KARAKTERISTIK PSIKOMETRIK TES BERDASARKAN PENDEKATAN TEORI TES KLASIK DAN TEORI RESPON AITEM

Ali Ridho
Fakultas Psikologi UIN Malang

ABSTRACT

The aim of this research study was to evaluate and compared psychometrics characteristics of achievement based on classical test theory (CTT) and item response theory (IRT) especially based on one (1PL), two (2PL), and three (3PL) parameters models. The data for the research consist of Senior High School students' responses to the Mathematics National Exit Examination Academic Year 2003/2004 in Yogyakarta. The subjects were 7000 (3500 male and 3 500 female students). The test has 40 multiple choice test items and is criterion referenced. By comparing the indices from CTT and IRT, the overall conclusion from this evaluation is that 2PL model is preferable to use when evaluating the test.

Keywords: *classical test theory, item response theory, multiple choice test*

Teori tes klasik (TTK) atau *classical test theory* (CTT) telah berkembang secara luas dan menjadi aliran utama di kalangan ahli psikologi dan pendidikan, serta bidang kajian perilaku (*behavioral*) yang lain, selama 20 dekade (Embretson & Reise, 2000). TTK memiliki kelemahan karena bersifat *examinee sample dependent* dan *item sample dependent* (Fan, 1998; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Hambleton, Robin, & Xing, 2000; Lord, 1980). Kelemahan tersebut memicu teori baru yang lebih memadai, yaitu teori tes modern, yang dikenal juga sebagai teori respon aitem (TRA) atau *item response theory* (IRT) dan dikenal pula dengan nama *latent traits theory* (LTT).

TRA memiliki beberapa kelebihan dibandingkan TTK. Secara terperinci Embretson & Reise (2000) mengemukakan 10 kelebihan TRA dibanding TTK, yaitu: (1) simpangan baku pengukuran atau *standard error of measurement* (SEM) memiliki nilai yang berbeda-beda antar skor (atau pola-pola respon), tetapi bersifat umum antar populasi; (2) tes yang lebih pendek bisa jadi lebih reliabel dibanding tes yang lebih panjang; (3) perbandingan skor-skor tes antar berbagai format akan optimal jika tingkat kesulitan tes bervariasi antar peserta; (4) estimasi-estimasi yang tidak bias bisa diperoleh dari sampel yang tidak representatif; (5) skor tes memiliki arti manakala dibandingkan dengan karakteristik aitem-aitem; (6) skala yang bersifat interval dicapai dengan menggunakan model pengukuran yang lebih logis; (7) tes dengan format aitem campuran dapat menghasilkan skor tes yang optimal; (8) skor-skor yang berubah dapat dibandingkan secara berarti jika tingkat skor awal berbeda; (9) hasil faktor analisis pada data skor kasar aitem menghasilkan sebuah *full information factor analysis*; dan (10) sifat-sifat aitem sebagai stimulus dapat secara langsung berhubungan den gan sifat-sifat psikometriknya.

Manfaat lain yang diperoleh dari TRA adalah efektivitasnya saat diterapkan pada administrasi berbasis komputer yang lebih dikenal dengan *computerized adaptive testing* (CAT) untuk tes-tes yang mengungkap kemampuan (McLeod, Lewis, & Thissen, 2003). Hal ini akan meningkatkan efektifitas waktu tes serta pengontrolan terhadap minimalisasi eror untuk tiap-tiap *testee*, kondisional terhadap kemampuan masing-masing (Xing & Hambleton, 2004).

Berbeda dengan TTK yang memfokuskan pada informasi pada level tes, TRA terutama memfokuskan pada informasi pada level aitem sehingga diharapkan dapat menutupi kekurangan yang terdapat pada TTK. Penerapan model IRT didasarkan atas beberapa asumsi berupa postulat, yaitu: (1) kinerja seorang peserta pada suatu aitem dapat diprediksikan oleh seperangkat faktor yang disebut *traits*, *latent traits*, atau kemampuan; dan (2) hubungan antara kinerja peserta pada suatu aitem dan seperangkat kemampuan (abilitas) laten yang mendasarinya dapat digambarkan oleh suatu fungsi yang menarik secara monotonik yang disebut *item characteristic Ffunction* atau *item characteristic curve* (ICC) (Hambleton, Swaminathan, & Rogers, 1991; Harvey & Hammer, 1999; Suryabrata, 2000). Jadi ICC adalah penggambaran dalam bentuk kurva yang menjelaskan hubungan antara *latent traits* dan kinerja subjek pada sebuah aitem.

Hambleton & Swaminathan (1985) menyatakan bahwa asumsi-asumsi yang mendasari TRA adalah unidimensi, independensi lokal, dan invariansi parameter. Sementara itu, Embretson & Reise (2000) menyebutkan bahwa asumsi yang paling pokok adalah: (1) masing-masing item memiliki bentuk kurva karakteristik aitem atau *item characteristic curves* (ICC) tertentu; dan (2) independensi lokal.

TRA adalah analisis aitem berdasarkan model. Ada 3 model dalam TRA yang terkenal, yaitu model: satu-parameter (1PL), dua-paramenter (2PL), dan tiga-parameter (3PL). Model matematik 3PL adalah:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}$$

Dimana *i* adalah aitem ke-*i*, c_i = faktor tebakan semu (*pseudo guessing*) aitem *i*, a_i = daya beda aitem *i*, b_i = tingkat kesukaran aitem *i*, dan θ adalah *traits-level* (dalam hal ini kemampuan) *examinee* atau para peserta tes. Jika c_i diasumsikan 0 ($c_i = 0$ untuk semua *i*), maka model 3PL menjadi 2PL:

$$P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}$$

Sementara, jika daya beda untuk semua aitem dalam model 2PL ditetapkan sama ($a_i = a$ untuk semua *i*), maka model tersebut menjadi model 1PL:

$$P_i(\theta) = \frac{e^{a(\theta - b_i)}}{1 + e^{a(\theta - b_i)}}$$

Meski secara teoritik-fundamental berbeda dengan TTK, TRA memiliki hubungan yang erat dengan TTK. Oleh sebab itu, bagi para pembaca yang telah mengenal TTK, hubungan tersebut dapat dijadikan dasar dalam periode awal untuk memahami TRA. Setelah mempelajari TRA secara lebih mendalam, barulah dapat diketahui manfaat

keunggulan TRA atas TTK. Sejauh pengamatan penulis, para ahli pengukuran psikologi dan pendidikan serta institusi yang terkait dengan tes dan hal yang terkait dengan pengembangan administrasinya, belum memberikan perhatian yang serius dalam menyadari dan menyambut gelombang perkembangan teori pengukuran. Oleh sebab itu, penulis tergerak untuk meneliti dan memaparkan analisis psikometrik tes berdasarkan metode TTK dan TRA serta hubungan antar konsep dalam kedua metode tersebut.

Studi yang mengkhususkan pada analisis perbandingan psikometrik berdasarkan TTK dan TRA belum banyak dilakukan di Indonesia. Studi yang berhasil penulis temukan adalah: *Using Classical Test Theory in Combination With Item Response Theory* (Bechger, Maris, Verstralen, & Beguin, 2003), *Item Response Theory and Classical Test Theory: An Empirical Comparison of Their Item/Response Pers on Statistics* (Fan, 1998), *Item Response Theory* (Harvey & Hammer, 1999), *A Monte Carlo Comparison of Item and Person Statistics Based on Item Response Theory Versus Classical Test Theory* (McDonald & Paunonen, 2002), dan *Some relationships between the information function of IRT and the signal/noise ratio and reliability coefficient of classical test theory* (Nicewander, 1993).

Tujuan penelitian ini adalah untuk mengungkap secara empirik karakteristik Tes UAN Matematika SMA tahun pelajaran 2003/2004 berdasarkan pendekatan TRA, yaitu: (1) invariansi *traits level* peserta θ berdasar model 1PL, 2PL dan 3PL, (2) invariansi parameter aitem pada model 1PL, 2PL dan 3PL, serta (3) membandingkan hasil pendekatan metode TTK dan TRA.

Manfaat penelitian ini adalah: (1) memberikan masukan bagi ilmuwan dan praktisi psikometri tentang bukti invariansi yang dapat ditegakkan dalam analisis hasil tes yang mengukur kinerja maksimum (dalam hal ini UAN), dan (2) diharapkan hasil ini mampu menggugah para ilmuwan dan praktisi dalam menggunakan TRA sebagai pendekatan analisis hasil tes sebagai pelengkap analisis hasil tes menggunakan TTK.

METODE PENELITIAN

Penelitian ini merupakan penelitian deskriptif, karena bertujuan untuk: (1) mengetengahkan karakteristik aitem-aitem tes UAN berdasarkan TTK dan TRA, dan (2) menelusuri bukti invariansi estimasi *traits-level* peserta (θ) serta invariansi parameter aitem berdasarkan model 1PL, 2PL dan 3PL.

Subjek Penelitian

Subjek penelitian ini adalah siswa SMA yang mengikuti UAN Matematika SMA tahun pelajaran 2003/2004 di Daerah Istimewa Yogyakarta. Jumlah subjek adalah 7000 orang (3500 laki-laki dan 3500 perempuan). Pengambilan subjek yang besar ini terkait dengan daya (*power*) statistik yang akan dihasilkan terkait dengan estimasi parameter aitem dan *latent traits* (Stone, 2003). Stone (2003) menyebutkan bahwa daya atau *power* statistik dalam uji kecocokan model atau *goodness of fit* (GOF) dalam model TRA tidak akan terpengaruh oleh ukuran sampel, asal seluruh aitem fit dengan model. Akan tetapi, jika terdapat satu saja aitem yang tidak fit dengan model, dalam replikasi 100 kali, daya statistik akan bertambah dengan berubahnya ukuran sampel dari 500 menjadi 2000. Makin besar ukuran sampel, makin besar pula daya statistik yang dapat diperoleh. Untuk itu penulis mengambil sampel dengan ukuran 7000 orang (masing-

masing 3500 laki-laki dan 3500 perempuan). Untuk mengestimasi parameter kemampuan (*traits-level*) dan parameter aitem, digunakan 7000 data respon tersebut.

Metode Pengumpulan Data

Data penelitian ini adalah data sekunder berupa hasil respon siswa terhadap perangkat tes UAN Matematika SMA tahun pelajaran 2003/2004 di Daerah Istimewa Yogyakarta yang diperoleh dari *scanning* Lembar Jawaban Komputer (LJK) siswa.

Metode Analisis Data

Metode Teori Tes Klasik (TTK)

Analisis deskriptif yang akan dipaparkan adalah mean dan deviasi standar skor, serta reliabilitas yang digunakan adalah reliabilitas internal Alpha. Pada level aitem, tingkat kesukaran ditunjukkan oleh r_{pbis} (Crocker & Algina, 1986) yang merupakan korelasi antara kinerja peserta tes pada sebuah aitem dibandingkan dengan kinerja peserta pada skor total. Selanjutnya, penyelidikan terhadap 10% peserta yang memiliki kinerja terendah dilakukan untuk memberikan isyarat bagaimanakah gambaran model TRA yang akan digunakan.

Metode Teori Respon Aitem (TRA)

Teknik yang digunakan:

1. untuk mengetahui karakteristik empirik aitem-aitem tes UAN matematika:
   a. memilih aitem-aitem yang memiliki $r_{pbis} \geq 0.2$ berdasarkan TTK untuk dilakukan analisis aitem dengan pendekatan TRA;
   b. mengestimasi parameter aitem menggunakan metode *marginal maximum likelihood* dengan bantuan program MULTILOG 7.03 (Thissen, 2003);
   c. menentukan karakteristik aitem dan melihat kecocokan seluruh data dengan model. Program komputer MULTILOG 7.03 pada tahap ini dapat menghasilkan berkas estimasi parameter-parameter: (1) daya beda *a*, (2) tingkat kesukaran *b*, dan (3) peluang tebakan semu *c*, serta (4) nilai –2 *Loglikelihood G* keseluruhan data, sesuai dengan model yang dipilih;
   d. menggambar fungsi informasi tes atau *information function* (IF) dan fungsi simpangan baku pengukuran atau *standard error of measurement* (SEM) dengan bantuan program MATHCAD 12;
   e. memilih model dengan mempertimbangkan kecocokan data, IF dan SEM;
   f. berdasarkan model terpilih, parameter aitem diterima jika: (1) $-2 \leq b_i \leq 2$; (2) $0 \leq a_i \leq 2$ (Hambleton, Swaminathan, & Rogers, 1991); dan (3) $0 \leq c_i \leq 0.35$ (Baker, 2001; Ridho, 2005; Risnawita, 2004);
2. untuk menguji invariansi estimasi parameter:
   a. invariansi estimasi parameter kemampuan para peserta dilakukan dengan melihat pola *scatter plot* dan linieritas berupa garis regresi antara estimasi kemampuan berdasarkan 10 aitem tersukar dengan estimasi kemampuan berdasarkan 10 aitem termudah. Pembandingan ini dilakukan pada ketiga model;
   b. invariansi estimasi parameter aitem dilakukan dengan membandingkan hasil estimasi parameter aitem tertentu dengan membatasi/mengontrol parameter aitem yang lain, relatif pada masing-masing model. Pembandingan dilakukan berdasarkan kalibrasi parameter aitem berdasarkan kelompok tinggi (*teta*

terestimasi $\hat{\theta} \geq 0$) dan kalibrasi aitem berdasarkan kelompok rendah (*teta terestimasi* $\hat{\theta} < 0$).

Membandingkan Teori Tes Klasik dan Teori Respon Aitem

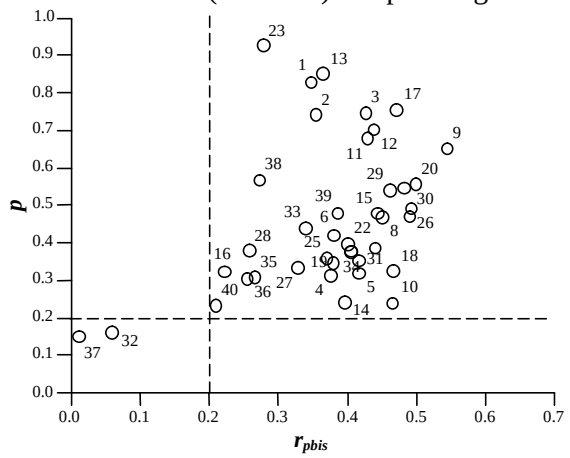
Pembandingan karakteristik aitem berdasarkan metode TTK dan TRA memiliki maksud memverifikasi teori tentang hubungan antara TTK dan TRA. Langkah-langkahnya adalah:

a. untuk melihat hubungan dalam hal tingkat kesukaran aitem melalui p ola *scatter plot* dan linieritas berupa garis regresi antara *b* (tingkat kesukaran aitem berdasarkan TRA) dan *p* (tingkat kesukaran berdasarkan TTK);
b. untuk melihat hubungan dalam hal daya beda aitem, dilihat melalui pola *scatter plot* dan linieritas berupa garis regresi antara *a* (daya beda aitem berdasarkan TRA) dan r_{pbis} (daya beda berdasarkan TTK).

HASIL DAN PEMBAHASAN

Pendekatan Teori Tes Klasik

Berdasarkan pendekatan TTK yang diterapkan, mean skor yang diperoleh adalah 18.628 dengan standar deviasi 6.910, range: 3-39. Reliabilitas berdasarkan Alpha adalah 0.844 dengan *standard error of measurement* SEM = 2.733. Tingkat kesukaran aitem *p* berkisar dari 0.152 (aitem 37) sampai dengan 0.928 (aitem 23). Sementara itu, korelasi *point biserial* r_{pbis} berkisar dari 0.011 (aitem 37) sampai dengan 0.543 (aitem 9).



Gambar 1. Korelasi *point biserial* r_{pbis} diplot dengan nilai *p* (40 aitem)

Untuk memahami lebih dalam, dibuat scatter plot antara korelasi *point biserial* r_{pbis} dan proporsi menjawab benar aitem *p*. Diagram tersebut dituangkan dalam Gambar . Sumbu horizontal menggambarkan r_{pbis} yang menunjukkan bagaimana variasi aitem-aitem dalam membedakan antar kemampuan para peserta tes. Jika diperhatikan lebih dalam, aitem nomor 37 dan 32 merupakan aitem yang bermasalah. Keduanya merupakan aitem yang sukar (p_{37} = 0.152; p_{32} = 0.163), namun memiliki daya beda yang rendah ($r_{pbis(37)}$ = 0.011; $r_{pbis(32)}$ = 0.059). Oleh karena itu, kedua aitem tersebut tidak diikutkan pada analisis selanjutnya. Hal ini dengan mengingat bahwa kedua aitem tersebut bersifat problematik sehingga menimbulkan permasalahan dalam proses

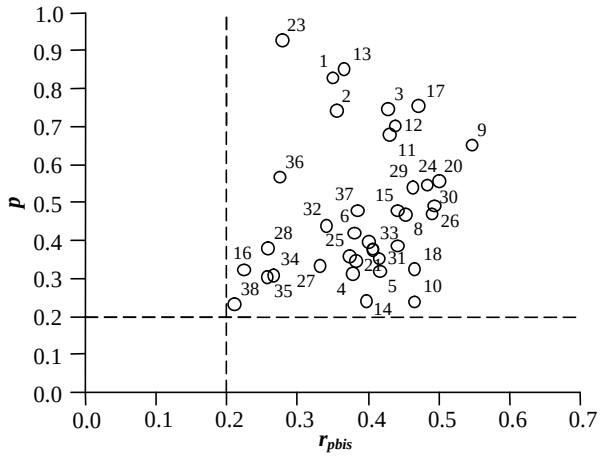
kalibrasi melalui pendekatan TRA. Informasi yang lebih detil tentang karakteristik aitem berdasarkan TTK dituangkan dalam Tabel.

Tabel 1. Nilai *p* dan r_{pbis} untuk 40 Aitem

Aitem	*p*	r_{pbis}	Aitem	*p*	r_{pbis}	Aitem	*p*	r_{pbis}
1	0.830	0.347	15	0.480	0.442	29	0.542	0.461
2	0.744	0.353	16	0.325	0.221	30	0.472	0.489
3	0.749	0.426	17	0.756	0.470	31	0.354	0.415
4	0.313	0.375	18	0.326	0.465	***32***	***0.163***	***0.059***
5	0.321	0.416	19	0.347	0.378	33	0.440	0.339
6	0.422	0.380	20	0.558	0.498	34	0.377	0.404
7	0.388	0.439	21	0.380	0.404	35	0.310	0.265
8	0.470	0.450	22	0.399	0.399	36	0.307	0.254
9	0.653	0.543	23	0.928	0.278	***37***	***0.152***	***0.011***
10	0.240	0.464	24	0.548	0.481	38	0.569	0.272
11	0.679	0.428	25	0.360	0.369	39	0.481	0.385
12	0.704	0.437	26	0.493	0.491	40	0.234	0.209
13	0.854	0.364	27	0.334	0.328			
14	0.244	0.395	28	0.381	0.257			

Keterangan: aitem yang menjadi perhatian tercetak tebal dan miring

Dieliminirnya aitem nomor 32 dan 37, menjadikan nilai-nilai *p* dan r_{pbis} model lebih rasional. Perhatikanlah Tabel 1 yang menuangkan korelasi *point biserial* r_{pbis} dengan nilai *p* pada 38 aitem. Sekarang mean skor = 18.312; standar deviasi = 6.903; skor minimal 2 dan maksimal 38. Reliabilitasnya pun meningkat menjadi 0.850 dengan SEM = 2.673. Sedangkan rentang nilai kore lasi *point biserial* r_{pbis} adalah 0.21 (aitem 40) sampai dengan 0.545 (aitem 9). Nilai *p* terentang dari 0.234 (aitem 40) sampai dengan 0.928 (aitem 23). Informasi yang lebih lengkap dituangkan dalam Tabel 1.

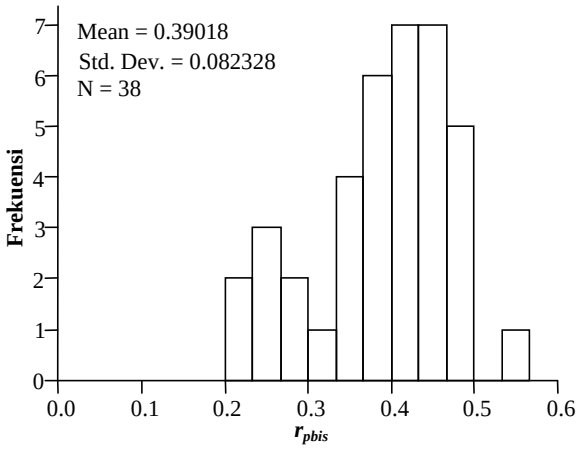


Gambar 2. Korelasi *point biserial* r_{pbis} diplot dengan nilai *p* (38 aitem)

Gambar 2. juga menunjukkan bahwa tidak ada hubungan antara daya beda dan tingkat kesukaran aitem. Kondisi ini mengarahka n pada satu keputusan bahwa kedua parameter tersebut memiliki peluang yang besar untuk dilibatkan dalam model yang dipilih. Dengan kata lain, model 2PL atau 3PL adalah alternatif pilihan yang lebih rasional yang dapat digunakan dibandingkan model 1PL. Seba ran daya beda aitem lebih jelas diamati dalam histogram yang memuat distribusi daya beda aitem pada Gambar .

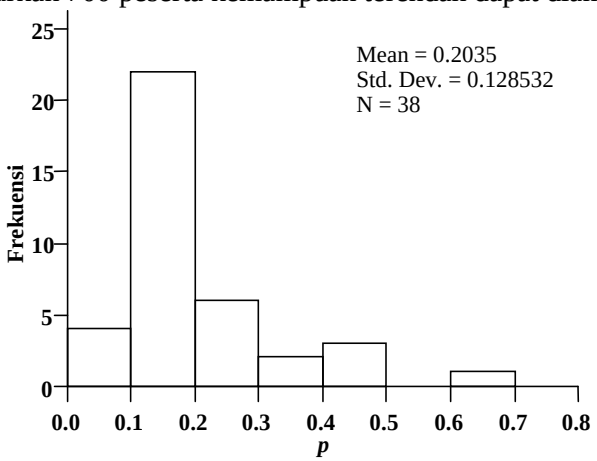
Tabel 1. Nilai *p* dan r_{pbis} untuk 38 aitem

Aitem	*p*	r_{pbis}	Aitem	*p*	r_{pbis}	Aitem	*p*	r_{pbis}
1	0.830	0.349	14	0.244	0.396	27	0.334	0.331
2	0.744	0.354	15	0.480	0.441	28	0.381	0.258
3	0.749	0.427	16	0.325	0.223	29	0.542	0.462
4	0.313	0.377	17	0.756	0.470	30	0.472	0.489
5	0.321	0.416	18	0.326	0.464	31	0.354	0.414
6	0.422	0.380	19	0.347	0.381	33	0.440	0.340
7	0.388	0.440	20	0.558	0.499	34	0.377	0.405
8	0.470	0.452	21	0.380	0.405	35	0.310	0.265
9	0.653	0.545	22	0.399	0.400	36	0.307	0.256
10	0.240	0.464	23	0.928	0.278	38	0.569	0.274
11	0.679	0.429	24	0.548	0.482	39	0.481	0.384
12	0.704	0.437	25	0.360	0.372	40	0.234	0.210
13	0.854	0.365	26	0.493	0.493			



Gambar 3. Sebaran daya beda 38 aitem

Untuk mengecek kemungkinan benarnya para peserta menjawab dengan cara menebak, maka diamati respon 10% para peserta yang memiliki kemampuan terendah (700 peserta dengan skor terendah) terhadap keseluruhan aitem. Sebaran nilai -*p* pada aitem-aitem berdasarkan 700 peserta kemampuan terendah dapat diamati pada Gambar .



Gambar 4. Sebaran nilai *p* pada 38 aitem berdasar 10% peserta skor terendah

Secara visual, tampak bahwa pada Gambar 4 histogram lebih bersifat juling ke kanan atau dengan kata lain frekuensi tinggi dimiliki oleh aite m-aitem dengan nilai-*p* yang rendah (sebelah kiri). Kondisi seperti ini mengisyaratkan bahwa sebagian besar para peserta dengan kemampuan rendah memiliki probabilitas menjawab benar dengan cara menebak, yaitu berkisar pada seputar nilai satu per banyaknya o psional jawaban (< 1/5 = 0.2). Dengan demikian, dugaan awal yang dapat ditarik berdasarkan fakta ini, yaitu model 2PL saja tidak cukup memadai untuk diterapkan, lebih baik menerapkan model 3PL.

Tes UAN pada dasarnya didesain sebagai tes yang bersifat *criterion referenced*, artinya lulus tidaknya para peserta ditentukan oleh suatu kriteria skor. Tes UAN juga dapat dikategorikan sebagai *power test* dimana waktu yang dialokasikan untuk menyelesaikan tes sudah cukup memadai. Walaupun dengan waktu yang cukup, bukan berarti para peserta telah memberikan respon dengan tanpa menebak dalam memilih jawaban benar. Selan itu, meskipun juga Tes UAN Matematika adalah *power test* di mana aspek kecepatan dalam menyelesaikan soal bukanlah salah satu aspek yang dipertimbangkan, akan tetapi dengan melihat kenyataan bahwa para peserta dengan kemampuan sangat rendah pun punya peluang yang memadai (sekitar 0.2) untuk menjawab benar dengan cara menebak maka dapat dikatakan bahwa model 3PL menjadi pilihan yang lebih rasional dibandingk an model 2PL. Informasi tentang sebaran nilai-*p* yang dihasilkan oleh para peserta yang berkemampuan rendah, secara lengkap disajikan dalam Tabel 2.

Tabel 2. Nilai *p* pada 38 aitem berdasar 10% peserta skor terendah

Aitem	*p*	Aitem	*p*	Aitem	*p*	Aitem	*p*
1	0.464	11	0.256	21	0.200	31	0.103
2	0.436	12	0.256	22	0.136	33	0.230
3	0.316	13	0.461	23	0.689	34	0.126
4	0.134	14	0.074	24	0.130	35	0.173
5	0.097	15	0.156	25	0.154	36	0.167
6	0.179	16	0.173	26	0.116	38	0.363
7	0.103	17	0.260	27	0.149	39	0.173
8	0.166	18	0.093	28	0.224	40	0.120
9	0.177	19	0.136	29	0.174		
10	0.071	20	0.151	30	0.147		

Tabel 3. Uji *Godness of Fit* Model

m	Model	G_m	Selisih ($G_m - G_{m+1}$)	Nilai Kritik χ^2 (5%, 38)	Keterangan
1	1PL	190405.0	-	-	-
2	2PL	187343.0	3062.0	53.384	2PL lebih informatif dibanding 1PL
3	3PL	186182.4	1160.6	53.384	3PL lebih informatif dibanding 2PL

Keterangan:
m = jumlah parameter tiap aitem
G = -2 loglikelihood

TRA adalah teori pengukuran berdasarkan model. Oleh karena itu perlu diadakan uji terhadap model yang diajukan. Untuk menguji dugaan awal tentang dipilihnya model 3PL, dilakukan uji kecocokan model. Dilihat dari uji kecocokan seluruh d ata respon

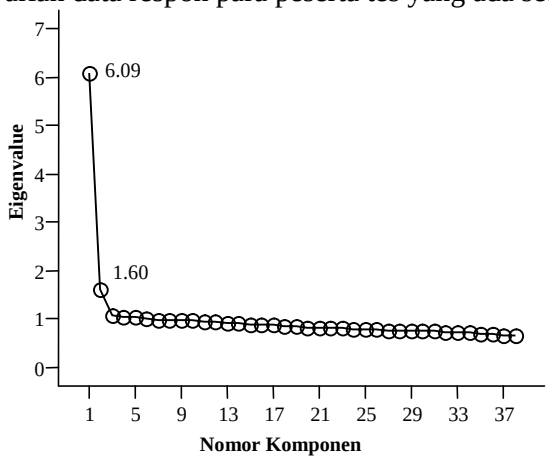
para peserta tes dengan model yang dipilih, model 3PL ternyata lebih mampu memberikan penjelasan secara lebih informatif dibanding dengan model 2PL. Artinya parameter peluang tebakan semu *c*—sebagai informasi tambahan setelah parameter daya beda *a* dan tingkat kesukaran *b*—memberikan kontribusi signifikan dalam menjelaskan data pola respon yang dimiliki para peserta tes. Secara statistik, hal ini terbukti dengan hasil uji kecocokan data atau *goodness of fit* dalam Tabel 3.

Pendekatan Teori Respon Aitem

Verifikasi Asumsi

Unidimensionalitas

Berdasarkan hasil analisis faktor, terdapat 6 nilai *eigenvalue* yang nilainya lebih dari 1. Secara lebih jelas dapat diperhatikan Gambar 5 dimana di dalamnya terdapat plot nomor komponen hasil ekstraksi dan nilai *eigenvalue*. Keenam faktor yang dominan ini mampu menjelaskan varian data respon para peserta tes yang ada sebesar 31.207%.



Gambar 5. Eigenvalue dari Analisis Faktor

Meski hanya 31.207%, jika diperhatikan lebih jauh, faktor pertama yang memiliki nilai *eigenvalue* sebesar 6.095 mampu menjelaskan varian sebesar 16.093%, paling dominan dibandingkan faktor yang lain. Dalam istilah lain dapat juga dikataka n terdapat satu faktor dominan yang mendasari para peserta memberikan respon pada aitem -aitem tes. Dominansi faktor pertama ini mampu memberi dukungan tentang bukti unidimensionalitas data respon yang dimiliki, di mana terdapat sebuah *latent traits* yang mendasari perilaku para peserta tes. *Latent traits* ini dapat disebut sebagai kemampuan matematika. Besarnya varian yang dapat dijelaskan masing -masing faktor tersebut tertuang dalam Tabel 4.

Tabel 4 Nilai Eigenvalue 6 Faktor dan %Varian yang Dijelaskan

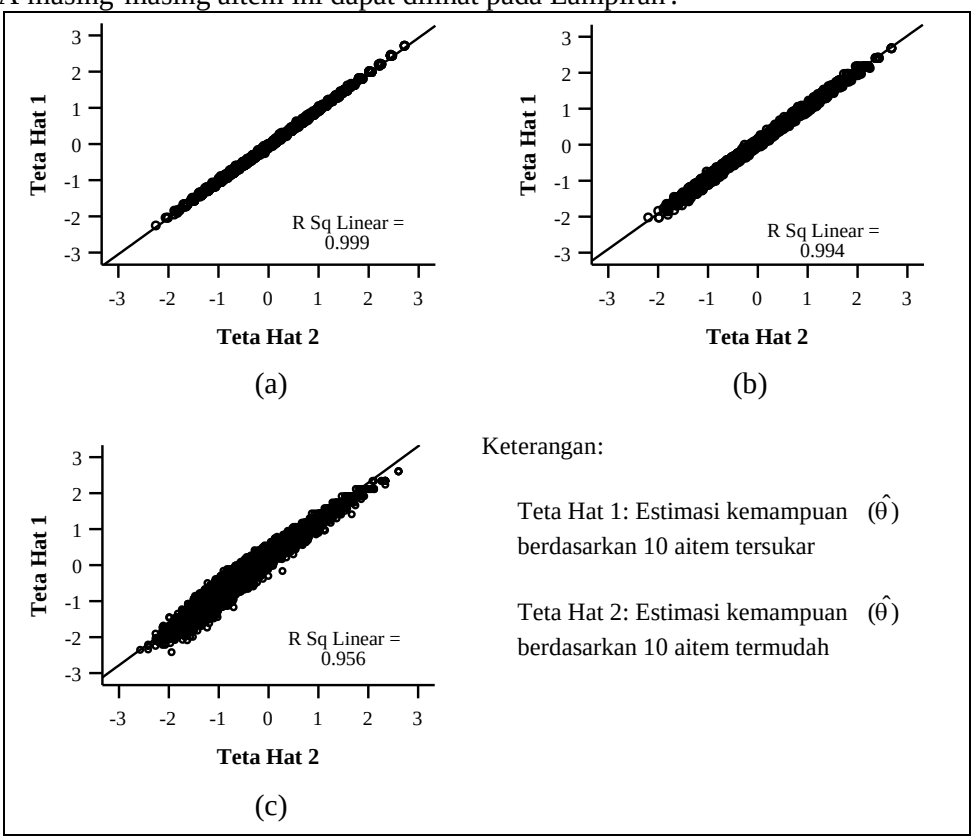
Komponen	Nilai Eigenvalue	% varian	Kumulatif % varian
1	6.095	16.039	16.039
2	1.597	4.204	20.243
3	1.074	2.826	23.070
4	1.051	2.765	25.835
5	1.040	2.737	28.572
6	1.001	2.635	31.207

Independensi Lokal

Independensi lokal berarti respon peserta terhadap sebuah aitem dan aitem yang lain bersifat independen setelah *latent traits* dikontrol (Hambleton, Swaminathan, & Rogers, 1991; Karabatsos & Sheu, 2004). *Latent traits* yang dimaksud di sini adalah kemampuan matematika. Dominansi satu faktor yang ada berdasarkan analisis faktor telah mengarahkan pada terpenuhinya bukti bahwa data yang dimiliki bersifat unidimensional, hanya terdapat satu faktor yang mempengaruhi para peserta untuk berperilaku. Berdasarkan fakta ini, dapat disebutkan juga bahwa karena data yang dimiliki bersifat unidimensional, maka respon yang diberikan para peserta tes bersifat independen, kondisional terhadap kemampuan mereka masing-masing. Jika kemampuan para peserta tes sudah diketahui, maka perilaku respon terhadap satu aitem tidak berpengaruh terhadap perilaku respon terhadap aitem yang lain.

Kurva Karakteristik Aitem

Asumsi ketiga dalam TRA yaitu masing-masing aitem memiliki kurva karakteristik aitem (KKA) atau *item characteristic curves* (ICC) yang mampu menggambarkan kinerja peserta yang memiliki kemampuan tertentu dengan probabilitas menjawab benar pada aitem yang dimaksud. Hal ini dapat dilakukan dengan menggambarkan masing-masing KKA aitem berdasarkan parameter-parameter yang dimiliki. Selengkapnya plot KKA masing-masing aitem ini dapat dilihat pada Lampiran.



Gambar 6. *Scatter plot* dan Garis Regresi antara Estimasi Kemampuan berdasarkan 10 aitem termudah dan 10 aitem tersukar pada Model: (a) 1PL , (b) 2PL, dan (c) 3PL

Harapan Terhadap Model

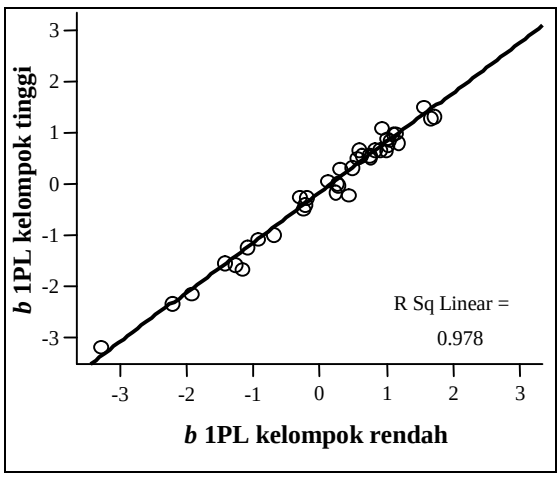
Invariansi Estimasi Kemampuan

Invariansi estimasi kemampuan berarti estimasi kemampuan tidak akan terpengaruh oleh kelompok aitem mana yang digunakan. Untuk menyelidiki invariansi estimasi kemampuan peserta tes, aitem-aitem dibagi menjadi dua, yaitu: satu kelompok 10 aitem termudah, dan satu kelompok 10 aitem yang tersukar. Pengelompokan ini didasarkan pada tingkat kesukaran aitem pada masing-masing model (1PL, 2PL, dan 3PL). Estimasi kemampuan para peserta berdasarkan kedua kelompok aitem tes tersebut kemudian diplot satu sama lain. Hasilnya dapat dilihat pada **Error! Reference source not found.**. Dengan melihat gambar tersebut, tampak bahwa estimasi kemampuan bersifat invarian berdasarkan aitem-aitem mudah atuapun aitem-aitem sukar.

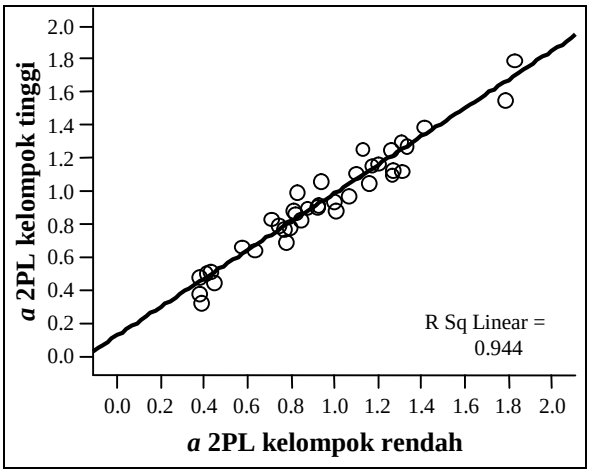
Invariansi Parameter Aitem

Sesuai dengan tujuan penelitian ini, akan diuji invariansi estimasi parameter aitem: (1) pada model 1PL, yaitu parameter tingkat kesukaran *b*; (2) pada model 2PL, yaitu parameter tingkat kesukaran *b* dan daya beda *a*; dan (3) pada model 3PL, yaitu tingkat kesukaran *b*, daya beda *a*, dan peluang tebakan semu *c*. Untuk itu respon para peserta dikelompokkan menjadi dua, yaitu: kelompok rendah dan kelompok tinggi. Kelompok rendah merupakan kelompok yang memiliki *teta terestimasi* atau *teta hat* $\hat{\theta} < 0$, dan kelompok tinggi adalah kelompok dengan $\hat{\theta} \geq 0$, kondisional pada masing-masing model.

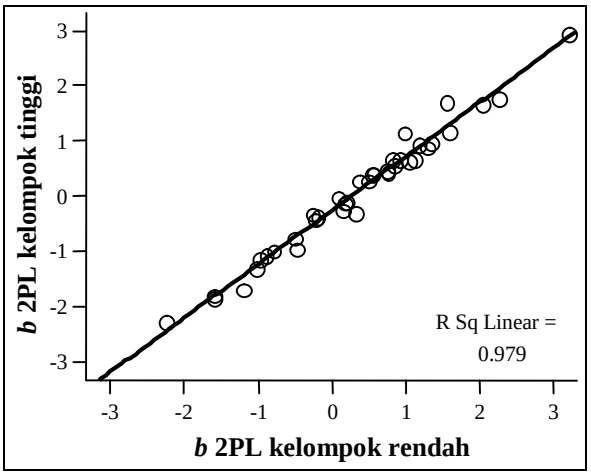
Hasil plot estimasi parameter aitem berdasar kelompok kemampuan tinggi dan kelompok kemampuan rendah dapat dilihat pada **Error! Reference source not found.**, Gambar 7, Gambar 8, Gambar 9, Gambar 10, Gambar 11, dan Gambar 12. Ketujuh gambar mengisyaratkan invariansi parameter aitem pada ke tiga model. Artinya estimasi parameter aitem tidak tergantung pada subjek-subjek mana yang digunakan untuk proses kalibrasinya.



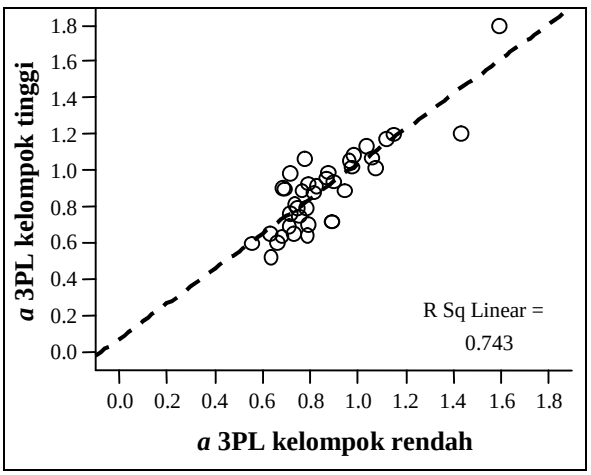
Gambar 7. *Scatter plot* dan Garis Regresi Estimasi *b* Model 1PL dengan mengontrol *a*



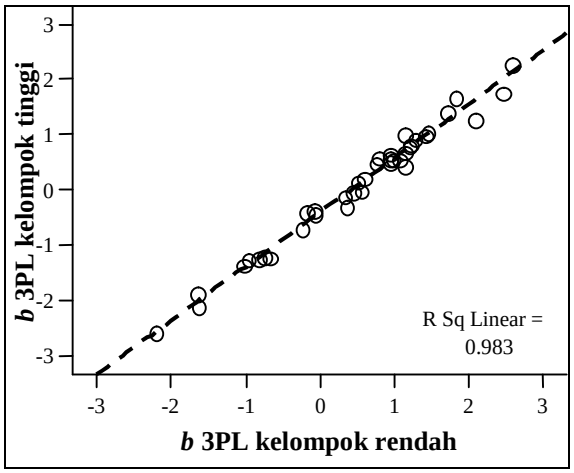
Gambar 8. ***Scatter plot*** **dan Garis Regresi Estimasi** ***a*** **Model 2PL dengan mengontrol** ***b***



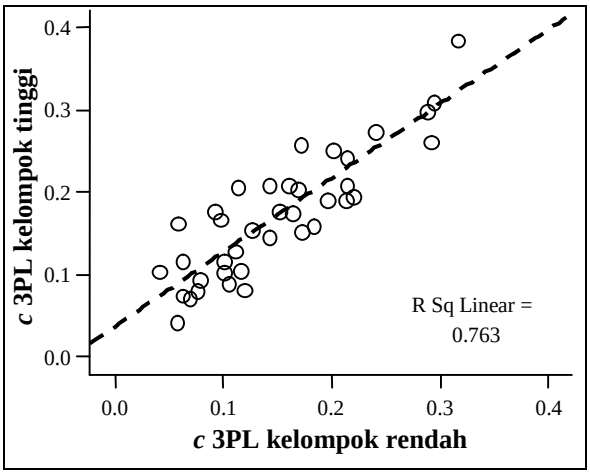
Gambar 9. ***Scatter plot*** **dan Garis Regresi Estimasi** ***b*** **Model 2PL dengan mengontrol** ***a***



Gambar 10. ***Scatter plot*** **dan Garis Regresi Estimasi** ***a*** **Model 3PL dengan mengontrol** ***b*** **dan** ***c***



Gambar 11. Scatter plot dan Garis Regresi Estimasi *b* Model 3PL dengan mengontrol *a* dan *c*



Gambar 12. *Scatter plot* dan Garis Regresi Estimasi *c* Model 3PL dengan mengontrol *a* dan *b*

Prediksi Model terhadap Hasil Tes

Goodness of Fit (GOF)

TRA merupakan pemodelan terhadap respon-respon para peserta tes. Berdasarkan model yang diajukan, model manakah yang paling mampu menjelaskan data respon tersebut? Oleh sebab itu perlu diadakan uji kecocokan data dengan model yang diajukan.

Sementara itu, uji kecocokan data atau *goodness of fit* (GOF) sangat tergantung dari ukuran sampel yang digunakan. Makin besar sampel pada level aitem, makin sensitif hasil uji tersebut sehingga hipotesis nol akan cenderung ditolak (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991) . Oleh karena belum adanya kesepakatan dari para ahli dalam menentukan GOF pada level aitem, maka penulis mengikuti saran mereka untuk menggunakan rasionalisasi dengan mendasarkan pada tujuan, format serta administrasi tes dalam memilih model yang digunakan.

Tes UAN merupakan *power* tes, dikembangkan untuk mengukur kinerja aktual berupa hasil belajar, dimana waktu yang dialokasikan sudah memadai untuk menyelesaikan seluruh aitem (40 aitem), serta format tes berbentuk *multiple choice*.

Dengan mendasarkan pada pertimbangan tersebut serta mengingat sampel yang digunakan adalah sampel yang besar (7000 peserta), maka model yang paling tepat adalah model 3PL di mana di dalamnya mengandung parameter peluang tebakan semu *c* sehingga terdapat parameter yang mampu menjelaskan probabilitas menjawab benar dengan cara menebak.

Fungsi Informasi dan Simpangan Baku

Sejauhmana masing-masing model tersebut memberikan informasi dijelaskan oleh fungsi informasi atau *information function* (IF) (Veerkamp & Berger, 1999). Dapat diperhatikan bahwa IF adalah sebuah fungsi sampai sejauh manakah model yang dipilih (1PL, 2PL, atau 3PL) mampu memberikan informasi tentang estimasi *traits-level* sepanjang skala *latent-traits*. Semakin tinggi puncak IF, makin informatif pula model yang dipilih mampu menjelaskan *traits-level* para peserta tes. Oleh karena itu, simpangan baku pengukuran atau *standard error of measurement* (SEM) merupakan fungsi yang berkebalikan dengan IF. Makin tinggi IF, makin rendah SEM.

Secara matematis, fungsi informasi aitem (IF) memenuhi persamaan:

$$I_i(\theta) = \frac{(P'_i(\theta))^2}{P_i(\theta) + (1 - P_i(\theta))}$$

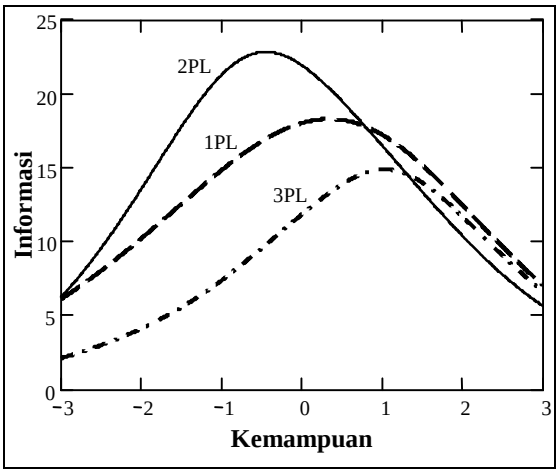
Sebagai akumulasi keseluruhan fungsi informasi aitem, maka akan diperoleh fungsi informasi tes atau *test information* (TI), yang secara matematis formulanya adalah:

$$TI(\theta) = \sum I_i(\theta)$$

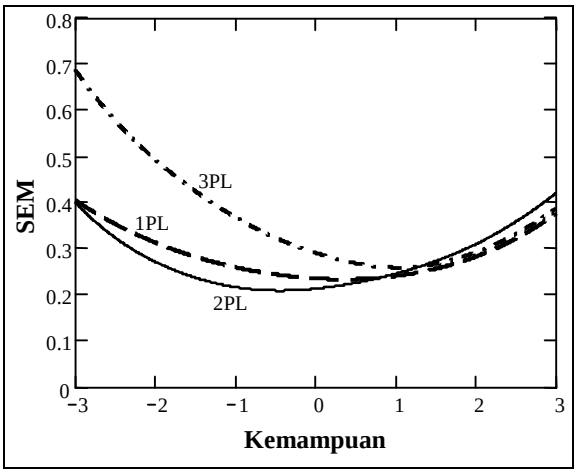
sedangkan SEM dapat dihitung untuk tiap-tiap kemampuan, θ, dengan formula

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta)}}$$

Perbandingan IF dan SEM yang mampu ditunjukkan oleh masing-masing model pada data respon para peserta tes UAN tertuang dalam Gambar 13 dan Gambar 14. Melalui gambar tersebut tampaklah bahwa dapat diurutkan puncak IF dari rendah menuju tinggi adalah: IF model 2PL, IF model 1PL, dan IF model 3PL. Melihat kenyataan seperti ini, model 2PL ternyata mampu memberikan informasi lebih tinggi dibanding model 1PL dan 3PL. Artinya, model 2PL dapat memberikan informasi yang lebih baik tentang hubungan antara pola respon para peserta tes dengan keseluruhan karakteristik masing-masing aitem. Hal ini pada gilirannya juga berimplikasi pada kepresisian estimasi kemampuan para peserta tes di mana makin tinggi IF maka makin presisi sebuah model dalam mengestimasi kemampuan para peserta.



Gambar 13. Fungsi Informasi berdasarkan Model 1PL, 2PL, dan 3PL



Gambar 14. Fungsi *Standard Error of Measurement* (SEM) berdasarkan Model 1PL, 2PL, dan 3PL

Tingkat presisi yang tinggi ini dapat dilihat pula dengan melandaskan pada SEM. Lihatlah Gambar 14, fungsi SEM model 2PL memiliki puncak terendah dibanding dua model yang lain sehingga dapat dikatakan bahwa model 2PL adalah model yang paling presisi dalam mengestimasi kemampuan para peserta tes.

IF merupakan salah satu kunci dalam mengambil keputusan tentang model mana yang digunakan, karena berdasarkan IF pula dapat diplot sebuah fungsi SEM. SEM inilah yang menentukan tingkat presisi hasil estimasi kemampuan para peserta tes. Mengingat tujuan akhir tes kemampuan aktual (dalam hal ini UAN) adalah menentukan perbedaan antar peserta, maka dapat disimpulkan model 2PL adalah model yang paling tepat. Oleh karena itu, pada pembahasan selanjutnya, model TRA yang digunakan adalah model 2PL.

Perbandingan TTK dan TRA

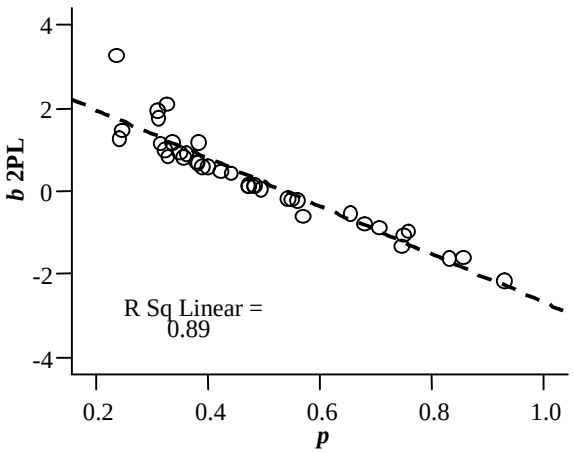
Ringkasan hasil analisis psikometrik berdasarkan TTK dapat dilihat kembali pada Tabel 1, sedangkan untuk TRA dituangkan dalam Tabel 5.

Tabel 5. Nilai *p* dan r_{pbis} untuk 38 aitem

Aitem	*a*	*b*	Aitem	*a*	*b*	Aitem	*a*	*b*

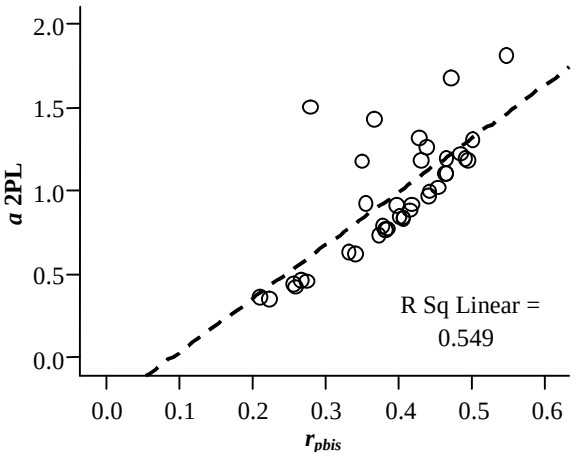
1	1.185	-1.598	14	.919	1.485	27	1.436	1.204
2	.929	-1.308	15	1.185	.133	28	.919	1.202
3	1.322	-1.042	16	.929	2.127	29	1.111	-.159
4	.794	1.176	17	1.322	-.945	30	1.200	.141
5	.925	1.004	18	.794	.857	31	.891	.827
6	.773	.502	19	.925	.953	33	.625	.446
7	.973	.596	20	.773	-.206	34	.846	.723
8	1.026	.167	21	.973	.713	35	.471	1.779
9	1.820	-.507	22	1.026	.605	36	.445	1.965
10	1.200	1.274	23	1.820	-2.154	38	.465	-.590
11	1.187	-.772	24	1.200	-.181	39	.779	.147
12	1.269	-.863	25	1.187	.910	40	.372	3.288
13	1.436	-1.580	26	1.269	.063			

Parameter tingkat kesukaran dalam TRA yang ditunjukkan dengan *b*, mengacu pada titik di sepanjang skala kemampuan dimana probabilitas menjawab benar adalah 0.5. Sementara pada TTK, parameter tingkat kesukaran ditunjukkan dengan proporsi menjawab benar *p* yang lebih mudah dimaknai sebagai tingkat kemudahan. Oleh karena itu, secara teoritik korelasi *b* dan *p* akan bersifat negatif. Gambar 15 menampilkan hubungan korelasi negatif tersebut dimana terlihat bahwa r_{bp}^2 = 0.89 atau r_{bp} = -0.943.



Gambar 15. Scatter plot dan Garis Regresi antara *b*-2PL dan *p*

Parameter daya beda dalam TRA ditunjukkan dengan *a*, yang pada dasarnya merupakan ukuran kemiringan *item characteristic curve* (ICC) pada masing-masing aitem. Dalam TTK daya beda aitem ditunjukkan oleh korelasi *point biserial* r_{pbis}, yaitu korelasi aitem-total atau tepatnya korelasi antara variabel dikotomi (aitem) dan variabel kuantitatif (skor total). Secara teoritik, hubungan antara *a* dan r_{pbis} adalah linier positif. Gambar 16 menunjukkan eksisnya hubungan tersebut, dimana $r_{(a)(pbis)}^2$ = .549 atau $r_{(a)(pbis)}$ = 0.741.



Gambar 16. Scatter plot dan Garis Regresi antara *a*-2PL dan r_{pbis}

SIMPULAN

Penelitian ini bertujuan menyelidiki karakteristik psikometrik tes UAN Matematika SMA baik pada level tes maupun pada level aitem. Tujuan utama penelitian ini adalah membandingkan model 1PL, 2PL, dan 3PL dalam TRA untuk kemudian dipilih model yang paling cocok. Lebih jauh, dilakukan pembandingan pula dengan TTK. Evaluasi menggunakan TTK menunjukkan bahwa tes UAN Matematika SMA memiliki reliabilitas internal sebesar 0.850. Tingkat kesukaran *p* terentang dari 0.234 sampai dengan 0.928 dan daya beda r_{pbis} terentang dari 0.210 sampai dengan 0.545. Satu kesimpulan penting yang dapat ditarik adalah bahwa masing-masing aitem memiliki daya beda yang berbeda-beda.

Evaluasi melalui pendekatan TRA didasarkan pada tiga kriteria. Kriteria pertama yaitu memverifikasi asumsi model. Hasil analisis faktor pada Gambar 5 menunjukkan bahwa terdapat satu faktor dominan yang mendasari para peserta dalam merespon keseluruhan aitem tes UAN. Daya beda yang tidak sama pada masing-masing aitem mengarahkan pada kesimpulan bahwa model 2PL atau 3PL lebih tepat digunakan dibanding model 1PL. Kemungkinan menjawab benar dengan cara menebak para peserta dengan kemampuan rendah mengarahkan pada kesimpulan bahwa model 3PL adalah model lebih baik dibanding model 2PL.

Kriteria kedua adalah sejauh mana harapan terhadap model yang dapat dipenuhi. **Error! Reference source not found.** mendeskripsikan bagaimana kinerja para peserta tes pada aitem-aitem yang mudah dan aitem-aitem yang sukar. Ketiga gambar tersebut mengisyaratkan bahwa estimasi kemampuan bersifat invarian pada model 1PL, 2PL, dan 3PL.

Gambar 7, Gambar 9, dan Gambar 11 menunjukkan bahwa estimasi tingkat kesukaran bersifat invarian pada ketiga model (1PL, 2PL, dan 3PL). Selanjutnya, Gambar 8 dan Gambar 10 mengarahkan pada kesimpulan bahwa estimasi daya beda juga bersifat invarian pada model 2PL dan 3PL. Lalu, Gambar 12 yang memuat sebaran estimasi parameter *c* dapat dijadikan dasar untuk mengatakan bahwa parameter peluang tebakan juga bersifat invarian pada model 3PL. Hal ini mengarahkan peneliti untuk memilih model 3PL.

Invariansi estimasi parameter-parameter aitem pada ketiga model di atas menunjukkan bahwa estimasi parameter aitem tidak tergantung sampel, dan estimasi kemampuan tidak tergantung pada aitem. Manfaat adanya sifat invarian yang dimiliki TRA tersebut akan tampak nyata manakala sebuah tes digunakan secara berulang kali pada kelompok sampel yang berbeda-beda.

Kriteria ketiga yaitu kecocokan data dengan model yang dipilih. Tabel 35 yang merangkum uji *Goodness of Fit* (GOF) dengan cara membandingkan ketiga model, mengarahkan peneliti untuk lebih menentukan model 3PL sebagai pilihan.

Akhirnya, *information function* (IF) dan SEM ketiga model dibandingkan. Keduanya tertuang pada Gambar 13 dan Gambar 14. Secara umum dapat dilihat bahwa model 2PL lebih mampu memberikan informasi dibandingkan model 1PL dan 3PL. SEM 2PL secara umum juga lebih rendah. Oleh karena itu, 2PL merupakan preferensi dibanding 1PL dan 3PL.

Mengingat tujuan terpenting sebuah tes adalah mengukur perbedaan para peserta tes dengan eror yang seminimal mungkin, maka IF dan SEM layak untuk dijadikan pertimbangan paling utama dalam menentukan sebuah model yang dipilih. Berdasarkan IF dan SEM yang dihasilkan masing-masing model maka dapat disimpulkan bahwa model 2PL adalah model yang paling tepat digunakan dalam menjelaskan data respon para peserta UAN Matematika.

Bagian akhir penelitian ini adalah membandingkan TTK dan TRA. Dari pembandingan tersebut dapat disimpulkan bahwa hasil estimasi kedua pendekatan tersebut sesuai dengan teori (Crocker & Algina, 1986). Daya beda aitem berkorelasi secara linier positif (Gambar 16), sedangkan tingkat kesukaran berkorelasi secara linier negatif (Gambar 15).

DAFTAR PUSTAKA

Baker, F. B. (2001). *The Basics of Item Response Theory.* New York: ERIC Clearinghouse on Assessment and Evaluation.

Bechger, T. M., Maris, G., Verstralen, H. H., & Beguin, A. A. (2003). Using Classical Test Theory in Combination With Item Response Theory. *Applied Psychological Measurement, 27* (5), 319–334.

Crocker, L. M., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* New York: Holt, Rinehart and Winston Inc.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologist.* NJ: Lawrence Erlbaum Associates Inc.

Fan, X. (1998). Item Response Theory and Classical Test Theory: An Empirical Comparison of Their Item/Response Person Statistics. *Educational and Psychological Measurement, 58* (3), 357-381.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Application.* Boston, MA: Kluwer Inc.

Hambleton, R. K., Robin, F., & Xing, D. (2000). Item Response Models for the Analysis of Educational and Psychological Test Data. Dalam H. E. Tinsley, & S. D. Brown, *Handbook of applied multivariate statistics and mathematical modeling* (hal. 553-581). San Diego, CA: Academic Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* CA: Sage Publication Inc.

Harvey, R. J., & Hammer, A. L. (1999). Item Response Theory. *The Counseling Psychologist, 27* (3), 353-383.

Karabatsos, G., & Sheu, C.F. (2004). Order-Constrained Bayes Inference for Dichotomous Models of Unidimensional Nonparametric IRT. *Applied Psychological Measurement, 28* (2), 110–125.

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problems.* Hillsdale, New Jersey: Lawrence Erlbaum Associates Publishers.

McDonald, P., & Paunonen, S. V. (2002). A Monte Carlo Comparison of Item and Person Statistics Based on Item Response Theory Versus Classical Test Theory. *Educational and Psychological Measurement, 62* (6), 921-943.

McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian Method for the Detection of Item Preknowledge in Computerized Adaptive Testing. *Applied Psychological Measurement, 27* (2), 121–137.

Nicewander, W. A. (1993). Some relationships between the information function of IRT and the signal/noise ratio and reliability coefficient of classical test theory. *Psychometrika, 58*, 139-141.

Ridho, A. (2005). *Keberfungsian Item Tes UAN Matematika SMA Tahun Pelajaran 2003/2004 di Propinsi DIY.* Yogyakarta: Sekolah Pascasarjana Universitas Gadjah Mada. Tesis. Tidak Diterbitkan.

Risnawita, R. S. (2004). *Karakteristik Butir Soal Tes Masuk Seleksi SLTPN 8 di Kotamadya Jogjakarta Tahun Ajaran 2001/2002 Berdasar Teori Respon Butir Model Logistik Tiga Parameter.* Yogyakarta: Program Pascasarjana Universitas Gadjah Mada. Tesis. Tidak Diterbitkan.

Stone, C. A. (2003). Empirical Power and Type I Error Rates for An IRT Fit Statistic That Considers the Precision of Ability Estimates. *Educational and Psychological Measurement, 63* (4), 566-583.

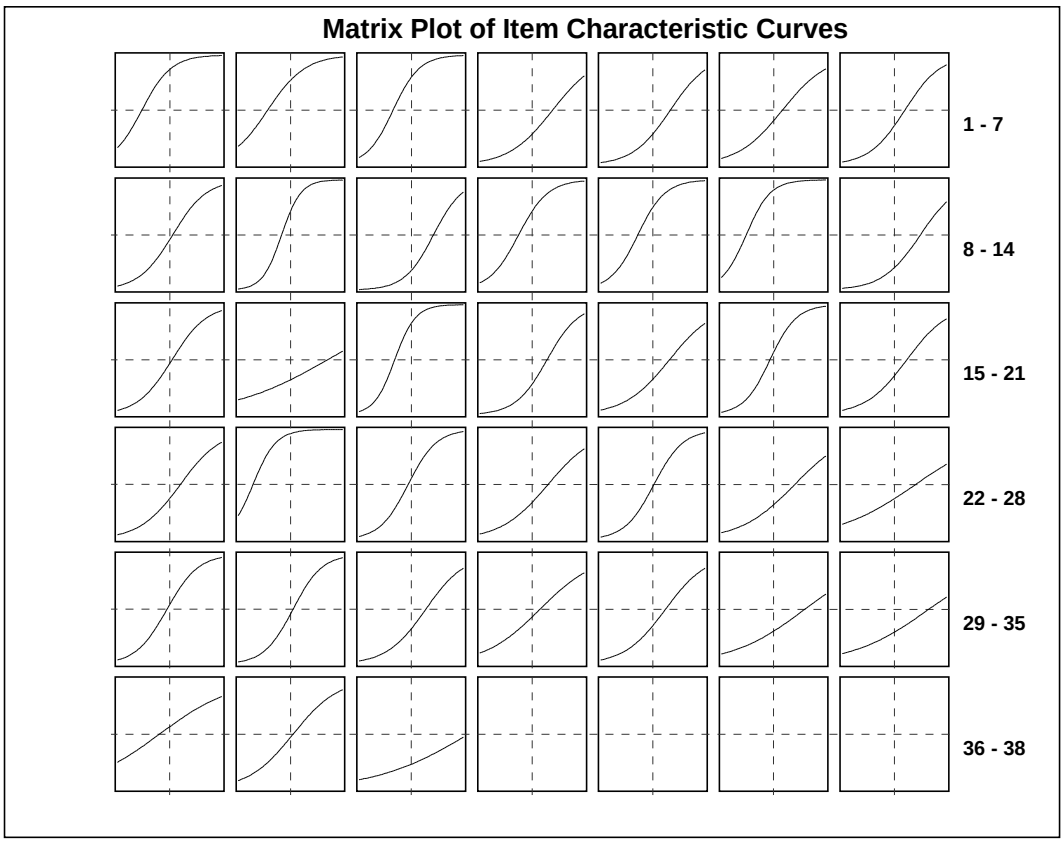
Suryabrata, S. (2000). *Pengembangan Alat Ukur Psikologi.* Yogyakarta: Andi.

Thissen, D. (2003). MULTILOG. Dalam M. du Toit, *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT* (hal. 345-409). North Lincoln: Scientific Software International.

Veerkamp, W. J., & Berger, M. P. (1999). Optimal Item Discrimination and Maximum Information for Logistic IRT Models. *Applied Psychological Measurement, 23* (1), 31–40.

Xing, D., & Hambleton, R. K. (2004). Impact of Test Design, Item Quality, and Item Bank Size on the Psychometric Properties of Computer-Based Credentialing Examinations. *Educational and Psychological Measurement, 64* (1), 5-21.

Lampiran


Matrix Plot of Item Characteristic Curves
1 - 7
8 - 14
15 - 21
22 - 28
29 - 35
36 - 38