JMPM: Jurnal Matematika dan Pendidikan Matematika
8 (2) September 2023 157-172
Tersedia secara online di https://www.journal.unipdu.ac.id/index.php/jmpm
Terakreditasi S3 – SK No. B/1796/E5.2/KI.02.00/2020

Penskoran yang *Fair* pada Tes Matematika Pilihan Ganda Menggunakan *Item Response Theory*

(*Fair Scoring on Multiple Choice Mathematics Tests Using Item Response Theory*)

Hari Purnomo Susanto[1*], Siti Irene Astuti D[2], Endang Mulyani[3]
[1, 2, 3] Penelitian dan Evaluasi Pendidikan, Universitas Negeri Yogyakarta-Sleman, DIY, Indonesia, 55281
[1]Sekolah Tinggi Keguruan dan Ilmu Pendidikan PGRI Pacitan-Pacitan, Jawa Timur, Indonesia, 63515
[*] email penulis korespondensi: haripurnomo.2021@student.uny.ac.id

Abstrak

Tujuan penelitian ini yaitu mengaplikasikan model IRT untuk melakukan penskoran yang lebih fair pada tes matematika berbentuk pilihan ganda. Tes pilihan ganda yang digunakan yaitu tes matematika pada konten Aljabar sebanyak 15 butir yang diujicobakan pada 65 siswa SMP kelas 8. Data ini digunakan untuk kalibrasi butir tes menggunakan IRT. Model fit ditentukan dengan metode M2. Instrumen cocok atau fit yang digunakan yaitu model 1PL dan 2PL. Banyak butir yang fit pada kedua model sama dengan banyaknya butir tes pilihan ganda. Selanjutnya, parameter-paramater butir digunakan sebagai dasar untuk melakukan penskoran. Hasil perbandingan menunjukkan, model IRT mampu memberikan penskoran yang lebih objektif dan tepat, berdasarkan karakteristik yang dimiliki oleh butir test. Model 2PL mampu memberikan penskoran yang paling fair dari pada model 1PL dan konvensional. Karakteristik parameter diskriminan pada model 2PL, membuatnya mampu membedakan kemampuan siswa. khususnya pada siswa-siswa yang memiliki jumlah jawaban benar yang sama, tetapi pola jawabannya berbeda.

Kata kunci: *penskoran tes; tes matematika; item response theory*

Abstract

This research aims to apply the IRT model to conduct fairer scoring on multiple choice mathematics tests. The multiple choice test used was a 15-item mathematics test on Algebra content which was tested on 65 grade 8 junior high school students. This data was used to calibrate the test items using IRT. Model fit was determined using the M2 method. The suitable or fit instruments used are the 1PL and 2PL models. The number of items that fit in both models is the same as the number of multiple choice test items. Next, the item parameters are used as a basis for scoring. The comparison results show that the IRT model is able to provide more objective and precise scoring, based on the characteristics of the test items. The 2PL model is able to provide the fairest scoring compared to the 1PL and conventional models. The characteristics of the discriminant parameters in the 2PL model make it capable of differentiating student abilities. especially for students who have the same number of correct answers, but different answer patterns.

Keywords: *test scoring; mathematics test; item response theory*

Cara mengutip dengan APA 7 style: Susanto, H. P., Astuti, S. I., & Mulyani, E. (2023). Penskoran yang Fair pada Tes Matematika Pilihan Ganda Menggunakan Item Response Theory. *JMPM: Jurnal Matematika dan Pendidikan Matematika*, *8*(2), 157-172.

https://dx.doi.org/10.26594/jmpm.v8i2.3293.

PENDAHULUAN

Tes berbentuk pilihan ganda telah menjadi andalan pengukuran dalam Pendidikan di Indonesia. Jenis tes ini dapat ditemui dalam tes-tes berskala nasional di Indonesia yaitu Ujian Akhir Nasional (UAN), Ujian Akhir sekolah (UAS), dan tes masuk perguruan tinggi, serta yang terkini yaitu Assesmen Kompetensi Minimum (AKM). Kelebihan menggunakan tes bentuk pilihan ganda yaitu bentuk tes ini dapat digunakan pada banyak mata pelajaran dan dapat digunakan untuk berbagai tujuan pengukuran Pendidikan (Stankous, 2016; Yuksel & Fidan, 2019). Bentuk tes ini memudahkan guru dalam melakukan tes dan melakukan penilaian (Brown & Abdulnabi, 2017; Scully, 2017), memiliki konsistensi (Ali dkk., 2016; Stankous, 2016), validitas isi (Bacon, 2003; Yuksel & Fidan, 2019), dan validitas konkuren yang tinggi (Bleske-Rechek dkk., 2007; Scully, 2017), hasil ujian yang sangat objektif (Lopes dkk., 2010; Scully, 2017), serta banyak informasi yang dapat diukur (Yuksel & Fidan, 2019). Bentuk tes ini akan lebih memudahkan ketika dipadukan dengan teknologi atau tes scara online seperti menggunakan moodle, sehingga dapat memberikan *feedback* secara langsung pada siswa (Lopes dkk., 2010; Scully, 2017; Torres dkk., 2011). Namun, bentuk tes pilihan ganda juga memiliki kelemahan, yaitu siswa memiliki peluang mejawaban benar, walaupun sebenarnya siswa tidak bisa mengerjakan tes tersebut. Tes ini tidak mampu mengukur kemampuan kognitif pada level sintesis dan evaluasi. Tes ini juga tidak banyak digunakan untuk mengukur perilaku tingkat lanjut (Yuksel & Fidan, 2019), misalnya, komunikasi dan artikulatif penjelasan, pengorganisasian informasi, dan kreativitas dalam menghasilkan ide orisinal (Lopes dkk., 2010).

Tujuan utama dari setiap tes yaitu untuk mengetahui level kemampuan kognitif berdasarkan jawaban dari siswa. Penskoran tes yang tepat harus menjadi prioritas utama, agar tes mampu mengukur kemampuan yang harus diukur. Terdapat beberapa jenis penskoran tes pilihan ganda yang sering digunakan yaitu pertama, *Number Right Scoring*, peskoran berdasarkan nomor benar. Jawaban benar diberi skor positif, dan jawaban salah atau tidak dijawab diberi skor nol. Total dari skor benar merupakan skor dari tes (Lesage dkk., 2013; Ypsilandis & Mouti, 2019). Kedua, *Negative Marking*, yang mana model penskoran ini memberikan skor satu (1) untuk jawaban benar dan negative satu (-1) pada siswa yang meberikan jawaban salah (Betts dkk., 2009; Ypsilandis & Mouti, 2019). Perbedaan model ini dengan model yang pertama yaitu butir yang tidak dijawab tidak diberi skor. Model ini dapat dijumpai pada tes masuk perguruan tinggi negeri, jika siswa menjawab benar diskor (4), tidak menjawab (0) dan menjawab salah (-1). Ketiga, *Omitted Answer Get Reward*. Pada model ini tidak hanya jawaban benar saja yang memperoleh skor, tetapi tidak memberikan jawaban juga mendapatkan skor. Model ini dilakukan untuk menghindari siswa mejawaban dengan cara menebak. Yang keempat yaitu *Weighted Scoring*. Model penskoran ini dapat dilihat pada Claudy (Ypsilandis & Mouti, 2019). Terdapat tiga metode pada bagian ini yaitu *Guttman Weights Scoring*, *Biserial Weights Scoring*, dan *Proportion Weights Scoring*. Model penskoran ini tidak jauh berbeda dengan model penskalaan dalam penskoran kuisionare.

Tes pilihan ganda juga sering digunakan dalam mata pelajaran matematika untuk mengetahui pencapaian hasil pembelajaran. Pada skala internasional dapat dijumpai pada tes PISSA(OECD, 2019) serta skala nasional yaitu Ujian Nasional (UN) dan AKM (Kemendikbud, 2021). Seperti penelitian yang dilakukan oleh Huntley dkk., (2009) dan Scully (2017), tes ini digunakan untuk mengukur *Higher order thinking and learning mathematic*. Bentuk tes ini juga digunakan untuk mengukur kemampuan literasi matemtaka (OECD, 2013, 2019). Akan tetapi, kita tahu bahwa untuk menyelesaikan tes matematika tidak hanya dibutuhkan ingatan dan pengetahuan saja, tetapi dibutuhkan

kemampuan dan keterampilan tertentu. Misalnya kemampuan dan keterampilan dalam teknik perhitungan, logika matematika, pemecahan masalah, memodelkan, dan menghubungkan konsep matematis. Kemampuan dan keterampilan ini akan digunakan berdasarkan level kognitif yang di ukur (Huntley dkk., 2009). Fakta ini seolah menunjukkan bahwa bentuk tes pilihan ganda tidak cocok untuk digunakan dalam tes matematika, terutama terkait penskorannya.

Penskoran tes matematika di Indonesia pada umumnya menggunakan metode *Number Right Score*. Skor total merupakan skor hasil tes yang menggambarkan kemampuan siswa pada level kognitif yang diukur. Metode penskoran seperti ini tidak cocok untuk penskoran tes matematika berbentuk pilihan ganda. Sesuai dengan karaktersitik objek dan komponen yang menjadi konsentrasi pengukuran dalam matematika, maka penskoran yang digunakan harus didasarkan pada karakteristik dari butir tes. Terutama pada perbedaan tingkat kesulitan dari setiap butir tes matematika. Sebagai contoh diberikan sepuluh butir tes matematika dengan tingkat kesulitan yang berbeda-beda. Siswa A dan B berturut-turut mejawab tes dengan pola jawaban (1,0,0,1,1,1,0,0,1,0) dan (0,1,1,0,0,1,0,0,1,1). dengan kata lain siswa A merespon benar untuk butir ke 1, 4, 5, 6, 9 saja dan siswa B merespon benar untuk butir ke 2, 3, 6, 9, 10 saja. Pada kasus ini, jika menggunakan metode penskoran konvensional maka skor akhir dari siswa A dan B sama yaitu 5. Metode penskoran seperti ini menjadi tidak adil (*unfair*) karena terlihat jelas bahwa siswa A dan B memiliki perbedaan kemampuan dalam memberikan respon jawaban benar pada nomor butir yang berbeda. Penskoran yang tidak adil ini akan menyebabkan adanya disparitas dalam interpretasi kemampuan matematika siswa yang sebenarnya. Kasus ini dapat diatasi dengan melakukan penskoran yang didasarkan pada karakateristik butir tes.

Karakteristik butir tes dapat ditentukan dengan kalibrasi inturmen tes. Dua pendekatan yang sering digunakan yaitu *Classical Test Theory* (CTT) dan *Item Response Theory* (IRT) (Cappelleri et al., 2014; Sudaryono, 2013). Kedua teori tersebut merupakan model matematika yang dapat digunakan untuk analisis butir tes. Model matematika CTT didasarkan pada model matematika sederhana; terutama rata-rata, proporsi, dan korelasi. Sedangkan IRT didasarkan pada pemodelan matematika yang lebih kompleks dalam menganlisis butir tes. CTT menggunakan penskoran konvensional yang menganggap total jawaban benar merupakan skor yang menggambarkan kemampuan siswa (Schaughency dkk., 2012). Sedangkan pada IRT skor dapat dilakukan dengan menghitung nilai *teta* (parameter kemampuan) yang memiliki hubungan langsung dengan karakteristik butir tes (lihat Rumus 1) (Retnawati, 2014; Thompson, 2021). Karakteristik ini ditunjukkan dengan adanya parameter-paramter butir yang dihasilkan dari proses kalibrasi IRT. Berdasarkan informasi tersebut, IRT menjadi fokus utama dalam penulisan studi ini dan menggunakannya untuk melakukan estimasi skor kemampuan matematika dengan instrumen berbentuk pilihan ganda.

$$P(\theta_j) = c_i + (1 - c_i)\frac{e^{1.7a_i(\theta_j - b_i)}}{1+e^{1.7a_i(\theta_j - b_i)}} \qquad (1)$$

Keterangan :

$P(\theta_j)$: Probabilitas peserta tes ke-j yang memiliki kemampuan θ untuk menjawab butir ke-i dengan benar.

θ_j : Tingkat Kemampuan Peserta tes ke-j

c_i : indeks *pseudo guessing* butir ke-i

a_i : Indeks *discriminant* butir ke-i

b_i : Indeks *difficulty* butir ke-i

Respon jawaban siswa pada tes pilihan ganda dalam konsep IRT akan dipandang

https://dx.doi.org/10.26594/jmpm.v8i2.3293

sebagai skor ordinal. Selanjutnya skor-skor tersebut ditranformasikan oleh Rumus (1) dan dijadikan sebagai dasar untuk mengestimasi parameter-paramter pada setiap butir tes. Terdapat beberapa jenis model IRT untuk data dikotomus (0 dan 1). Pertama, Model dengan satu *parameter logistic* (1PL) atau 1PL, model ini akan terpenuhi ketika rumus 1 memiliki nilai $a = 1$, dan $c = 0$. Karakteristik pada model 1PL ditandai dengan parameter *difficulty* (Tingkat kesulitan). Kedua, Model 2PL, karakteristik dari model 2PL ditandai dengan adanya parameter *discriminant* (daya beda) dan *difficulty*. Pada Rumus (1) kondisi model 2PL akan terpenuhi ketika $c = 0$, dan ketiga, model 3PL, karakteristik model ini ditandai dengan adanya parameter *discriminant*, *difficulty* tan dan *psudo guessing* (tebakan semu) (Hambleton dkk., 1991; Retnawati, 2014). Parameter *discriminant* memiliki fungsi dalam membedakan kemampuan, Parameter *difficulty* memiliki peran menggambarkan tingkat kesulitan setiap butir test, dan parameter *psudo guessing* berperan dalam menjelaskan kemungkinan siswa dengan kemampuan rendah menjawab dengan benar pada suatu butir (Retnawati, 2014). Pada contoh siswa A dan B diatas, dengan model 1 PL akan memberikan estimasi skor yang sama. Walaupun terkesan hampir sama dengan metode konvensional, tetapi dapat dibandingkan kelogisan penskorannya melalui parameter *difficulty*. Selanjutnya pada model 2PL dan 3PL untuk contoh diatas, siswa A dan B akan memiliki skor yang berbeda. Perbedaan skor kedua siswa yang dihasilkan oleh model 2PL dan 3PL terjadi karena berfungsinya parameter *discriminant* pada setiap butir. Lebih jelasnya dapat dilihat pada bagaian ilustrasi penskoran pada artikel ini.

Penjelasan di atas memberikan gambaran bahwa penskoran pada CTT fokus pada banyak jawaban yang benar. Estimasi skor tidak memperhatikan parameter butir karena bobot semua butir sama (Huda & Mardapi, 2015). Parameter butir hanya digunakan untuk menentukan karakteristik atau kualitas butir saja (Kasanova & Sulistiyono, 2023; Khaerudin, 2016), tetapi tidak digunakan sebagai dasar penskoran. Sebaliknya, pada IRT, penskoran sangat memperhatikan pola respons. Pola ini menjadi dasar IRT dalam memanfaatkan parameter dari butir dalam mengestimasi skor siswa. Dua pola respons dengan banyak jawaban benar yang sama tetapi pola respon berbeda akan memiliki skor yang berbeda. Secara tidak langsung parameter dari butir berperan sebagai bobot penskoran. Fungsionalitas dari parameter butir pada IRT menunjukkan bahwa penskoran benar-benar menginterpretasikan kemampuan siswa. Perbedaan inilah yang menjadikan IRT lebih *fair* jika digunakan dalam penskoran tes pilihan ganda.

Berdasarkan informsi di atas, artikel ini bertujuan untuk melakukan penskoran tes matematika berbentuk pilihan ganda dengan menggunakan konsep IRT. Hasil penskoran dengan IRT selanjutnya akan dibandingkan dengan hasil penskoran menggunakan metode konvensional.

METODE PENELITIAN

Studi ini menggunakan desain penelitian deskriptif kuantitatif, yang bertujuan untuk memberikan gambaran penskoran yang *fair* menggunakan konsep IRT pada tes matematika berbentuk pilihan ganda. Penulis meyakini bahwa penskoran dengan menggunakan konsep IRT akan memberikan hasil penskoran yang lebih logis dan mampu memberikan interpretasi skor kemampuan yang lebih baik jika dibandingkan penskoran konvensional.

Responden yang berpartisipasi dalam penelitian ini sebanyak 65 siswa SMP kelas 8. Instrumen yang digunakan terdiri dari 15 butir tes matematika berbentuk pilihan ganda. Adapun konten matematika yang diujikan dalam tes yaitu konten Aljabar (Kemendikbud, 2021). Konstruk dari butir tes dapat dilihat pada Tabel 1. Tes terdiri 4 respon pilihan dengan 1 jawaban kunci dan 3 distraktor. Berdasarkan 65 data yang diperoleh, dilakukan analisis validitas kontruk menggunakn analisis faktor konfirmatori atau *confirmatory*

https://dx.doi.org/10.26594/jmpm.v8i2.3293

factor Analysis (CFA). Hasil perhitungan CFA menghasilkan nilai *p-value* χ^2, RMSEA, dan CFI berturut-turut 0.26>0.05, 0.041< 0.05 dan 0.949> 0.9. Menurut Hair et al. (2019) hasil tersebut menjelaskan bahwa instrumen tes yang digunakan memiliki model kontruk yang fit dengan data empiris yang diperoleh atau dengan kata lain memiliki model konstruk yang valid. Selanjutnya hasil perhitungan reliabilitas instrumen menghasilakn nilai *Alpha Cronbac* sebesar 0.78. Nilai ini menunjukkan bahwa instrumen tes matematika yang digunakan memiliki konsistensi pada level baik.

Tabel 1. Konstruk Instrumen Tes Pada Konten Alajabar

Sub Konstruk	Banyak Butir	No Butir
Bentuk Aljabar	4	3,5,8,15
SPLDV	7	1,2,10,11,12,13,14
Proporsi	4	4,6,7,9

Sesuai dengan tujuan dari studi ini, untuk dapat melakukan penskoran menggunakan IRT, instrumen tes harus melalui proses kalibrasi terlebih dahulu. Kalibrasi digunakan untuk menentukan parameter-parameter dari butir yang fit. Adapun proses kalibrasi dapat dilakukan dengan 6 langkah yaitu (1) menentukan metode M2 digunakan untuk menentukan model Fit, (2) menentukan model fit yang digunakan (Chalmers, 2012; Maydeu-Olivares, 2013). (3) membuktikan asumsi Unidimensi, (4) membuktikan asumsi Independesi lokal. (5) menentukan butir fit, dan (6) mengestimasi parameter butir (Chalmers, 2012; Paek & Cole, 2019). Interpretasi parameter yang diperoleh dari hasil kalibrasi dapat dikategorisasikan menggunakan Tabel 2 (Bichi & Talib, 2018). Perhitungan kalibrasi instrument tes pada artikel dilakukan dengan menggunakan program R (R Core Team, 2022) dengan bantuan *package irtawsi* (Susanto dkk., 2023). *Package* ini memiliki kelebihan dapat digunakan dalam Bahasa Indonesia dan dapat memberikan interpretasi hasil analisis secara langsung. Package *irtawsi* dikembangkan berdasarkan *package mirt* (Chalmers, 2012) sehingga kedua *package* tersebut akan memiliki hasil analisis yang sama. Perbedaan kedua *package* tersebut yaitu untuk menggunakan mirt dalam kalibrasi diperlukan kemampuan dan pemahaman akan kode-kode bahasa R, sebaliknya *irtawsi* memiliki tampilan user interface seperti pada kebanyakan *software* statistika, sehingga tidak memerlukan kode-kode dalam bahasa R. Perbedaan kedua yaitu pengguna mirt harus melakukan interpretasi sendiri terhadap hasil analisis yang diperoleh, sedangkan jika menggukan irtawsi hasil analisis dan interpretasinya secara otomatis dapat dimunculkan. Selain itu, pada *irtawsi* bahasa dapat diatur ke dalam Bahasa Indonesia, pada mirt tidak bisa. Paparan di atas yang menjadi alasan mengapa penulis menggunakan *irtawsi* pada kalibrasi dengan IRT.

Tabel 2. Interpretasi Parameter Butir

Parameter	Interval parameter	Interpretasi
Difficulty (b)	$-3 < b < -2$	Sangat Mudah
	$-2 < b < -1$	Mudah
	$-1 < b < 1$	Sedang
	$1 < b < 2$	Sulit
	$2 < b < 3$	Sangat Sulit
Discriminant *index* (a)	$a > 1.7$	Sangat tinggi
	$1.35 < a \leq 1.69$	Tinggi
	$0.65 < a \leq 1.34$	Sedang
	$0.35 < a \leq 0.64$	Rendah
	$0.01 < a \leq 0.34$	Sangat Rendah

https://dx.doi.org/10.26594/jmpm.v8i2.3293

Pada studi ini interval skor yang digunakan pada konsep IRT berada pada rentang -3 sampai 3 (Baker, 2001; Bichi & Talib, 2018). Kategori pada Tabel 2 diadaptasi dari (Bichi & Talib, 2018). Tabel 2 digunakan untuk menyeleksi butir yang memiliki butir yang dapat menyebabkan terjadinya bias dalam estimasi skor kemampuan. Berdasarkan kategorisasi tersebut butir yang memiliki parameter *difficulty* < -3 atau >3 harus di-*drop* dan tidak digunakan. Selanjutnya, butir yang memiliki parameter *discriminant* dibawah 0.5 juga tidak digunakan. Pembatasan nilai *discriminant* dilakukan untuk menghindari terjadinya bias estimasi skor kemampuan yang dihasilkan. Pengecekan batas-batas parameter ini diberlakukan untuk butir-butir yang *fit* saja, sedangkan untuk butir yang tidak *fit* secara otomatis tidak digunakan.

Berdasarkan butir-butir yang digunakan, akan ditentukan fungsi informasi tes atau *tes information function* (TIF). Fungsi tes ini berperan dalam menentukan rentang interval kemampuan siswa yang cocok dites menggunakan butir-butir instrumen yang digunakan. Selain itu, melalui TIF dapat juga diketahui standar kesalahan pengukuran atau *Standard Error Measurement* (SEM) (Retnawati, 2014). Isntrumen tes dikatakan baik jika memiliki SEM < 0.3. Pada artikel ini, rentang kemampuan yang diperoleh digunakan untuk memeriksa apakah skor estimasi minimal dan maksimal berada pada rentang ini. Perhitungan TIF pada studi ini dilakukan dengan menggunakan *package irtawsi* (Susanto dkk., 2023).

Estimasi skor dapat dilakukan dengan menggunakan informasi pada langkah ke-6 pada prosedur kalibrasi di atas. Berdasarkan parameter-paramter tersebut, estimasi skor tes matematika dapat dilakukan dengan banyak metode salah satunya yaitu *Expected A Posteriori* (EAP), untuk metode estimasi yang lain dapat di lihat pada (Chalmers, 2012; Magis & Barrada, 2017). Pada studi ini metode EAP digunakan untuk estimasi skor kemampuan matematika siswa. Perhitungan estimasi skor kemampuan matematika pada artikel ini menggunakan program R (R Core Team, 2022) dengan *package irtawsi* (Susanto dkk., 2023).

HASIL DAN PEMBAHASAN

Pada bagian ini akan ditampilkan hasil dari setiap prosedur kalibrasi instrumen menggunakan IRT di atas. Selanjutnya hasil kalibrasi diaplikasikan untuk mengestimasi skor tes matematika siswa.

Kalibrasi Butir Tes Matematika

Langkah 1. Menentukan model fit dengan metode M2

Kriteria model fit yang digunakan pada artikel ini yaitu *p-value* >0.01, RMSEA <0.8, dan CFI >0.9 (Maydeu-Olivares, 2013, 2014). Hasil analisis dari *package irtawsi* dapat dilihat pada tabel 3. Berdasarkan kriteria model fit tersebut, maka setiap model dikotomus pada Tabel 3 dapat digunakan.

Tabel 3. Memilih Model Fit

Model	M2	df	p	RMSEA	CFI	Keputusan	Banyak Butir fit
1PL	116.869	104	0.183	0.044	0.954	Fit	15
2PL	116.869	104	0.183	0.044	0.954	Fit	15
3PL	81.199	75	0.292	0.036	0.978	Fit	14

Langkah 2. Menentukan model yang digunakan

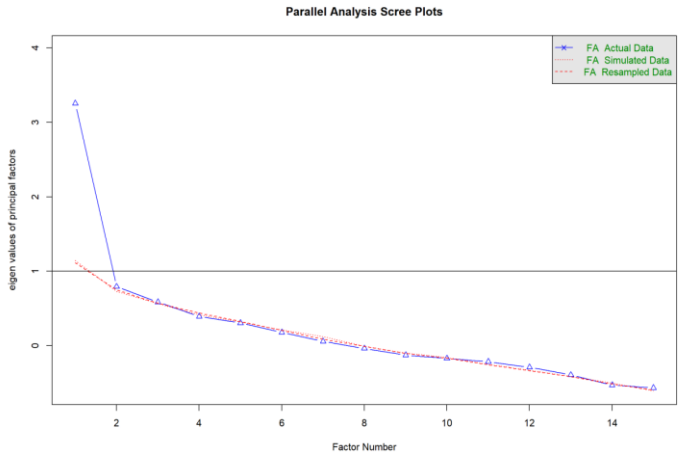
Penentuan model yang digunakan pada artikel ini mengacu pada model yang fit. Asalkan statusnya fit, maka ketiga model dapat digunakan. Pada tabel 3 semua model dapat digunakan. Normalnya, hanya dipilih satu model untuk digunakan untuk kalibrasi dan penentuan model yang digunakan, biasanya menggunakan nilai dari *Akaike*

https://dx.doi.org/10.26594/jmpm.v8i2.3293

Information Criterion (AIC), atau *Bayesian Information Criterion* (BIC) yang terkecil dari setiap model (Chalmers, 2016; Paek & Cole, 2019). Mengingat sebagai ilustrasi untuk dibandingkan dengan metode penskoran konvensional, maka pada artikel ini digunakan dua model IRT sekaligus yaitu model 1PL dan 2PL. Alasan kedua model ini dipilih yaitu keduanya memiliki banyak butir fit sama.

Langkah 3. Membuktikan asumsi unidimensi

Pada artikel ini asumsi unidimensi dibuktikan dengan menggunakan *scree plot* analisis faktor ekploratori dengan metode pararel analisis. *Scree plot* hasil analisis *package irtawsi* dapat dilihat pada Gambar 1. Gambar *Scree plot* ini dapat juga ditentukan dengan *package lavaan*(Rosseel, 2012). Pada Gambar 1 dapat dilihat bahwa banyaknya lereng yang curam hanya satu, yaitu dari nilai eigen pertama (segitiga pertama) dan nilai eigen kedua (segitiga kedua). Banyak lereng yang curang mengindikasikan banyak dimensi(Retnawati, 2014). Gambar tersebut membuktikan bahwa instrumen yang digunakan benar-benar hanya mengukur satu dimensi variabel laten saja. Variabel tersebut yaitu kemampuan siswa pada konten aljabar. Uji asumsi ini tidak terikat dengan model IRT yang diipilih, karena asumsi ini dapat dihitung bahkan sebelum model IRT dilakukan.



Gambar 1. *Scree Plot* Unidimensi

Langkah 4. Membuktikan asumsi independensi lokal

Asumsi independensi local dapat dibuktikan menggunakan metode Q3(Paek & Cole, 2019). Namun dalam studi ini, pembuktian independensi lokal didasarkan pada hasil uji asumsi unidimensi. Jika asumsi unidimensi terbukti benar hanya mengukur satu dimensi saja, maka secara otomatis asumsi idenpendensi lokal juga terpenuhi (Retnawati, 2015). Terbuktinya asumsi independensi lokal ini mengindikasikan bahwa siswa dalam menjawab suatu butir tes tidak dipengaruhi oleh butir tes yang lain (Sudaryono, 2013). Pelanggaran terhadap asumsi ini akan mempengaruhi estimasi parameter *ability* atau kemampuan (Edwards dkk., 2018).

Langkah 5. Menentukan butir fit. Butir dikatakan fit jika memiliki nilai *p-value* χ^2 lebih besar dari 0.05 (Paek & Cole, 2019; Retnawati, 2014). Ringkasan hasil perhitungan dengan menggunakan *package irtawsi* dapat dilihat pada tabel 4. Setiap butir pada model 1PL dan 2 PL fit.

Berdasarkan kriteria penentuan butir fit, maka dapat dilihat bahwa nilai $p - value$ ^{2} dari setiap butir pada Tabel 4 > 0.05, sehingga semua butir pada masing-masing model dapat dikatakan fit atau cocok untuk digunakan untuk digunakan dalam estimasi skor kemampuan. Sebagai catatan, sebelum menggunakan semua butir fit untuk

https://dx.doi.org/10.26594/jmpm.v8i2.3293

estimasi skor kemampuan, perlu dilakukan pengecekan terhadap parameter -parameter dari setiap butir. Pengecekan ini selengkapnya dijelaskan setelah dilakukan Langkah 6.

Tabel 4. Butir Fit Pada Model 1PL dan 2PL

Butir	1PL		2PL	
	$p-value$	Keputusan	$p-value$	Keputusan
1	0.70	Fit	0.44	Fit
2	0.08	Fit	0.06	Fit
3	0.10	Fit	0.19	Fit
4	0.56	Fit	0.53	Fit
5	0.62	Fit	0.44	Fit
6	0.17	Fit	0.52	Fit
7	0.59	Fit	0.59	Fit
8	0.82	Fit	0.76	Fit
9	0.16	Fit	0.10	Fit
10	0.45	Fit	0.73	Fit
11	0.52	Fit	0.40	Fit
12	0.30	Fit	0.76	Fit
13	0.41	Fit	0.46	Fit
14	0.49	Fit	0.49	Fit
15	0.24	Fit	0.34	Fit

Langkah 6. Mengestimasi parameter butir

Parameter butir dari model 1PL dan 2PL dapat dilihat pada Tabel 5. Parameter-parameter ini menjadi kunci utama dalam estimasi skor kemampuan matematika siswa pada konten aljabar. Selanjutnya, parameter butir yang diperoleh dikategorisasikan dengan menggunakan informasi pada Tabel 2 diatas. Hasil kategorisasi dapat dilihat pada Tabel 5.

Tabel 5. Parameter Model 1PL dan 2PL

Butir	Parameter model 1PL		Parameter model 2PL			
	b	Kriteria *b*	*a*	Kriteria *a*	*b*	Kriteria *b*
1	-0,59	Sedang	1,89	Sangat Tinggi	-0,40	Sedang
2	-1,50	Mudah	1,52	Tinggi	-1,12	Mudah
3	-1,11	Mudah	0,87	Sedang	-1,18	Mudah
4	0,44	Sedang	1,08	Sedang	0,39	Sedang
5	-0,93	Sedang	0,92	Sedang	-0,95	Sedang
6	0,04	Sedang	0,68	Sedang	0,05	Sedang
7	-0,59	Sedang	1,11	Sedang	-0,53	Sedang
8	-0,76	Sedang	1,62	Tinggi	-0,55	Sedang
9	-0,59	Sedang	0,84	Sedang	-0,65	Sedang
10	-0,43	Sedang	0,69	Sedang	-0,55	Sedang
11	0,52	Sedang	0,98	Sedang	0,50	Sedang
12	-0,51	Sedang	2,74	Sangat Tinggi	-0,31	Sedang
13	-0,19	Sedang	1,53	Tinggi	-0,14	Sedang
14	-0,85	Sedang	1,15	Sedang	-0,74	Sedang
15	-1,21	Mudah	0,92	Sedang	-1,23	Mudah

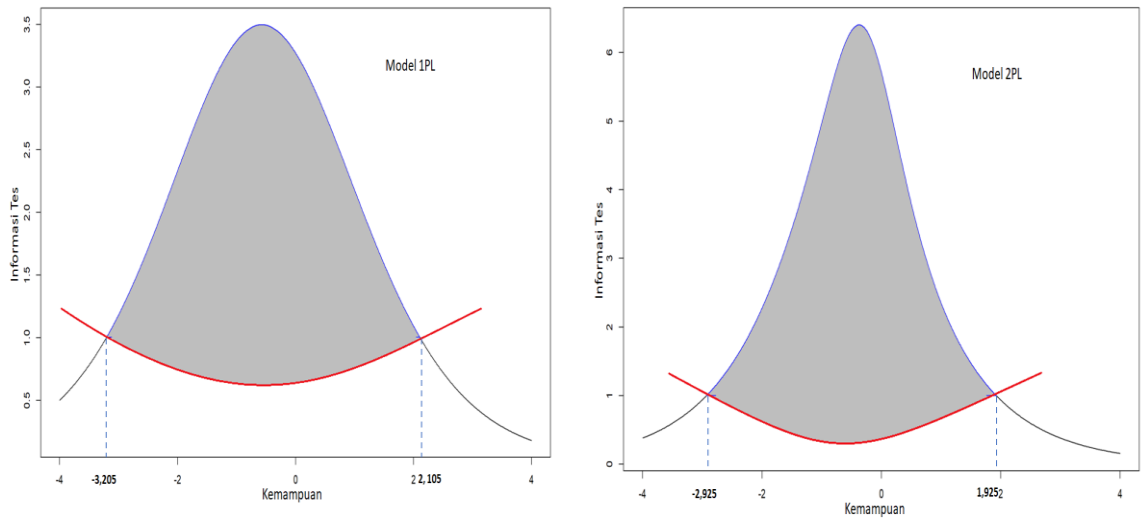
https://dx.doi.org/10.26594/jmpm.v8i2.3293

Pada Tabel 5 terlihat jelas bahwa untuk parameter *difficulty* (*b*) untuk setiap model tidak ada butir yang memiliki nilai parameter *b* lebih kecil dari – 3 atau lebih besar dari 3. Selanjutnya, pada model 2PL tidak terdapat nilai parameter *discriminant* (*a*) yang kurang dari 0.5. Hasil ini menjelaskan bahawa semua butir baik terindikasi model 1PL atau 2PL tidak ada yang di drop, atau dengan kata lain digunakan semua.

Enam Langkah kalibrasi telah dipenuhi berdasarkan kriteria yang digunakan dan pengecekan terhadap parameter-parmeter butir juga telah dlakukan. Maka, hasil kalibrasi instrumen tes matematika dengan model 1PL dan 2PL siap untuk digunakan dalam mengestimasi skor kemampuan matematika siswa.

Fungsi Informasi Tes (TIF)

Hasil analisis TIF pada masing-masing model dengan menggunakan *pacakage irtawsi* ditampilkan dalam bentuk visual serperti yang dapat dilihat pada Gambar 2. Kurva dengan warna biru meunjukkan bentuk visual dari TIF dan kurva dengan warnya menunjukkan SEM. Daerah arsiran menunjukkan daerah perptongan TIF dengan SEM. Berdasarkan gambar 2, TIF dari model 1PL menjelaskan bahwa instrumen yang digunakan cocok untuk digunakan untuk mengukur siswa yang memliki kemampuan -3.205 sampai 2.105 dengan SEM = 0.26. Selanjutnya, TIF dari model 2PL menggambarkan bahwa instrumen tes cocok jika digunakan untuk mengukur kemampuan matematika siswa yang memiliki kemampuan antara -2.925 sampai 1.925 dengan SEM = 0.23.



Gambar 2. TIF dari Model 1PL dan 2PL.

Estimasi Skor (Peskoran dengan IRT)

Hasil perhitungan skor matematika menggunakan model 1PL dan 2PL ditampilkan dalam Tabel 6. Pada kedua model tersebut tidak terdapat satupun butir yang tidak fit, sehingga mempermudah dalam menentukan estimasi skor kemampuan matematika siswa. Perhitungan estimasi skor cukup dihitung berdasarkan hasil analisis *package irtawsi*. Namun, Jika terdapat butir yang tidak fit, maka perhitungan estimasi skor dapat dilakuakn dengan menggunakan *package CatR* (Chalmers, 2016), dan sintak dapat dilihat pada https://cran.r-project.org/web/packages/catR/catR.pdf. Pada Tabel 6

tidak ditampilkan hasil estimasi peskoran untuk keseluruhan siswa.

Selanjutnya agar memudahkan dalam membandingkan penskoran model konvensional dengan model IRT, maka penskoran ditransformasikan kedalam rentang 0-100. Pada model konvensioanl dihitung dengan menggunakan rumus (2) dan model IRT menggunakan rumus (3). Penentuan rumus skor akhir pada IRT ini didasarkan pada interval dari tingkat kesulitan antara -3 sampai 3 (Sudaryono, 2013).

$$skor\ akhir = \frac{jawaban\ benar}{15} x100 \quad (2)$$

$$skor\ akhir = \frac{skor\ estimasi+3}{6} x100 \quad (3)$$

Pada rumus (2) nilai 15 menunjukkan banyaknya butir tes yang digunakan. Selanjutnya pada rumus (3) nilai 3 pada pembilang menunjukkan batas maksimum skor IRT yang digunakan. Selanjutnya nilai 6 pada bagian penyebut menunjukkan rentang skor terendah dengan skor tertinggi pada konsep IRT. Hasil perhitungan dengan menggunakan rumus (2) dapat dilihat pada tabel 6 kolom ke-4, dan rumus (3) pada kolom ke-6 untuk model 1PL dan kolom ke-8 untuk model 2PL.

Data pada Tabel 6 menjelaskan bahwa penskoran dengan metode konvensional sangat mudah dilakukan. Penskoran hanya dihitung berdasarkan banyak jawaban benar dan salah saja. Namun metode ini tidak memperhatikan karakteristik dari butir tes yang digunakan. Karakteristik yang dimaksud yaitu parameter *difficulty* dan *discriminant*, sehingga skor yang digunakan kurang merepresentasikan variabel laten yang sedang diukur. Variable tersebut yaitu kemampuan matematika siswa dalam konten Aljabar. Pada Tabel 6 terlihat jelas penskoran ini tidak mampu membedakan siswa yang memiliki jawaban benar yang sama banyaknya, tetapi pola jawabannya berbeda. Penskoran dengan metode konvensional dapat menjadi lebih *fair* ketika terdapat pembototan yang disesuaikan dengan hasil kalibrasi instrument menggunakan konsep *Conventional Test Theory* (CTT). Namun pembobotan yang dihasilkan didasarkan pada pengetahuan atau pengalaman dari guru atau pengembang tes, sehingga penskoran menjadi lebih subjektif. Selain itu, model matematis dari konsep CTT tidak memfasilitasi adanya hubungan secara langsung parameter butir dengan parameter kemampuan.

Hasil penskoran dengan menggunakan model 1PL tidak jauh berbeda dengan penskoran konvensional. Kesamaan ini dapat dilihat pada Tabel 6, siswa yang memiliki banyak jawaban benar yang sama memiliki skor yang sama, walaupun pola jawaban yang terjadi berbeda. Sehingga penskoran dengan model 1PL tidak mampu membedakan kemampuan matematika dari siswa yang satu dengan siswa yang lain yang memiliki banyak benar sama tetapi pola jawaban berbeda. Sebagai contoh, lihat data ke 10 dan 11, model 1 PL akan mengestimasi skor yang sama yaitu -0,57 atau 40,43. Hal ini disebabkan pada model 1PL tidak memiliki parameter daya beda atau *discriminant*. Parameter ini menjadi penting karena dapat mempengaruhi akurasi penilaian dalam membedakan berbedaan *point* pada skala abilitas atau kemampuan (Sudaryono, 2013). Perbedaan penskoran dengan 1PL dan konvensional terletak pada siswa yang menjawab benar semua. Model konvensional menghasilkan nilai 100, sedangkan pada model 1PL tidak, siswa yang menjawab benar semua soal memperoleh nilai 82.48.

Penskoran yang lebih *fair* ditunjukkan oleh model 2PL. Karakteristik yang dimiliki oleh model 2 PL tidak hanya berupa parameter *difficulty* tetapi juga parameter *discriminant* (DeMars, 2010; Hambleton dkk., 1991; Retnawati, 2014; Sudaryono, 2013) yang membuatnya mampu membedakan siswa yang memiliki banyak jawaban benar sama, tetapi pola jawaban berbeda. Disini parameter *discriminant* menunjukkan perannya secara langsung dalam membedakan kemampuan siswa yang rendah, sedang, maupun

https://dx.doi.org/10.26594/jmpm.v8i2.3293

tinggi (Bichi & Talib, 2018). Semakin tinggi nilai parameter *discriminant* dari suatu butir, maka butir tersebut semakin mampu membedakan kemampuan siswa (Hays dkk., 2000). Begitu juga sebaliknya, semakin rendah parameter *discriminant* dari suatu butir, maka butir tersebut semakin sulit dalam membedakan kemampuan siswa. Hal ini disebabkan ketika parameter *discriminant* kecil peluang untuk menjawab benar suatu butir soal tersebut akan semakin tinggi untuk semua level kemampuan (Baker, 2001). Sebagai contoh, pada Tabel 6 untuk data ke 24 dan 25, masing-masing siswa memiliki banyak jawaban yang benar sama banyak, namun hasil penskoran menunjukkan siswa ke-24 dan ke 25 berturut-turut memperoleh skor 68,55 da 77,57.

Tabel 6. Estimasi Skor Konvensioanl dan IRT.

No	Pola jawaban siswa	Konvensional		1PL		2PL	
		Skor	Skor (0-100)	Skor IRT	Skor (0-100)	Skor IRT	Skor (0-100)
1	(0,0,0,0,0,0,0,0,0,0,0,0,1,1,0)	2	13,33	-1,91	18,20	-1,52	24,62
2	(0,1,0,0,0,0,0,0,0,1,0,0,0,0,0)	2	13,33	-1,91	18,20	-1,64	22,75
3	(0,0,0,0,1,0,0,0,0,0,1,0,0,0,0)	2	13,33	-1,91	18,20	-1,71	21,47
4	(0,0,0,0,1,1,0,1,0,0,0,0,0,0,0)	3	20,00	-1,61	23,25	-1,40	26,62
5	(0,0,0,0,0,0,0,0,0,0,0,1,0,1,1)	3	20,00	-1,61	23,25	-1,09	31,78
6	(0,0,0,1,0,0,0,0,0,1,1,0,0,0,0)	3	20,00	-1,61	23,25	-1,51	24,85
7	(0,0,0,0,0,0,1,0,1,1,0,0,0,0,0)	3	20,00	-1,61	23,25	-1,54	24,42
8	(1,1,0,0,0,0,1,0,0,0,0,0,0,0,0)	3	20,00	-1,61	23,25	-1,15	30,90
9	(0,0,1,0,1,1,0,0,0,0,0,0,0,0,1)	4	26,67	-1,33	27,87	-1,37	27,20
10	(1,0,0,0,1,0,1,0,0,0,0,0,0,1,1)	5	33,33	-1,07	32,20	-0,89	35,17
11	(0,1,1,0,0,1,0,0,0,0,0,0,1,0,1)	5	33,33	-1,07	32,20	-0,97	33,87
12	(0,0,1,0,0,0,1,1,1,1,0,0,0,0,1)	6	40,00	-0,82	36,35	-0,88	35,32
13	(0,0,1,0,0,1,1,0,1,1,0,0,0,0,1)	6	40,00	-0,82	36,35	-1,04	32,65
14	(0,0,1,0,1,0,0,1,1,0,0,0,0,1,1)	6	40,00	-0,82	36,35	-0,84	36,08
15	(0,1,0,0,1,0,0,1,0,1,0,0,1,1,0)	6	40,00	-0,82	36,35	-0,66	38,98
16	(0,0,1,1,1,0,1,0,1,0,0,0,0,0,1)	6	40,00	-0,82	36,35	-0,93	34,48
17	(1,1,1,0,0,0,0,1,0,1,0,1,0,1,0)	7	46,67	-0,57	40,43	-0,21	46,50
18	(0,1,1,0,1,1,0,1,1,0,0,0,0,0,1)	7	46,67	-0,57	40,43	-0,67	38,82
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
49	(0,1,1,1,1,0,1,1,1,1,0,1,1,1,1)	12	80,00	0,75	62,57	0,59	59,75
50	(1,1,1,1,1,0,1,0,1,0,1,1,1,1,1)	12	80,00	0,75	62,57	0,71	61,90
51	(1,1,1,0,1,1,1,1,1,1,0,0,1,1,1)	12	80,00	0,75	62,57	0,33	55,47
52	(1,1,1,1,1,0,1,1,1,0,1,1,1,1,1)	13	86,67	1,09	68,18	1,165	69,42
53	(1,1,1,1,0,1,0,1,1,1,1,1,1,1,1)	13	86,67	1,09	68,18	0,97	66,08
54	(1,1,1,0,1,1,1,1,1,1,0,1,1,1,1)	13	86,67	1,09	68,18	0,96	65,93
55	(0,1,1,0,1,1,1,1,1,1,1,1,1,1,1)	13	86,67	1,09	68,18	0,72	61,98
56	(1,1,1,0,1,1,1,1,0,1,1,1,1,1,1)	13	86,67	1,09	68,18	1,00	66,62
57	(1,1,1,1,1,1,1,1,1,0,1,1,1,0,1)	13	86,67	1,09	68,18	1,02	67,00
58	(1,1,0,1,1,1,1,1,0,1,1,1,1,1,1)	13	86,67	1,09	68,18	1,06	67,65

https://dx.doi.org/10.26594/jmpm.v8i2.3293

No	Pola jawaban siswa	Konvensional		1PL		2PL	
		Skor	Skor (0-100)	Skor IRT	Skor (0-100)	Skor IRT	Skor (0-100)
59	(1,1,1,0,1,1,1,1,0,1,1,1,1,1,1)	13	86,67	1,09	68,18	1,00	66,62
60	(1,0,1,0,1,1,1,1,1,1,1,1,1,1,1)	13	86,67	1,09	68,18	0,81	63,48
61	(1,1,1,1,1,1,1,1,1,1,1,1,1,1,0)	14	93,33	1,48	74,68	1,32	71,92
62	(1,1,1,1,1,1,1,1,1,1,1,1,0,1,1)	14	93,33	1,48	74,68	1,11	68,55
63	(1,1,1,1,1,1,1,1,1,1,0,1,1,1,1)	14	93,33	1,48	74,68	1,29	71,57
64	(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)	15	100,00	1,95	82,48	1,67	77,80
65	(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)	15	100,00	1,95	82,48	1,67	77,80

Perbedaan model konvensional dan model IRT dalam penskoran juga terlihat jelas pada skor siswa yang mampu menjawab semua butir dengan benar. Model konvensional memberikan skor sempurna yaitu 100. Sebaliknya, model konvensional akan memberikan skor 0 untuk siswa yang menjawab semua butir dengan salah, karena tidak didasarkan pada parameter butir tes. Pada IRT tidak demikian. Parameter kemampuan berbanding lurus dengan parameter *difficulty* (Hambleton dkk., 1991; Sudaryono, 2013). Siswa yang memiliki kemampuan tinggi akan memiliki kemungkinan yang lebih tinggi untuk menjawab benar butir dengan *difficulty* yang sulit, bahkan sangat sulit, dan begitu juga sebaliknya. Sifat ini juga dapat dibuktikan melalui Rumus 1 di atas. Sifat inilah yang menyebabkan penskoran menggunakan IRT pada Tabel 6 tidak memberikan skor sempurna pada siswa yang memiliki jawaban benar semua. Dengan kata lain, pada Tabel 5 pada model 1PL dan 2PL memiliki *difficulty* pada level mudah dan sedang saja, tidak ada butir pada level sangat mudah, sulit ataupun sangat sulit. Oleh karena itu, penskoran menggunakan IRT sangat ditentukan oleh karakteristik butir yaitu parameter-parameter yang melekat pada setiap butir. Bahkan ketika terdapat siswa memiliki jawaban salah semua, skor yang diperoleh tidak nol. Siswa dengan jawaban semua salah, pada model 1PL dan 2PL berturut-turut akan memiliki skor -2.41 (9.83) dan -2.28 (12). Sebaliknya, siswa yang menjawab semua butir dengan benar akan memiliki skor 1,95 (82.48) untuk model 1PL dan 1.67 (77.8) untuk model 2PL. Ini membuktikan bahwa penskoran tes matematika dengan pilihan ganda benar-benar didasarkan pada karakteristik dan kemampuan dari butir tes sehingga hasil penskoran menjadi sangat objektif dan fair.

Seperti yang dijelaskan pada bagian metode, bahwa tes yang digunakan pada penelitian ini sebanyak 15. Dilihat secara dikotomi, siswa hanya memiliki dua kemungkinan respon jawaban yaitu benar (1) atau salah (0), sehingga banyaknya pola respon jawaban yang mungkin terjadi sebanyak 2^{15} = 32768. Sedangkan data yang digunakan hanya berasal dari 65 siswa saja, anggap saja pola respon yang dihasilkan dari 65 siswa ini berbeda, berarti masih terdapat 32708 pola respon yang tidak digunakan dalam estimasi parameter butir menggunakan masing-masing model IRT diatas. Uniknya, berdasarkan parameter butir yang diperoleh diatas, kedua model IRT masih mampu mengestimasi skor siswa yang memiliki pola respon yang tidak digunakan untuk estimasi parameter (Brown & Croudace, 2014). Dengan demikian, instrumen matematika pada konten aljabar ini dapat digunakan untuk estimasi skor siswa yang tidak menjadi sampel dalam penelitian. informasi Ini menunjukkan bahwa parameter-parameter butir yang terbentuk menjadi acuan standar dalam menentukan skor kemampuan matematika siswa.

Selanjutnya berdasarkan TIF dari masing-masing model mengkonfirmasi keberfungsian dari tes yang digunakan. Rentang estimasi skor dari model 1PL -2,41 sampai 1,95 masih berada dalam rentang kemampuan -3.205 sampai 2.105. Begitu

https://dx.doi.org/10.26594/jmpm.v8i2.3293

pula pada model 2PL rentang estimasi skor mulai dari -2,28 sampai 1.67 masih berada pada rentang kemampuan -2.925 sampai 1.925. Secara tidak langsung ini membuktikan presisi pengukuran kemampuan matematika siswa dengan menggunakan instrumen tes yang digunakan. Nilai SEM yang dihasilkan model 1PL dan 2PL berturut-turut sebesar 0.26 dam 0.23, keduanya nilai SEM tersebut <0.3. Hasil ini menjelaskan bahwa instrumen tes memiliki kualitas yang baik (Retnawati, 2014).

Sebagai catatan, secara teknis penskoran menggunakan konsep IRT tidak mudah digunakan. Beberapa faktor penyebabnya yaitu 1) proses kalibrasi instrumen tidak mudah dilakukan secara manual, sehingga tanpa bantuan *software* IRT metode ini tidak akan digunakan. 2) proses kalibrasi yang ketat, terutama pada uji asumsi yang sangat mungkin tidak terpenuhi. Namun, kendala teknis ini akan menjadi mudah ketika dibiasakan atau ada sosialisasi yang masif terkait aplikasi penggunaan IRT di lapangan, khususnya di sekolah.

Hasil analisis data dan diskusi di atas membuktikan bahwa penskoran dengan konsep IRT menawarkan penskoran yang objektif dan *fair*. Hasil perbandingan menunjukkan bahwa model 2PL mampu mengestimasi skor dan sekaligus mampu membedakan kemampuan siswa. Selanjutnya, model 1PL mampu mengestimasi skor kemampuan, tetapi tidak mampu membedakan kemampuan siswa. Sedangkan pada metode konvensional penskoran yang dihasilkan tidak mampu merepresentasikan kemampuan siswa, karena tidak melibatkan parameter butir dalam penskoran, walaupun proses ketat harus dialului untuk kalibrasi dalam menghasilkan parameter-parameter butir. Melalui studi ini, kami menyarankan bahwa penskoran tes matematika berbentuk pilihan ganda yang fair dapat dilakukan dengan menggunakan konsep IRT.

KESIMPULAN DAN SARAN

Sesuai dengan tujuan dari artikel ini, maka dapat disimpulkan bahwa model IRT mampu memberikan penskoran yang lebih objektif dan *fair*. Penskoran dengan model IRT benar-benar didasarkan pada karakteristik dari butir tes. Karakteristik ini ditunjukkan oleh parameter-parameter butir pada instrumen yang secara langsung memiliki peran dalam mengestimasi skor kemampuan siswa. Model 2 PL mampu mberikan penskoran yang paling *fair* dari pada model 1PL dan konvensional. Karakteristik parameter diskriminan pada model 2PL, membuatnya mampu membedakan kemampuan siswa, khususnya pada siswa-siswa yang memiliki jumlah jawaban benar yang sama, tetapi pola jawabannya berbeda.

Kelemahan dalam studi ini antara lain yaitu jumlah responden yang digunakan masih kurang, sehingga dalam penelitian selanjutnya responden perlu ditambahkan dengan tujuan menghasilkan parameter-parameter butir instrumen yang memiliki akurasi yang lebih bagus dalam estimasi pramter kemampuan siswa. Jumlah butir yang mewakili *difficulty* pada level sangat mudah, sulit dan sangat sulit juga belum ada, sehingga menyebabkan tidak terdapat skor 0 dan 100. Pada penelitian selanjutnya diharapkan mampu menggunakan butir tes yang lebih banyak, sehingga dapat memberikan parameter-parameter butir yang lebih bervariasi dengan level *difficulty* dan *discriminant*, bahkan *pseudo guessing* yang lebih bervariasi. Penskoran ini juga lebih cocok digunakan pada sekolah-sekolah yang memiliki banyak siswa. Selain itu, penskoran ini masih dilakukan dengan menggunakan kode-kode bahasa R yang dan tidak mudah dipahami oleh pengguna yang tidak terbiasa dengan program R. Selanjutnya perlu dikembangkan *package* atau aplikasi penskoran dengan *user interface* yang dapat memudahkan penskoran dengan konsep IRT menggunakan program R dan tentunya tanpa harus menggunakan kode-kode dalam bahasa pemrograman R.

https://dx.doi.org/10.26594/jmpm.v8i2.3293

DAFTAR RUJUKAN

Ali, S. H., Carr, P. A., & Ruit, K. G. (2016). Validity and reliability of scores obtained on multiple-choice questions: Why functioning distractors matter. *Journal of the Scholarship of Teaching and Learning*, *16*(1), 1–14. https://doi.org/10.14434/josotl.v16i1.19106

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education*, *25*(1), 31–36. https://doi.org/10.1177/0273475302250570

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

Betts, L. R., Elder, T. J., Hartley, J., & Trueman, M. (2009). Does correction for guessing reduce students' performance on multiple-choice examinations? yes? no? sometimes?. *Assessment & Evaluation in Higher Education*, *34*(1), 1–15. https://doi.org/10.1080/02602930701773091

Bichi, A. A., & Talib, R. (2018). Item response theory: An introduction to latent trait models to test and item development. *International Journal of Evaluation and Research in Education (IJERE)*, *7*(2), 142. https://doi.org/10.11591/ijere.v7i2.12900

Bleske-Rechek, A., Zeug, N., & Webb, R. M. (2007). Discrepant performance on multiple-choice and short answer assessments and the relation of performance to general scholastic aptitude. *Assessment & Evaluation in Higher Education*, *32*(2), 89–105. https://doi.org/10.1080/02602930600800763

Brown, A., & Croudace, T. (2014). Scoring and estimating score precision using multidimensional irt models. in S. P. Reise & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling* (pp. 325–351). Routledge. https://doi.org/10.4324/9781315736013-26

Brown, G. T. L., & Abdulnabi, H. H. A. (2017). Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades. *Frontiers in Education*, *2*. https://doi.org/10.3389/feduc.2017.00024

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing Patient-Reported Outcomes Measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Chalmers, R. P. (2012). Mirt : A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, *48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, *71*(5). https://doi.org/10.18637/jss.v071.i05

DeMars, C. (2010). *Item response theory*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001

Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*, *23*(1), 138–149. https://doi.org/10.1037/met0000121

Hair, J. F. J., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis. Cengage.

Hambleton, R., Swaminathan, H., & Rogers, H. J. (1991). *Fundamental of item response theory*. SAGE Publications.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care*, *38*(9 Suppl), 1128-1142. https://doi.org/10.1097/00005650-200009002-00007

Huda, N., & Mardapi, D. (2015). Komparasi model penskoran berdasarkan teori respons butir pada soal ujian nasional mata pelajaran matematika. *Jurnal Evaluasi Pendidikan*, *3*(1), 56-66. https://journal.student.uny.ac.id/ojs/index.php/jep/article/view/1225

Huntley, B., Engelbrecht, J., & Harding, A. (2009). Can multiple choice questions be successfully used as an assessment format in undergraduate mathematics? *Pythagoras*, *69*, 3-16. https://doi.org/10.4102/pythagoras.v0i69.41

Kasanova, R., & Sulistiyono, R. (2023). Evaluasi butir soal pilihan ganda penilaian tengah semester dalam pembelajaran tematik untuk kelas v di sdn gladak anyar 4 pamekasan. *Journal on Education*, *6*(1), 5820–5834. https://jonedu.org/index.php/joe/article/view/3773

Kemendikbud. (2021). *Asesmen nasional: lembar tanya jawab*. Jakarta: Pusat Asesmen dan Pembelajaran.

Khaerudin. (2016). Teknik penskoran tes obyektif model pilihan ganda. *Jurnal Madaniyah*, 6(2), 183-200. https://journal.stitpemalang.ac.id/index.php/madaniyah/article/view/27/14

Lesage, E., Valcke, M., & Sabbe, E. (2013). Scoring methods for multiple choice assessment in higher education – Is it still a matter of number right scoring or negative marking? *Studies in Educational Evaluation*, *39*(3), 188–193. https://doi.org/10.1016/j.stueduc.2013.07.001

Lopes, A. P., Babo, L., Azevedo, J., & Torres, C. (2010). Multiple-choice tests - a tool in assessing knowledge. *Proceedings of INTED 2010 - 4th International Technology, Education and Development Conference*. https://doi.org/978-84-613-5538-9

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with r : recent updates of the package catr. *Journal of Statistical Software*, *76*(Code Snippet 1). https://doi.org/10.18637/jss.v076.c01

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research & Perspective*, *11*(3), 71–101. https://doi.org/10.1080/15366367.2013.831680

Maydeu-Olivares, A. (2014). Evaluating the fit of irt models. in s. p. reise & d. a. revicki (eds.), *Handbook of Item Response Theory Modeling* (pp. 129–145). Routledge. https://doi.org/10.4324/9781315736013-15

OECD. (2013). *PISA 2012 assessment and analytical framework: mathematics, reading, science, problem solving and financial literacy.* OECD Publishing.

OECD. (2019). What students know and can do. *OECD Multilingual Summaries PISA 2018 Results (Volume I)*. OECD Publishing.

Paek, I., & Cole, K. (2019). *Using r for item response theory model applications*. London: Routledge. https://doi.org/10.4324/9781351008167

R Core Team. (2022). *R: a language and environment for statistical computing*. Vienna : R Foundation for Statistical Computing. https://www.r-project.org/

Retnawati, H. (2014). *Teori respons butir dan penerapannya*. Yogyakarta: Nuha Medika.

Retnawati, H. (2015). The comparison of accuracy scores on the paper and pencil testing vs. computer-based testing. *Turkish Online Journal of Educational Technology*, *14*(4), 135-142. http://www.tojet.net/articles/v14i4/14413.pdf

Rosseel, Y. (2012). Lavaan : an r package for structural equation modeling. *Journal of Statistical Software*, *48*(2). https://doi.org/10.18637/jss.v048.i02

Schaughency, E., Smith, J. K., Meer, J. van der, & Berg, D. (2012). *Classical test theory and higher education: five questions*. In C. Secolsky & D. B. Denison (Eds.), Handbook on measurement, assessment, and evaluation in higher education. (pp. 117-131). New York: Routledge.

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking.

Practical Assessment, Research and Evaluation, *22*(4), 1-13. https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1355&context=pare

Stankous, N. V. (2016). Constructive response vs. multiple-choice tests in math: american experience and discussion (review). *European Scientific Journal (Special Edition)*,308-316. https://core.ac.uk/download/pdf/328025438.pdf

Sudaryono. (2013). *Teori responsi butir (edisi pertama)*. Yogyakarta: Graha Ilmu.

Susanto, H. P., Retnawati, H., Abadi, A. M., Haryanto, H., & Ali, R. M. (2023). *Irtawsi: items response theory analysis with steps and interpretation (r package version 0.3.4)*. CRAN R Pgroam. https://cran.r-project.org/package=irtawsi

Thompson, N. (2021). *Classical test theory vs. item response theory*. https://assess.com/classical-test-theory-vs-item-response-theory/

Torres, C., Lopes, A. P., Babo, L., & Azevedo, J. (2011). Improving multiple-choice questions. *US-China Education Review*, *B*(1), 1-11. https://files.eric.ed.gov/fulltext/ED522219.pdf

Ypsilandis, G. S., & Mouti, A. (2019). Fairness and ethics in multiple choice (mc) scoring: An empirical study. *Journal of Language and Education*, *5*(1), 85–97. https://doi.org/10.17323/2411-7390-2019-5-1-85-97

Yuksel, M. E., & Fidan, H. (2019). A decision support system using text mining based grey relational method for the evaluation of written exams. *Symmetry*, *11*(*11*), *1426*, 1-24. https://doi.org/10.3390/sym11111426