Widyadari DOI: 10.5281/zenodo.5575879
Vol. 22 No. 2 (Oktober 2021)
e-ISSN : 2613-9308 p-ISSN : 1907-3232
Hlm. 608 - 619

KARAKTERISTIK TES PRESTASI BELAJAR BERDASARKAN PENDEKATAN KLASIK DAN *ITEM RESPONSE THEORY*

Anak Agung Purwa Antara[1*], Ni Made Serma Wati[2]
[1,2]IKIP Saraswati
purwa.antara@gmail.com ; serma.wati93@gmail.com

ABSTRACT

This research is empirical study in order to determine (1) the characteristics of achievement test which analyzed with classical approach, (2) the characteristics of achievement test which analyzed with Item Response Theory (IRT) approach and, (3) the comparison of slope that analyzed with classical and IRT approaches. This research was conducted in an elementary school in Tabanan with data retrieval using stratified random sampling technique. Characteristics of the test analyzed using Parscale Program (Muraki & Bock, 1977) with Marginal Maximum likelihood estimation. The results of the analyzed based on the classical approach shows that the average value of coefficient Pearson correlation 0.373 and the average of coefficient Polyserial correlation 0.460 more than 0.2 That means the general tests that arranged have good slope.. The results of analysis IRT approach indicated that the minimum value of probability 0.198 (greater than 0.05), which means fit with the model. The mean of slope 0.644, greater than 0.2. That means the tests have good slope. Similarly, the value of location of the test -.0.430, that means the tests have moderate level of difficulty. Analysis of test with IRT approach have an average slope greater than tests that analyze with classical approach. That means the analysis test with IRT approach more careful in distinguishing abilities of students with one another.

Keywords: *Slope, Classical Approach, IRT Approach*

ABSTRAK

Penelitian ini adalah penelitian empirik dengan tujuan untuk mengetahui (1) karakteristik tes prestasi belajar yang dianalisis dengan pendekatan teori tes klasik, (2) karakteristik tes prestasi belajar yang dianalisis dengan pendekatan Item Response Theory (IRT), (3) perbandingan tingkat daya beda butir tes prestasi belajar yang dianalisis dengan pendekatan Klasik dan IRT. Penelitian ini dilakukan di Sekolah Dasar di Kabupaten Tabanan dengan pengambilan data menggunakan teknik stratified random sampling. Karakteristik tes yang disusun dianalisis menggunakan Program Parscale dengan estimasi Marginal Maximum likelihood. Hasil analisis berdasarkan pendekatan klasik menunjukkan bahwa nilai rerata koefisien korelasi Pearson 0.373 dan rerata koefisien korelasi Polyserial 0.460 lebih dari 0.2 yang berarti secara umum tes yang disusun memiliki daya beda yang baik. Hasil analisis dengan pendekatan IRT menunjukkan bahwa nilai minimum probability 0.198 (lebih dari 0.05) yang berarti fit dengan model. Rerata slope 0.644, lebih besar dari 0.2. Hal ini berarti tes yang disusun memiliki daya beda butir yang baik. Demikian pula nilai location dari tes sebesar -.0.430, yang berati tes yang disusun memiliki tingkat kesukaran butir yang sedang. Analisis tes dengan pendekatan IRT memiliki rerata slope lebih besar dari tes yang dianalisis dengan pendekatan klasik. Hal ini berarti analisis tes dengan pendekatan IRT lebih teliti dalam membedakan kemampuan siswa yang satu dengan yang lainnya.

***Kata kunci*:** Daya beda, Pendekatan Klasik, Pendekatan IRT

PENDAHULUAN

Fokus utama penyelenggaraan pendidikan nasional seperti diamanatkan oleh Undang-Undang Sistem Pendidikan Nasional No 20 Tahun 2003 adalah peningkatan kompetensi lulusan dari suatu lembaga pendidikan. Peningkatan kompetensi lulusan tersebut tidak terlepas dari aspek materi yang harus dipelajari atau diajarkan, proses pembelajaran (meliputi metode atau strategi yang digunakan dalam pembelajaran sehingga materi atau pengetahuan yang mesti dikuasai siswa dapat diserap secara maksimal oleh siswa) dan evaluasi terhadap program pembelajaran yang telah dilakukan sehingga diperoleh informasi yang akurat apakah tujuan pembelajaran telah tercapai sesuai dengan harapan.

Menurut Mardapi (2005:5) peningkatan kompetensi lulusan dari suatu lembaga pendidikan tersebut dapat dimulai dari peningkatan kualitas program pembelajaran dan peningkatan kualitas penilaian yang dilakukan guru di dalam kelas. Pernyataan tersebut dapat dimaknai bahwa penilaian yang berkualitas akan mampu memberikan informasi yang akurat tentang program pembelajaran yang telah dilakukan, apakah tujuan pembelajaran yang diharapkan telah dapat dicapai atau tidak. Informasi yang akurat akan berdampak pada pengambilan keputusan yang tepat dalam rangka perbaikan kualitas hasil belajar atau kompetensi lulusan.

Penilaian hasil belajar pada suatu program pembelajaran tidak terlepas dari program pengukurannya. Artinya penilaian yang berkualitas hanya dapat diperoleh melalui hasil pengukuran yang berkualitas. Pengukuran yang berkualitas memerlukan instrument atau alat ukur yang berkualitas pula. Untuk mendapatkan alat ukur atau tes yang berkualitas diperlukan analisis yang akurat dan cermat. Analisis tes selain dilakukan secara teori yang meliputi telaah butir berdasarkan aspek isi, konstruksi, dan bahasa, perlu juga dilakukan analisis butir secara empirik.

Analisis butir secara empirik dapat dilakukan dengan dua pendekatan yaitu pendekatan teori tes klasik dan pendekatan teori respons butir (*Item Response Theory*, *IRT*). Teori tes klasik (*classical test theory;* CTT) berkembang dan digunakan secara luas di Indonesia dan menjadi teori utama di kalangan ahli psikologi dan pendidikan, serta bidang kajian perilaku (*behavioral*) yang lain, selama 20 dekade (Embretson & Reise, 2000). Teori tes klasik atau disebut teori skor murni klasik (Allen & Yen, 1979:57) didasarkan pada suatu model aditif, yakni skor amatan merupakan penjumlahan dari skor sebenarnya dan skor

kesalahan pengukuran. Secara matematis dituliskan dengan: X = T + E, di mana X adalah skor amatan, T skor murni, dan E skor kesalahan pengukuran

Analisis tes menggunakan teori tes klasik memiliki kelemahan karena bersifat *examinee sample dependent* dan *item sample dependent* (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991). Artinya statistik yang digunakan dalam model tes klasik seperti tingkat kesukaran dan daya pembeda soal sangat tergantung pada sampel yang dipergunakan dalam analisis. Rerata tingkat kemampuan, rentang, dan sebaran kemampuan siswa yang dijadikan sampel dalam analisis sangat mempengaruhi nilai statistik yang diperoleh. Sebagai contoh, tingkat kesukaran soal akan tinggi apabila sampel yang akan digunakan mempunyai kemampuan lebih tinggi dari rerata kemampuan siswa dalam populasinya. Daya pembeda soal akan tinggi apabila tingkat kemampuan sampel bervariasi atau mempunyai rentang kemampuan yang besar. Demikian pula dengan reliabilitas tes. Skor siswa yang diperoleh dari suatu tes sangat terbatas pada tes yang digunakan. Kesimpulan hasil tes tidak dapat digeneralisasikan di luar tes yang digunakan. Skor perolehan seseorang sangat tergantung pada pemilihan tes yang digunakan bukan pada kemampuan peserta tes tersebut. Karena keterbatasan penggunaan skor tes, teori tes klasik tidak mempunyai dasar untuk mempelajari perkembangan kemampuan siswa dari waktu ke waktu, kecuali jika siswa tersebut menempuh tes yang sama dari waktu ke waktu. Konsep reliabilitas tes dalam konteks teori tes klasik didasarkan pada kesejajaran perangkat tes yang dalam prakteknya sangat sukar untuk dipenuhi. Jika prosedur tes retes digunakan, sampel yang diambil sangat tidak mungkin berperilaku sama pada saat tes dikerjakan untuk yang kedua kalinya.

Teori tes klasik tidak memberikan landasan untuk menentukan bagaimana respons seseorang peserta tes apabila diberikan butir tertentu. Tidak adanya informasi ini tidak memungkinkan melakukan desain tes yang bervariasi sesuai dengan kemampuan peserta tes (*adaptive or tailored testing*). Indeks kesalahan baku pengukuran diasumsikan sama untuk setiap peserta tes. Padahal seseorang peserta tes mungkin berperilaku lebih konsisten dalam menjawab soal dibandingkan peserta tes lainnya. Demikian sebaliknya, dan masih ada kelemahan-kelemahan lainnya.

Kelemahan-kelemahan yang muncul dalam teori tes klasik memicu munculnya teori baru yang lebih memadai, yaitu teori respon butir (*Item Response Theory; IRT*), yang dikenal pula dengan nama *latent traits theory*. Menurut Embretson & Reise (2000) kelebihan *IRT* dibandikan *CTT* antara lain (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 pes erta; (4) estimasi-estimasi yang tidak bias bisa diperoleh dari sampel yang tidak representatif; (5) skor tes memiliki arti manakala dibandingkan dengan karakteristik item-itemnya; (6) skala yang bersifat interval dicapai dengan menggunakan model pengukuran yang lebih logis; (7) tes dengan format item campuran dapat menghasilkan skor tes yang optimal; (8) skor -skor yang berubah dapat dibandingkan secara berarti jika tingkat skor awal berbeda; (9) hasil analisis faktor pada data skor kasar item menghasilkan sebuah *full information factor analysis* ; dan (10) sifat-sifat item sebagai stimulus dapat secara langsung berhubungan dengan sifat-sifat psikometriknya.

Prinsip dasar Teori Respons Butir (*IRT*) menurut Hambleton, Swaminathan, & Rogers (1991: 5) adalah

> The desirable features of an alternative test theory would include (a) item characteristics that are not group-dependent, (b) scores describing examinee proficiency that are not test-dependent, (c) a model that is expressed at the item level rather than at test level, (d) a model that does not require strictly parallel tests for assessing reliability, and (e) a model that provides a measure of precision for each ability score.

Pernyataan tersebut menunjukkan bahwa tujuan teori respons butir adalah membentuk statistik butir yang tidak bergantung pada kelompok, membentuk skor tes yang dapat menggambarkan kemampuan subjek tanpa bergantung pada indeks kesukaran butir tes, membentuk model tes yang asumsi-asumsinya mempunyai dukungan kuat, membentuk model tes yang tidak memerlukan asumsi paralel dalam pengujian reliabilitasnya, dan membentuk model tes yang dapat memberikan dasar pencocokan antara butir tes dan tingkat kemapuan subjek.

Landasan pemikiran pada teori respons butir didasarkan pada dua postulat, yaitu: (a) performansi subjek pada suatu butir dapat diprediksi oleh seperangkat faktor yang disebut *latent traits* atau kemampuan laten, dan (b) hubungan antara performansi subjek pada suatu butir dan perangkat kemampuan laten yang mendasarinya digambarkan oleh fungsi monoton naik yaitu kurva karakteristik butir (Hambleton, Swaminathan & Rogers, 1991: 7). Pernyataan tersebut menunjukkan bahwa performansi peserta tes dalam merespons suatu butir tes tergantung dari kemampuan yang dimilikinya. Semakin tinggi kemampuan yang dimiliki semakin baik performansi yang ditampilkan peserta tes, sebagai mana digambarkan dengan kurva monoton naik.

Model matematis *IRT* mengekspresikan probabilitas menjawab benar oleh peserta tes bergantung kepada kemampuan yang dimiliki dan karakteristik butir soal. Model *IRT* memuat sekumpulan asumsi tentang data terhadap model yang digunakan yaitu unidimensi, independensi lokal dan invariansi parameter (Hambleton, Swaminathan & Rogers, 1991: 8).

Asumsi unidimensi dimaksudkan bahwa setiap tes hanya mengukur satu kemampuan. Jika butir-butir soal mengukur lebih dari satu kemampuan, maka respons terhadap butir soal tersebut merupakan kombinasi dari berbagai kemampuan peserta tes. Kontribusi setiap kemampuan terhadap respons peserta tidak diketahui, dan hal ini bertentangan dengan tujuan *IRT*.

Asumsi independensi lokal menyatakan bahwa kinerja seseorang pada suatu butir soal tidak mempengaruhi kinerja pada butir soal lain. Jika kemampuan yang mempengaruhi kinerja dibuat konstan, maka respons subjek terhadap butir soal manapun akan bebas secara statistik.

Asumsi invariansi parameter menyatakan bahwa karakteristik butir soal tidak berubah meskipun subjek yang menjawab butir tersebut berubah-ubah dan berbeda tingkat kemampuannya. Fungsi logistik digunakan untuk mengembangkan fungsi probabilitas model *IRT* yang terdiri dari: (a) model 1-P dengan parameter tingkat kesulitan, (b) model 2-P dengan parameter tingkat kesulitan dan daya beda, (c) model 3-P dengan parameter tingkat kesulitan, daya beda dan tebakan, untuk butir soal dikotomus. Model 1-P yang dikenal dengan sebutan model Rasch dikembangkan menjadi model politomus *Partial Credit Model* (*PCM*) yaitu dengan menjabarkan lokasi butir menjadi beberapa katagori. Muraki (1992) mengembangkan kembali *PCM* yang memungkinkan butir

dalam skala memiliki perbedaan dalam hal parameter lereng. Model ini kemudian diberi nama *Generalized Partial Credit Model* (*GPCM*).

Pendekatan teori tes klasik dan *IRT* secara fundamental memang berbeda, walaupun demikian teori klasik memiliki hubungan erat dengan *IRT*. Hubungan tersebut dapat dijadikan dasar dalam tahap awal untuk memahami *IRT*. Hasil pengamatan di Bali, sebagian besar guru masih terpaku pada penggunaan teori tes klasik dalam analisis tes untuk mengukur hasil belajarnya, sehingga informasi yang diperoleh masih sangat terbatas. Oleh karena itu penelitian tentang perbandingan karakteristik tes berdasarkan pendekatan teori tes klasik dan IRT perlu dilakukan. Guru perlu memperoleh informasi bagaimana melakukan analisis tes hasil belajar berdasarkan teori yang ada sehingga dapat diperoleh informasi hasil belajar yang telah dilakukan secara cermat dan akurat.

Studi yang relevan telah dilakukan oleh beberapa ahli seperti, *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), dan *A Monte Carlo Comparison of Item and Person Statistics Based on Item Response Theory Versus Classical Test Theory* (McDonald & Paunonen, 2002).

Penelitian ini dilakukan di Sekolah Dasar, pada pelajaran matematika dengan menggunakan tes model campuran. Tujuan penelitian adalah untuk mengetahui (1) karakteristik tes prestasi belajar matematika jika dilakukan analisis dengan pendekatan klasik; (2) karakteristik tes prestasi belajar matematika jika dianalisis dengan pendekatan *IRT*; (3) perbandingan tingkat daya beda butir tes jika dianalisis menggunakan pendekatan klasik dan *IRT*.

Secara teoretis, penelitian ini diharapkan dapat memberikan kontribusi bagi perkembangan pengukuran dalam pendidikan matemaika, antara lain: (1) memberikan informasi tentang analisis karakteristik tes prestasi belajar dengan pendekatan klasik dan *IRT* yang lebih dikembangkan untuk jenjang SMP dan SMA/SMK; (2) memberikan informasi tentang analisis karakteristik tes dengan pendekatan Klasik dan *IRT* untuk bidang-bidang strudi IPA lainnya.

Secara praktis hasil penelitian ini dapat dimanfaatkan dalam: (1) melakukan seleksi butir dalam menyusun tes yang berkualitas; (2) memberikan petunjuk bagi guru, dalam menyusun tes secara professional sehingga

mampu melakukan pembelajaran lebih profesional dan bertanggung jawab.

METODE PENELITIAN

Penelitian ini adalah penelitian empirik yang diawali dengan pengembangan intrumen (tes) prestasi belajar matematika model campuran untuk kelas VI Sekolah Dasar yang diujikan pada semester 2 (tes sumatif) yang disusun berdasarkan pokok bahasan bilangan, geometri, pengukuran dan pengolahan data. Penyusunan kisi-kisi dan penulisan soal dilakukan oleh tim yang terdiri dari dua orang guru senior mata pelajaran matematika Sekolah Dasar. Validitas isi dan keterbacaan soal melibatkan dua ahli (*expert*) dalam bidang pendidikan matematika dan pengukuran, 10 guru dan 20 orang siswa kelas VI Sekolah Dasar. Instrumen (tes) yang telah diperbaiki berdasarkan analisis *expert* diujicoba di lima belas Sekolah Dasar. Data hasil ujicoba dianalisis menggunakan pendekatan Klasik dan *IRT* menggunakan program *Parscale* (Muraki & Bock, 1977) dengan estimasi *Marginal Maximum Likelihood* (*MML*).

Pengumpulan data penelitian dilakukan dengan *stratified random sampling* melibatkan sampel sebanyak 260 siswa kelas VI. Penerapan random dilakukan pada tingkat sekolah, sedangkan penentuan strata sekolah dengan memperhatikan letak sekolah dan katagori sekolah. Respon siswa dikoreksi oleh dua orang *rater* untuk mendapatkan skor yang baik. Untuk menjamin konsistensi penilaian, skor dari dua orang *rater* tersebut diuji relabilitasnya dengan menggunakan uji reliabilitas inter *rater* dengan pendekatan *Hoyt* (Mardapi, 2012: 86).

Pemenuhan asumsi unidimensi dan validitas konstruk dari tes dilakukan dengan analisis faktor exploratori dan konfirmatori. Banyaknya dimensi yang diukur oleh tes, dilihat dari *sree plot* nilai *Eigen*. Hal ini sesuai dengan pendapat Demars (2010: 39) bahwa *eigenvalue* dari inter-item matrik korelasi adalah salah satu metode yang *simple* untuk uji dimensionalitas. Pengujian kecocokan model hipotetik pengukuran terhadap data empiris menggunakan analisis faktor konfirmatori. Program yang digunakan adalah Lisrel 8.54 dengan indikator *goodness of fit* (Joreskog & Sorbom, 1996: 27)

HASIL DAN PEMBAHASAN

Koefisien reliabilitas (r_{11}) inter-*rater* skor diperoleh sebesar 0.926. (lebih dari 0.700 sesuai kriteria). Hal tersebut berarti bahwa kedua *rater* memberikan penilaian yang konsisten. Dengan demikian skor yang diberikan oleh kedua *rater* dapat

digunakan secara acak. Hal ini juga berarti unsur *subjektivitas* dari masing-masing *rater* (penilai) tidak berpengaruh terhadap pemberian skor pada tes, sehingga skor yang diberikan dapat digunakan sebagai data penelitian.

Tabel 1
Ringkasan Reliabilitas Hoyt

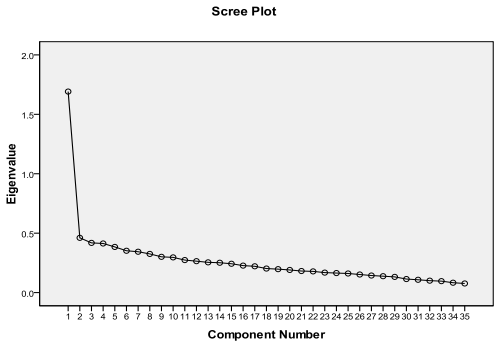
Sumber Variasi	*JK*	*dł*	*RJK*	r_{11}
Antar-penilai	11366.629		11366.629	0.926
Antar-	841523.800	34	24750.70	

Nilai *Chi-Square* pada uji Bartlet tes kelas VI sebesar 1786.476 dengan derajat kebebasan 596 dan niali-p kurang dari 0.01. Hasil ini menunjukkan bahwa ukuran sampel sebesar 260 yang digunakan pada penelitian telah cukup.

Tabel 2
KMO and Bartlett's Test[a]

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		865
Bartlett's Test of Sphericity	Approx. Chi-Square	1786.476
	df	596
	Sig.	.000

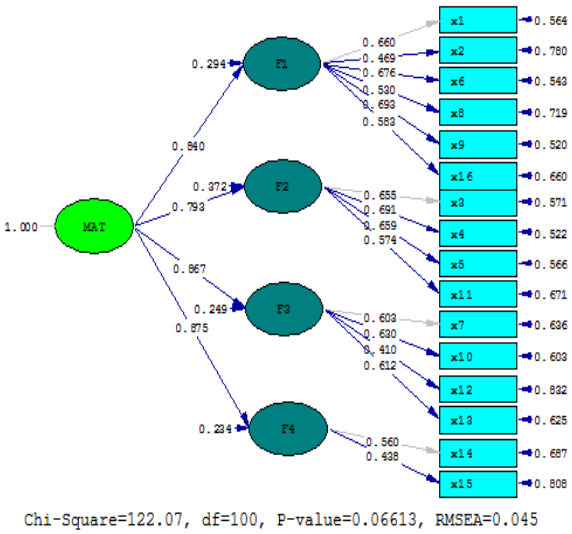
Hasil *scree plot* tes kelas VI (Gambar1) menunjukkan bahwa nilai *Eigen* tampak mulai landai pada faktor ke dua. Hal ini menunjukkan terdapat satu faktor yang dominan yaitu kemampuan matematika. Faktor-faktor yang lain berkaitan dengan kemampuan matematika tersebut. Dengan demikian tes yang disusun hanya mengukur satu dimensi atau satu kemampuan saja yaitu kemampuan matematika.



Gambar 1
Scree Plot Nilai *Eigen* Tes Kelas VI

Hasil *running* lisrel (Gambar 2) mendapatkan nilai *chi-square* sebesar 122.07 dengan df = 100 dan *p* =0.066, Nilai *Root Mean Square Error Approximation* (*RMSEA*) = 0.045, nilai *Non-Normed Fit Index* (*NNFI*) = 0.963, *Comparative Fit Index* (*CFI*) = 0.969 dan χ^2/df = 122.07/100=1.22< 3. Hal tersebut menunjukkan hipotesis nol diterima artinya model yang digunakan *fit* dengan data, dimana nilai *p-valuae* = 0.066 lebih besar dari nilai α=0,05. Dukungan terhadap model yang dikembangkan oleh data empirik (sampel) dapat dilihat juga dari besarnya *RMSEA* =0.045 yang lebih kecil dari α=0.08 dan nilai index kesesuaian yang diperoleh menggambarkan kesesuaian

model dengan data sebesar 0.879. Hasil ini menggambarkan bahwa tes matematika yang dikonstruk atas 4 variabel laten dengan 16 indikator *fit* dengan model.



F1=Bilangan, F2=Geometri,
F3=Pengukuran, F4=Pengolahan Data
Gambar 2
Diagram Path Hasil *Runing* Lisrel

Karakteristik Butir Tes Berdasarkan Teori Tes Klasik

Karakteristik butir tes dengan pendekatan Klasik dibaca pada *output Parscale PH1*. Daya beda butir yaitu sejauh mana butir tesebut mampu membedakan antara siswa yang mampu dan kurang mampu ditunjukkan oleh korelasi *Pearson & polyserial.* Butir yang memiliki nilai korelasi *Pearson* (r_p) & *Polyserial* (r_{ps}) >0.2 adalah butir yang baik.

Tabel 3 menunjukkan bahwa rerata skor dari tes kelas VI adalah 27.660 berarti sedikit di atas rerata ideal (24.50). Varians skor keempat tes cukup besar ini berarti distribusi skor cukup besar. Hal ini juga ditunjukkan oleh rentang skor yang cukup lebar. Rentang skor tes terletak antara 10 dan 42. Distribusi skor siswa membentuk *curve* sedikit juling ke kiri, karena nilai reratanya lebih rendah dari median. *Curve* yang juling ke kiri menunjukkan sebagian besar siswa mendapat skor tinggi, sedang *curve* yang juling ke kanan sebaliknya. Nilai rerata koefisien korelasi *Pearson* adalah 0. 373 dan rerata koefisien korelasi *Polyserial* adalah 0. 460. Semuanya berada di atas 0.2 yang berarti tes yang disusun (tes kelas IV) secara umum memiliki daya beda yang baik.

Table 3
Karakteristik Tes Berdasarkan Pendekatan Teori Tes Klasik

Component	Characteristics Test
Mean score	27.660
Variance	46.822
Standard Deviation	6.843
Skewness	-0.090
Curtosis	-0.614
Minimum Score	10
Maximum Score	42
Median	28
Mean korelasi pearson (r_p)	0.373
Mean korelasi point biserial(r_{ps})	0.460
Test length	35
Subject	260

Karakteristik Butir Tes Berdasarkan Pendekatan *IRT*

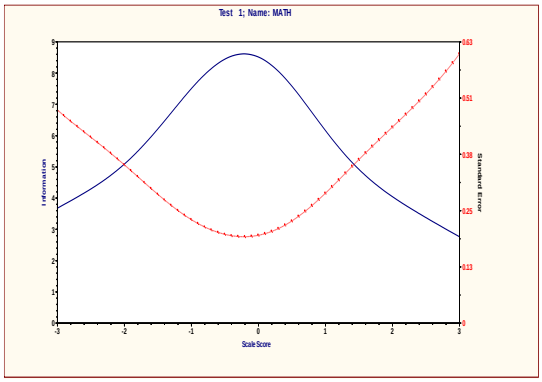
Analisis butir tes menurut *IRT* menggunakan model campuran dikotomus dan politomus *GPCM*. Karakteristik butir pada masing-masing tes meliputi *slope* (daya beda), *location* (tingkat kesukaran), dan *items fit statisticts (probability)* yang dibaca pada *output Parscale PH2*. Suatu butir dikatakan baik jika memiliki nilai *slope (a)* terletak pada interval $0.2 < a < 2$, nilai *location (b)* antara -2 dan 2 dan nilai *probability (p)* lebih dari 0.05 (Hambleton & Swaminathan (1985: 37). Berikut disajikan nilai rerata, standar deviasi, varians, nilai minimum, nilai maksimum, median, *skewnwss*, dan *kurtosis* dari parameter-parameter tersebut.

Table 4
Karakteristik Tes Berdasarkan Pendekatan IRT

Component	Characteristics		
	Slope	Location	Probability
Mean	0.654	-0.430	0.590
Standard Deviation	0.317	0.309	0.218
Variance	0.100	0.095	0.047
Minimum	0.255	-1.243	0.198
Maximum	1.502	0.029	0.962
Skewness	1.364	-0.881	-0.001
Curtosis	1.242	0.575	-0.893
Median	0.533	-0.351	0.586
Test Length	35		
Respondents	260		

Nilai probabilitas (*probability*) dari butir pada tes kelas VI yang ditunjukkan oleh tabel 4 memiliki nilai lebih besar dari 0.05 yang berarti semua butir pada tes tersebut *fit* dengan model. Hal itu dapat dilihat dari nilai minimum *probability* tes sebesar 0.198. Nilai rerata parameter *slope* adalah 0.654, lebih besar dari 0.2. Hal ini berarti tes yang disusun memiliki daya beda butir yang baik. Demikian pula nilai *location* dari tes sebesar -.0.430, berada di sekitar titik nol yang berati tes memiliki tingkat kesukaran butir yang sedang.

Informasi hubungan antara fungsi informasi tes dengan kesalahan baku pengukuran (*Standard Error of Measurement*) ditunjukkan Gambar 3.



Gambar 3
Grafik Hubungan skor dan *Standard Error* Tes Kelas VI

Keterangan:
------ = Kesalahan BakuPengukuran
——— = Informasi Tes

Gambar 3 menunjukkan bahwa tes yang disusun memiliki *error* yang rendah pada rentang skala berkisar dari -2 dampai dengan +2 skala logit, artinya tes akan memberikan informasi yang optimal jika digunakan untuk mengukur kemampuan siswa pada rentang

kemampuan antara -2 sampai dengan +2 skala logit. Hal ini sesuai dengan pendapat Hambleton, Swaminathan, & Rogers, (1991: 13) yang menyatakan bahwa parameter *b* akan diterima pada nilai yang berkisar antara -2.0 hingga +2.0 skala logit. Nilai *b* kurang dari -2.0 dikatakan butir tersebut sangat mudah atau memiliki *probability of endorsement* sangat tinggi dan di atas +0.2 dikatakan sulit atau probabilitas kemendukungan sangat rendah.

Perbandingan Tingkat Daya beda butir Tes yang dianalisis dengan Pendekatan Klasik dan IRT

Hasil analisis tes dengan menggunkan pendekatan teori tes klasik memperoleh rerata daya beda sebesar 0.373 atau 0.460. Nilai tesebut ditunjukkan oleh koefisien korelasi Pearson atau koefisien korelasi polyserial yang dapat dibaca pada output parscale PH1. Hali ini menunjukkan bahwa tes yang disusun mampu mengestimasi perbedaan kemampuan siswa menjadi 4 kelompok atau lima kelompok. Sedangkan hasil analisis tes menggunakan pendekatan *IRT* menunjukkan nilai rerata daya beda tes sebesar 0.644 yang ditunjukkan oleh nilai rerata slope yang dapat dibaca pada output parscale PH2. Hali ini berarti daya beda tes yang dianalisis dengan pendekatan IRT mampu mengestimasi perbedaan kemampuan siswa menjadi tujuh kelompok. Jika hasil analisis kedua pendekatan tersebut dibandingkan tampak bahwa penggunaan pendekatan *IRT* dalam analisis daya beda tes lebih teliti dibandingkan pendekatan teori tes klasik.

SIMPULAN DAN SARAN

Berdasarkan pembahasan hasil penelitian di atas dapat diambil kesimpulan bahwa (1) analisis tes prestasi belajar matematika bentuk campuran menggunakan teori tes klasik menghasilkan daya beda sebesar 0.373 atau 0.460 yang berarti bahwa tes yang disusun mampu mengelompokkan kemampuan siswa menjadi 4 kelompok atau lima kelompok; (2) analisis tes prestasi belajar matematika bentuk campuran menggunakan pendekatan *IRT* memperoleh rerata *slope* sebesar 0.654 yang berarti tes yang disusun mampu mengelompokkan kemampuan siswa menjadi tujuh kelompok, dan (3) analisis tes dengan pendekatan *IRT* lebih teliti dalam menentukan perbedaan kemampuan siswa.

Hasil penelitian ini diharapkan dapat memberikan sumbangan pemikiran bagi guru dalam menyusun tes untuk mengukur prestasi belajar siswanya. Tidak terpaku pada penggunaan teori tes klasik tetapi juga

menggunakan pendekatan *IRT* sehingga diperoleh informasi yang lebih akurat.

DAFTAR PUSTAKA

Allen, M.J. & Yen, W.M. (1979). *Introduction to measurement theory*. Belmont, CA : Wadsworth, MC.

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

Depdiknas, (2003). Undang-Undang RI Nomor 20, Tahun 2003, tentang Sistem Pendidikan Nasional.

DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML estimation. *Applied Measurement in Education*, 15, 15-31.

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*. Boston, MA: Kluwer Inc.

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

Joreskog, K.G. & Sorbom, D. (1996). *LISREL 8: structural equation modeling.* Chicago: Scientific Software International.

Mardapi, D. (2008). *Teknik penyusunan tes dan nontes*. Yogyakarta: Mitra Cendikia(2012). *Pengukuran, penilaian, & evaluasi pendidikan*. Yogyakarta: Nuha Medika

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), 921-94

Muraki, E. (1992). A generalized partial credit model. Aplication of an algorithm. *Applied Psychological Measurement*, 16, 159-176.

Muraki,E., & Bock, R.D. (1997). *Parscale* : *IRT item analysis and test scoring for rating-scale data.* Chicago: Scientific Software International.

Ridho A. (2005). *Karakteristik Psikometrik Tes Berdasarkan Pendekatan Teori TesKlasik dan Teori Respon Aitem*. Fakultas Psikologi UIN. Malang

Ridho A. (2007). Karakteristik Psikometrik Tes Berdasarkan Pendekatan Teori TesKlasik dan Teori Respon Aitem *Jurnal Psikologi INSAN, 2 (2), 1-statslab-rshiny.fmipa.unej.ac.i*