HUMANITAS, Vol.15, No.1, Februari 2018, Hal. 46 - 61
ISSN 1693-7236, Terakreditasi **B** oleh DIKTI, No: 36a/E/KPT/2016

EVALUASI KARAKTERISTIK PSIKOMETRIK TES BAKAT DIFFERENSIAL DENGAN TEORI KLASIK

Farida Agus Setiawati[1], Rita Eka Izzaty[2], Veny Hidayat[3]
Universitas Negeri Yogyakarta
Jl. Colombo No.1, Yogyakarta 55281
[1]farida_as@uny.ac.id, [2]rita_ekaizzaty@uny.ac.id, [3]venyh@uny.ac.id

Abstract

The aims of the research were to analyze the psychometric characteristics of a set Differential Aptitude Test (DAT) consisting of five subtests, they are: numerical, abstract reasoning, space relation, and mechanics. This research using quantitative research method. The data ware collected from documentation pschological testing in UNY. The total data of the study are 2118 youth students in Yogyakarta. The analysis of psychometric characteristics of instrument using classical approach. The characteristic instrument are index of difficulty, index of disciminancy, effectifitas distractor, and reliability coefficient. The data were analyzed with Microcat ITEMAN 3.0 program. The result of analysis can be concluded that there are many information about item characteristic. The items characteristik have variation in index difficulty. Majority items have good index discimination, and many items are bad or need improvement. The items of abstract reasoning subtes have most distractor option which are not functioning properly, and all subtests in DAT instruments are reliable.

Keyword: classical theory, differential aptitude test (DAT), psychometric characteristics

Abstrak

Penelitian ini bertujuan untuk menganalisis karakteristik psikometrik seperangkat instrumen tes bakat diferensial (DAT) yang terdiri dari lima subtes, yakni subtes verbal, numerikal, berpikir abstrak, relasi ruang, dan mekanik. Penelitian ini menggunakan metode penelitian kuantitatif. Data dikumpulkan dari dokumentasi biro tes psikologi yang ada di UNY. Total data yang terkumpul adalah 2118 siswa di Daerah Istimewa Yogyakarta (DIY). Analisis karakteristik psikometrik instrumen dengan menggunakan pendekatan klasik. Karakteristik instrumen yang dianalisis adalah indeks kesukaran soal, indeks daya beda, keberfungsian pengecoh dan koefisien reliabilitas. Data dianalisis dengan bantuan program Microcat ITEMAN 3.0. Hasil analisis penelitian ini mendapatkan informasi karakteristik psikometris pada tiap butir. Indeks kesulitan butir-butir instrumen ini bervariasi, Sebagian besar indeks daya beda tergolong baik dan beberapa butir tergolong rendah atau perlu perbaikan. Butir-butir pada subtes berpikir abstrak memiliki banyak pengecoh yang tergolong tidak berfungsi efektif, dan semua subtes DAT tergolong reliabel.

Kata kunci: karakteristik psikometrik, teori tes klasik, tes bakat diferensial

Pendahuluan

Setiap manusia memiliki serangkain kemampuan yang dimiliki. Salah satu kemampuan yang sering dikaitkan dengan kesuksesannya adalah bakat atau *aptitude*. Bakat didefinisikan sebagai seperangkat karakteristik yang berhubungan dengan kemampuan individu untuk memperoleh pengetahuan atau keterampilan Salkind & Rashmussen (2007). Kubiszyn & Borich (2003) juga mendefinisikan bakat sebagai nama lain untuk potensi atau kemampuan. Berdasarkan kedua pendapat di atas, dapat disimpulkan bahwa bakat merupakan salah satu bentuk kemampuan khusus pada manusia yang berupa potensi, yakni dapat tumbuh dan berkembang secara lebih besar lagi seiring dengan latihan pengetahuan maupun ketrampilan yang dijalani.

Penelusuran tentang bakat awalnya terjadi bidang kerja atau jabatan, tetapi kemudian merambah ke bidang pendidikan (Anastasi, 1982; Goslin, 1963). Hingga saat ini, penelusuran bakat di berbagai bidang pendidikan telah banyak dilakukan karena bakat memiliki peran penting dalam memberikan pertimbangan pada penjurusan atau kelanjutan studi (Mankar & Chavan, 2013; Chatterjee, 2007 dalam Hashmi, dkk, 2012). Kemampuan dalam mempelajari satu bidang dengan bidang yang lain ada yang memerlukan bakat yang sama namun memiliki intensitas (kadar) yang berbeda dan ada juga yang berbeda. Kesuksesan pada setiap bidang studi juga memerlukan dukungan dari berbagai bentuk bakat yang mendukungnya. Sebagai contoh, kesuksesan studi bidang arsitek di perguruan tinggi perlu didukung oleh kemampuan skolastik (Stickler, 2007; Oyetunde, dalam Ballado, Morales, & Ortiz, 2014) dan bakat relasi ruang dan mekanik (Pearson Assessment, 2009). Oleh sebab itu, upaya untuk melakukan deteksi bakat yang dimiliki setiap siswa diperlukan untuk pengembangan dirinya.

Salah satu bentuk tes yang sering digunakan untuk mendeteksi bakat siswa adalah Tes bakat differensial atau *Differential Aptitude Test* (DAT). Tes ini disusun oleh George K. Bennet, Harold G. Seashore, & Alexander G. Wesman pada tahun 1947. Lahirnya tes ini dilatarbelakangi oleh pemikiran ahli psikologi bahwa kemampuan mental tidak hanya terdiri dari satu faktor saja melainkan banyak faktor sehingga dibutuhkan suatu tes yang dapat mengukur bermacam-macam faktor dengan beberapa skor sesuai dengan kemampuan yang diukur (Mankar & Chavan, 2013; Hashmi, dkk, 2012; Anastasi, 1982).

Pada umumnya, tes bakat dapat dibagi menjadi dua, yakni *Test Special Aptitude* yang terfokus pada satu bakat saja, misal mengukur bakat dibidang teknik mekanik, bakat pekerjaan tertentu (klerikal); dan *Multiple Aptitude Batteries* yang terdiri dari sejumlah tes dan digunakan untuk menganalisis profil individu (Anastasi & Urbina, 1997). Tes DAT termasuk dalam jenis *multiple aptitude batteries*, yakni tes yang mampu mengukur bermacam-macam bakat seseorang. Dalam instrumen DAT terdapat tujuh subtes yang sering disebut seri multiple bakat yakni *verbal reasoning (vr), numerical ability (na), abstract reasoning (ar), clerical speed and accuracy (csa), mechanical reasoning (mr), space relations (sr), dan language usage* yang terdiri atas *spelling and sentences* (Mankar & Chavan, 2013; Bennett, Seashore, & Wesman, 1948). Dalam aplikasinya, tes ini dapat disajikan secara keluruhan (1 seri) atau terpisah (tiap subtes).

Penelitian ini menggunakan lima subtes dalam tes bakat yang akan dianalisis karakteristik psikometriknya yakni subtes verbal, numerikal, penalaran, relasi ruang, dan mekanik. Subtes verbal digunakan untuk mengukur kemampuan menggunakan kata-kata (kosa kata). Subtes numerikal digunakan untuk mengukur kemampuan

berpikir dengan angka dan penguasaan hubungan numerik, misalnya penjumlahan atau pengurangan sederhana. Subtes penalaran digunakan untuk mengukur kemampuan penalaran yang bersifat nonverbal, meliputi kemampuan memahami hubungan logis dari figur-figur abstrak. Subtes relasi ruang digunakan untuk mengukur seberapa jauh kemampuan sesorang mengenal ruang tiga dimensi. Subtes mekanik digunakan untuk mengukur daya penalaran di bidang kerja mekanis dan prinsip fisika.

Alat ukur yang baik haruslah memiliki kualitas butir-butir yang baik pula (Kaplan & Saccuzzo, 2005). Kualitas butir soal sering dilihat dari karakteristik psikometrik butir-butir pada instrumen. Evaluasi karakteristik butir soal dilakukan mengetahui kualitas butir-butir soal mana saja yang bagus, sehingga bisa tetap dipertahankan berdasarkan hasil analisis terhadap butir-butir soal yang menyusun sebuah instrumen. Dengan demikian alasan dilakukannya evaluasi karakteristik butir soal sesuai dengan tujuan dilakukannya analisis butir soal yaitu untuk meningkatkan kualitas soal, yaitu apakah suatu soal: (1) dapat diterima karena telah didukung oleh data statistik yang memadai, (2) diperbaiki karena terbukti terdapat beberapa kelemahan, dan (3) tidak digunakan sama sekali karena terbukti secara empiris (Boopathiraj & Chellamani, 2013; Mitra, et.al, 2009; Gronlund, 1998; Lange, Lehmann, & Mehrens, 1967).

Karakteristik psikometrik suatu instrumen sangat penting untuk diteliti karena merupakan atribut yang terkait dengan tes psikologi (Furr & Bacharach, 2008:8). Oleh sebab itu, berbagai penelitian tentang karakteristik psikometrik menggunakan teori tes klasik pada instrumen tes bakat banyak dilakukan. Ballado, Morales, & Ortiz (2014) menganalisis instrumen tes bakat berbentuk *multiple choice* yang terdiri dari 150 butir dan dikenakan pada 130 mahasiswa baru di University of Eastern Philippine. Analisis dilakukan menggunakan metode klasik yang hasilnya adalah 78,67% butir memiliki tingkat kesukaran sedang, 11,33% butir berfungsi baik dalam membedakan abilitas siswa, dengan koefisien reliabilitas sebesar 0,78. Dengan menggunakan metode analisis yang sama, Hashmi, et.al (2012) meneliti tes bakat dalam bidang matematika yang terdiri dari subtes aritmatik, aljabar, dan geometri. Subjek penelitian yang digunakan sebanyak 288 siswa laki-laki dan 166 siswa perempuan kelas 10 yang tersebar dalam 13 sekolah di distrik Bahawalpur dan Multan. Hasil analisis menunjukkan bahwa tingkat kesukaran butir bergerak dari angka 0,13-0,83; indeks diskriminasi butir bergerak dari angka 0,06-0,70; dengan koefisien reliabilitas sebesar 0,82. Koefisien reliabilitas yang memuaskan juga didapatkan oleh Elbokai (2012). Instrumen *School and Collage Ability Test* (SCAT) yang ditelitinya menghasilkan koefisien reliabilitas sebesar 0,88.

Penelitian lain yang cukup menarik dilakukan di negara Bangladesh. Rahman (2014) menyusun instrumen *teacher aptitude test* berdasarkan *standardized test*: *GATB, DAT, SATB, TATB, TAT, Wellesley Spelling Scale, Yale Educational Aptitude Test Battery, Multiple Aptitude Test*, dan *Online Aptitude test*. Instrumen yang disusun terdiri dari 259 butir dan terdistribusi kedalam lima subtes yakni subtes penalaran, subtes verbal, numerikal, bahasa Inggris, dan bahasa Bangle. Subjek yang dilibatkan dalam penelitian tersebut adalah 100 orang guru dari 11 sekolah yang ada di kota Dakka. Hasil penelitian menunjukkan bahwa 90% butir memiliki tingkat kesukaran dengan interval 0,40 sampai 0,76 dan 46% butir memiliki daya beda yang baik; serta reliabilitas subtes bahasa Bangle, bahasa Inggris, numerikal, penalaran, dan verbal berturut-turut adalah 0,701; 0,604; 0,598; 0,724; dan 0,506.

Secara garis besar, analisis butir secara empirik ini dapat dibedakan menjadi dua, yaitu dengan pendekatan teori tes klasik (*Classical Test Theory, CTT*) dan teori

respons butir (*Item Response Theory, IRT*) (Ojerinde, 2013; Thorpe & Favia, 2012; DeMars, 2010; Crocker & Algina, 2008; Van der Linden & Hambleton, 1997; Hambleton, Swaminathan, & Rogers, 1991; Hambleton & Swaminathan, 1985). Penggunaan kedua pendekatan memiliki kelebihan dan kekurangan masing-masing. Sebagai penelitian awal, analisis karakteristik butir dilakukan berdasarkan pada pendekatan teori tes klasik.

Penggunaan teori tes klasik dalam konstruksi dan interpretasi skor telah berjalan beberapa dekade lamanya. Hal ini dikarenakan berbagai kelebihan yang ditawarkan oleh analisis butir secara klasik itu sendiri, yakni: murah, dapat dilaksanakan sehari-hari dengan cepat menggunakan komputer, sederhana, familier dan dapat menggunakan data dari beberapa peserta didik atau sampel kecil (Linn, 1989). Keunggulan lain dari teori ini terletak pada kemudahan dalam pemahaman konsepnya, ukuran sampel tidak perlu besar dan dilakukan dengan perhitungan sederhana.

Teori tes klasik menggunakan model pengukuran skor yang tampak, yaitu merupakan penjumlahan dari skor murni dan kesalahan pengukuran (Awopeju & Afolabi, 2016; Ojerinde, 2013; Junker, 2012; Mardapi, 2008; Traub, 1997). Hubungan antara skor tampak atau *observed score* (X), skor murni atau *true score* (T) dan kesalahan pengukuran atau *erro*r (E) dapat ditulis dalam persamaan 1. Teori klasik merupakan sebuah teori yang mudah dalam penerapannya serta model yang cukup berguna dalam mendeskripsikan bagaimana kesalahan dalam pengukuran dapat mempengaruhi skor tampak.

$$X = T + E \qquad (1)$$

Berdasarkan persamaan (1), maka terdapat enam macam asumsi dalam teori tes klasik, yaitu: (1) instrumen hanya mengukur satu dimensi, (2) ρ_{ET} = 0, tidak ada hubungan antara skor kesalahan dan skor murni, (3) ρ_{E1E2} = 0, tidak ada hubungan antara kesalahan pada pengukuran pertama dengan kesalahan pada pengukuran kedua, (4) ρ_{E1T2} = 0, tidak ada hubungan kesalahan pada pengukuran pertama dengan skor murni pada pengukuran kedua, (5) ρ_{E2T1} = 0, tidak ada hubungan kesalahan pada pengukuran kedua dengan skor murni pada pengukuran pertama, dan (6) μ_E, rata-rata kesalahan pengukuran pada populasi adalah 0 (Zoghi & Valipour, 2014; Ojerinde, 2013; Allen & Yen, 1979). Asumsi-asumsi teori tes klasik tersebut dikembangkan dalam berbagai formula yang berguna dalam melakukan pengukuran psikologis. Formula tersebut diantaranya adalah tingkat kesukaran, indeks daya beda, efektivitas distraktor, reliabilitas tes, dan kesalahan pengukuran (Murphy & Davidshofer, 2003).

Di Indonesia, DAT banyak digunakan di bidang pendidikan yang digunakan untuk penelusuran bakat terkait pemilihan program di SMA dan penelusuran karir. Terkait penelitian DAT, masih langka dilakukan khususnya terkait analisis psikometrik secara klasik. Dukungan empirik sangat dibutuhkan untuk dilakukannya analisis alat tes tersebut guna mengungkap bakat yang sebenarnya. Terkait dengan hal ini, maka diperlukan sebuah kajian empirik dan teoritik untuk mengevaluasi karakteriktik psikometrik DAT dengan subjek penelitian di Indonesia sehingga kesesuaian dan ketepatan hasil pengukuran dapat dipertanggungjawabkan.

Metode Penelitian

Penelitian ini secara keseluruhan menggunakan pendekatan kuantitatif, yang dalam pelaksanaannya terdiri dari beberapa bagian penelitian yang saling terkait, yaitu penelitian analisis karakteristik psikometrik instrumen dengan menggunakan subjek terbaru serta analisis dilakukan dengan menggunakan pendekatan klasik. Setelah karakteristik butir pada tiap butir diketahui maka evaluasi dan seleksi pada butir-butir tersebut dapat dilakukan.Instrumen peneltian ini adalah instrumen tes bakat diferensial yang diadaptasi dari instrumen yang disusun oleh Bennet, Harold G. Seashore, & Alexander G. Wesman pada tahun 1947. Instrumen ini berbentuk *multiple choice* dan terdiri dari lima subtes, dengan spesifikasi seperti yang tersaji dalam Tabel 1.

Tabel 1
Spesifikasi Tes Bakat Differensial

Subtes	Jumlah Butir	Alokasi Waktu (menit)
Verbal	50	30
Numerikal	40	30
Penalaran	50	25
Relasi Ruang	60	30
Mekanik	68	30

Data didapatkan dari hasil dokumentasi hasil tes psikologi yang dilakukan oleh Unit Pelayanan Bimbingan dan Konseling dan Program Studi Psikologi. Total data yang terkumpul sejumlah 2118. Data yang diperoleh selanjutnya dianalisis secara kuantitatif melalui pendekatan teori tes klasik. Bentuk karakteristik psikometrik dari instrumen DAT yang dianalisis adalah indeks kesulitan soal, daya beda, efektifitas pengecoh, reliabilitas dan kesalahan pengukuran.

1. Indeks Kesulitan Soal

$$P = B/T \quad (1)$$

Keterangan :
P : Indeks kesukaran soal
B : Jumlah peserta tes yang menjawab butir dengan benar
T : Jumlah seluruh peserta tes yang mengerjakan butir

2. Indeks Daya Beda

Indeks daya beda dihitung dengan menggunakan formula korelasi poinbiserial (ρ_{pbis})

$$\rho_{pbis} = \frac{\mu_+ - \mu_x}{\sigma_x}\sqrt{\frac{p}{q}} \quad (2)$$

Keterangan :
ρ_{pbis}: indeks daya beda (korelasi point biserial)
μ_+ : rata-rata skor dari peserta tes yang menjawab benar butir soal
μ_x : rata-rata skor total
σ_x : simpangan baku skor total
p : proporsi jawaban yang benar terhadap semua jawaban
q : $1 - p$

3. Efektivitas Distraktor

$$ED = \frac{Jumlah\ peserta\ tes\ yang\ memilih\ distraktor}{Jumlah\ peserta\ tes\ keseluruhan} \quad (3)$$

4. Reliabilitas
reliabilitas dihitung dengan formula alpha, sebagai berikut:

$$r_{xx}^{'} = [\frac{n}{n-1}][1 - \frac{\sum s_i^2}{s_x^2}] \quad (4)$$

Keterangan :
$r_{xx}^{'}$: reliabilitas
s_x^2 : varians skor tes
n : banyaknya item dalam tes
s_i^2 : varians skor butir

5. Kesalahan Baku Pengukuran

$$S_E = S_X\sqrt{1 - r_{xx}^{'}} \quad (5)$$

Keterangan :
S_E : kesalahan baku pengukuran
S_X : simpangan baku

Hasil dan Pembahasan

Analisis karakteristik psikometrik instrumen tes bakat diferensial dilakukan secara terpisah pada setiap subtesya. Analisis dilakukan menggunakan pendekatan teori tes klasik dengan bantuan program Microcat ITEMAN 3.0. Adapun karakteristik yang dimaksud meliputi tingkat kesukaran (TK), daya beda (DB), efektivitas distraktor (ED), reliabilitas, dan kesalahan pengukuran.

1. Tingkat Kesukaran

Tingkat kesukaran (TK) didefinisikan sebagai proporsi peserta tes yang menjawab benar pada butir soal tertentu yang nilainya berkisar antara 0 sampai 1 (Aiken, 1994: 66). Semakin besar nilai TK yang diperoleh dari perhitungan berarti semakin mudah butir soal tersebut. Butir soal yang memiliki TK = 0 berarti tidak ada siswa yang mampu menjawab butir soal itu dengan benar, dan bila butir soal memiliki TK = 1 berarti semua siswa menjawab benar butir soal tersebut. Tingkat kesulitan butir soal biasanya dikaitkan dengan tujuan tes. Misalnya, pada ujian semester digunakan butir soal yang memiliki TK sedang, untuk seleksi digunakan butir soal yang memiliki TK sukar, dan untuk diagnostik digunakan butir soal yang memiliki TK mudah (Kusaeri & Suprananto, 2012).

Tabel 2 menyajikan statistik deskriptif tingkat kesukaran butir hasil analisis karakteristik psikometrik dari masing-masing subtes pada instrumen DAT yang diteliti. Berdasarkan tabel tersebut terlihat bahwa subtes penalaran memiliki rerata tingkat kesukaran butir tertinggi yakni 0,774 yang artinya subtes penalaran adalah subtes paling

mudah karena dapat dijawab benar oleh 77,4% peserta tes. Sementara itu, subtes verbal adalah subtes yang paling sukar karena hanya 48,9% peserta tes yang mampu menjawab subtes ini dengan benar. Tingkat kesukaran maksimum atau butir paling mudah berada pada subtes penalaran dengan harga indeks kesukaran sebesar 0,961 sedangkan tingkat kesukaran minimum atau butir paling sukar berada pada subtes verbal dengan harga indeks kesukaran sebesar 0,489.

Tabel 2
Statistik Deskriptif Tingkat Kesukaran Instrumen DAT

Subtes	TK Minimum	TK Maksimum	Rerata TK
Verbal	0,029	0,933	0,489
Numerikal	0,109	0,905	0,531
Penalaran	0,088	0,961	0,774
Relasi Ruang	0,075	0,946	0,596
Mekanik	0,066	0,894	0,542

Hasil analisis karakteristik psikometrik pada Tabel 2 menunjukkan bahwa tingkat kesukaran butir pada masing-masing subtes memiliki rentang yang bervariasi, yakni: 0,489-0,933 untuk subtes verbal; 0,531-0,905 untuk subtes numerikal; 0,774-0,961 untuk subtes penalaran; 0,596-0,946 untuk subtes relasi ruang; dan 0,542-0,894 untuk subtes mekanik. Meskipun begitu, keseluruhan tes memiliki rerata tingkat kesukaran butir yang termasuk dalam kategori sedang. Hal ini didasarkan pada kriteria yang dipaparkan oleh Allen & Yen (1979: 121); Nitko (1983); Hingorjo & Jaleel (2012); Sayyah, dkk (2012); dimana indeks kesukaran butir yang dapat diterima adalah 0,30-0,70 karena pada interval tersebut informasi kemampuan siswa diperoleh dengan maksimal. Jadi, butir yang memiliki tingkat kesukaran di bawah 0,30 dapat digolongkan ke dalam butir sukar sedangkan butir yang memiliki tingkat kesukaran di atas 0,70 dapat digolongkan ke dalam butir mudah. Berdasarkan kriteria tersebut, maka identifikasi tingkat kesukaran butir pada masing-masing subtes disajikan pada Tabel 3.

Tabel 3 menunjukkan bahwa subtes penalaran memiliki butir mudah paling banyak yakni 62% (31 butir dari total 50 butir), sedangkan subtes numerikal memiliki butir mudah paling sedikit yakni 20% (8 butir dari total 40 butir). Subtes verbal memiliki butir sukar paling banyak yakni 28% (14 butir dari total 50 butir), sedangkan subtes penalaran memiliki butir sukar paling sedikit yakni 2% (1 butir dari total 50 butir). Ditinjau dari tujuan pelaksanaan tes, perlu diperhatikan bahwa butir soal yang terlalu mudah atau terlalu sukar terkadang kurang memberikan informasi yang berguna bagi peserta tes pada umumnya. Butir yang terlalu mudah mengakibatkan pengecoh tidak berfungsi dengan baik. Hal ini terjadi karena soal yang mudah akan membuat testee mudah memilih kunci jawaban sebagai jawabannya, sehingga pengecoh menjadi tidak berfungsi. Soal yang terlalu sulit memungkinkan kunci jawaban tidak dijawab oleh para testee atau apabila dijawab hanya menebak saja. Pada soal yang terlalu mudah atau terlalu sukar ini, apabila daya bedanya dan pengecohnya berfungsi dengan baik, maka soal tersebut masih dapat digunakan.

Tabel 3
Ringkasan Tingkat Kesukaran Instrumen DAT

Subtes	Tingkat Kesukaran	Butir	Jumlah	Persentase (%)
Verbal	Mudah	1, 3, 4, 5, 7, 8, 10, 18, 22, 24, 31, 43	12	24,00
	Sedang	6, 11, 12, 15, 17 , 19, 20, 21, 23, 25, 27, 28, 30, 33, 34, 35, 37, 39, 40, 41, 42, 46, 47, 48	24	48,00
	Sukar	2, 9, 13, 14, 16, 26, 29, 32, 36, 38, 44, 45, 49, 50	14	28,00
Numerikal	Mudah	1, 2, 3, 5, 13, 15, 17, 19	8	20,00
	Sedang	4, 6, 7, 8, 9, 10, 11, 14, 16, 18, 20, 21, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 35, 36, 37, 40	27	67,50
	Sukar	12, 23, 34, 38, 39	5	12,50
Penalaran	Mudah	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 20, 21, 22, 23, 24 , 26, 27, 28, 29, 31, 32, 34, 38, 44	31	62,00
	Sedang	16, 19, 25, 30, 33, 35, 36, 37, 39, 40, 41, 42, 43, 45, 46, 47, 48, 50	18	36,00
	Sukar	49	1	2,00
Relasi Ruang	Mudah	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 18, 19, 20, 21, 22, 24, 25, 28, 31, 32, 33, 34, 35, 36, 40, 57	31	51,67
	Sedang	16, 17, 23, 26, 27, 29, 30, 37, 38, 39, 42, 43, 44, 45, 47, 48, 50, 54	18	30,00
	Sukar	41, 46, 49, 51, 52, 53, 55, 56, 58, 59, 60	11	18,33
Mekanik	Mudah	1, 3, 4, 6, 7, 9, 10, 13, 14, 17, 20, 26, 29, 31, 32, 38, 40, 56, 59	19	27,94
	Sedang	2, 8, 11, 12, 16, 18, 19, 21, 23, 24, 25, 27, 28, 30, 33, 34, 35, 36, 39, 41, 42, 43, 45, 46, 47, 49, 51, 52, 53, 58, 60, 62, 63, 64, 65, 66, 67, 68	38	55,88
	Sukar	5, 15, 22, 44, 48, 50, 52, 54, 55, 57, 61	11	16,18

2. Daya Beda

Daya beda adalah kemampuan butir soal tes untuk membedakan peserta didik yang memiliki kemampuan tinggi (menguasai materi yang ditanyakan) dan rendah (belum menguasai materi yang ditanyakan). Daya beda berkaitan dengan derajat kemampuan butir soal dalam membedakan dengan baik karakteristik peserta tes dalam tes yang diujikan (Anastasi & Urbina, 1997:179). Indeks daya beda nilainya berkisar antara -1,00 sampai dengan 1,00. Semakin tinggi dan positif indeks daya beda berarti semakin baik butir soal tersebut dalam membedakan peserta tes kelompok atas (skornya tinggi) dan peserta tes kelompok bawah (skornya rendah). Jika daya beda bernilai negatif berarti lebih banyak peserta tes kelompok bawah (tidak memahami materi yang diajarkan) menjawab dengan benar butir soal tersebut dibandingkan dengan kelompok atas (memahami materi yang diajarkan).

Tabel 4 menyajikan statistik deskriptif daya beda hasil analisis karakteristik psikometrik dari masing-masing subtes pada instrumen DAT yang diteliti. Berdasarkan tabel tersebut terlihat bahwa subtes penalaran juga memiliki rerata daya beda butir tertinggi yakni 0,543 yang artinya subtes penalaran adalah subtes paling dapat membedakan bakat siswa. Sementara itu, subtes mekanik adalah subtes yang memiliki

rerata daya beda terendah yakni 0,301. Indeks daya beda maksimum berada pada subtes penalaran yakni sebesar 0,525 sedangkan indeks daya beda minimum berada pada subtes verbal yakni sebesar -0,029.

Tabel 4
Statistik Deskriptif Karakteristik Daya Beda Instrumen DAT

Subtes	DB Minimum	DB Maksimum	Rerata DB
Verbal	-0,029	0,513	0,416
Numerikal	0,096	0,521	0,497
Penalaran	0,043	0,525	0,543
Relasi Ruang	0,018	0,502	0,469
Mekanik	-0,123	0,397	0,301

Apabila suatu butir soal tidak dapat membedakan dengan baik karakteristik peserta tes, maka butir soal tersebut dapat dicurigai kemungkinannya: (a) kunci jawaban butir soal itu tidak tepat, (b) butir soal itu memiliki dua atau lebih kunci jawaban yang benar, (c) kompetensi yang diukur tidak jelas, (d) pengecoh tidak berfungsi, (e) materi yang ditanyakan terlalu sulit sehingga banyak peserta tes yang menebak jawaban, dan (f) sebagian besar peserta tes yang memahami materi yang ditanyakan berpikir ada yang salah informasi dalam butir soalnya (Kusaeri & Suprananto, 2012).

Hasil analisis karakteristik psikometrik pada Tabel 3 menunjukkan bahwa daya beda butir pada masing-masing subtes juga memiliki rentang yang bervariasi, yakni: -0,029-0,513 untuk subtes verbal; 0,096-0,521 untuk subtes numerikal; 0,043-0,525 untuk subtes penalaran; 0,018-0,502 untuk subtes relasi ruang; dan -0,123-0,397 untuk subtes mekanik. Meskipun begitu, keseluruhan tes memiliki rerata daya beda butir yang termasuk dalam kategori baik. Hal ini didasarkan pada kriteria yang dipaparkan oleh Mitra, dkk, 2009; Reynolds, Livingston, & Wilson, 2009; Mardapi, 2008; Kartowagiran, 2012; dimana butir berfungsi baik dalam membedakan kemampuan siswa manakala memiliki indeks daya beda minimal 0,30. Berdasarkan kriteria tersebut, maka identifikasi daya beda butir pada masing-masing subtes disajikan pada Tabel 5.

Tabel 5 menunjukkan bahwa subtes penalaran memiliki butir dengan daya beda baik yang paling banyak yakni 76% (38 butir dari total 50 butir) sedangkan subtes mekanik memiliki butir dengan daya beda baik yang paling sedikit yakni 29,41% (20 butir dari total 68 butir). Sementara itu, subtes mekanik memiliki butir dengan daya beda tidak baik yang paling banyak yakni 70,59% (48 butir dari total 68 butir) sedangkan subtes penalaran memiliki butir dengan daya beda tidak baik yang paling sedikit yakni 24% (12 butir dari total 50 butir).

Tingkat kesukaran butir turut menjadi penyebab suatu butir memiliki daya beda tidak baik atau rendah. Thorndike, dkk (1991) memberikan alasan bahwa butir-butir soal yang memiliki daya beda rendah dapat disebabkan oleh tingkat kesukaran butir soal yang terlalu rendah (butir soal terlalu sukar) atau terlalu tinggi (butir soal terlalu mudah). Koefisien reliabilitas pun akan meningkat manakala butir-butir yang daya bedanya tidak cukup memuaskan tidak diikutsertakan dalam analisis. Selain itu, pengecoh yang tidak masuk akal meskipun butir soal tersebut memiliki tingkat kesukaran yang diterima dapat menjadi penyebab rendahnya daya beda. Keberadaan pengecoh yang tidak masuk akal ini akan memudahkan siswa untuk memutuskan bahwa pengecoh tersebut salah sehingga kemungkinan siswa menjawab benar dengan menebak sangat tinggi dan menyebabkan butir soal menjadi terlalu mudah. Sebaliknya, pengecoh

yang terlalu dekat nilai kebenarannya dengan kunci dapat menyebabkan butir soal menjadi terlalu sukar.

Tabel 5
Ringkasan Indeks Daya Beda Instrumen DAT

Subtes	Daya Beda	Butir	Jumlah	Persentase (%)
Verbal	Baik	4, 6, 7, 12, 14, 15, 17, 18, 19, 20, 24, 25, 27, 28, 29, 30, 31, 32, 33, 34, 35, 37, 39, 41, 42, 44, 47	27	54, 00
	Tidak Baik	1, 2, 3, 5, 8, 9, 10, 11, 13, 16, 21, 22, 23, 26, 36, 38, 40, 43, 45, 46, 48, 49, 50	23	46,00
Numerikal	Baik	4, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 35, 36, 37, 40	30	75,00
	Tidak Baik	1, 2, 3, 5, 8, 22, 23, 34, 38, 39	10	25,00
Penalaran	Baik	3, 4, 5, 6, 7, 9, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 25, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 39, 40, 41, 42, 43, 44, 45, 46, 47	38	76,00
	Tidak Baik	1, 2, 8, 10, 11, 13, 24, 30, 38, 48, 49, 50	12	24,00
Relasi Ruang	Baik	3, 5, 7, 8, 9, 11, 12, 13, 14, 17, 18, 19, 20, 21, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 40, 42, 43, 44, 47, 50, 52, 53, 54, 55, 57, 59	41	68,33
	Tidak Baik	1, 2, 4, 6, 10, 15, 16, 22, 27, 39, 41, 45, 46, 48, 49, 51, 56, 58, 60	19	31,67
Mekanik	Baik	4, 8, 13, 17, 23, 25, 26, 31, 32, 36, 40, 42, 45, 59, 60, 62, 64, 65, 67, 68	20	29,41
	Tidak Baik	1, 2, 3, 5, 6, 7, 9, 10, 11, 12, 14, 15, 16, 18, 19, 20, 21, 22, 24, 27, 28, 29, 30, 33, 34, 35, 37, 38, 39, 41, 43, 44, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 61, 63	48	70,59

3. Efektivitas Pengecoh

Penyebaran jawaban peserta tes pada tiap pilihan jawaban (*pengecoh*) yang disediakan pada soal bentuk pilihan ganda juga menjadi dasar dalam penelaahan butir soal. Pengecoh adalah pilihan jawaban yang bukan merupakan kunci jawaban. Analisis efektifitas pengecoh bertujuan untuk mengetahui berfungsi tidaknya pengecoh atau distraktor yang telah dibuat. Pengecoh dibuat dengan tujuan mengecoh peserta tes yang: (a) kurang begitu memahami materi, (b) kurang teliti dalam melakukan teknis perhitungan dalam soal yang menyangkut perhitungan matematis. Agar dapat melakukan fungsinya untuk mengecoh dengan baik, pengecoh harus dibuat yang homogen atau mirip dengan kunci jawaban. Distribusi respon jawaban dikatakan efektif bila tiap pilihan jawaban minimal ada yang menjawab paling sedikit 5% dari peserta tes (Mehta & Mokhasi, 2014; Chauhan, dkk, 2015; Mukherjee & Lahiri, 2015; Quaigrain & Arhin, 2017; Mardapi, 2012). Dalam penelitian ini pengecoh dikatakan efektif apabila terdapat 5% atau lebih peserta memilih pilihan jawaban tersebut dan tidak efektif apabila yang memilih pilihan jawaban tersebut kurang dari 5%.

Tabel 6 menyajikan ringkasan efektivitas pengecoh hasil analisis karakteristik psikometrik dari masing-masing subtes pada instrument DAT yang diteliti. Berdasarkan

tabel tersebut terlihat bahwa subtes penalaran memiliki butir terbanyak yang pengecoh-pengecohnya tidak berfungsi dengan baik yakni 48 butir. Sementara itu, subtes mekanik hanya memiliki sedikit butir yang pengecoh-pengecohnya tidak berfungsi dengan baik yakni 8 butir.

Tabel 6
Ringkasan Karakteristik Efektivitas Pengecoh Instrumen DAT

Subtes	Banyaknya Butir yang Memiliki Pengecoh Tidak Efektif
Verbal	43
Numerikal	23
Penalaran	48
Relasi Ruang	39
Mekanik	8

Berdasarkan Tabel 6, dapat disimpulkan bahwa subtes penalaran memiliki banyak butir yang pengecoh-pengecohnya tidak cukup efektif dalam mengecoh siswa sedangkan subtes mekanik hanya memiliki sedikit butir yang pengecoh-pengecohnya tidak cukup efektif dalam mengecoh siswa. Dalam butir-butir tersebut, tidak hanya termuat satu pengecoh saja yang tidak efektif, namun bisa memuat beberapa pengecoh sekaligus. Bahkan dalam beberapa butir, muncul notifikasi atau peringatan untuk melakukan pengecekan kembali terhadap kunci jawaban (*check the key*). Peringatan tersebut mengindikasikan bahwa mayoritas siswa yang kemampuannya tinggi lebih memilih pengecoh dan tidak memilih kunci jawaban. Apabila kunci jawaban telah diperiksa dan memang sudah benar, maka kemungkinan kesalahan terletak pada adanya perbedaan konsep (miskonsepsi) siswa. Selanjutnya, apabila soal ini tetap ingin digunakan, maka pilihan jawaban yang menjadi kunci maupun pengecoh perlu diperbaiki atau direvisi. Perbaikan butir perlu dilakukan selain dengan memperhatikan distribusi respon jawaban, juga memperhatikan daya beda pada masing-masing pengecoh. Daya beda yang paling tinggi dan positif pada pilihan kunci jawaban menunjukkan siswa yang pintar (skor totalnya tinggi) cenderung menjawab soal ini dengan benar, sementara siswa yang kurang pintar cenderung menjawab salah. Daya beda positif dan tinggi pada pilihan jawaban selain kunci jawaban tidak dikehendaki karena siswa yang pintar lebih memilih pengecoh disbanding kunci jawaban.

4. Reliabilitas dan Kesalahan Pengukuran

Reliabilitas dan kesalahan pengukuran adalah dua hal yang saling terkait. Mehrens & Lehmann (1973) menyatakan bahwa reliabilitas merupakan derajat keajegan (*consistency*) antara dua buah hasil pengukuran pada objek yang sama. Allen & Yen (1979) menambahkan bahwa suatu tes dikatakan reliabel jika skor amatan mempunyai korelasi yang tinggi dengan skor yang sebenarnya. Koefisien reliabilitas yang baik adalah diatas 0,70 (Linn, 1989; Mardapi, 2014) meskipun secara teoritik besarnya koefisien reliabilitas berkisar dari 0,00 sampai 1,00 (Azwar, 2013).

Hasil perhitungan reliabilitas pada tiap subtes, subtes relasi ruang memiliki indeks reliabilitas yang paling tinggi yakni 0,862 dengan kesalahan pengukuran sebesar 3.075. Sementara itu, subtes mekanik memiliki indeks reliabilitas yang paling rendah yakni 0,737 dengan kesalahan pengukuran sebesar 3,579. Secara keseluruhan, hasil perhitungan reliabilitas dan kesalahan baku pengukuran disajikan pada Tabel 7.

Tabel 7
Ringkasan Reliabilitas dan Kesalahan Pengukuran Instrumen DAT

Subtes	Reliabilitas	Kesalahan Pengukuran
Verbal	0,809	2,965
Numerikal	0,850	2,709
Penalaran	0,858	2,655
Relasi Ruang	0,862	3,075
Mekanik	0,737	3,579

Hasil analisis yang disajikan pada Tabel 7 menunjukkan bahwa semua subtes memiliki koefisien reliabilitas yang baik karena telah melampaui kriteria minimal yang ditentukan. Artinya, semua subtes terbukti reliabel untuk digunakan sebagai alat pengukuran bakat individu. Secara berturut-turut reliabilitas subtes verbal, numerikal, penalaran, relasi ruang, dan mekanik yakni 0,809; 0,850; 0,858; 0,862; dan 0,862. Dikarenakan kesemua subtes dalam instrumen DAT terbukti reliabel maka apabila dilakukan pengukuran bakat, hasil pengukuran bakat akan memberikan hasil informasi yang ajeg atau konsisten.

Sementara itu, berdasar skor reliabilitas dapat diketahui kesalahan baku pengukuran (SEM). SEM pada umumnya digunakan untuk memahami kesalahan yang bersifat acak yang mempengaruhi skor peserta tes. Hasil SEM secara berturut-turut adalah subtes verbal, numerikal, penalaran, relasi ruang, dan mekanik yakni 2,965; 2,709; 2,655; 3,075; dan 3,579. Semakin kecil nilai kesalahan baku pengukuran, maka skor hasil pengukuran dengan tes tersebut semakin tepat atau akurat. Dengan demikian hasil pengukuran dengan menggunakan DAT dapat dipertanggungjawabkan.

Simpulan

Berdasarkan hasil penelitian di atas dapat disimpulkan adanya berbagai bentuk karakteristik psikometris yang melekat pada DAT. Butir-butir pada DAT memiliki tingkat kesukaran butir yang bervariasi, ada yang mudah, sedang dan sulit. Indeks daya beda pada sebagian besar butir tergolong baik, meskipun masih ada butir yang memiliki daya beda rendah atau negatif sehingga perlu diperbaiki. Dari kelima subtes yang ada pada seperangkat tes DAT, subtes penalaran memiliki beberapa butir yang pengecohnya tidak efektif. Dilihat dari masing-masing subtes, instrumen DAT memiliki reliabilitas yang tinggi sehingga hasil pengukuran dengan tes ini ajeg dan konsisten, serta akurat karena didapatkan SEM yang rendah. Penelitian lebih lanjut diperlukan untuk memperbaiki butir-butir yang memiliki indeks daya beda yang rendah dan beberapa butir yang memiliki pengecoh yang kurang efektif dan melakukan analisis karakteristik psikometrik lagi agar didapatkan butir-butir yang bagus. Hasil penelitian ini bermanfaat dalam memberikan informasi karakteristik psikometris DAT. Informasi ini bermanfaat bagi pengguna tes terkait dengan pertimbangannya ketika akan menggunakan tes ini. Bagi pengembang tes, hasil penelitian ini bermanfaat dalam upaya untuk melakukan seleksi butir, membuat butir yang lebih bagus, membuat bank soal, maupun upaya untuk membuat tes berbasis komputer.

Daftar Pustaka

Aiken, L.R. (1994). *Psychological testing and assessment (8th ed.).* Boston: Allyn and Bacon.

Allen, M.J., & Yen, W.M. (1979). *Introduction to measurement theory*. California: Brooks/Cole Publishing Company.

Anastasi, A. (August 1982). *Aptitude and achievement tests: the curious case of the indestructible strawperson*. Paper presented in Invited Symposia: State of the Art Series, Achievement Testing, at the meeting of the American Psychological Association, Washington, D.C.

Anastasi, A. & Urbina, S. (1997). *Psychological testing*. Indiana: Prentice Hall, Inc.

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.

Azwar, S. (2013). *Reliabilitas dan validitas (Edisi keempat).* Yogyakarta: Pustaka Pelajar.

Ballado, R.S., Morales, R.A., & Ortiz, R.M. (2014). Development and validation of a teacher education aptitude test. *International Journal of Interdisciplinary Research and Innovations, 2* (4) 129-133.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1948). Differential aptitude tests: some comments by the authors. *Journal Counseling & Development, 27* (1), 20–22. doi: 10.1002/j.2164-5892.1948.tb01449.x.

Boopathiraj, C. & Chellamani, K. (2013). Analysis of tests items on difficulty level and discrimination index in the test for research in education. *International Journal of Science and Interdisciplinary Research*. 2(2). 189-193.

Chauhan, P., Chauhan, G. R., Chauhan, B. R., Vaza, J. V., & Rathod, S. P. (2015). Relationship between difficulty index and distracter effectiveness in single best-answer stem type multiple choice questions. *International Journal of Anatomy and Research*.3(4). 1607-1610. doi: 10.16965/ijar.2015.299.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. New York: Holt, Reinhart, and Winston, Inc.

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

Elbokai, H.F. (2012). Reliability and validity study of the school and college ability test (SCAT) advanced form. *International Journal of Humanity and Social Sciences*. 2 (11). 89-96.

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

Goslin, D.A. (1963). *The search for ability: standardized testing in social perspective*. New York: Russell Sage.

Gronlund, N.E. (1998). *Assessment of student achievement (6th Ed.)*. Boston, MA: Allyn and Bacon.

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

Hambleton, R.K., Swaminathan, H., & Rogers H.J. (1991). *Fundamental of item response theory*. London: Sage Publication.

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.

Hingorjo, R.M. & Jaleel, F. (2012). Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *Journal of Pakistan Medical Association.* 62(2). 142-147.

Junker, B.W. (2012). *Some aspects of classical reliability theory & classical test theory*. Pittsburgh, PA: Carniege Mellon University.

Kalton G. (1983). *Introduction to survey sampling*. New Delhi: Sage Publication Inc

Kaplan, R. M. & Saccuzzo, D.P. (2005). *Psychological testing: principles, application, and issues* (6th Ed.). Belmont: Thomson Wadsworth.

Kartowagiran, B. (Oktober 2012). *Penulisan butir soal*. Makalah disajikan dalam Pelatihan penulisan dan analisis butir soal bagi Sumber daya PNS Dik-Rekinpeg, di Hotel Kawanua Aerotel, Jakarta.

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

Kusaeri & Suprananto. (2012). *Pengukuran dan penilaian pendidikan.* Yogyakarta: Graha Ilmu.

Lange, A., Lehmann, I.J., & Mehrens, W.A. (1967). Using item analysis to improve tests. *Journal of Educational Measurement, 4* (2), 65–68. doi**:** 10.1111/j.1745-3984.1967.tb00572.x

Linn, R.L. (1989). *Education measurement (3th ed.)*. New York: MacMillan Publishing Company.

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

Mardapi, D. (2008). *Teknik penyusunan instrumen tes dan non tes*. Yogyakarta: Mitra Cendekia Offset.

Mardapi, D. (2012). *Pengukuran, penilaian, dan evaluasi pendidikan*. Yogyakarta: Nuha Litera.

Mardapi, D. (September 2014). *Bahan pelatihan penilaian otentik*. Makalah disajikan dalam Konferensi HEPI, di Denpasar Bali.

Mehta, G., & Mokhasi, V. (2014). Item analysis of multiple choice question-an assessment of the assessment tool. *International Journal of Health Sciences and Research*, *4* (7), 197-202.

Mehrens, W.A., & Lehmann, J.L. (1973). *Measurement and evaluation in education and psychology*. New York: Holt, Rinehart, and Winston, Inc.

Mitra, N. K, Nagaraja, H. S., Ponnudurai, G., & Judson, J. P. (2009). The levels of difficulty
and discrimination indices in type a multiple choice questions of pre-clinical semester
1 multidisciplinary summative tests. International E-Journal of Science, Medicine, and Education. 3 (1). 2-7. Retrieved from http://web.imu.edu.my/ejournal/approved/iMEC_2.original_p02-07.pdf

Mukherjee, P., & Lahiri, S.K. (2015). Analysis of multiple choice questions (MCQs): item and test statistics from an assessment in a medical college of kolkata west bengal. *IOSR Journal of Dental and Medical Sciences*, *14* (12), 47-52.

Murphy, K.R. & Davidshofer, C.O. (2003). *Psychological testing: principles and application*. New Jersey : Prentice-Hall Inc.

Nitko, A. (1983). *Educational test and measurement: an introduction*. New York: Harcourt Brace Jovanovich, Inc.

Ojerinde, D. (May 2013). *Classical test theory (ctt) vs item response theory (irt): an evaluation of the comparability of item analysis results*. Lecture Presentation at the Institute of Education University of Ibadan, Ibadan, Oyo, Nigeria.

Pearson Assessment. (2009) The differential aptitude test. *Report for simon sample*. Upper Sadle River: Pearson Education.

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-development test in educational measurement and evalution. *Research Article*. Cogent Education. 1-11.

Rahman, A. (2014). Developing teaching aptitude test: a perspective of Bangladesh. *Green University Review of Social Sciences*, *1* (1), 75-89.

Reynolds, C.R., Livingston, R.B., & Wilson, V. (2009). *Measurement and assessment in education (2nd ed)*. Boston: Pearson Education Inc.

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

Sayyah, M., Vakili, Z., Alavi, N. M., Bigdeli, M., Soleymani, A., Assarian, M.,& Azarbad, Z.
(2012). An item analysis of written multiple-choice questions: Kashan university of
medical sciences. *Nursing and Midwifery Studies*, *1* (2), 83-87. doi: 10.5812/nms.8738.

Scheaffer, R. L., Mendenhall, W., & Lyman, O. (1990). *Elementay survey sampling* (*4th Ed*). Boston: PSW KENT Publishing Company.

Stickler. L. (2007). A critical review of the SAT: menace or mild-mannered measure?. *The College of New Jersey. TCNJ Journal of Student Scholarship*, *9*, 1-9.

Thorndike, et.al (1991). *Measurement and evaluation in psychology and education (5th ed.).* New York: Macmillan Publishing Company.

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.

Traub, R.E. (1997). Classical test theory in historical perspective. *Educational Measurement and Practice*. 8-14.

Van der Linden, W.J., & Hambleton, R.K. (1997). *Handbook of modern item response theory*. New York: Springer.

Varma, S. (n.d.). *Preliminary item statistics using point-biserial correlation and p-values*. Morgan Hill, CA: Educational Data Sistem, Inc.

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 Sciences, 4*(4). 424-435.