JURNAL STATISTIKA DAN SAINS DATA
Volume 2 Nomor 2, April, 2025, 164-175
https://journal.student.uny.ac.id/index.php/jssd
http://dx.doi.org/10.21831/jssd.v2i2.21895

PENYETARAAN PERANGKAT TES IPA SMP BERDASARKAN TEORI TES KLASIK DAN TEORI RESPONS BUTIR

Isnu Bima Saputra*, Universitas Negeri Yogyakarta
Heri Retnawati, Universitas Negeri Yogyakarta
*e-mail: Isnubima.2019@student.uny.ac.id

Abstrak. Penelitian ini bertujuan untuk 1) Mendeskripsikan karakteristik butir perangkat tes IPA dengan berdasarkan parameter yang digunakan dalam penelitian ini; 2) Mengetahui hasil penyetaraan perangkat tes IPA dengan menggunakan teori tes klasik dan modern (teori respons butir); 3) Mengetahui perbedaan keakuratan hasil penyetaraan perangkat tes UNBK IPA SMP tahun 2016 dengan pendekatan teori tes klasik dan modern (teori respons butir). Data yang digunakan adalah jawaban dari peserta tes UNBK IPA SMP tahun 2016 dengan total peserta sebanyak 16.765 dan terdapat 5 paket soal. Karakteristik butir dan penyetaraan tes dilakukan dengan pendekatan teori tes klasik dan teori respons butir. Pada pendekatan teori respons butir, analis data menggunakan model 3PL. Penyetaraan tes dengan teori tes klasik menggunakan 2 metode yaitu linear dan *equipercentile*, sedangkan dengan teori respons butir menggunakan metode *mean-mean*, *mean-sigma*, haebara, dan *stocking-lord*. Pada pendekatan teori tes klasik dan teori respons butir, karakteristik butir UNBK IPA SMP se-DIY memiliki rata - rata tingkat kesukaran butir, daya beda, dan *guessing* berkategori baik. Penyetaraan tes menghasilkan skor konversi yang bervariasi terdapat skor konversi yang naik ataupun turun jika dibandingkan dengan skor awal. Hasil penyetaraan tes metode haebara merupakan metode yang paling akurat dibandingkan metode lainnya.

Kata kunci: Teori tes klasik, Teori respons butir, UNBK IPA SMP se-DIY.

Abstract. *This study aims to 1) Describe the characteristics of the science test items based on the parameters used in this study; 2) Determine the results of the equivalence of the science test using classical and modern test theories (item response theory); 3) Determine the differences in the accuracy of the equivalence results of the 2016 SMP Science UNBK test with the classical and modern test theory approaches (item response theory). The data used are the answers of the 2016 SMP Science UNBK test participants with a total of 16,765 participants and there are 5 question packages. Item characteristics and test equivalence are carried out using the classical test theory and item response theory approaches. In the item response theory approach, data analysts use the 3PL model. Test equivalence with classical test theory uses 2 methods, namely linear and equipercentile, while with item response theory using the mean-mean, mean-sigma, haebara, and stocking-lord methods. In the classical test theory approach and item response theory, the characteristics of the UNBK IPA SMP items throughout DIY have an average level of item difficulty, discrimination power, and guessing in the good category. The test equivalency produces varying conversion scores, there are conversion scores that increase or decrease when compared to the initial score. The results of the haebara method test equivalency are the most accurate method*

compared to other methods.

Keywords: Classical test theory, Item response theory, UNBK Science Junior High School in throughout DIY.

PENDAHULUAN

Evaluasi merupakan salah satu rangkaian kegiatan untuk meningkatkan kualitas, kinerja, atau produktivitas suatu lembaga selama melaksanakan program sehingga diperoleh informasi yang telah ataupun belum tercapai, dan dari informasi tersebut dapat dilakukan perbaikan (Djemari Mardapi, 2008: 8). Salah satu evaluasi dalam dunia pendidikan adalah dengan adanya tes hasil belajar. Tes hasil belajar merupakan media yang digunakan untuk mengetahui kemampuan dari peserta didik selama proses pembelajaran yang telah dilakukan sebelumnya. Menurut Sudijono (2009: 67) tes merupakan cara yang dapat dipergunakan atau prosedur yang harus ditempuh dalam rangka pengukuran dan penilaian di bidang pendidikan karena perannya yang dapat menggambarkan ukuran kemampuan peserta didik selama mengikuti pembelajaran. Oleh karena itu, Pengukuran dan penilaian dalam dunia pendidikan bisa dibilang memiliki peranan yang cukup penting karena hasil dari pengukuran dan penilaian tersebut merupakan cerminan kemampuan dari peserta didik. Pada tahun 2016, tes hasil belajar yang dilakukan untuk mengetahui hasil evaluasi peserta didik di akhir pembelajaran dilakukan menggunakan computer (UNBK). UNBK untuk SMP dan sederajat menggunakan 4 perangkat tes yang disiapkan meliputi Matematika, Bahasa Indonesia, Bahasa Inggris, dan Ilmu Pengetahuan Alam (IPA) serta pada setiap perangkat tes terdapat beberapa paket soal yang digunakan.

Demikian pula dengan perangkat tes IPA, terdapat 5 paket soal yang digunakan. Paket-paket soal tersebut diharapkan memiliki bobot sama dikarenakan dikembangkan dari standar kompetensi yang sama. Namun pada kenyataannya, bisa terjadi antara perangkat tes 1 dengan perangkat yang lain memiliki bobot tes yang berbeda seperti perangkat tes 1 lebih mudah daripada perangkat tes 2 ataupun sebaliknya. Hambleton & Swaminathan (1991) mengatakan bahwa meskipun perangkat tes yang disusun berdasarkan kisi-kisi yang sama, hampir tidak terdapat perangkat tes yang mempunyai tingkat kesukaran yang sama. Penggunaan beberapa paket soal dalam tes memiliki keunggulan dan kelemahan. Keunggulan menggunakan beberapa paket soal yaitu dapat mengurangi potensi kecurangan siswa selama tes berlangsung. Sementara itu kelemahan dari menggunakan beberapa paket soal yaitu jaminan bahwa bobot ataupun perangkat tes yang digunakan adalah setara. Kesetaraan pada perangkat tes ini dapat dilakukan dengan menggunakan metode penyetaraan skor tes.

Brennan & Kolen (2014) mengungkapkan untuk menghubungkan skor-skor tes yang memiliki bentuk-bentuk lain yang dibangun dengan spesifikasi yang sama (penyetaraan tes). Penyetaraan merupakan proses statistik yang digunakan untuk mengatur beberapa perangkat tes sehingga masing-masing perangkat tes memiliki bobot yang sama. Selain itu, penyetaraan sebagai proses mengkonversi soal pada beberapa perangkat tes yang setara atau paralel. Hal serupa juga dikemukakan oleh Hambleton, Swaminathan, dan Rogers (1991) bahwa penyetaraan merupakan perbandingan hasil tes satu dari perangkat tes yang berbeda dengan penyetaraan skor pada kedua tes. Penyetaraan dapat diartikan suatu proses statistik yang digunakan untuk menyetarakan skor tes dari beberapa perangkat tes yang berbeda.

Penyetaraan terdiri dari dua jenis, yaitu jenis horizontal dan vertikal (Retnawati, 2014: 95-96). Penyetaraan jenis horizontal berarti proses penyetaraan menggunakan beberapa perangkat soal dan diberikan terhadap kelompok yang memiliki tingkat kemampuan yang sama. Sedangkan penyetaraan jenis vertikal berarti proses penyetaraan menggunakan beberapa

perangkat soal dan diberikan terhadap kelompok yang memiliki tingkat kemampuan yang berbeda. Proses penyetaraan dilakukan dengan 2 pendekatan, yaitu pendekatan teori tes klasik dan teori respons butir. Pendekatan teori tes klasik dapat dilakukan dengan menggunakan dua metode yaitu metode linear dan *equipercentile.* Sementara, penyetaraan menggunakan teori respons butir dilakukan dengan 4 metode yaitu mean-mean, mean-sigma, haebara, dan stocking-lord.

Penelitian ini dimaksudkan untuk 1) Mendeskripsikan karakteristik butir perangkat tes IPA dengan berdasarkan parameter yang digunakan dalam penelitian ini. 2) Mengetahui hasil penyetaraan perangkat tes IPA dengan menggunakan teori tes klasik dan modern (teori respons butir). 3) Mengetahui perbedaan keakuratan hasil penyetaraan perangkat tes UNBK IPA SMP tahun 2016 dengan pendekatan teori tes klasik dan modern (teori respons butir).

METODE

Data yang digunakan dalam penelitian ini adalah data hasil jawaban UNBK SMP se-DIY mata pelajaran IPA tahun 2016. Berdasarkan data yang ada, terdapat 529 lembaga pendidikan dengan sampel hasil jawaban UNBK sebanyak 16.765 yang digunakan dalam penelitian ini. Analisis perangkat soal IPA meliputi uji asumsi, karakteristik butir, dan penyetaraan tes yang akan dilakukan dengan *software* RStudio.

Pada pendekatan teori tes klasik, karakteristik butir meliputi tingkat kesukaran butir (b) dan daya beda (a) akan dianalisis menggunakan *package CTT*. Parameter tingkat kesukaran butir (b) di bawah 0,3 dikategorikan sulit, untuk nilai 0,3 – 0,7 dikategorikan sedang, dan untuk di atas 0,7 dikategorikan mudah. Sementara itu, nilai daya beda (a) jika memiliki nilai negatif dikategorikan tidak memenuhi (TM), 0 – 0,3 akan dikategorikan rendah, 0,3 – 0,7 akan dikategorikan sedang, dan 0,7 ke atas akan dikategorikan tinggi.

Berbeda dengan teori tes klasik, pendekatan dengan teori respons butir akan menggunakan data dengan nilai daya beda negatif (dari karakteristik butir teori tes klasik di atas). Setelah itu, estimasi parameter akan dilakukan untuk mencari model parameter logistik yang akan digunakan. Model parameter logistik yang akan diuji kecocokannya terdiri dari model rasch, 2PL, 3PL, dan 4PL. Hambleton, Swaminathan, dan Rogers (1991: 12-17) mengatakan pada model rasch ini berasumsi bahwa indeks kesukaran butir (b) merupakan satu-satunya yang berpengaruh terhadap kemampuan peserta tes, dengan nilai b_i yang baik berkisar antara $-2 < b_i < 2$. berasumsi bahwa terdapat 2 parameter butir yang berpengaruh terhadap kemampuan peserta tes, yaitu indeks kesukaran butir (b) dan daya pembeda (a), dengan nilai a_i yang baik berkisar antara $0 < a_i < 2$. Model 3PL berasumsi bahwa terdapat 3 parameter butir yang berpengaruh terhadap kemampuan peserta tes, yaitu indeks kesukaran butir (b), daya pembeda (a), dan faktor menebak (*pseudo guessing*/ c). Nilai c_i yang baik berkisar antara $0 < c_i < 1/k$, dengan k= opsi jawaban = 4, maka c_i dikatakan baik jika $0 < c_i < 0{,}25$. Model 4PL berasumsi bahwa terdapat 4 parameter butir yang berpengaruh terhadap kemampuan peserta tes, yaitu indeks kesukaran butir (b), daya pembeda (a), faktor menebak (*pseudo guessing*, c), dan faktor ceroboh (*careless*). Faktor ceroboh membuat peserta tes dengan kemampuan tinggi dapat menjawab dengan salah untuk butir soal dengan tingkat kesukaran yang lebih rendah Hambleton et al (1991: 48-49)

Hidayati (2005) menjelaskan bahwa terdapat 3 (tiga) asumsi dasar dalam teori respons butir, yaitu unidimensi, independensi lokal, dan invariansi parameter. Unidimensi berarti butir-butir soal pada perangkat tes hanya mengukur satu kemampuan. Uji asumsi unidimensi akan terpenuhi jika hasil menunjukkan pada dimensi 1 nilai eigen lebih tinggi dari dimensi yang lain. Independensi lokal berarti kemampuan-kemampuan peserta yang berhubungan dengan tes

dianggap konstan, maka respons peserta terhadap butir soal pada perangkat tes secara statistik tidak saling terkait (independen). Retnawati (2014:3) mengatakan bahwa asumsi indepedensi lokal akan terpenuhi jika asumsi unidimensi terpenuhi. Sementara itu, invariansi parameter berarti parameter butir bergantung pada parameter kemampuan peserta tes dan begitu juga sebaliknya.

Penyetaraan dengan pendekatan teori tes klasik dilakukan dengan metode linear dan metode *equipercentile*. Menurut Kolen dan Brennan (2014: 31) pada penyetaraan linear skor yang jaraknya sama dari rata-rata (standar deviasi) ditetapkan sama. konversi linear dilakukan dengan menetapkan skor deviasi standar (nilai-z) yang sama pada perangkat tes yang digunakan. Persamaan dengan menggunakan metode linear adalah sebagai berikut.

$$l_Y(x) = y = \frac{\sigma(Y)}{\sigma(X)}[x - \mu(X)] + \mu(Y) \quad (1)$$

Sementara pada metode *equipercentile*, perbedaan isi dan tingkat kesukaran antar tes dijelaskan menggunakan transformasi non linear dengan persamaan:

$$eq_Y(x) = G^{-1}[F(x)] \quad (2)$$

Metode *equipercentile* memungkinkan hasil penyetaraan berada pada kisaran skor perangkat tes yang di setarakan. Meskipun begitu, pada metode *equipercentile* data diasumsikan sebagai variabel acak kontinu bukan sebagai variabel diskrit. Pendekatan yang dilakukan agar metode *equipercentile* dapat digunakan adalah dengan menggunakan pendekatan pengkontinuan variabel diskrit (Kolen dan Brennan, 2014: 38).

Kesalahan pengukuran standar (*standard error of Equating, SEE*) berfungsi sebagai tolak ukur keakuratan metode-metode di atas. Menurut Kolen dan Brennan (2014), SEE pada desain kelompok yang ekuivalen ditentukan dengan persamaan:

$$var[\hat{l}_y(x_i)] \cong \frac{\sigma^2(Y)[1-\rho(X,Y)]}{N}\left\{2 + [1+\rho(X,Y)]\left[\frac{x_i - \mu(X)}{\sigma(X)}\right]^2\right\} \quad (3)$$

Keakuratan metode dapat dilihat dari nilai SEE. semisal metode menghubungkan tes C_1 lebih akurat daripada metode menghubungkan tes C_2 jika $SE(C_1) < SE(C_2)$.

Pada pendekatan teori respons butir, Retnawati (2014: 105-106) mengungkapkan jika model pada teori respons butir sesuai dengan data yang digunakan, maka transformasi linear pada skala θ juga akan sesuai dengan data tersebut. Hubungan antara parameter butir dengan kemampuan peserta tes jika dilakukan transformasi linear pada penyetaraan teori respons butir model 3PL dapat diuraikan sebagai berikut (Kolen dan Brennan, 2014: 178).

$$\theta_{Ii} = A\theta_{Ii} + B \quad (4)$$

Selanjutnya, parameter butir pada skala di atas dapat ditulis sebagai berikut.

$$a_{Jj} = \frac{a_{Ij}}{A} \quad (5)$$

$$b_{Ij} = Ab_{Ij} + B \quad (6)$$

$$a_{Ij} = a_{Ij} \quad (7)$$

Metode penyetaraan yang terdapat pada teori respons butir, berdasarkan metode kalibrasi, dapat dibagi menjadi 2 yaitu metode momen dan metode grafik. Pada metode momen, terdapat beberapa metode antara lain metode mean-mean serta metode mean-sigma. Sementara untuk metode grafik, terdapat metode kurva karakteristik dari Haebara serta kurva karakteristik dari Stocking & Lord.

Pada metode Mean-Mean, parameter yang digunakan dalam menentukan konstanta penyetaraan (α dan β) adalah rerata dari parameter daya pembeda (α) dan indeks kesukaran butir (b). Hubungan antara daya pembeda dengan indeks kesukaran butir dari dua kelompok tes dapat dituliskan dalam persamaan berikut:

$$b_2 \alpha b_1 + \beta, a_2 = \frac{a_1}{\alpha} \tag{8}$$

Pada metode Mean-sigma, parameter yang digunakan dalam menentukan konstanta penyetaraan (α dan β) adalah rerata parameter indeks kesukaran butir (b) dan simpangan baku indeks kesukaran butir (S). Hubungan antara indeks kesukaran dengan simpangan baku dari indeks kesukaran butir adalah hubungan linear dengan,

$$b_2 = \alpha b_1 + \beta \tag{9}$$

Penyetaraan pada metode haebara didasarkan pada fungsi karakteristik butir. konstanta penyetaraan diperoleh dari selisih nilai fungsi absis yang sama pada setiap kurva karakteristik dari skala yang telah di setarakan kemudian dikuadratkan dan dijumlah. Bentuk persamaan pada metode ini dapat dituliskan sebagai berikut.

$$H(\theta_i) = \sum_{j=1}^{n} (T_{ij} - T_{ij}^*)^2 \tag{10}$$

Sama halnya seperti metode Haebara, metode Stocking & Lord juga didasarkan fungsi karakteristik butir. konstanta penyetaraan diperoleh dari selisih nilai fungsi untuk absis yang sama pada setiap kurva karakteristik dari skala yang telah di setarakan kemudian dikuadratkan, dengan kriteria tertentu konstanta α dan β didapatkan dari meminimumkan fungsi tertentu yang memuat variabel α dan β.

$$SL(\theta_i) = (T_i - T_i^*)^2 \tag{11}$$

HASIL DAN PEMBAHASAN

Hasil

Teori tes klasik

Karakteristik butir dengan pendekatan teori tes klasik masing-masing paket soal dapat dilihat pada Tabel 1.

Tabel 1. Hasil karakteristik butir dengan teori tes klasik

Kriteria	Paket 1	Paket 2	Paket 3	Paket 4	Paket 5
Banyak butir	40	40	40	40	40
Banyak peserta tes	3501	3321	3357	3279	3307
Reliabilitas	0,915	0,914	0,908	0,913	0,906
Validitas	0,479	0,479	0,470	0,475	0,457
Rerata tingkat kesukaran butir	0,674	0,719	0,679	0,681	0,690
Rerata daya beda	0,430	0,440	0,428	0,435	0,415
SEM	2,414	2,380	2,504	2,490	2,419

Tabel 1 menunjukan bahwa paket 1 sampai paket 5 memiliki koefisien reliabilitas yang tinggi, yaitu berkisar antara 0,908 – 0,915 dan validitas yang berkisar antara 0,457 – 0,479 yang juga berkategori baik. Sementara, parameter tingkat kesukaran butir pada masing-masing paket memiliki rata-rata yang baik, yaitu 0,674 – 0,719 yang berkategori sedang dan mudah. rata-rata daya beda juga tergolong baik (sedang) yang berkisar antara 0,415 – 0,440. *Standard Error Measurment* (SEM) menunjukan nilai 2,380 – 2,504, relatif rendah, yang dapat diartikan bahwa hasil dari pengukuran masing-masing paket cenderung stabil dan konsisten.

Teori respons butir

Pada pendekatan teori respons butir estimasi parameter butir/ penentuan model parameter logistik dilakukan terlebih dahulu. Data yang akan digunakan sesuai dengan hasil karakteristik butir pada teori tes klasik untuk butir-butir yang memenuhi kriteria.

Tabel 2. Rangkuman hasil analisis estimasi parameter butir

Paket soal	Model	Jumlah kecocokan model	AIC	BIC
Paket 1	Rasch	4	122838,9	123079,2
	2PL	14	119946,4	120414,6
	3PL	26	117205,5	117907,8
	4PL	28	117041,0	117977,4
Paket 2	Rasch	5	119846,6	120097,0
	2PL	25	116744,3	117233,0
	3PL	30	114957,1	115690,1
	4PL	29	114189,8	115167,1
Paket 3	Rasch	4	133178,6	133429,5
	2PL	23	129129,4	129618,9
	3PL	29	127332,3	128066,5
	4PL	30	127340,9	128319,9
Paket 4	Rasch	6	129340,9	129585,6
	2PL	21	125311,0	125788,3
	3PL	31	123437,8	124153,7
	4PL	29	122491,9	123446,4
Paket 5	Rasch	6	111974,9	112212,9
	2PL	20	108633,0	109096,9
	3PL	26	106026,6	106722,5
	4PL	25	106123,8	107051,6

Butir-butir yang tidak memenuhi kriteria karena memiliki daya beda negatif terdapat pada butir 3 dan 8 pada paket 1 dan paket 5 serta butir 39 pada paket 4. Pada masing-masing paket soal akan dilakukan analisis dengan membandingkan model *Rasch*, model 2PL, model 3PL, dan model 4PL dengan melihat dari banyaknya butir soal yang cocok pada masing-masing model, dan dari nilai AIC serta BIC. Kriteria butir soal dapat dikatakan cocok apabila *p-value* > 0,05 dan untuk nilai AIC serta BIC, semakin kecil nilainya maka menunjukkan model terbaik. Analisis data akan menggunakan *software* Rstudio dengan *package MIRT*.

Berdasarkan Tabel 2, model parameter logistik terbaik adalah model 3PL. Pada model 3PL memiliki jumlah model cocok terbanyak yaitu pada paket soal 2, 4, dan 5, memiliki nilai AIC terkecil terbanyak pada paket soal 3, 4, dan 5, serta memiliki nilai BIC terkecil terbanyak pada paket soal 1, 3, dan 5. Uji asumsi teori respons butir terdiri dari unidimensi, indepedensi lokal, dan invariansi parameter. Uji unidimensi dilakukan dengan analisis faktor menggunakan *software* Rstudio dengan *package get_eigenvalue()*, dengan hasil pada Tabel 3.

Tabel 3. Nilai *eigen*

Dimensi	Paket 1	Paket 2	Paket 3	Paket 4	Paket 5
Dim.1	10,278	9,775	9,480	9,990	10,037
Dim.2	1,896	1,938	1,918	1,788	2,044
Dim.3	1,432	1,531	1,762	1,576	1,369
Dim.4	1,280	1,470	1,490	1,352	1,276
Dim.5	1,200	1,306	1,347	1,302	1,191
Dim.6	1,153	1,205	1,169	1,253	1,092

Terlihat dari Tabel 3 bahwa pada semua paket soal nilai eigen pada dimensi 1 kurang lebih 5x dari nilai eigen terdekat dari dimensi 2. Sesuai dengan hal tersebut, maka dapat dikatakan bahwa asumsi unidimensi terpenhi. Asumsi indepensi lokal akan terpenuhi jika asumsi unidimesi terpenuhi sesuai pada bagian metode di atas.

Uji asumsi invariansi parameter dilakukan dengan membagi data menjadi 2 yaitu peserta dengan urutan nomor ganjil dan genap. Asumsi invariansi parameter yang akan diuji menggunakan model 3PL dengan parameter daya beda (a), tingkat kesukaran butir (b), *pseudo guessing*, dan kemampuan peserta tes (θ).

Tabel 4. Korelasi invariansi parameter

Paket soal	Daya beda (a)	Tingkat kesukaran butir (b)	Pseudo guessing (c)	Kemampuan (θ)
Paket 1	0,999	0,928	0,837	0,827
Paket 2	0,857	0,967	0,898	0,902
Paket 3	0,882	0,970	0,874	0,898
Paket 4	0,807	0,966	0,830	0,908
Paket 5	0,999	0,886	0,827	0,816

Berdasarkan Tabel 7, terlihat bahwa invariansi parameter daya beda memiliki di atas 0,8 yang dapat diartikan invariansi daya beda terpenuhi. Sama dengan hal tersebut invariansi parameter tingkat kesukaran butir memiliki nilai di atas 0,8 yang dapat diartikan invariansi tingkat kesukaran butir terpenuhi. Pada parameter *guessing* invariansi juga berada di atas 0,8 yang dapat diartikan invariansi parameter *guessing* terpenuhi. Selain itu, pada invariansi

kemampuan nilai yang dihasilkan juga berada di atas 0,8 sehingga dapat dikatakan bahwa invariansi parameter kemampuan terpenuhi.

Analisis karakteristik butir pada model 3PL meliputi 3 parameter yaitu, tingkat kesukaran butir (b), daya beda (a), dan *guessing* (c).

Tabel 5. Karakteristik butir model 3PL

Kriteria	Paket 1	Paket 2	Paket 3	Paket 4	Paket 5
Banyak butir	38	40	40	39	38
Banyak peserta tes	3501	3321	3357	3279	3307
Reliabilitas	0,924	0,909	0,914	0,910	0,925
Validitas	0,506	0,479	0,470	0,488	0,491
Rerata tingkat kesukaran butir	-0,453	-0,457	-0,235	-0,317	-0,482
Rerata daya beda	1,428	1,199	1,124	1,124	1,397
Rerata *guessing*	0,169	0,176	0,147	0,146	0,177

Berdasarkan Tabel 5 terlihat bahwa semua paket soal memiliki koefisien reliabilitas yang tinggi, berkisar antara 0,909 – 0,925 dan validitas yang berkisar antara 0,470 - 0,506. Selain itu, rata-rata parameter tingkat kesukaran butir diperoleh sebesar -0,453 - -0,235 yang berkategori baik ($-2 < b < 2$). Rata-rata daya beda diperoleh sebesar 1,199 – 1,428 yang berkategori baik ($0 < a < 2$). Sedangkan, rata-rata parameter *guessing* diperoleh sebesar 0,146 - 0,177 yang berkategori baik($c < 0,25$).

Penyetaraan dengan menggunakan metode linear dan metode *equipercentile*. Rancangan penyetaraan ini akan menggunakan desain grup ekuivalen dengan penyetaraan paket 1 ke paket 2, paket 1 ke paket 3, paket 1 ke paket 4, dan paket 1 ke paket 5. Oleh karena itu pada metode linear dan metode *equipercentile* akan didapatkan distribusi kumulatif.

Tabel 6. Distribusi kumulatif penyetaraan linear

Skor	Paket 1 ke 2	Paket 1 ke 3	Paket 1 ke 4	Paket 1 ke 5
0	-1,587	0,469	-3,020	0,405
1	-0,513	1,441	-1,906	1,370
2	0,561	2,414	-0,791	2,336
3	1,635	3,386	0,323	3,301
4	2,709	4,359	1,438	4,267
5	3,783	5,331	2,552	5,233
6	4,856	6,303	3,667	6,198
7	5,930	7,276	4,781	7,164
8	7,004	8,248	5,896	8,130
9	8,078	9,220	7,010	9,095
10	9,152	10,193	8,125	10,061

Tabel 7. Distribusi kumulatif penyetaraan metode Equipercentile

Skor	Paket 1 ke 2	Paket 1 ke 3	Paket 1 ke 4	Paket 1 ke 5
0	-0,046	-0,015	-0,093	0,047
1	0,866	0,969	0,735	1,121
2	1,774	1,971	1,562	2,173
3	2,680	2,987	2,372	3,208
4	3,590	4,012	3,181	4,227
5	4,505	5,042	4,004	5,234
6	5,416	6,074	4,841	6,230
7	6,338	7,105	5,693	7,219
8	7,271	8,133	6,561	8,200
9	8,217	9,158	7,439	9,175
10	9,176	10,177	8,331	10,146

Pada penyetaraan menggunakan teori respons butir akan dilakukan dengan 4 metode yaitu mean-mean, mean-sigma, haebara, dan stocking-lord. Paket soal akan disetarakan dengan paket 1 sehingga menghasilkan nilai alpha dan beta yang selanjutnya akan digunakan sebagai persamaan penyetaraan. Hasil analisis dari teori respons butir disajikan pada Tabel 8.

Tabel 8. Alpha dan Beta penyetaraan teori respons butir

Paket	Alpha/Beta	Mean-Mean	Mean-Sigma	Haebara	Stocking-Lord
Paket 2 ke 1	Alpha	0,692	1,245	0,861	0,869
	Beta	-0,148	0,095	0,078	0,073
Paket 3 ke 1	Alpha	0,692	1,056	0,911	0,924
	Beta	-0,324	-0,256	-0,045	-0,187
Paket 4 ke 1	Alpha	0,711	1,153	0,935	0,989
	Beta	-0,282	-0,159	-0,100	-0,256
Paket 5 ke 1	Alpha	0,986	0,986	0,983	0,986
	Beta	0,022	0,019	0,015	0,027

Berdasarkan penyetaraan dengan alpha dan beta pada tabel 8, maka didapatkan hasil konversi penyeetaraan. Perbandingan rerata kemampuan dari peserta tes sebelum dilakukan penyetaraan dan sesudah penyetaraan dapat dilihat pada Tabel 9.

Tabel 9. Rangkuman rerata tingkat kemampuan peserta tes

Paket soal	Awal	MM	MS	Hb	SL
Paket 2 ke 1	-0,118	-0,223	-0,051	-0,004	-0,016
Paket 3 ke 1	-0,069	-0,367	-0,325	-0,109	-0,186
Paket 4 ke 1	-0,015	-0,287	-0,174	-0,082	-0,172
Paket 5 ke 1	-0,090	-0,057	-0,069	-0,073	-0,063

Evaluasi keakuratan penyetaraan dilakukan dengan menghitung nilai *Standart error of equating* (SEE) dari masing-masing metode penyetaraan dan semakin kecil nilai SEE maka keakuratan penyetaraan semakin baik.

Tabel 10. SEE penyetaraan teori tes klasik

Paket soal	Linear	*Equipercentile*	Ket
paket 2 to 1	0,076	0,304	Linear
paket 3 to 1	0,064	0,363	Linear
paket 4 to 1	0,081	0,303	Linear
paket 5 to 1	0,064	0,482	Linear
Rata-rata	0,071	0,363	Linear

Berdasarkan Tabel 10 terlihat bahwa nilai SEE dari penyetaraan menggunakan teori tes klasik dengan metode linear untuk masing-masing paket soal adalah nilai SEE yang lebih kecil jika dibandingkan dengan metode equipercentile, sehingga dapat dikatakan penyetaraan dengan metode linear adalah yang paling baik/ akurat pada penyeteraan teori tes klasik. Selanjutnya, pada penyetaraan teori respons butir, rangkuman nilai SEE untuk empat metode penyetaraan seperti yang terlihat pada Tabel 11.

Tabel 11. SEE penyetaraan teori respons butir

Paket	MM	MS	HB	SL	Ket
Paket_2	0,013	0,014	0,010	0,010	HB
Paket_3	0,018	0,019	0,014	0,016	HB
Paket_4	0,018	0,022	0,018	0,019	MM
Paket_5	0,009	0,009	0,009	0,009	SL
Rata-rata	0,015	0,016	0,013	0,013	HB

Berdasarkan Tabel 11 terlihat bahwa nilai SEE dari penyetaraan menggunakan teori respons butir dengan metode haebara untuk paket 2 ke 1 dan paket 3 ke 1 adalah yang paling kecil, untuk paket 4 ke 1 yang paling kecil adalah metode mean-mean, serta untuk paket 5 ke 1 yang paling kecil adalah metode *stocking&lord*.

Pembahasan

Karateristik butir tes dengan menggunakan pendekatan teori tes klasik secara keseluruhan memiliki rata-rata tingkat kesukaran butir termasuk kategori baik. Hal ini didasarkan pada Retnawani (2016:114-115) yang mengatakan butir soal yang baik memiliki tingkat kesukaran butir dengan interval 0,3 – 0,7. Sementara itu, pendapat lain dijelaskan Djemari Mardapi (2008:116) bahwa tingkat kesukaran butir yang dapat diterima sebesar 0,3 - 0,8. Hal tersebut mendukung penelitian yang telah dilakukan oleh Nusrotus Sa'idah (2012) dimana butir soal yang diterima (baik) adalah butir dengan tingkat kesukaran butir 0,3 – 0,8. Sementara pada pendekatan teori respons butir rata-rata tingkat kesukaran butir berkategori baik ($-2 < b < 2$). Wahyuni dan Kusrini (2017: 15) mengungkapkan bahwa tingkat kesukaran butir $-2 \leq b < -0,5$ kategori mudah, $-0,5 \leq b < 0,5$ kategori sedang, dan $0,5 \leq b \leq 2$ kategori sulit. Mengacu akan kriteria tersebut b berada pada $-0,5 \leq b < 0,5$, semua paket soal memiliki rata-rata tingkat kesukaran butir berkategori sedang.

Parameter daya beda dengan pendekatan teori tes klasik memiliki rata-rata yang baik sehingga dapat digunakan sebagai alat ukur kemampuan peserta tes. Pendapat tersebut didukung oleh Sumadi Suryabrata (2005:131) dan Wibawa (2019) indeks daya beda butir yang diterima adalah $\geq 0,3$. Sejalan dengan hal tersebut, Nusrotus Sa'idah (2012) mengkategorikan daya butir dengan nilai kurang dari 3 sebagai butir yang kurang. Pendapat berbeda diungkapkan Ebel & Frisbie (1986) dan Frisbie (2005) dalam Retnawati (2016:116) selama indeks daya butir bernilai positif, maka tidak perlu diperhatikan. Sependapat dengan ungkapan tersebut, Nauli T.

Siregar (2010) pada penelitiannya butir dengan nilai dari daya beda yang positif tetap dipakai pada penelitiannya. Selanjutnya, pendekatan teori respons butir menunjukan bahwa rata-rata daya beda berkategori baik ($0 < a < 2$). Meskipun begitu, terdapat beberapa butir dengan daya beda yang lebih besar dari 2 setelah dianalisis menggunakan model 3 PL. Serupa dengan penelitian Nusrotus Sa'idah (2012) yang mengasilkan nilai daya beda lebih dari 2 dan selanjutnya butir-butir tersebut dikatakan tidak baik. Sementara, rata-rata parameter guessing menunjukkan bahwa rata-rata parameter guessing berkategori baik ($c < 0,25$). Meskipun begitu, terdapat beberapa butir soal dengan parameter guessing yang memiliki nilai lebih dari 0,25 yang berkategoti tidak baik. Butir-butir tersebut berarti terdapat kemungkinan yang besar peserta tes menjawab benar butir tes hanya dengan menebaknya saja.

Bandalos (2017) mengungkapkan jika tingkat kesukaran butir diperoleh dari peserta dengan kemampuan yang lebih tinggi, butir dari paket yang lain akan terlihat lebih mudah. Karakteristik butir akan berubah jika paket soal dikerjakan oleh peserta tes dengan kemampuan yang berbeda. Faktor tersebut juga bisa disebabkan oleh peserta tes yang tidak siap, maka akan kesulitan saat mengerjakan paket soal sehingga akan berdampak pada tingkat kesukaran butir (Agus Santoso, 2018)

Pada Tabel 6 dan 7 terlihat bahwa konversi skor, baik dari metode linear dan *equipercentile*, dari paket 1 ke paket 2 dan paket 1 ke paket 4 akan membuat skor paket 1 menjadi lebih tinggi jika skornya semakin besar. Sementara, pada paket 1 ke paket 3 dan paket 1 ke paket 5 akan membuat skor paket 1 menjadi lebih rendah jika skornya semakin besar.

Pada Tabel 9 terlihat bahwa rerata kemampuan peserta tes pada paket 2 menurun dari kemampuan awal pada penyetaraan *Mean-mean*, Sedangkan pada metode *Mean-Sigma*, Haebara, dan Stocking-Lord rerata kemampuan peserta naik. Rerata kemampuan peserta tes pada paket 3 menurun dari kemampuan awal pada penyetaraan Mean-mean, Mean-Sigma, Haebara, dan Stocking-Lord. rerata kemampuan peserta tes pada paket 4 menurun dari kemampuan awal pada penyetaraan Mean-mean, Mean-Sigma, Haebara, dan Stocking-Lord. rerata kemampuan peserta tes pada paket 5 naik dari kemampuan awal pada penyetaraan *Mean-mean, Mean-Sigma*, Haebara, dan Stocking-Lord.

Sesuai dengan apa yang diungkapkan oleh Lahner (2020) bahwa penyetaraan dengan menggunakan pendekatan teori respons butir lebih akurat daripada menggunakan pendekatan teori tes klasik. Setelah dilakukan keakuratan penyetaraan dengan pendekatan teori tes klasik dan teori respons butir, dapat diartikan bahwa pada penyetaraan menggunakan metode linear dan *equipercentile* tidak akurat. Hal ini disebabkan karena hasil dari nilai rata-rata SEE yang paling kecil metode linear $0,071 > 0,05$. Sedangkan pada penyetaraan dengan pendekatan teori respons butir pada semua metode penyetaraan memiliki rata-rata MSE yang lebih kecil dari 0,05. Akan tetapi, pada penentuan metode penyetaraan terbaik pada penelitian ini, akan dipilih dengan metode dengan rata-rata SEE terkecil, yaitu metode Haebara $0,0129 < 0,05$.

SIMPULAN

Pada pendekatan teori tes klasik dan teori respons butir, karakteristik butir UNBK IPA SMP se-DIY memiliki rata - rata tingkat kesukaran butir, dan daya beda yang baik. Selain itu, pada teori respons butir rata-rata parameter *guessing* berkategori baik. Penyetaraan tes menghasilkan skor konversi yang bervariasi terdapat skor konversi yang naik ataupun turun jika dibandingkan dengan skor awal. Sementara hasil penyetaraan tes metode haebara merupakan metode yang paling akurat dibandingkan metode lainnya. Berdasarkan penelitian ini, sebaiknya dilakukan analisis menggunakan pendekatan teori respons butir dengan metode

selain 3PL dan dengan metode yang lebih atau pengujian penyetaraan vertikal atau multidimensi.

DAFTAR PUSTAKA

Agus Santoso. (2018). Karakteristik butir tes pengantar statistika sosial berdasarkan teori respon butir. *Jurnal Pendidikan Matematika dan Sains*, 6(2), 158-168.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. New York: The Guilford Press.

Hambleton R.K, Swaminathan H., & Rogers H. J. (1991). *Fundamentals of item respons theory*. Newbury Park: SAGE Publications.

Hidayati, Kana. (2005). *Penerapan teori respons butir dalam penyetaraan tes.* Makalah. Diakses dari https://eprints.uny.ac.id/11757/1/PM-15%20Kana%20UNY.pdf.

Kolen M. J., & Brennan R. L. (2014). *Test equating, scaling, and linking methods and practices*. New York: Springer.

Lahner; FM., Schauber, S., Lorwald, A.C. et al. (2020). Measurement precision at the cut score in medical multiplechoice exams: Theory matters. *Perspect Med Educ*, 9:220–228. https://doi.org/10.1007/s40037-020-00586-0

Mardapi, Djemari. (2008). *Teknik penyusunan instrumen tes dan non tes*. Yogyakarta: Mitra Cendikia.

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

Retnawati, Heri. (2016). *Analisis kuantitatif instrumen penelitian*. Yogyakarta: Parama Publishing. Tesis. Yogyakarta. Universitas Negeri Yogyakarta.

Sa'idah, Nusrotus. (2012). *Penyetaraan horisontal tes ujian sekolah kimia SMA di provinsi Yogyakarta*. Tesis. Yogyakarta. Universitas Negeri Yogyakarta.

Siregar, Nauli T. (2010). *Penyetaraan horisontal tes uas bahasa inggris tingkat SMP di provinsi Daerah Istimewa Yogyakarta*. Skripsi. Yogyakarta. Universitas Negeri Yogyakarta.

Sudijono, Anas. (2009). *Pengantar evaluasi pendidikan*. Jakarta: Raja Grafindo Persada.

Suryabrata, Sumadi. (2005). *Pengembangan alat ukur psikologis*. Yogyakarta: Andi Offset.

Wahyuni dan Kusrini. (2017). Penerapan computerized adaptive test pada tes online menggunakan algoritma teori respon butir model 3 PL. *METIK Jurnal*, 1(2), 13-17.

Wibawa, E. A. (2019). Karakteristik butir soal tes ujian akhir semester hukum bisnis. *Jurnal Pendidikan Akuntansi Indonesia*, 17(2), 87-96.