar subjek uji, sedangkan butir dengan tingkat kesukaran ektrim sukar mengindikasikan bahwa butir ini tidak mampu dijawab oleh sebagian besar subjek uji (Shenoy, Sayeli, & Rao, 2016; Gajjar, Sharma, Kumar, & Rana, 2014). Butir-butir yang memiliki tingkat kesukaran ekstrim ini sebaiknya diperbaiki karena dapat memberikan dampak buruk pada indeks daya beda butir (Gajjar, dkk., 2014; Nwadinigwe & Naibi, 2013). Meskipun begitu, butir ekstrim mudah masih dapat disiasati dengan ditempatkan di awal tes sebagai pertanyaan pemanasan, sedangkan butir esktrim sukar sangat perlu ditinjau ulang apakah masalahnya terletak pada penggunaan bahasa yang membingungkan, merupakan pertanyaan yang kontroversional, atau kemungkinan kesalahan kunci jawaban (Shenoy, dkk., 2016; Gajjar, dkk., 2014).

Indeks Daya Beda Butir

Tabel 5 menyajikan ringkasan indeks daya beda butir (a) hasil estimasi pada subtes verbal dan numerikal. Berdasarkan tabel tersebut terlihat bahwa untuk kedua subtes

yang diteliti, baik model IRT 2-PL maupun 3-PL memberikan kesimpulan bahwa 100% butir-butir yang termuat dalam instrumen SAT dapat membedakan kemampuan seseorang yang memiliki bakat skolastik yang tinggi dan rendah.

Tabel 5.
Ringkasan Daya Beda Butir

Subtes	Model	*A*	Nomor Butir	Jumlah	Persentase
Verbal	2-PL	Baik	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50	50	100 %
		Tidak Baik	-	0	0 %
	3-PL	Baik	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50	50	100 %
		Tidak Baik	-	0	0 %
Numerikal	2-PL	Baik	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40	40	100 %
		Tidak Baik	-	0	0 %
	3-PL	Baik	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40	40	100 %
		Tidak Baik	-	0	0 %

Parameter indeks daya beda butir-butir pada tes verbal maupun tes numerikal memiliki rerata indeks daya beda butir yang baik, sehingga dapat disimpulkan bahwa butir-butir instrumen SAT telah berfungsi dengan baik dalam membedakan bakat skolastik siswa. Baik pada CTT maupun IRT, indeks daya beda butir merupakan kemampuan butir untuk membedakan peserta didik yang memiliki kemampuan tinggi dan peserta didik yang memiliki kemampuan rendah. [ada perbedaan mendasar dari CTT dan IRT, pada CTT tidak mempunyai satuan dan tidak menghasilkan pengukuran yang intervalnya sama]. Apabila dikaitkan dengan konteks tes bakat, maka indeks daya beda butir dapat dimaknai sebagai kemampuan butir untuk membedakan subjek yang memiliki bakat tinggi dan rendah pada tes skolastik. Butir-butir pada subtes SAT sebagian besar memiliki indeks daya beda yang baik, meskipun jika dilihat lebih lanjut ternyata indeks daya beda butir memiliki rentang 0,117 - 1,446 (verbal 2-PL); 0,272 - 2,204 (verbal 3-PL); 0,272 - 1,655 (numerikal 2-PL); serta 0,127 - 2,200 (numerikal 3-PL). Berdasarkan rentang tersebut terlihat tidak ada butir yang memiliki indeks daya beda yang negatif. Apabila ada butir yang memiliki indeks daya beda negatif sebaiknya dipertimbangkan penggunaannya (direvisi atau dibuang) oleh pengembang tes, hal ini dikarenakan butir ini cenderung direspons salah oleh siswa yang memiliki kemampuan yang tinggi atau sebaliknya direspons benar oleh orang yang memiliki kemampuan rendah.

Indeks Tebakan Semu

Tabel 6 menyajikan ringkasan indeks tebakan semu (c) hasil estimasi subtes verbal dan numerikal. Berdasarkan kriteria nilai maksimum tebakan semu yang baik adalah satu banding jumlah distraktor yaitu 0.2, maka indeks tebakan semu pada semua butir dikelompokkan sebagaimana disajikan pada Tabel 6.

Tabel 6.
Ringkasan Indeks Tebakan Semu Butir

Subtes	Model	Tebakan Semu	Nomor Butir	Jumlah	Persentase
Verbal	3-PL	Baik	2, 3, 4, 6, 7, 9, 10, 12, 13, 14, 16, 17, 18, 20, 22, 23, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39, 41, 42, 43, 44, 45, 47, 49, 50	37	75,51 %
		Tidak Baik	1, 5, 8, 11, 15, 19, 21, 24, 25, 30, 40, 48	12	24,49 %
Numerikal	3-PL	Baik	1, 3, 4, 6, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 24, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40	31	77,50 %
		Tidak Baik	2, 5, 7, 14, 20, 21, 22, 23, 25	9	22,50 %

Parameter indeks tebakan semu hasil analisis yang disajikan pada tabel 6 menunjukkan bahwa subtes numerikal memiliki butir-butir dengan indeks tebakan semu baik yang lebih banyak (77,50%) dibandingkan subtes verbal (75,51%). Dengan adanya indeks tebakan semu (c) pada model logistik tiga parameter, memungkinkan bagi peneliti untuk mendeteksi subjek yang menjawab dengan acak atau subjek yang memiliki kemampuan rendah yang menjawab butir soal dengan benar dengan batasan atau *lower asymptote* yang wajar (Yang & Kao, 2014; Adedoyin & Mokobi, 2013; DeMars, 2010). Indeks tebakan semu dalam instrumen SAT dapat dikatakan telah memiliki rerata yang baik. Hal ini dikarenakan rerata indeks tebakan semu pada subtes verbal maupun subtes numerikal yakni berturut-turut sebesar 0,160 dan 0,164 telah berada di dalam range yang diacu. Meskipun begitu, ternyata masih ada butir yang memiliki nilai indeks tebakan semu yang cukup tinggi yakni 0,303 pada subtes verbal dan 0,264 pada subtes numerikal. Tinggi rendahkan tebakan semu terkait dengan kualitas pengecoh yang digunakan, kualitas pengecoh yang kurang baik dapat membuat subjek memilih kunci jawaban dengan menebak. Indeks tebakan semu hanya memberikan pengaruh yang kecil terhadap indeks daya beda butir karena besarnya indeks ini tergantung pada kualitas pengecoh yang digunakan (Haladyna, 2004).

Test Information Function dan *Standard Error*

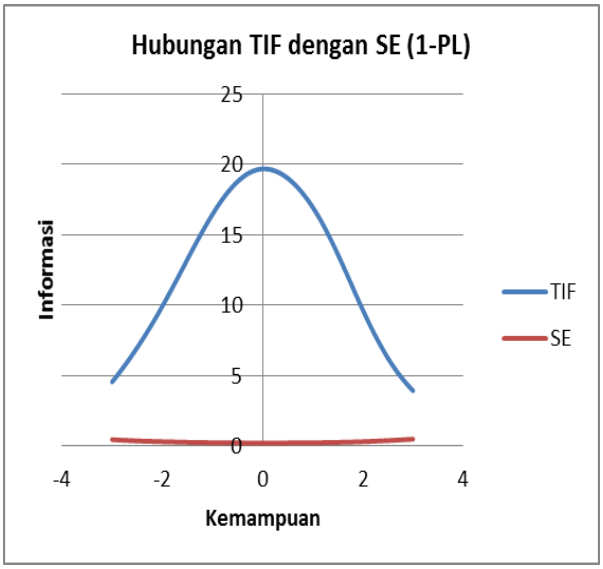
Kekuatan suatu butir menjelaskan informasi hasil pengukuran disebut dengan *Item Informasi Function* (IIF), dan jika melekat pada tes disebut *Test Informasi Fungtion* (TIF). Fungsi informasi tes ini menunjukkan *laten trait level* mana yang memberikan respons paling andal dan seberapa banyak informasi psikometrik yang dapat diukur dari *laten trait level* tersebut (Zieba, 2013; Markon, 2013). TIF dapat dimaknai setara dengan konsep reliabilitas dalam teori tes klasik, namun lebih akurat untuk mengestimasi *latent trait* peserta tes dibandingkan koefisien reliabilitas (Samejima, 1994; Hambleton & Swaminathan, 1985). TIF memiliki hubungan yang terbalik dengan dengan *Standard Error* (SE). Fungsi informasi tes disajikan secara grafis yang merupakan hubungan antara nilai informasi dari butir-butir penyusun tes terhadap tingkat sifat laten (Zieba, 2013). Analisis grafik akan mengarah pada kesimpulan bahwa semakin sempit bentuk grafis dari fungsi informasi tes, maka semakin sempit pula cakupan *latent trait* yang diukur melalui tes tersebut (Zieba, 2013).

Pada subtes verbal hubungan tersebut untuk semua model disajikan pada Gambar 1-3. Sedangkan hubungan TIF dengan SE pada subtes numerikal untuk semua model disajikan pada Gambar 4-6. Keenam gambar tersebut menunjukkan bahwa kurva TIF dan SE yang terbentuk memiliki titik potong yang cukup jauh, sehingga dapat diinterpretasikan bahwa baik subtes verbal maupun subtes numerikal mampu mengukur bakat verbal dan bakat numerik siswa pada rentang kemampuan yang cukup luas. Di samping itu, dari gambar diketahui bahwa

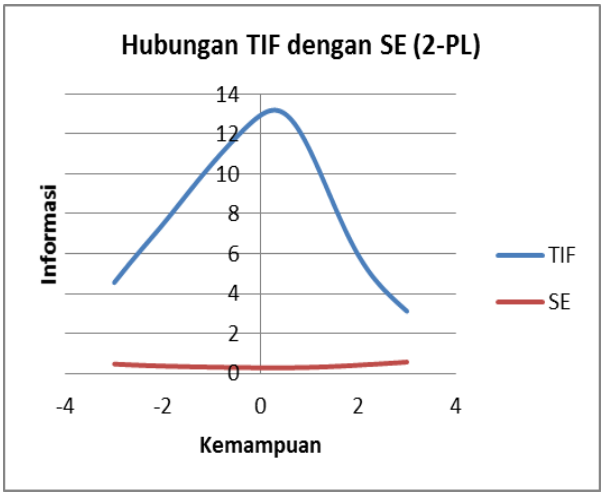
garis fungsi informasi tes berada di atas kesalahan bakunya, sehingga hasil pengukuran dengan instrumen tersebut akurat.

Nilai fungsi informasi maksimum untuk subtes verbal pada semua model dicapai pada kemampuan (teta) sebesar 0,05 (TIF = 19,690; SE = 0,225) untuk model IRT 1-PL; 0,05 (TIF = 19,690; SE = 0,225) untuk model IRT 2-PL; serta 0,60 (TIF = 40,335; SE = 0,157) untuk model IRT 3-PL. Sementara itu, nilai fungsi informasi maksimum untuk subtes numerikal pada semua model dicapai pada kemampuan (teta) sebesar 0,05 (TIF = 18,848; SE = 0,230) untuk model IRT 1-PL; -0,50 (TIF = 16,191; SE = 0,248) untuk model IRT 2-PL; serta 0,80 (TIF = 33,479; SE = 0,173) untuk model IRT 3-PL. Berdasar kriteria yang dikemukakan oleh Hambleton & Lam (2009), yaitu nilai TIF 5 sama dengan estimasi reliabilitas dalam teori tes klasik sebesar 0,80; sedangkan Nilai TIF 10 sama dengan 0,90, maka melalui gambar satu hingga 6 menunjukkan TIF pada tes verbal maupun numerikal cenderung tinggi.

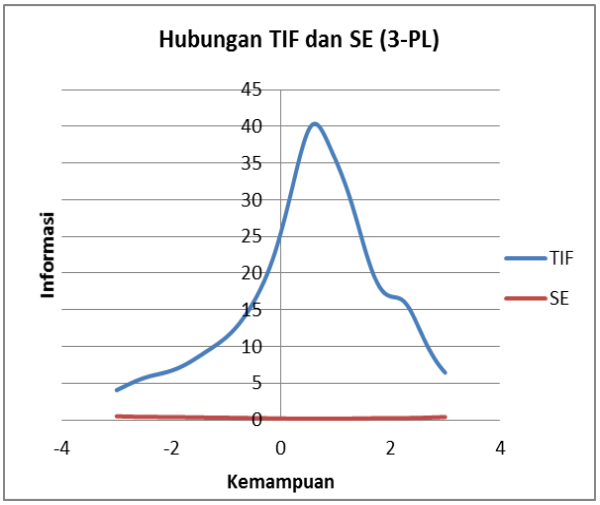
Selanjutnya, dikarenakan fungsi informasi tes juga dapat menunjukkan *latent trait level* mana yang memberikan respons paling andal, maka untuk subtes verbal, fungsi infomasi maksimum dapat dicapai apabila diujikan pada subjek dengan kemampuan 0,60 (TIF = 40,335; SE = 0,157) sedangkan untuk subtes numerikal, fungsi infomasi maksimum dapat dicapai pada subjek yang memiliki kemampuan 0,80 (TIF = 33,479; SE = 0,173).



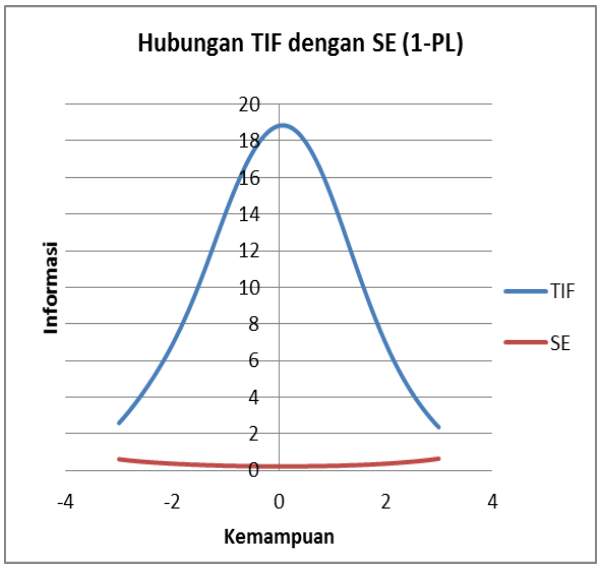
Gambar 1. Hubungan TIF dengan SE (1-PL) pada subtes verbal



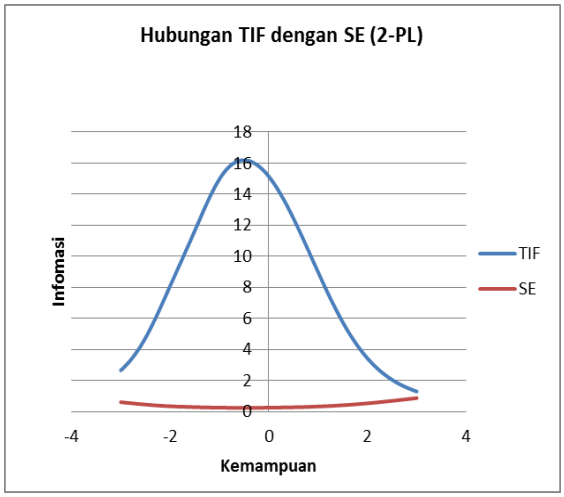
Gambar 2. Hubungan TIF dengan SE (2-PL) pada subtes verbal



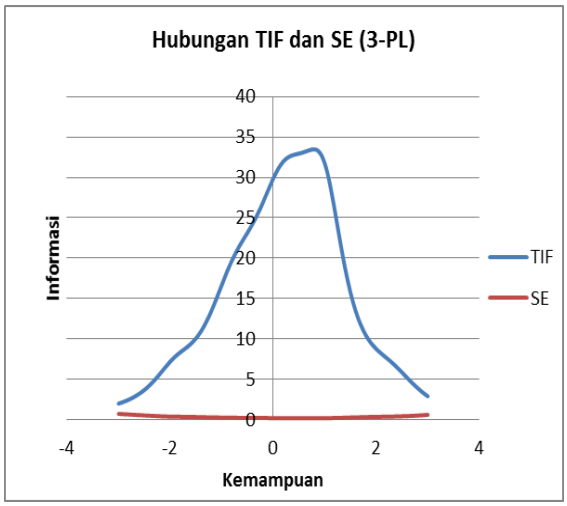
Gambar 3. Hubungan TIF dengan SE (3-PL) pada subtes verbal



Gambar 4. Hubungan TIF dengan SE (1-PL) pada subtes numerikal



Gambar 5. Hubungan TIF dengan SE (2-PL) pada subtes numerikal



Gambar 6. Hubungan TIF dengan SE (3-PL) pada subtes numerikal

Hasil penelitian ini menunjukkan model IRT 3-PL untuk semua subtes yang diuji memberikan fungsi informasi tes yang paling tinggi dan kesalahan pengukuran yang paling rendah dibandingkan model 2-PL Hal ini disebabkan rerata indeks daya beda butir pada model 3-PL ($\bar{a}$ verbal=1,108 dan $\bar{a}$ numerikal=1,157) lebih tinggi daripada indeks daya beda butir pada model 2-PL ($\bar{a}$ verbal=0,777 dan $\bar{a}$ numerikal=0,883). Pada model IRT yang mengakomodasi adanya indeks daya beda butir, apabila indeks daya beda butir semakin besar maka akan semakin besar pula nilai fungsi informasi tes yang diperoleh (Yang & Kao, 2014; Zieba, 2013). Adanya indeks daya beda inilah yang mengakibatkan informasi butir pada model 2-PL lebih tinggi dari 3-PL, demikian pula pada model 1-PL menjadi paling rendah karena model ini tidak mengakomodasi parameter indeks daya beda.

SIMPULAN

Berdasarkan hasil analisis yang telah dilakukan, dapat disimpulkan bahwa subtes verbal memiliki kecocokan model yang tinggi pada model 2-PL dan 3-PL, sedangkan pada subtes numerikal kecocokan model yang paling tinggi pada model 2-PL. Berdasar hasil analisis parameter butir, baik indeks kesukaran soal, indeks daya beda dan tebakan semu pada setiap butir berada pada kategori yang cenderung sama pada model 1-PL, 2-PL dan 3-PL, meskipun ada beberapa butir yang bergeser kategorinya, terutama yang memiliki skor pada skor perbatasan. Hasil analisis indeks kesukaran soal masih didapatkan butir yang sangat sulit dan mudah, indeks daya beda semua butir sudah bagus. dan terdapat beberapa butir yang memiliki tebakan semu yang tinggi. Berdasar TIF ketiga model parameter, butir-butir pada instrumen SAT dapat digunakan untuk mengukur kemampuan verbal dan numerikal pada semua level kemampuan, meskipun demikian model 3-PL dapat memberikan informasi hasil yang paling tinggi dibanding model yang lain.

DAFTAR PUSTAKA

Abed, E. R., Al-Absi, M. M., & Shindi, Y. A. A. (2016). Developing a numerical ability test for students of education in Jordan: An application of item response theory. *International Education Studies*, *9*(1), 161-174. http://dx.doi.org/10.5539/ies.v9n1p161

Abedalaziz, N., & Leng, C. H. (2013). The relationship between CTT and IRT approaches in analyzing item characteristics. *The Malaysian Online Journal of Education Science*, *1*(1), 64-70.

Adedoyin, O. O., & Mokobi, T. (2013). Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. *International Journal of Asian Social Science*, *3*(4), 992-1011.

Adedoyin, O. O., Nenty, H. J., & Chilisa, B. (2008). Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. *Educational Research and Review*, *3*(2), 83-93.

Agronow, S., & Studley, R. (November, 2007). *Prediction of college GPA from new SAT test scores: A first look*. Paper presented at the Annual Meeting of the California Association for Institutional Research (CAIR), Monterey, CA.

Ahnaldi, G. H. (2015). Aptitude test and successful college students: The predictive validity of the general aptitude test (GAT) in Saudi Arabia. *International Education Studies*, *8*(4), 1-6. http://dx.doi.org/10.5539/ies.v8n4p1

Anderson, P., & Morgan, G. (2008). *Developing tests and questionnaires for a national assessment of educational achievement*. Washington DC: The World Bank.

Awasthy, S., & Kaur, G. (2009). Aptitude battery for personnel below officer rank in Indian army. *Journal of the Indian of Applied Psychology*, *35*, 148-153.

Awopeju, O. A., & Afolabi, E. R. I. (2016). Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal*, *12*(28), 263-284. http://dx.doi.org/10.19044/esj.2016.v12n28p263

Baker, F. B. (2001). *The basics of item response theory (2nd Ed*). USA: ERIC Clearinghouse on Assessment and Evaluation.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1948). Differential Aptitude Tests. *Journal of Consulting Psychology*, *12*(1), 62.

Brzezinska, J. (2016). Latent variable modelling and item response theory analyses in marketing research. *Folia Oeconomica Stetinensia*, *16*(2), 163-174. https://doi.org/10.1515/foli-2016-0032

Chung, H. (2005). *Calibration and validation of the body self-image questionnaire using the rasch analysis* (*Master's Thesis)*. University of Georgia, Athens Georgia.

Courville, T. G. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics* (*Dissertation)*. Texas A&M University, Texas.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*.

New York: Holt, Reinhart, and Winston, Inc.

Curabay, M. (2016). *Meta-analysis of the predictive validity of scholastic aptitude test (SAT) and American college testing (ACT) scores for college GPA* (*Master's Thesis)*. University of Denver, Denver.

DAT for Selection General Abilites Battery: Technical Manual and User Guide. (2013). London: Pearson Education Ltd. ISBN 978 0 749104 54.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

DeMars, C. (2010). *Item response theory: Understanding statistics measurement*. New York: Oxford University Press, Inc.

Ehrman, M. E., Leaver, B. L, & Oxford, R. L. (2003). A brief overview of individual differences in second language learning. *System*, *31*, 313-330. https://doi.org/10.1016/S0346-251X(03)00045-9

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists multivariate applications book series*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Ereme, M. D. (2005). Use of psychological test in guidance and counseling. *Journal of Vocational, Science, and Educational Development*, *5*(1), 8-14.

Erguven, M. (2014). Two approaches to psychometric process: Classical test theory and Item response theory. *Journal of Education*, *2*(2), 23-30.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357-381.

Fox, J. P. (2007). Multilevel IRT modeling in practice with the package MLIRT. *Journal of Statistics Software*, *20*(5), 1-16.

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics an introduction*. Los Angeles: Sage Publication.

Gajjar, S., Sharma, R., Kumar, P., & Rana, M. (2014). Item and test analysis to identify quality multiple choice questions (MCQs) from an assessment of medical students of Ahmedabad, Gujarat. *Indian Journal of Community Medicine*, *39*(1), 17-20. http://doi.org/10.4103/0970-0218.126347

Geiser, S., & Studley, R. (2002). UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. *Educational Assessment*, *8*(1), 1-26. https://doi.org/10.1207/S15326977EA0801_01

Grove, W. A., Wasserman, T., & Grodner, A. (2006). Choosing a proxy for academic aptitude. *The Journal of Economic Education*, *37*(2), 131-47. https://doi.org/10.3200/JECE.37.2.131-147

Guler, N., Uyanik, G. K., & Teker, G. T. (2013). Comparison of classical test theory and item response theory in terms of item parameters. *European Journal of Research on Education*, *2*(1). 1-6.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items (3rd Ed.)*. New Jersey: Lawrence Erlbaum Associates Publishers.

Haley, D. C. (1952). *Estimation of the dosage mortality when the dose is subject to error (Technical Report, No. 15)*. Stanford, CA: Stanford University, Applied Mathematics and Statistics Laboratory.

Hambleton, R.K., & Lam, W. (2009). *Redesign of MCAS tests based on a consideration of information functions (MCAS Validity Report No. 18; CEA-689)*. Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Hambleton, R. K., & Swaminathan, H. (1985). *Items response theory: principles and application.* Boston: Kluwer-Nijjhoff Publish.

Hashmi, M. A., Zeeshan, A. Saleem, M., & Akbar, R. A. (2012). Development and validation of an aptitude test for secondary school mathematics students. *Bulletin of Educational and Research, 34*(1), 65-76.

Jung, R. E., Ryman, S. G., Vakhtin, A. A., Carraso, J., Wertz, C., & Flores, R. A. (2014). Subcortical correlates of individual differences in aptitude. *PLOS ONE, 9*(2), 1-9.

Kobrin, J. L., & Michel, R. S. (2006). *The SAT as a predictor of different levels of college performance (College Board Research Report No.2006-3).* New York: College Entrance Examination Board.

Kubiszyn, T., & Borich, G. (2003). *Educational testing and measurement (7th Ed.).* Singapore: John Wiley & Sons, Inc.

Lalor, J. P., Wu, H., & Yu, H. (2016). Building an evaluation scale using item response theory. In J. Su, K. Duh, & X. Carreras (Eds). *Proceedings of the conference on empirical methods in natural language process, Austin-Texas, 1-5 November 2016* (pp. 648-657). Austin, Texas: Association for Computational Linguistics.

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*, 509-525. https://doi.org/10.1348/000711009X474502

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*(4), 989–1020.

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement, 34*(3), 304-315.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment, 1*(1), 1-11.

Mahakud. (2013). Is it essential to measure intelligence along with aptitude test for career guidance. *International Refereed Research Journal, 4*(1), 92-102.

Mankar. J. & Chavan. D. (2013). Differential aptitude testing of youth. *International Journal of Scientific and Research Public, 3*(7), 1-6.

Markon, K. E. (2013). Information utility: Quantifying the total psychometric information provided by measure. *Psychological Methods, 18*(1), 15-35. http://DOI.org:/ 10.1037/a0030638

Mehrens, W. A., & Lehmann, I. J. (1987). *Using standardized tests in education (4th Ed.)*. New York: Longman Inc.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models (2nd Ed.)*. Mooresville: Scientific Software Inc.

Nazimuddin, S. K. (2015). A study of individual differences in educational situations. *International Journal of Scientific Engineering and Research*, *3*(7), 180-184.

Nwadinigwe, P. I., & Naibi, L. (2013). The number of options in a multiple-choice test item and the psychometric characteristics. *Journal of Education and Practice, 4*(28), 189-196.

Olufemi, A. S. (2013). Item response theory as a basis for measuring latent trait of interest. *Greener Journal of Social Sciences, 3*(7), 378-382.

Oyetunde, A. A. (2007). Construction and validation of a general science aptitude test (GSAT) for Nigerian junior secondary school graduate. *Iorin Journal of Education*, *27*, 22-23.

Philip, A., & Ojo, B. O. (2017). Application of item characteristic curve (ICC) in the selection of test items. *British Journal of Education*, *5*(2), 21-41.

Pollard, B., Dixon, D., Dieppe, P., & Johnston, M. (2009). Measuring the ICF components of impairment, activity limitation and participation restriction: An item analysis using classical test theory and item response theory. *Health and Quality of Life Outcomes*, *7*, 41. http://doi.org/10.1186/1477-7525-7-41

Pyari, P., Mishra, K., & Dua, B. (2016). Astudy of impact of aptitude in mathematics as stream selection at higher secondary level. *Issues and Ideas in Education*, *4*(2), 141-149. http://doi.org/10.15415/iie.2016.32011.

Robinson, P. (2012). Individual differences, aptitude, complexes, SLA processes, and aptitude test development. In M. Pawlak (Ed). *New perspective on individual differences in language learning and teaching* (pp. 57-75). Verlag, Berlin: Springer.

Retnawati, H. (2014). *Teori respons butir dan penerapannya, untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarja*. Yogyakarta: Parama Publishing.

Salkind, N. J., & Rasmussen, K. (2007). *Encyclopedia of measurement and statistics*. London: SAGE Publications, Inc.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement*, *18*(3), 229-244.

Sattler, J. M. (1988). *Assessment of children (3rd Ed.)*. San Diego, CA: Jerome M. Sattler, Publisher.

Sauce, B., & Matzel, L. D. (2013). The causes of variation in learning and behavior: why individual difference matter. *Frontiers in Psychology*, *4*(396), 1-10. https://doi.org/10.3389/fpsyg.2013.00395

Shenoy, P. J., Sayeli, V., & Rao, R. R. (2016). Item-analysis of multiple choice questions: A pilot attempt to analyze formative assessment in pharmacology. *Research Journal of Pharmaceutical, Biological, and Chemical Sciences*, *7*(2), 1683-1687.

Stumpf, H., & Stanley, J. C. (2002). Group data on high school grade point averages and scores on academic aptitude tests as predictors of institutional graduation rates. *Educational and Psychological Measurement*, *62*(6), 1042-1052.

Thorpe, G. L., & Favia, A. (2012). Data analysis using item response theory methodology: an instroduction to selected programs and applications. *Psychology Faculty Scholarship*, *20*, 1-33.

van der Aa, N., Bartels, M., te Velde, S. J., Boomsma, D. I., de Geus, E. J. C., & Brug, J. (2012). Genetic and environmental influences on individual differences in sedentary behaviour during adolescence. *Archives of Pediatrics and Adolescent Medicine*, *166*(6), 509-514. doi: 10.1001/archpediatrics.2011.1658

van der Linden, W. J., & Hambleton, R. K. (1996). *Handbook of modern item response theory*. New York: Springer.

Yang, F. M., & Kao, S. T. (2014). Item respons theory for measurement validity, *Shanghai Archieves of Psychiatry*, *26*(3), 171-177. http://doi.org/10.3969/j.issn.1002-0829.2014.03.010

Vosloo, H. N., Coetzee, N., & Claassen, N. C. W. (2000). *Manual for the differential aptitude tests form S*. Pretoria: Human Sciences Research Council.

Zieba, A. (2013). The item information function in one and two-parameter logistic models-a comparison and use in the analysis of the results of school test. *Didactics of Mathematics*, *10*(14), 87-96. http://doi.org/10.15611/dm.2013.10.08

Zoghi, M., & Valipour, V. (2014). A comparative study of classical test theory and item response theory in estimating test item parameters in a linguistics test. *Indian Journal of Fundamental and Applied Life Science*, *4*(4), 424-435.