Vol. 9, No. 2, pp. 100 – 115 e-ISSN: 2715–856X p-ISSN:2338-1183

Jurnal Pendidikan Matematika

http://jurnal.fkip.unila.ac.id/index.php/MTK

Development of Higher Order Thinking Skills (HOTS) Based Mathematics Test using the Rasch Model for First Middle School Students

Zubaidah R[1], Dona Fitriawan[2], Halini, Dwi Astuti[3]

[1,2,3]Program Studi Pendidikan Matematika FKIP Universitas Tanjungpura

[1]Email: zubaidah.r@fkip.untan.ac.id

Received: 6 April, 2021 *Accepted: 11 June, 2021* *Published:30 June, 2021*

Abstract

Specifically, the research objectives are (1) Describe the steps for preparing a HOTS-based mathematics learning outcome test instrument using the Rasch Model, (2) Describe the results of the analysis of test items that meet the characteristics of the items that are feasible at the limited trial stage, (3) Describe the results of the analysis test items that meet the characteristics of the proper items based on the Rasch Model, (4) Describe the characteristics of a good test instrument according to the Rasch Model. This type of research is a 4-D model development research with the steps (1) Define, namely front end analysis, student analysis, concept analysis, task analysis and goal determination, (2) Design (design), namely compiling a grid, determine the form and number of items, formulate questions, determine the length of the test, (3) Develop (development), namely review tests, test trials, analyze test results, improve tests, assemble tests, (4) Disseminate (spread), namely carry out. The research sample was students of class VIII SMPN 3, SMPN 10, and SMPN 20 Pontianak, totaling 188 students. The conclusions are (1) The test development step consists of four stages, namely: the defining stage namely; front end analysis, learner analysis, concept analysis, learning objectives analysis (2) The HOTS-based junior high school mathematics learning outcomes test has the suitability validity of the panelists as many as 30 items and has a very high level of panelist reliability suitability. (3) A total of 25 items have met all the characteristics of the items. (4) Characteristics of the items on the test instrument are feasible according to the Rasch model in this study.

Keywords: *higher order thinking skills (hots); rasch model; test development*

Abstrak

Secara spesifik tujuan penelitian ini adalah (1) Mendeskripsikan langkah-langkah penyusunan instrumen tes hasil belajar matematika berbasis HOTS dengan menggunakan Model Rasch, (2) Mendeskripsikan hasil analisis butir soal yang memenuhi karakteristik butir soal layak. pada tahap uji coba terbatas, (3) Mendeskripsikan hasil analisis butir soal yang memenuhi karakteristik butir soal layak berdasarkan Model Rasch, (4) Mendeskripsikan ciri-ciri instrumen tes yang baik menurut Model Rasch. Jenis penelitian ini adalah penelitian pengembangan model 4-D dengan tahapan (1) Define yaitu analisis pendahuluan, analisis siswa, analisis konsep, analisis tugas dan penentuan tujuan, (2) design yaitu menyusun kegiatan rencana penelitian, menentukan bentuk dan jumlah soal, merumuskan soal, menentukan lamanya tes, (3) Develop yaitu review tes, tes uji coba, menganalisis hasil tes,

memperbaiki tes, merakit tes, (4) Deseminate, yaitu melaksanakan penelitian. Sampel penelitian adalah siswa kelas VIII SMPN 3, SMPN 10 dan SMPN 20 Pontianak yang berjumlah 188 siswa. Kesimpulannya adalah (1) Tahap pengembangan tes terdiri dari empat tahap, yaitu: tahap pendefinisian yaitu; Analisis front end, analisis peserta didik, analisis konsep, analisis tujuan pembelajaran (2) Tes hasil belajar matematika SMP berbasis HOTS memiliki validitas kesesuaian panelis sebanyak 30 item dan memiliki tingkat kesesuaian reliabilitas panelis yang sangat tinggi. (3) Sebanyak 25 item telah memenuhi semua karakteristik item. (4) Karakteristik butir pada instrumen tes layak menurut model Rasch dalam penelitian ini.

Kata Kunci: higher order thinking skills (hots); model rasch; pengembangan tes

---

PENDAHULUAN

Evaluasi dan penilaian sangat diperlukan dalam mengukur kualitas proses dan hasil belajar yang dilakukan peserta didik. Penilaian merupakan proses pengumpulan informasi tentang kinerja peserta didik untuk digunakan sebagai dasar dalam membuat keputusan (Wyat-Smith dan Cumming, 2002). Black dan Wiliam (1998) mengatakan bahwa penilaian adalah aktivitas yang dilakukan oleh guru dan peserta didik untuk menilai diri mereka sendiri, yang memberikan informasi yang dapat digunakan sebagai umpan balik untuk memodifikasi aktivitas belajar dan mengajar. Standar Nasional Pendidikan Tinggi menyatakan penilaian dalam hasil belajar oleh pendidik bertujuan untuk memantau, mengevaluasi, kemajuan, dan perbaikan proses dan hasil belajar peserta didik secara berkala dan berkesinambungan (Kemendikbud, 2020).

Penilaian menyediakan data untuk memberi nilai dan menentukan tingkat pencapaian sasaran dan tujuan yang ingin dicapai, baik untuk yang ditujukan kepada siswa, orang tua, pengurus, institusi bidang pendidikan yang lebih tinggi maupun *potential employers* (Johnson & Johnson, 2002). Peranan guru dan tes eksternal terstandarisasi yang telah memenuhi kriteria validitas dan reliabilitas menjadi sangat penting dalam pelaksanaan penilaian. Tes eksternal berfungsi untuk mengukur sejauh mana proses pembelajaran yang dilakukan selama periode tertentu tercapai.

Tes hasil belajar merupakan salah satu instrument dalam pelaksanaan penilaian kognitif yang didesain untuk mengukur hasil proses belajar peserta didik. Agar hasil tes dapat mengambarkan hasil pengukuran sesuai dengan tujuan apa yang semestinya diukur, maka tes yang digunakan harus tes yang berkualitas yaitu tes yang telah teruji secara empiris standar keterukurannya seperti kevalidan dan reliabelitas, tingkat kesukaran, daya pembeda soal dan tingkat keterbacaan soal.

Mata pelajaran matematika tidak hanya membekali peserta didik dengan berhitung ataupun menggunakan rumus dalam mengerjakan soal tes tetapi juga menggunakan kemampuan analitisnya dalam memecahkan berbagai masalah keseharian

(Permendikbud, 2013). Berdasarkan hasil tes dan survey yang dilakukan oleh PISA pada tahun 2015, hasil untuk matematika siswa Indonesia masih tergolong rendah yaitu pada peringkat 63 dari 69 negara yang di evaluasi. Siswa-siswa Indonesia masih rendah dalam penguasaan materi dan kesulitan dalam menjawab soal yang membutuhkan penalaran. Hal ini disebabkan karena siswa cenderung belajar menghafal rumus tanpa memahami konsepnya. Oleh karena itu, munculkan soal-soal bertipe *Higher Order Thinking Skill (HOTS)* yang menuntut kemampuan berpikir tingkat tinggi dan proses dalam bernalar, sehingga diharapkan mampu mengasah kemampuan berpikir kritis, logis dan kreatif (Tofade *et al.*, 2013; Collins, 2014). Menurut Conklin (2012) ada dua karakteristik keterampilan berpikir tingkat tinggi (HOTS) yaitu berpikir kritis dan berpikir kreatif. Pendapat lain menyatakan, terdapat tiga cakupan level berpikir tingkat tinggi yaitu (a) sebagai transfer, di mana siswa dapat menghubungkan pengetahuan awal dengan pengetahuan baru. (b) sebagai berpikir kritis (penalaran, mengamati, mempertanyakan dan menyelidiki) di mana sikap siswa dalam mengambil keputusan. (c) sebagai pemecahan masalah (Brookhart, 2014). Sementara Resnick (1997), menyatakan, HOTS adalah bersifat non-algorithmic, kompleks, variasi solusi, variasi interpretasi, mengandung banyak kriteria, dan membutuhkan banyak usaha.

Menurut Anderson, Krathwohl dan Bloom (2002) dalam revisi Taksonomi Bloom, menyatakan proses berpikir kognitif terbagi menjadi kemampuan berpikir tingkat rendah (*Lower Order Thinking*) dan kemampuan berpikir tingkat tinggi (*Higher Order Thinking*). Yang termasuk di dalam LOT meliputi kemampuan mengingat (*remember*), memahami (*under-stand*), dan menerapkan (*apply*), sedangkan HOT meliputi kemampuan menganalisis (*analyze*), mengevaluasi (*evaluate*), dan menciptakan (*create*). Soal-soal HOTS yang baik merupakan instrument-instrumen pengukuran yang mampu mengasah kemampuan berpikir yang diharapkan tidak sekedar mampu mengingat (*recall*), mampu menyatakan kembali (*restate*) atau mampu merujuk tanpa melakukan pengolahan (*recite*). Pada konteks assesmen mengukur kemampuan: 1) transfer satu konsep ke konsep lainnya, 2) memproses dan menerapkan informasi, 3) mencari kaitan dari berbagai informasi yang berbeda-beda, 4) menggunakan informasi untuk menyelesaikan masalah, dan 5) menelaah ide dan informasi secara kritis (Sumar dan Sumar, 2020).

Rendahnya kemampuan peserta didik tidak hanya semata-mata disebabkan oleh materi yang sulit, metode atau pendekatan yang kurang bervariasi, tetapi juga sangat ditentukan oleh alat evaluasi yang digunakan untuk mengukur kemampuan peserta didik sesuai dengan keterampilan yang akan dicapai. Oleh karena itu, untuk menguji kualitas suatu instrumen seperti tes diperlukan analisis butir tes, agar butir-butir item yang

dikembangkan dapat menjalankan fungsinya sebagai alat ukur yang baik (Sudijono, 2012; Saputro *et al*., 2015).

Dalam teori pengukuran, terdapat dua pendekatan yang sering digunakan untuk menganalisis butir item yaitu Teori Tes Klasik (*Classical Test Theory*) dan teori respon butir (*Item Response Theory*) (Djemari, 2012). Selanjutnya Mardapi menyatakan bahwa ada beberapa kelemahan dalam teori klasik antara lain; (a) tingkat kesukaran dan daya pembeda butir tergantung pada kelompok peserta yang di tes, (b) teknik analisis tes dengan cara membandingkan peserta tes kelompok atas, tengah dan bawah, (c) dasar teori yang menentukan bahwa peserta tes memperoleh hasil tes sesuai dengan kemampuan yang bersangkutan belum ada, (d) *Standar Error Measurement* (SEM) berlaku pada seluruh peserta tes. Berdasarkan kelemahan teori tes klasik, pendekatan *Item Response Theory* (IRT) muncul untuk mengatasi kelemahan tersebut. Salah satu model *IRT* yang terkenal adalah Model Rasch (Qasem, 2013).

Pemodelan Rasch merupakan pengembangan pengukuran yang obyektif. Hasil pengukuran tergantung pada siapa yang diukur (*Test-Dependent-Scoring*). Jumlah jawaban benar pada sebuah tes tergantung pada subyek yang diukur dan bersifat deskriptif serta berlaku untuk semua subyek yang diukur (Sumintoro dan Widhiarso, 2015). Selanjutnya dinyatakan bahwa tedapat lima kriteria yang membuat pemodelan pengukuran Rasch merupakan suatu pengukuran menjadi obyektif yaitu; (1) memberikan ukuran yang linier, (2) data yang hilang bisa teratasi, (3) proses estimasi dilakukan secara tepat, (4) sesuatu yang tidak tepat dan tidak umum dapat ditemukan, (5) merupakan instrumen pengukuran bersifat independen.

Berapa penelitian terkait pengembangan soal-soal HOTS sudah banyak dilakukan khusus untuk meteri tertentu. Seperti yang dilakukan oleh (Novinda, Silitonga dan Hamdani, 2019); (Fitriawan, 2020), khusus untuk materi Aljabar Siswa kelas VII SMP dengan kesimpulan penelitian telah memperoleh produk soal yang valid, praktis, dan efektif. Penelitian ini hanya mengukur kevalidan, kepraktisan dan keefektifan instrumen, tetapi tidak menganalisis butir. Selanjutnya penelitian yang dilakukan oleh (Wadiyani, 2019), khusus pada materi geometri menyimpulkan bahwa soal yang dikembangkan sebanyak 8 butir soal yang memiliki reliabilitas tinggi, daya beda butir tergolong baik, dan taraf kesukaran tergolong sedang dan sukar. Dalam penelitian ini proses analisis butir dilakukan secara manual menggunakan rumus atau formula menurut para ahli. Dari beberapa penelitian pengembangan soal HOTS konten matematika belum banyak ditemukan analisis butir menggunakan Pemodelan Rasch.

Berdasarkan alasan yang telah diuraikan di atas, penelitian ini dianggap penting dalam pembelajaran matematika khususnya dalam pengembangan instrument untuk mengukur kemampuan tingkat tinggi (HOTS) bagi peserta didik sesuai dengan tuntutan

kurikulum yang berlaku saat ini dengan menggunakan analisis Model Rasch. Melalui instrumen yang dihasilkan diharapan hasil pengukuran terhadap peserta didik yang dikenakan tes tidak dirugikan dan diperlakukan adil karena sesuai dengan kemampuan mereka dalam belajar. Instrumen yang telah disusun ini juga dapat dijadikan sebagai acuan bagi guru dalam mengkonstruksi soal-soal baik untuk keperluan penilaian formatif maupun sumatif sebagai upaya untuk melatih dan meningkatkan kemampuan berpikir kritis dan kreatif. Dengan melatih peserta didik berpikir tingkat tinggi berarti menyiapkan pribadi berkualitas agar mereka lebih tangguh dalam menghadapi kehidupan yang lebih komplek dan penuh tantangan di masa depan (Syahwaludi dan Suratman, 2016).

METODE

Penelitian ini mengikuti alur Sivasailam Thiagarajan, dkk dalam (Syahwaludi dan Suratman, 2016), maka jenis penelitiannya adalah penelitian pengembangan. Langkah-langkah penelitiannya sebagai berikut: a) *Define* (pendefinisian) yaitu analisis ujung depan, analisis peserta didik, analisis konsep, analisis tugas dan penentuan tujuan, b) *Design* (perancangan) yaitu menyusun kisi-kisi, menentukan bentuk dan jumlah item, merumuskan soal, menentukan panjang tes, c) *Develop* (pengembangan) yaitu telaah tes, uji coba tes, analisis hasil uji coba, perbaiki tes, merakit tes, d) *Deseminate* (penyebaran) yaitu melaksanakan tes, menafsirkan hasil tes dengan menggunakan analisis Model Rasch.

Pada tahap *Deseminate* (penyebaran), pelaksanaan tes melibatkan peserta didik berjumlah 188 yang berasal dari SMP Negeri 3 Pontianak, SMP Negeri 10 Pontianak, dan SMP Negeri 20 Pontianak. Teknik pelaksanaan tes disesuaikan dengan aturan yang diberlaku di sekolah selama masa pandemic dan dibantu oleh guru matematika dan mahaisiswa PPL di sekolah di mana tes dilaksanakan.

Data yang diperoleh berupa data dikotomi dan dianalisis menggunakan Model Rasch yaitu model probabilistic yang didefinisikan sebagai individu yang memiliki tingkat kemampuan yang lebih besar dibandingkan individu lainnya seharusnya memiliki peluang yang lebih besar untuk menjawab satu butir soal dengan benar. Dengan prinsif yang sama, butir yang lebih sulit menyebabkan peluang individu untuk mampu menjawabnya makin kecil (Bond dan Fox., 2015). Untuk data yang berbentuk dikotomi, pemodelan Rasch menggabungkan suatu algoritma yang menyatakan hasil ekspektasi probabilistik dari aitem ‘i’ dan responden ‘n’, yang secara matematis:

$$P_{ni}\,(x_{ni} = 1|b_n, d_i) = \frac{e^{(bn-di)}}{1+e^{(bn-di)}}$$

Di mana P_{ni} ($x_{ni} = 1|b_n, d_i$) adalah probabilitas dari responden ndalam butir iuntuk menghasilkan jawaban benar (x = 1) dengan kemampuan responden, ßn, dan tingkat kesulitan item δi (Cavanagh dan Waugh, 2011).

Rumus yang digunakan untuk mengetahui validitas kesesuaian panelis adalah rumus Aiken dalam (Dali, 2012).

$$V = \frac{\sum ni\,|\,i - r\,|}{N(t-1)}$$

Hasil penilaian validitas instrumen oleh panelis diperoleh 30 butir diputuskan valid. Untuk mengetahui reliabilitas kesesuaian panelis digunakan rumus Hoyt dalam (Dali, 2012) yaitu:

$$R_{kk} = \frac{RJK_p - RJK_e}{RJK_p}$$

Berdasarkan hasil perhitungan dengan menggunakan rumus Hoyt, diperoleh Koefisien reliabilitas kesesuaian panelis instrument hasil belajar sebesar r = 0,968. Hal ini menunjukkan bahwa konsistensi hasil penilaian antar panelis tergolong sangat tinggi.

HASIL DAN PEMBAHASAN

Hasil Penelitian

Butir soal yang layak digunakan dapat dilihat dari beberapa analisa data.:

Tingkat Keterbacaan

Tes yang diberikan kepada peserta didik di tingkat SMP/sederajat minimal memiliki nilai *readability index* sama dengan atau lebih besar dari 6 (RI ≤ 6). Tingkat keterbacaan pada penelitian ini dapat dilihat pada Tabel 1.

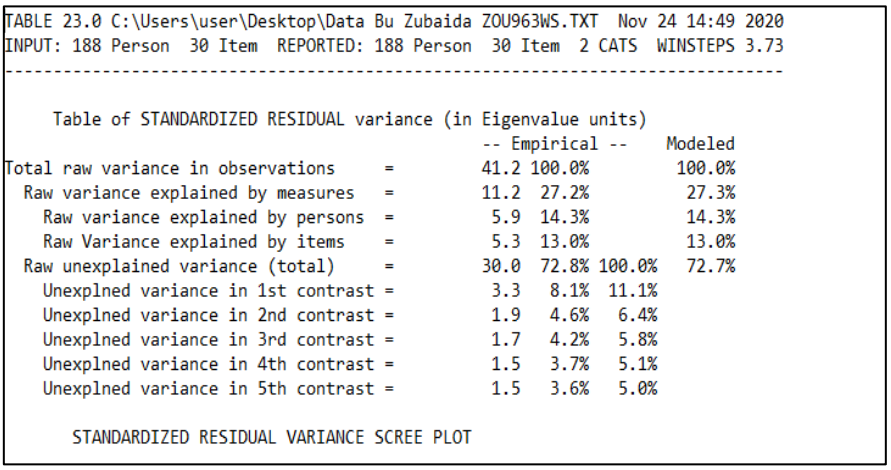
Tabel 1. Tingkat Keterbacaan Butir Soal

No	$\bar{St}$	$\bar{Wt}$	*RI*	No	$\bar{St}$	$\bar{Wt}$	*RI*
1	8	5.375	3.415	16	13	5.077	3.9
2	13	5.846	5.1	17	10.67	5.406	3.9704
3	14.5	4.276	2.9353	18	8.25	5.818	4.1539
4	21	3.905	3.5914	19	8	5.063	2.9275
5	13.25	4.491	3.0328	20	11.5	6.478	5.8011
6	11	4.909	3.2582	21	9	6.556	5.4467
7	10.67	5.563	4.2142	22	11	5.545	4.2509
8	11.33	5.324	3.968	23	7.5	5.667	3.775
9	10	5.867	4.562	24	10	6.3	5.238

10	10.5	5.905	4.7164	25	11.67	6.571	5.9781
11	9.333	6.571	5.5348	26	12.5	5.8	4.933
12	11.67	5.714	4.641	27	10,56	5.87	4.345
13	8	5.333	3.35	28	9.5	5.684	4.1824
14	9	5.148	3.2511	29	8,5	6	4,48
15	11.5	4.696	3.0202	30	7,5	5,33	3,205

Unidimensionalitas

Unidimensionalitas menyatakan apakah instrumen tes yang dikembangkan mampu mengukur apa yang seharusnya diukur. Nilai minimum yang harus dicapai untuk memenuhi standar unidimensionalitas adalah 20%. *Raw variance explained by measures* 27,2%. Syarat unidimensionalitas sudah terpenuhi karena nilainya lebih dari 20%, Nilai unidimensionalitas pada penelitian ini sebagai berikut:



```
TABLE 23.0 C:\Users\user\Desktop\Data Bu Zubaida ZOU963WS.TXT  Nov 24 14:49 2020
INPUT: 188 Person  30 Item  REPORTED: 188 Person  30 Item  2 CATS  WINSTEPS 3.73
-------------------------------------------------------------------------------

    Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                        -- Empirical --    Modeled
Total raw variance in observations     =         41.2 100.0%         100.0%
  Raw variance explained by measures   =         11.2  27.2%          27.3%
    Raw variance explained by persons  =          5.9  14.3%          14.3%
    Raw Variance explained by items    =          5.3  13.0%          13.0%
  Raw unexplained variance (total)     =         30.0  72.8% 100.0%   72.7%
    Unexplned variance in 1st contrast =          3.3   8.1%  11.1%
    Unexplned variance in 2nd contrast =          1.9   4.6%   6.4%
    Unexplned variance in 3rd contrast =          1.7   4.2%   5.8%
    Unexplned variance in 4th contrast =          1.5   3.7%   5.1%
    Unexplned variance in 5th contrast =          1.5   3.6%   5.0%

      STANDARDIZED RESIDUAL VARIANCE SCREE PLOT
```

Gambar 1. Nilai Unidimensional

Reliabilitas

Dari hasil analisis program Winstep diperoleh nilai *real item realibility* sebesar 0,94 yang masuk ke dalam kategori istimewa.

SUMMARY OF 188 MEASURED (EXTREME AND NON-EXTREME) Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	21.6	29.9	1.46	.60				
S.D.	6.1	.5	1.52	.36				
MAX.	30.0	30.0	4.93	1.84				
MIN.	6.0	26.0	-1.56	.39	.70	-2.0	.19	-1.8

REAL RMSE	.72 TRUE SD	1.34	SEPARATION	1.88	Person RELIABILITY .78
MODEL RMSE	.71 TRUE SD	1.35	SEPARATION	1.91	Person RELIABILITY .78
S.E. OF Person MEAN = .11					

Person RAW SCORE-TO-MEASURE CORRELATION = .94
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .88

SUMMARY OF 30 MEASURED (NON-EXTREME) Item

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	135.3	187.6	.00	.20	1.00	.0	.98	-.1
S.D.	21.9	1.0	.81	.03	.13	1.6	.26	1.6
MAX.	172.0	188.0	2.12	.28	1.28	3.2	1.67	4.0
MIN.	71.0	185.0	-1.65	.17	.76	-3.1	.59	-2.6

REAL RMSE	.20 TRUE SD	.79	SEPARATION	3.85	Item RELIABILITY .94
MODEL RMSE	.20 TRUE SD	.79	SEPARATION	3.95	Item RELIABILITY .94
S.E. OF Item MEAN = .15					

Gambar 2. Output Hasil Reliabilitas

Tingkat Kesukaran Butir Soal

Tingkat kesulitan butir pada model rasch merupakan nilai peluang yang kemudian diskalakan dengan memasukkan nilai logaritma. Hasil estimasi logit dari *odd-ratio* (teori tes klasik) disebut logit atau W-score atau nilai measure terbagi menjadi empat kategori, yakni:

Nilai measure < -1 = item sangat mudah
Nilai measure -1 s.d. 0 = item mudah
Nilai measure 0 s.d. 1 = item sulit
Nilai measure > 1 = item sangat sulit

Berdasarkan hasil analisis Model Rasch (table 1), memberi keterangan bahwa 3 soal dikategirikan sangat sukar yaitu soal nomor 8, 12, 24. Sebanyak delapan soal yang dikategorikan sulit yaitu soal nomor 1, 2, 9, 10, 13, 14, 19, 26, 28. Soal yang dikategorikan mudah sebanyak 10 yaitu soal nomor 4, 7, 11, 15, 17, 18, 20, 21, 23, 27. Sementara soal yang dikategorikan sangat mudah adalah nomor 3, 6, 16 dan 22.

Tingkat Kesesuian Butir

Tingkat kesesuaian item ini digunakan untuk melihat ketepatan item dengan model atau *item fit*. Item fit menjelaskan apakah item soal yang dikembangkan berfungsi normal melakukan pengukuran atau tidak. Jika ada item yang tidak fit, hal ini mengindikasikan adanya miskonsepsi subjek dalam menjawab soal tersebut. Untuk memeriksa mana butir yang *fit* dan *misfit* bisa digunakan nilai INFIT MNSQ dari setiap

butir. Nilai rata-rata dan deviasi standar dijumlahkan, kemudian dibandingkan, nilai logit yang lebih besar dari nilai tersebut mengindikasikan butir yang *misfit*. Jumlah logit butir dari MEAN dan S.D : 1,00 + 0,13 = 1,13 maka dari kriteria ini terdapat lima butir dengan nilai INFIT MNSQ yang lebih besar dar 1,13 yaitu: 5, 24, 25, 29, dan 30.

Daya Diskriminasi Butir

Daya Diskriminasi Rasch atau nilai korelasi skor butir ditunjukkan oleh *Pt Measure Corr*. Nilai *Pt Measure Corr* 1,0 mengindikasikan bahwa semua peserta tes dengan abilitas rendah menjawab butir dengan salah dan semua peserta tes dengan abilitas tinggi menjawab butir dengan benar. Sementara nilai *Pt Measure Corr* negative mengindikasikan butir soal yang menyesatkan karena peserta tes dengan kemampuan rendah mampu menjawab butir dengan benar dan peserta tes dengan kemampuan tinggi justru menjawab salah. Klasifikasi nilai tersebut yakni sangat bagus (>0,40), bagus (0,30–0,39), cukup (0,20-0,29), tidak mampu mendiskriminasi (0,00-0,19), dan membutuhkan pemeriksaan terhadap butir (< 0,00). Dari hasil analisis model rasch (table 4.3) diperoleh keterangan bahwa semua butir memiliki *Pt Measure Corr* > 0,40. Dengan demikian ke 30 butir memilki daya pembeda yang sangat bagus.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
8	71	188	2.12	.18	1.04	.5	1.03	.2	.55	.57	72.7	75.0	08
24	94	188	1.39	.18	1.21	2.6	1.38	3.1	.42	.54	67.6	72.2	24
12	102	188	1.14	.17	1.05	.7	1.03	.3	.51	.53	70.5	71.9	12
9	113	188	.81	.18	.78	-3.1	.71	-2.6	.63	.51	82.4	72.2	09
19	113	185	.76	.18	.87	-1.8	.81	-1.5	.58	.51	76.3	72.6	19
1	121	188	.56	.18	.93	-.9	1.01	.1	.52	.49	77.8	73.1	01
10	122	188	.53	.18	.81	-2.5	.74	-2.0	.59	.49	81.8	73.2	10
29	124	188	.46	.18	1.23	2.7	1.67	4.0	.34	.49	67.0	73.4	29
30	128	188	.33	.18	1.28	3.2	1.48	2.8	.32	.48	66.5	74.1	30
2	132	188	.20	.18	.93	-.8	.77	-1.5	.52	.46	76.7	74.9	02
5	132	188	.20	.18	1.14	1.6	1.23	1.4	.38	.46	73.3	74.9	05
14	133	188	.17	.18	.83	-2.1	.72	-1.8	.56	.46	78.4	75.1	14
25	133	188	.17	.18	1.20	2.3	1.35	2.0	.35	.46	69.3	75.1	25
13	134	188	.13	.19	1.00	.0	.94	-.3	.46	.46	75.0	75.3	13
28	136	188	.06	.19	1.08	.9	1.19	1.1	.40	.45	76.1	75.8	28
26	137	188	.03	.19	1.05	.6	.99	.0	.43	.45	74.4	76.1	26
18	136	185	-.01	.19	.76	-2.9	.62	-2.3	.58	.45	82.7	76.5	18
20	140	188	-.08	.19	1.00	.0	.92	-.3	.44	.44	75.6	76.8	20
21	141	188	-.12	.19	1.04	.5	.85	-.7	.43	.44	73.9	77.1	21
7	142	188	-.15	.19	.93	-.8	.79	-1.1	.48	.43	79.0	77.3	07
17	145	185	-.35	.20	1.01	.2	.89	-.4	.42	.42	77.5	79.0	17
11	148	188	-.38	.20	.78	-2.3	.59	-2.1	.54	.41	83.5	79.2	11
4	150	188	-.46	.20	1.00	.0	.98	.0	.40	.40	80.1	79.9	04
15	152	188	-.55	.21	.98	-.1	.78	-.9	.42	.39	81.8	80.8	15
23	152	188	-.55	.21	1.07	.6	1.04	.3	.35	.39	80.7	80.8	23
27	159	188	-.87	.22	1.11	.9	1.14	.6	.29	.36	83.5	84.0	27
22	164	188	-1.13	.24	1.01	.1	1.16	.6	.31	.33	88.1	86.5	22
6	168	188	-1.37	.25	1.06	.4	1.07	.3	.26	.30	88.1	88.7	06
16	166	185	-1.41	.26	.85	-.8	.89	-.1	.38	.30	89.6	89.1	16
3	172	188	-1.65	.28	.92	-.4	.67	-.7	.34	.27	90.9	90.9	03
MEAN	135.3	187.6	.00	.20	1.00	.0	.98	-.1			78.0	77.7	
S.D.	21.9	1.0	.81	.03	.13	1.6	.26	1.6			6.4	5.2	

Gambar 3. Output Data Hasil Tingkat Kesukaran, Tingkat Kesesuaian, dan Daya Diskriminasi Butir

Separasi

Peneglompokan dari responden dan butir dapat diketahui dari nilai separasi. Makin besar nilai separasi, maka kualitas instrument dalam hal keseluruhan responden dan butir akan semakin bagus, karena bisa mengindikasikan kelompok responden dan kelompok butir. Persamaan lain yang dapat digunakan untuk melihat secara lebih teliti disebut dengan pemisahan strata:

$$H = \frac{[(4xSEPARATION) + 1]}{3}$$

(Sumintono & Widhiarso, 2013)

Dengan nilai butir separation 3,85 maka H = [(4x3,85)+1]/3 = 5,47 dibulatkan menjadi 5, yang bermakna terdapat lima kelompok butir soal.

Butir soal yang dinyatakan layak adalah butir soal yang memenuhi seluruh kriteria yaitu tingkat keterbacaan, validitaas isi, unidimensionalitas, reliabilitas, peta person-butir, dan tingkat kesulitan butir. Tabel 2 menunjukkan hasil rekapitulasi butir yang layak menurut model rasch.

Tabel 2. Rekap Kelayakan Butir

No	Kriteria						Ket.
	1	2	3	4	5	6	
1.	3.41	0.94	27 %	0,94	✓	✓	PAKAI
2.	5.1	0.93	27%	0,94	✓	✓	PAKAI
3.	2.93	0.96	27%	0,94	✓	✓	PAKAI
4.	3.59	0.94	27%	0,94	✓	✓	PAKAI
5.	3.03	0.94	27%	0,94			REVISI
6	3.26	0.90	27 %	0,94	✓	✓	PAKAI
7	4.21	0.86	27%	0,94	✓	✓	PAKAI
8	3.97	0.90	27 %	0,94	✓	✓	PAKAI
9	4.56	0.83	27 %	0,94	✓	✓	PAKAI
10	4.72	0.86	27 %	0,94	✓	✓	PAKAI
11	5.53	0.86	27 %	0,94	✓	✓	PAKAI
12	4.64	0.66	27 %	0,94	✓	✓	PAKAI
13	3.35	0.88	27 %	0,94	✓	✓	PAKAI
14	3.25	0.93	27 %	0,94	✓	✓	PAKAI
15	3.02	0.90	27 %	0,94	✓	✓	PAKAI
16	3.9	0.96	27 %	0,94	✓	✓	PAKAI
17	3.97	0.89	27 %	0,94	✓	✓	PAKAI
18	4.15	0.94	27 %	0,94	✓	✓	PAKAI
19	2.93	0.92	27 %	0,94	✓	✓	PAKAI
20	5.80	0.92	27 %	0,94	✓	✓	PAKAI
21	5.45	0.90	27 %	0,94	✓	✓	PAKAI

No	Kriteria						Ket.
	1	2	3	4	5	6	
22	4.25	0.92	27 %	0,94	✓	✓	PAKAI
23	3.78	0.94	27 %	0,94	✓	✓	PAKAI
24	5.24	0.97	27 %	0,94			REVISI
25	5.98	0.89	27 %	0,94	✓	✓	REVISI
26	4.93	0.86	27 %	0,94	✓	✓	PAKAI
27	4.35	0.90	27 %	0,94	✓	✓	PAKAI
28	4.18	0.97	27 %	0,94	✓	✓	PAKAI
29	3.42	0,97	27 %	0,94	✓	✓	REVISI
30	5.1	0,87	27 %	0,94	✓	✓	REVISI

Butir soal yang layak digunakan adalah butir soal *fit* atau sesuai dengan model Rasch dapat dilihat pada Tabel 3.

Tabel 3. Butir Tes Layak Digunakan

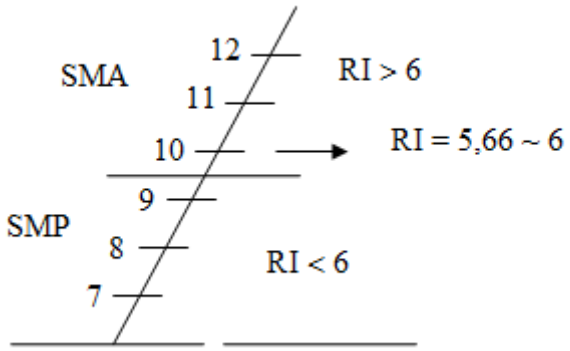
Nomor soal	**Kategori soal**
8, 12	Sangat sulit
9, 19, 1, 10, 2, 14, 13, 28, 26	Sulit
18, 20, 21, 7, 17, 11, 4, 15, 23, 27	Mudah
22, 6, 16, 3	Sangat mudah

Sementara butir soal *misfit* dapat dibuang atau diperbaiki yaitu butir no: 5, 24, 25, 29, dan 30.

Instrumen tes yang dikembangkan adalah butir tes matematika berbasis HOTS berdasarkan kurikulum 2013 dengan jenjang kognitif berdasarkan taksonomi Blom yang direvisi oleh Anderson, Krathwohl dan Bloom (2002) pada tingkatan C4 (menganalisis), C5 (mengevaluasi) dan C6 (mencipta). Hasil penelitian adalah untuk mendapatkan instrument untuk mengukur tingkat abilitas peserta didik pada materi matematika berbasis HOTS yang memenuhi karakteristik instrument yang baik yaitu tingkat validitas dan reliabilitasnya telah terukur. Selain itu, tujuan penelitian adalah untuk memperoleh butir-butir tes hasil belajar berbasis HOTS yang memenuhi karakteristik butir yang layak menurut Model Rasch.

Penelitian menggambarkan tingkat validitas kesesuaian panelis yang berjumlah 18 panelis terhadap 30 butir tes berbasis HOTS diyatakan valid dengan kategori tinggi sebanyak 29 butir dan kategori sedang sebanyak 1 butir. Koefisien reliabilitas kesesuaian panelis sebesar r = 0,968. Nilai koefesien reliabilitas menunjukkan konsistensi hasil penilaian antar panelis tergolong sangat tinggi. Dari hasil analisis validitas dan reliabilitas kesesuaian panelis menunjukkan bahwa tes yang dikembangkan secara isi dinyatakan valid dan reliabel.

Tingkat keterbacaan butir, digambarkan dengan *readability indeks* (RI) untuk peserta didik tingkat SMP < 6. Berdasarkan hasil analisis tingkat keterbacaan butir menggambarkan bahwa sebanyak 30 butir memiliki RI < 6 dengan rata-rata tingkat keterbacaan sebesar 4.1906. Indikataor keterbacaan yang diperoleh menunjukkan bahwa tes ini dapat diberikan untuk siswa jenjang SMP (Sutrisno, 2008). Kriteria tingkat keterbacaan dapat dilihat pada Gambar 4.



Gambar 4. Kriteria Readability Index

Ditinjau dari segi waktu, peserta didik dapat menyelesaikan 30 soal selama 100 menit. Berarti dengan waktu 120 menit peserta didik dapat menyelesaikan sebanyak 30 butir pada saat uji coba dalam skala besar. Alternatif jawaban yang digunakan tidak menghambat peserta didik dalam menjawab seperangkat tes yang diberikan, sehingga tidak perlu diperbaiki. Demikian juga dalam penggunaan bahasa, semua butir dapat dimengerti oleh peserta didik dengan baik, sehingga tidak perlu ada perbaikan dari aspek bahasa.

Berikut beberapa contoh soal pada tiap tingkatan butir yang layak.

Contoh butir yang tergolong sangat sukar, butir nomor 8

"Madi dapat mengisi suatu drum air sampai penuh dalam waktu 20 menit. Randi dapat mengisi drum tersebut sampai penuh dalam waktu 15 menit. Jika mereka bekerja sama, waktu untuk mengisi drum tersebut sampai 2/3 bagian adalah ……"

A. 12 menit C. 22 menit
B. 15 menit D. 25 menit

Contoh butir tergolong sukar, butir nomor 26 dan 28

"Ada dua kantong yang berisi jeruk . Kantong pertama berisi 30 jeruk, dengan 10% diantaranya busuk. Kantong kedua berisi 70 jeruk, dengan 20% diantaranya busuk. Semua jeruk dikedua kantong itu kemudian dijadikan satu lalu dimasukkan ke dalam peti .Peluang jeruk yang tidak busuk dalam peti itu adalah …."

A. 0,82 C. 0,84
B. 0,83 D. 0,85

"Data pada tabel berikut merupakan nilai matematika 40 siswa.

Nilai	Frekwensi
50-54	8
55-59	9
60-64	x
65-69	8
70-74	Y

Jika x - 3y = 3, maka modus data tersebut adalah....”

A. 56,15 C. 61,64

B. 61,51 D. 69,5

Contoh butir tergolong mudah, butir nomor 4

“Perhatikan table berikut:

Jenis barang	Jumlah (kg)	Harga (Rp)
Beras	2,5	30.000
Gula	3,5	52.500
Kopi	1,25	37.500

Apabila ibu membeli 4,25 kg beras dan 2,5 kg kopi, maka uang yang ibu bayarkan adalah…

A. Rp 126.000 C. Rp 82.500

B. Rp 120.000 D. Rp 67.500

Contoh soal yang tergolong sangat mudah, butir nomor 3

“Jika x + y + z = 25, x + y = 19 dan y + z = 18, maka nilai y adalah …. “

A. 16 C. 13

B. 16 D. 12

Penelitian pada tahap penyebaran (*Disseminate*) telah dilakukan uji coba pada skala luas dengan jumlah sampel minimal 5 kali jumlah item. Hasil uji coba pada skala luas akan digunakan untuk menentukan validitas item/ Unidimensionalitas, reliabilitas item, tingkat kesukaran butir dan daya pembeda soal dan kesesuaian butir melalui analisis Model Rasch. Penafsiran apakah butir tes yang telah dirancang dan telah diujikan kepada sejumlah subyek dengan berbagai tingkatan kemampuan memiliki kualitas dan kelayakan sesuai dengan analisis Model Rasch. Bentuk soal yang dikembangkan berupa pilihan ganda agar cakupan materi lebih luas (1 semester). Oleh karena itu data yang diperoleh berupa data dikotomi (benar dan salah) dan teknik pengambilan data hanya menggunakan teknik pengukuran tidak disertai dengan teknik komunikasi langsung (wawancara) sehingga data ini tidak dapat menggambarkan jawaban (penyelesaian) peserta tes.

Hasil analisis menggunakan software Winstep akan diinterpretasikan sesuai dengan tujuan penelitian ini yaitu memilih butir yang layak yang memiliki tingkat kesukaran yang tidak lebih kecil dari kemampuan peserta tes yang paling lemah dan tidak lebih besar dibandingkan kemampuan peserta tes yang paling tinggi.

SIMPULAN

Dari hasil penelitian dan pembahasan dapat diperoleh kesimpulan bahwa: 1) Hasil analisis dapat disimpulkan bahwa tes hasil belajar matematika siswa sekolah menengah pertama berbasis HOTS memiliki Validitas kesesuaian panelis sebanyak 30 butir dan memiliki tingkat kesesuaian reliabilitas panelis sangat tinggi; 2) Sebanyak 25 butir telah memenuhi seluruh karakteristik butir berdasarkan model Rasch yaitu: a) Memiliki nilai unidimensionalitas yang sudah memenuhi standar nilai minimum yaitu 27,2% (*Raw variance explained by measure*s > 20%), b) Memiliki nilai *Item Reliability* istimewa, c) Setiap item memiliki validitas isi tinggi; 3) Karakteristik butir pada instrumen tes yang layak menurut model Rasch pada penelitian ini, yaitu: a) Tingkat keterbacaan tes memiliki nilai rata-rata *readability index* (RI) ≤ 6). Dengan demikian instrument tes yang dikembangkan layak diberikan kepada anak SMP, b) Memiliki 25 butir soal yang berada dalam kemampuan peserta didik tertinggi maupun terendah (butir soal tidak terlalu mudah dan tidak terlalu sukar), c) Memiliki 26 butir soal dengan nilai tingkat kesukaran yang layak (nilai *measure* antara -1 sampai dengan 1 dan) dan sesuai dengan model Rasch (nilai *infit* dan *outfit MNSQ* antara 0,5 sampai dengan 1,5) sehingga dapat digunakan untuk mengukur kemampuan peserta didik.

REFERENSI

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2002). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, 41(4): 352. http://books.google.com/books?id=JPkXAQAAMAAJ&pgis=1

Black, P. & Wiliam, D. (1998). Assessment and Classroom Learning. In *Assessment in Education: Principles, Policy & Practice*, 5(1): 7-74.

Bond, G, T., & Fox C., M. (2015). *Applying the Rasch Model Fundamental Measurement in the Human Sciences, Third Edition.* New York: Routledge.

Brookhart, S. M. (2014). *How to design questions and tasks to assess student thinking.* United States of Amerika: ASCD Member Book.

Cavanagh, R. F., Waugh, R, F. (2011). *Applications of Rasch Measurement in Learning Environments Research.* Netherlands: Sense Publisers.

Collins, R. (2014). *Skills for the 21st Century: Teaching Higher-Order Thinking.* Curriculum & Leadership Journal, 12(14): 1-8.

Conklin, W. (2012). *Higher-Order Thinking Skills to Develop 21st Century Learners.* Huntingon Beach: Shell Educationl Publishing, Inc.

Dali, N. S. (2012). *Teori Sekor Pada Pengukuran Mental*. Jakarta: PT Nagarani Citrayasa.

Djemari, M. (2012). *Pengukuran Penilaian & Evaluasi Pendidikan. Yogyakarta : Nuha Medika.*

Fitriawan, D. (2020). Pengembangan Bahan Ajar Aljabar Linear Elementer Berdasarkan Kemampuan Koneksi Matematis. *Jurnal Pendidikan Matematika dan Sains*, 6(1): 93–104. http://dx.doi.org/10.26418/jpmipa.v11i2.37476

Johnson, D. W., & Johnson, R. T. (2002). Learning Together and Alone: Overview and Meta-Analysis. Asia Pacific. *Journal of Education*, 22: 995-1005. https://doi.org/10.1080/0218879020220110

Kemendikbud. (2020). *Peraturan Menteri Pendidikan Dan Kebudayaan Nomor 03 Tahun 2020 Tentang Standar Nasional Perguruan Tinggi*.

Novinda, M. R. R., Silitonga, H. T. M., & Hamdani. (2019). Pengembangan Tes Pilihan Ganda Menggunakan Model Rasch Materi Gerak Lurus Kelas X Pontianak Artikel Penelitian. *Jurnal Pendidikan Dan Pembelajaran*, 8(6): 1-11.

Permendikbud. (2013). *Kurikulum 2013*. 309–316.

Qasem, M. A. N. (2013). A Comparative Study of Classical Theory (Ct) and Item Response Theory (Irt) In Relation To Various Approaches of Evaluating the Validity and Reliability of Research Tools. *IOSR Journal of Research & Method in Education (IOSRJRME)*, 5(5):77-81. http://dx.doi.org/10.9790/7388-0357781

Resnick, M.D., *et al.* (1997). Protecting Adolescents from Harm: Findings from the National Longitudinal Study on Adolescent Health. *Journal of the American Medical Association,* 278, 823-832. https://doi.org/10.1001/jama.278.10.823

Saputro, M., Yadi, A., & Dona, F. (2015). Faktor-Faktor yang Mempengaruhi Prestasi Belajar (Studi Korelasi Pada Mahasiswa Pendidikan Matematika IKIP PGRI Pontianak). *Jurnal Pendidikan Informatika dan Sains*, 4(2): 233–246. http://dx.doi.org/10.31571/saintek.v4i2.73

Sudijono, A. (2012). *Pengantar Evaluasi Pendidikan.* Jakarta: Rajawali Press.

Sumintono, B., & Widhiarso, W. (*2013*). *Aplikasi Model Rasch untuk Penelitian Ilmu-ilmu Sosial*. Cimahi: Trim Komunikata Publishing House.

Sumintoro, B., & Widhiarso. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan.* Cimahi: Trim Komunikata Publishing House.

Sutrisno, L. (2008). *Remediation of Weaknesses of Physics Concepts.* Pontianak: Untan

Press.

Syahwaludi, M., R, Z., & Suratman, D. (2016). *Higher Order Thingking Skills Siswa Pada Materi Statistika Kelas Xi Ipa Man 2 Pontianak*, 5(11): 1–12.

Tofade, T., Elsner, J., & Haines, S. T. (2013). Best Practice Strategies for Effective use of Questions as a Teaching Tool. *American Journal of Pharmaceutical Education*, 77(7): 155. https://dx.doi.org/10.5688%2Fajpe777155

Sumar, W. T., & Sumar, S. T. (2020). Implementasi Program Pengembangan Keprofesian Berkelanjutan Guru melalui Peningkatan Kompetensi Pembelajaran Berbasis Zonasi. *Pedagogika*, 10(2): 84–94. https://doi.org/10.37411/pedagogika.v10i2.60

Wadiyani, D. (2019). Pengembangan Soal Higher Order Tinking Skills untuk Pengkategorian Kemampuan Pemecahan Masalah Geometri Siswa SMP. *Jurnal Pendidikan Dan Pembelajaran Matematika Indonesia*, 8(2): *161-170*. https://doi.org/10.23887/jppm.v8i2.2854

Wyat-Smith, & Cumming, J. J. (2002). *Educational Assessment in the 21st Century: Connecting Theory and Practice.* London: Springer.