JOURNAL OF KOMODO SCIENCE EDUCATION
Vol. 01 No. 01 November 2018
ISSN : 2621-1238 (online)
http: ejournal.stkipsantupaulus.ac.id/index.php/jkse

PENGEMBANGAN SOAL TES BERPIKIR TINGKAT TINGGI MATERI FLUIDA UNTUK SMA

Nova Liana[1], Wayan Suana, Feriansyah Sesunan[3], Abdurrahman[4]

[1]Universitas Lampung, novaliana167@gmail.com
[2]Universitas Lampung, wsuane@gmail.com
[3]Universitas Lampung, feriansyah_sesunan@yahoo.co.id
[4]Universitas Lampung, abe@unila.ac.id

Abstract

The purpose of this research is to design two-tier multiple choice (TTMC) test on fluid material by Rasch Model analysis. Development was adopted from Adams and Wieman (2011) including the format of item and construction of item setting, assessment guidelines assigning, expert test, and a revision of item. The subject of research trials were 57 students of grade XI MIA in SMAN 1 Kotaagung. The data obtained were analyzed using Rasch Model with Winstep 3,73 applications. Based on the results of research, it was concluded that (1) questions of HOTS developed was valid {achieve the acceptable ranges of PT-Measure, Outfit Mean Square (MNSQ), and value of Outfit z-standardized (ZSTD)}, (2) questions of HOTS developed have excellent reliability with alpha chonbrach's of 0.94, (3) there are 5 very difficult questions, 10 difficult questions, 11 easy questions, and 4 very easy questions, (4) respondent has good consistency of answer, (5) deception options on all item was valid, and (6) all items have a discrimination very good with value of Pt Measure Corr more than from 0,40. Questions of HOTS that developed can increase higher order thinking skills of students so they have good mastery of concept.

Kata Kunci: *fluida, HOTS, model rasch, TTMC.*

PENDAHULUAN

Pembelajaran abad 21 telah mengubah paradigma belajar yakni dari paradigma *teaching* menjadi *learning*. Pada abad 21, guru bukan lagi menjadi pusat belajar melainkan siswa. Selain itu, pembelajaran abad 21 juga menuntut siswa lebih aktif dalam pembelajaran agar siswa dapat mengembangkan kemampuan berpikir kritis dan berpikir kreatif untuk keberhasilan siswa khususnya di bidang pendidikan. Hal ini sejalan dengan pendapat ahli yang mengungkapkan bahwa *characteristics of higher order thinking skills: higher order thinking skills encompass both critical thinking and creative thinking* (Conklin, 2012: 14). Kemampuan berpikir kritis dan kreatif merupakan kemampuan berpikir tingkat tinggi. Dengan demikian, maka kondisi ideal pembelajaran abad 21 pada pelajaran fisika akan terwujud jika kemampuan berpikir tingkat tinggi diterapkan.

Pada kenyataannya, prestasi fisika tahun 2011 yang diukur pada aspek *reasoning*, Indonesia berada pada ranking 40 dari 42 negara (Micheal & Ina, 2013). Artinya, hasil pencapain kognitif siswa di bidang fisika pada aspek *reasoning* keterampilan berpikir tingkat tinggi (*Higher Order Thinking Skills*) masih tergolong rendah. Oleh karena itu, sebaiknya guru mengarahkan siswanya untuk berpikir tingkat tinggi dalam suatu pembelajaran agar pembelajaran tersebut menjadi jauh lebih bermakna. Pembelajaran dengan me-nerapkan kemampuan berpikir tingkat tinggi akan membuat siswa tidak hanya mampu mengingat dan memahami suatu konsep, tetapi siswa juga mampu menganalisis, mengevaluasi, dan meng-kreasikan suatu konsep dengan baik, sehingga penting sekali bagi siswa untuk memiliki keterampilan berpikir tingkat tinggi.

Pada suatu pembelajaran, salah satu hal yang dapat menyebabkan rendahnya penguasaan konsep yaitu siswa mengalami kesulitan belajar sains khususnya pelajaran fisika. Salah satu materi fisika SMA yang dianggap sulit bagi siswa adalah materi fluida (fluida statis dan fluida dinamis). Hal ini sesuai dengan analisis data lapangan yang telah dilakukan yang menunjukkan bahwa mereka kesulitan dalam mem-pelajari materi fluida (fluida statis dan fluida dinamis), dan soal materi fluida (fluida statis dan fluida dinamis) tergolong sulit karena guru pernah memberikan soal-soal yang menantang. Soal-soal yang menantang dapat meningkatkan kemampuan berpikir tingkat tinggi siswa. Kemampuan berpikir tingkat tinggi dapat membangun pemahaman konsep siswa sehingga siswa mampu memecahkan suatu masalah. Hal ini diperkuat oleh Trilling dan Fadel (2009) yang menyatakan bahwa pada setiap subjek dan tingkatan pendidikan, proses pem-belajaran perlu mengintregasikan pembelajaran *content knowladge* dengan kegiatan-kegiatan yang dapat mem-bentuk kemampuan berpikir tingkat tinggi dan pemecahan masalah.

Siswa dengan kemampuan berpikir tingkat tinggi tidak hanya membutuhkan kemampuan mengingat, tetapi juga kemampuan lain yang lebih tinggi yang meliputi kemampuan menganalisis, mengevaluasi, dan mencipta, sehingga sangat penting dalam dunia pendidikan agar serumit apapun masalah yang diberikan akan mudah untuk diselesaikan. Krathworl dan Andrerson (2001) mengungkapkan bahwa kemampuan berpikir tingkat tinggi dalam Taksonomi Bloom yang telah direvisi melibatkan kemampuan analisis (C4), mengevaluasi (C5), dan mencipta (C6). Keterampilan berpikir tingkat tinggi mampu temukan cara baru untuk memecahkan masalah sehari-hari mereka dan menyelesaikan sesuai keputusan (Yee, 2015).

Keberhasilan suatu pendidikan dapat diketahui dengan adanya evaluasi. Salah satu cara untuk mengetahui apakah siswa sudah memiliki keterampilan berpikir tingkat tinggi yaitu dengan cara melakukan penilaian. Penilaian yang berupa tes dapat digunakan untuk mengasah kemampuan berpikir siswa, dan berpengaruh dalam menentukan keterampilan berpikir siswa. Jika siswa diharapkan dapat berpikir tingkat tinggi, maka penilaian tes yang digunakan harus

merepresentasikan kemampuan berpikir tingkat tinggi siswa, dimana instrumen penilaian tersebut dapat berupa soal-soal tes berpikir tingkat tinggi, artinya jenis-jenis soal tersebut merupakan suatu instrumen yang dapat mengukur kemampuan berpikir tingkat tinggi siswa. Hal tersebut diperkuat dengan pernyataan bahwa pertanyaan yang digunakan untuk menguji keterampilan berpikir tingkat tinggi dapat mendorong siswa supaya dapat berpikir mendalam terhadap suatu materi (Bernett & Francis, 2012).

Nyatanya, guru pernah menerapkan kemampuan berpikir tingkat tinggi pada materi fluida namun hanya sebatas level kognitif analisis (C4). Hal ini dikarenakan soal-soal yang diberikan guru kebanyakan soal-soal kategori kemampuan berpikir kritis, dan juga kurang tersedianya soal-soal kategori kemampuan berpikir tingkat tinggi pada level evaluasi (C5), dan kreasi (C6), sehingga anak-anak kurang terlatih dalam menyelesaikan soal berpikir tingkat tinggi. Selain itu, instrumen tes yang digunakan oleh guru dalam bentuk uraian dan pilihan ganda. Hal tersebut dikarenakan penilaian soal lebih objektif dan penskorannya lebih mudah. Padahal instrumen yang mengukur kemampuan berpikir tingkat tinggi siswa dapat menggunakan berbagai tipe soal seperti *modified multiple choice*, konstruksi jawaban singkat, dan konstruksi jawaban panjang yang telah dikembangkan oleh Ramirez dan Ganaden (2008). Bentuk soal *modified multiple choice* yang telah dikembangkan untuk mengukur kemampuan berpikir tingkat tinggi adalah *two-tier multiple choice question* (pilihan ganda bertingkat) (Treagust, 2006).

Two-tier multiple choice question dapat juga digunakan untuk menguji pemahaman siswa dan mengidentifikasi miskonsepsi siswa. Tingkatan kedua pada soal *two-tier multiple choice* dapat digunakan untuk melihat kemampuan berpikir tingkat tinggi siswa dan melihat kemampuan siswa dalam memberikan alasan. Penyertaan tingkatan kedua untuk mengurangi terjadinya keberuntungan yang sering menjadi kelemahan dari bentuk soal pilihan ganda pada umumnya (Cullinane *et al.*, 2011). Ketepatan dan keajegan soal *two-tier multiple choice* dalam pengembangan ini dianalisis menggunakan Model *Rasch* dengan aplikasi *Winstep* 3,73. Tabatabaee-yazdi *et al.* (2017) telah menggunakan Model *Rasch* dengan aplikasi *Winstep* 3,73 untuk melihat kevalidan dan keajegan dari 40 soal kuesioner untuk mengukur keberhasilan guru. Arsad *et al.* (2013) juga telah menggunakan Model *Rasch* untuk menghasilkan pertanyaan-pertanyaan yang efektif dalam menilai tingkat kemampuan siswa.

Analisis *Rasch* merupakan suatu alternatif metode pengukuran modern yang menciptakan dasar pengukuran yang sesuai dengan kriteria satuan SI dan bertindak sebagai instrumen dengan satuan pengukuran yang jelas dan dapat berfungsi sebagai model yang baik (Saidfudin & Ghulman, 2009). Analisis dilakukan menggunakan data empiris yang diperoleh langsung dari penilaian terhadap soal yang diujikan kepada siswa dan kemudian dikonversi ke skala logit. Logit digunakan sebagai satuan pengukuran. Kemudian hasil diperkirakan korelasi linier. *Rasch*

memungkinkan perubahan dari mendefinisikan konsep reliabilitas baris data yang paling kompatibel untuk menghasilkan suatu alat pengukuran berulang yang dapat dipercaya. Itu lebih terfokus pada pembuatan alat pengukuran daripada mencari data untuk menyesuaikan model pengukuran (Osman *et al.*, 2011).

Berdasarkan permasalahan di atas, untuk dapat melengkapi tuntutan dalam pembelajaran kurikulum 2013 maka telah dilakukan pengembangan soal tes *two-tier multiple choice* untuk mengukur kemampuan berpikir tingkat tinggi pada materi fluida Fisika SMA. Adapun Tujuan penelitian ini adalah (1) menghasilkan soal tes HOTS yang valid pada materi fluida fisika SMA, (2) mendeskripsikan reliabilitas soal tes HOTS, (3) menganalisis tingkat kesulitan butir soal, (4) menganalisis kesesuaian individu, (5) menganalisis opsi pengecoh, dan (6) menganalisis daya beda butir soal.

METODE PENELITIAN

Penelitian ini menggunakan metode penelitian dan pengembangan (*Research and Development*). Penelitian ini mengembangkan soal tes *two-tier multiple choice* sebanyak 30 butir soal yang dapat mengukur kemampuan berpikir tingkat tinggi dan penguasaan konsep siswa. Pengembangan dilaksanakan pada materi fisika dengan tema "Fluida (fluida statis dan dinamis)" kelas XI MIA SMAN 1 Kotaagung semester ganjil, tahun ajaran 2017/2018. Penelitian ini menggunakan model pengembangan soal yang diadaptasi dari Adams dan Wieman (2011) yang tahapannya meliputi menentukan format butir soal, menentukan konstruksi butir soal, menentukan pedoman penilaian, uji ahli, dan revisi butir soal.

Menentukan format butir soal. Adapun format butir soal yang dapat diterapkan dalam soal tes diantaranya berupa pilihan ganda (*multiple choice*), essay, pilihan benar salah, dan lain sebagainya. Pada penelitian ini menggunakan format butir soal *two-tier test* pilihan ganda (*multiple choice*). Hal ini diperkuat oleh penelitian Shidiq dkk. (2014), yang mengungkapkan bahwa instrumen *two-tier multiple choice* dapat digunakan untuk mengukur kemampuan berpikir tingkat tinggi siswa.

Menentukan konstruksi butir soal. Dalam membangun atau mengkonstruksi soal tes yang sesuai dan mencerminkan keterampilan kemampuan berpikir tingkat tinggi, maka dalam menyusun butir soal tes harus sesuai dengan indikator-indikator kemampuan berpikir tingkat tinggi yang telah ditetapkan, selain itu bahasa yang digunakan harus jelas dan mudah dipahami.

Menentukan pedoman penilaian. Pedoman penilaian harus disesuaikan dengan tiap butir soal yang telah dibuat. Pedoman penilaian ini digunakan untuk menentukan dan mengetahui pencapaian keterampilan kemampuan berpikir tingkat tinggi siswa yaitu jika *first-tier* benar dan

scond-tier benar diberi skor 3, *first-tier* benar dan *scond-tier* salah diberi skor 2, *first-tier* salah dan *scond-tier* benar diberi skor 1, serta *first-tier* salah dan *scond-tier* salah diberi skor 0.

Uji ahli. Pada tahap ini dilakukan uji validitas dan reliabilitas hasil rancangan soal tes melalui uji ahli terhadap aspek materi, bahasa, dan konstruk. Soal tes yang dinyatakan valid dan reliabel adalah soal yang memiliki nilai koefisien validitas dan reliabilitas pada kategori cukup hingga kategori tinggi dimana validitas instrumen dilakukan uji oleh ahli dan uji reliabilitas isntrumen diperoleh dengan menggunakan rumus *alpha cronbach*. Uji ahli dalam penelitian pengembangan ini dilakukan oleh tiga dosen yang ahli pada bidangnya.

Revisi butir soal. Berdasarkan dari hasil uji ahli (uji validitas dan reliabilitas), maka butir soal-soal yang kurang baik akan direvisi kembali dan soal-soal yang tidak layak akan diganti dengan soal yang baru. Hasil dari revisi tersebut melalui uji ahli akan menghasilkan butir soal yang layak dan dapat digunakan sebagai soal yang valid dan reliabel dalam mengukur keterampilan kemampuan berpikir tingkat tinggi siswa.

Teknik analisis data pada penelitian ini menggunakan Model *Rasch* dengan aplikasi *Winstep 3,73* untuk menganalisis validitas, reliabilitas, tingkat kesukaran butir soal, kesesuaian responden terhadap jawaban, dan tingkat abilitas siswa.

HASIL DAN PEMBAHASAN

Penelitian ini menggunakan 3 kelompok level kognitif dengan mengembangkan sub level kognitif tiap indikator. Indikator-indikator yang digunakan untuk mengembangkan soal tes kemampuan berpikir tingkat tinggi (HOTS) ini didalamnya terdapat stimulus yang berupa wacana, grafik, ataupun gambar, kata kerja operasional yang menggambarkan level kognitif tingkat tinggi, dan materi yang akan dicapai. Indikator-Indikator pada soal HOTS ini menggunakan KKO Anderson dan Krathworl yang telah direvisi dan disesuaikan dengan level kognitif yang diukur. Secara rinci dapat dilihat dalam Tabel 1.

Tabel 1
Indikator Berpikir Tingkat Tinggi yang Dikembangkan

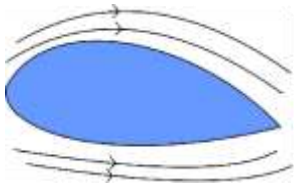
Level Kognitif	Sub Level Kognitif	Jumlah
C_4	Menguji Ide	8
	Menganalisis Argumen	2
	Mengidentifikasi Argumen	2
C_5	Menilai Pernyataan	5
	Menilai Argumen	4
C_6	Merancang Desain	3
	Memberikan Solusi	6

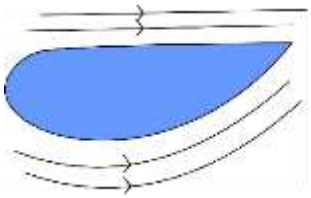
Soal tes HOTS yang dikembangkan telah memenuhi standar untuk penilaian karena intrumen tersebut memiliki reliabilitas, validitas, serta skala peringkat yang baik. Soal tes HOTS yang dikembangkan tersebut merupakan soal yang kontekstual dan mengambarkan kejadian-kejadian yang nyata dalam kehidupan namun tetap mengaitkan pada konsep fisika yang telah dipelajari. Adapun spesifikasi soal tes HOTS materi fluida yang dikembangkan adalah sebagai berikut:

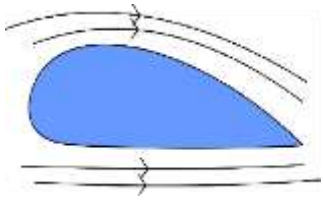
a. Soal tes HOTS pada materi fluida (statis dan dinamis) yang dikembangkan valid dan reliabel sehingga dapat digunakan untuk mengukur kemampuan berpikir tingkat tinggi siswa yaitu menganalisis (C4), mengevaluasi (C5), dan mengkreasi (C6) serta KKO yang digunakan pada C4 (manganalisis, menyimpulkan, menentukan, membandingkan, dan mengoreksi), C5 (mengkritisi dan memprediksi), dan C6 (melakukan, membuat, dan merancang) (Krathworl & Anderson, 2001).
b. Soal yang dikembangkan berbasis masalah kontekstual, terdapat stimulus yang menarik, tidak familiar, dan kebaruan.
c. Jenis soal yang dikembangkan adalah *two-tier multiple choice* dengan tiga pilihan jawaban pada kedua tingkat soal tersebut. Dimana siswa memilih jawaban beserta alasan jawaban tersebut yang dianggap paling benar.
d. Soal tes dikerjakan secara individu dan bersifat *close book* dengan waktu pengerjaan 30 butir soal tes HOTS selama 90 menit.

Contoh soal No. 27, level kognitif C6 dengan sub level kognitif merancang desain.

Sayap pesawat terbang, selain berfungsi sebagai tumpuan kesetimbangan, juga berfungsi memberikan gaya angkat pada pesawat tersebut. Agar sayap pesawat mampu menghasilkan gaya angkat pesawat yang besar, maka desain penampang pesawat yang tepat adalah….

A. 

B. 

C. 

Alasan yang tepat memilih jawaban di atas adalah….

1) Kecepatan udara di atas sayap pesawat besar sehingga tekanan udara di bawah sayap pesawat akan semakin besar.
2) Kecepatan udara di atas sayap pesawat kecil sehingga tekanan udara di bawah sayap pesawat akan semakin besar.
3) Kecepatan udara di atas sayap pesawat kecil sehingga tekanan udara di bawah sayap pesawat akan semakin kecil.

Solusi menyelesaikan soal tersebut adalah dengan menggunakan prinsip gaya angkat pada pesawat terbang. Penampang sayap pesawat terbang memiliki bagian belakang yang lebih tajam dan sisi bagian atas yang lebih melengkung daripada sisi bagian bawahnya. Bentuk sayap pesawat tersebut menyebabkan kecepatan aliran udara di bagian atas lebih besar daripada di bagian bawah sehingga tekanan udara di bawah sayap pesawat lebih besar daripada di bagian atas.

Selanjutnya, soal yang telah dikembangkan diberikan kepada tiga dosen ahli dalam bidang pendidikan fisika dan evaluasi pembelajaran untuk menilai aspek konstruk, materi, dan bahasa. Berdasarkan hasil uji ahli diperoleh bahwa rata-rata nilai validitas instrumen dari aspek konstruk sebesar 84% yang dikategorikan valid dengan revisi kecil, aspek materi sebesar 85% yang dikategorikan valid dengan revisi kecil, dan aspek bahasa sebesar 90% yang dikategorikan valid yang tinggi, sehingga soal yang dikembangkan layak digunakan untuk mengukur kemampuan berpikir tingkat tinggi siswa. Selanjutnya, soal diuji keterbacaannya kepada 3 orang siswa dan hasilnya menunjukkan bahwa siswa sangat mudah memahami maksud pertanyaan pada setiap butir soal yang berkaitan dengan kejelasan bahasa, huruf, angka, gambar, simbol, dan grafik. Setelah dilakukan uji ahli dan keterbacaan soal, soal diujicobakan secara terbatas kepada 57 siswa kelas XI MIA. Uji coba dilakukan pada tanggal 7 Mei 2018. Soal yang digunakan sebanyak 30 butir soal, terdiri dari 14 soal fluida statis dan 16 soal fluida dinamis. Berdasarkan hasil uji coba, data yang diperoleh dianalisis menggunakan Model *Rasch* melalui program *Winstep* 3.73 untuk mengukur validitas, reliabilitas, tingkat kesukaran butir soal, daya beda soal, kesesuaian responden terhadap jawaban, dan tingkat abilitas siswa.

Validitas

Tingkat kesesuaian butir soal (*item fit*) digunakan untuk menjelaskan apakah butir soal berfungsi normal melakukan pengukuran atau tidak. Apabila didapati suatu soal yang tidak fit, berarti terdapat indikasi bahwa terjadi miskonsepsi pada siswa terhadap butir soal tersebut. Menurut Boone *et al*. (2014) serta Bond dan Fox (2015), nilai o*utfit mean-square*, *outfit z-standard*, dan *point measure correlation* adalah kriteria yang digunakan untuk melihat tingkat kesesuaian butir (*item fit*). Dimana nilai *outfit mean square* (MNSQ) yang diterima yaitu 0,5 < MNSQ < 1,5, nilai *outfit Z-standard* (ZSTD) yang diterima yaitu -2,0 < ZSTD < +2,0, dan nilai *point measure correlation* (Pt Mean Corr) yang diterima yaitu 0,4 < Pt Mean Corr < 0,85. Berdasarkan ketiga kriteria tersebut, semua butir soal yang dikembangkan telah memenuhi kriteria sehingga dapat disimpulkan bahwa semua soal valid dan tidak perlu diubah atau diganti. Semua soal fit, artinya tidak terjadi miskonsepsi pada siswa terhadap butir soal.

Reliabilitas

Hasil analisis data diperoleh bahwa 30 soal dinyatakan reliabel dengan nilai *alpha Cronbach's* sebesar 0,94 yang termasuk dalam kategori soal tes dengan reliabilitas bagus sekali. Nilai *alpha Cronbach's* ini digunakan untuk mengukur reliabilitas yaitu interaksi antara *person* dan *item* secara keseluruhan seperti yang ditunjukkan pada tabel 2 dan 3.

Tabel 2
Person Reliability

	TOTAL SCORE	COUNT	MEASURE	MODEL EROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	39.4	30.0	-.27	.24	1.01	.1	.99	.0
S.D.	17.6	.0	.90	.02	.26	1.0	.22	.8
MAX.	78.0	30.0	1.89	.31	1.59	2.3	1.47	1.7
MIN.	12.0	30.0	-1.88	.21	.62	-1.9	.58	-1.5
REAL RMSE .25 TRUE SD .87 SEPARATION 3.49 Person RELIABILITY .92								
MODEL RMSE .24 TRUE SD .87 SEPARATION 3.69 Person RELIABILITY .93								
S.E. OF Person MEAN = .12								

Person RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY= .94

Tabel 2 menunjukkan bahwa nilai rata-rata INFIT MNSQ dan OUTFIT MNSQ secara berurutan yaitu 1,01 dan 0,99 artinya nilainya mendekati nilai sempurnya yaitu 1,00. Nilai rata-rata INFIT ZSTD dan OUTFIT ZSTD secara berurutan yaitu 0,1 dan 0,0 artinya kualitas *person* makin baik karena nilainya mendekati sempurna yaitu 0,0. Nilai *person reliability* sebesar 0,92 yang menunjukkan bahwa konsistensi jawaban dari responden bagus sekali. Nilai *sparation* sebesar 3,49 sehingga diperoleh nilai pemisahan stratanya sebesar 4,99 (dibulatkan menjadi 5) artinya bahwa terdapat lima kelompok responden. Ini menunjukkan bahwa semakin besar nilai *sparation* maka semakin bagus kualitas responden secara keseluruhan. Hal ini mencerminkan beragamnya abilitas (heterogen) yang menunjukkan keterwakilan abilitas siswa yang mengikuti tes.

Tabel 3
Item Reliability

	TOTAL SCORE	COUNT	MEASURE	MODEL EROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	74.9	57.0	.00	.17	1.00	.1	.99	.1
S.D.	28.2	.0	.79	.02	.16	.9	.18	.9
MAX.	130.0	57.0	1.59	.22	1.27	1.5	1.28	1.4
MIN.	25.0	57.0	-1.50	.16	.71	-1.4	.65	-1.5
REAL RMSE .18 TRUE SD .77 SEPARATION 4.38 Person RELIABILITY .95								
MODEL RMSE .17 TRUE SD .77 SEPARATION 4.52 Person RELIABILITY .95								
S.E. OF Person MEAN = .15								

Tabel 3 menunjukkan bahwa nilai rata-rata INFIT MNSQ dan OUTFIT MNSQ secara berurutan yaitu 1,00 dan 0,99 artinya nilainya mendekati nilai sempurnya yaitu 1,00. Nilai rata-

rata INFIT ZSTD dan OUTFIT ZSTD secara berurutan yaitu 0,99 dan 0,1 artinya kualitas *item* makin baik karena nilainya mendekati sempurna yaitu 0,0. Nilai *item reliability* sebesar 0,95 yang menunjukkan bahwa kualitas *item-item* dalam soal istimewa. Nilai *sparation* sebesar 4,38 sehingga diperoleh nilai pemisahan stratanya sebesar 6,17 (dibulatkan menjadi 6), artinya bahwa terdapat enam kelompok *item*. Hal ini menunjukkan bahwa semakin besar nilai *sparation* maka semakin bagus kualitas *item* secara keseluruhan.

Tingkat Kesulitan Butir Soal

Untuk mengetahui tingkat kesulitan butir soal (*item measure*) dilihat dari nilai *logit* tiap butir soal yang dapat dilihat pada kolom *measure*. Nilai *logit* yang tinggi menunjukkan tingkat kesulitan soal yang paling tinggi. Tingkat kesulitan butir soal pada penelitian ini dikelompokkan dengan menggunakan nilai *logit* masing-masing *item*. Penelitian ini mengkategorikan tingkat kesulitan soal dalam 4 kategori berdasarkan nilai *logit*, sebagaimana ditampilkan pada Tabel 4.

Tabel 4
Jumlah Kategori Soal

Measure	Kategori	No. Soal	Jumlah	Persentase
> +1	Sangat Sulit	26, 8, 21, 27, dan 25	5	17%
0 - 1	Sulit	28, 10, 20, 17, 29, 3, 15, 30, 7, dan 19	10	33%
-1 - 0	Mudah	6, 23, 18, 5, 9, 1, 24, 15, 22, 12, dan 11	11	37%
< -1	Sangat Mudah	14, 2, 13, dan 4	4	13%
		Jumlah	**30**	**100%**

Tabel 4 menunjukkan bahwa terdapat 5 soal sangat sulit, 10 soal sulit, 11 soal mudah, dan 4 soal sangat mudah. Berdasarkan nilai *logit*, soal yang paling sulit dikerjakan oleh responden adalah S26, ini ditunjukkan dari nilai *logit* sebesar 1,59 yang merupakan nilai *logit* tertinggi dibandingkan dengan soal-soal yang lainnya. Selain itu, soal yang paling mudah dikerjakan oleh responden adalah S4, ini ditunjukkan dari nilai *logit* sebesar -1,50 yang merupakan nilai *logit* terendah dari soal-soal lainnya. Dengan kata lain, *item* S26 memiliki tingkat kesulitan tertinggi (1,59 *logit*) dan *item* S4 memiliki tingkat kesulitan yang rendah (-1,50 *logit*).

Daya Beda Soal

Butir soal tes dikatakan bagus jika memiliki daya beda yang bagus pula. Daya beda soal dalam penelitian ini dikenal dengan istilah daya diskriminasi *Rasch* atau nilai korelasi skor butir dan skor *Rasch* (*Pt Measure Corr*) pada prinsipnya sama dengan daya diskriminasi *item* yang diukur dengan pendekatan teori tes klasik. Hanya saja pada teori tes klasik komputasinya menggunakan skor mentah, pada *Pt Measure Corr* yang digunakan adalah skor *measure*. Nilai *Pt Measure Corr* 1,0 mengindikasikan bahwa semua peserta tes dengan abilitas rendah menjawab butir soal dengan salah dan abilitas tinggi menjawab butir soal dengan benar. Sementara nilai *Pt*

Measure Corr negatif mengindikasikan butir soal yang menyesatkan karena peserta tes dengan kemampuan yang rendah mampu menjawab butir soal dengan benar dan peserta tes dengan kemampuan tinggi menjawab salah. Soal-soal dengan nilai korelasi negatif harus diperiksa untuk dilihat apakah kunci jawaban salah, perlu direvisi atau dihapus dari tes (Smiley, 2015).

Hasil analisis pada *Pt Measure Corr* menunjukkan bahwa nilai *Pt Measure Corr* positif, artinya tidak ada butir soal yang menyesatkan sehingga tidak perlu merevisi atau pun menghapus soal. Selain itu, nilai *Pt Measure Corr* pada semua butir soal lebih besar dari 0,40 artinya semua soal mempunyai daya beda yang sangat bagus. Hal ini sesuai dengan pendapat Alagumalai (2005) yang mengklasifikasikan nilai *Pt Measure Corr* > 0,40 (sangat bagus), 0,30-0,39 (bagus), 0,20-0,29 (cukup), 0,00-0,19 (tidak mampu mendeskriminasikan, dan < 0,00 (membutuhkan pemeriksaan terhadap butir).

Tingkat Kesesuaian Individu

Untuk melihat tingkat kesesuaian individu (*Person Fit*) yang menunjukkan tingkat kesesuaian pola respons. Kriteria yang digunakan untuk melihat tingkat kesesuaian individu (*person fit*) sama seperti yang digunakan pada *item fit*. Berdasarkan ketiga kriteria tersebut, semua responden telah memenuhi kriteria sehingga dapat disimpulkan bahwa semua responden memiliki konsistensi jawaban yang baik.

Tingkat Abilitas Siswa

Tingkat abilitas siswa digunakan untuk mengidentifikasihan tingkat kemampuan siswa dalam menjawab soal. Tingkat abilitas siswa telah diurutkan dari yang tertinggi hingga terendah berdasarkan nilai *logit* tiap *person* yang dapat dilihat pada kolom *measure*. Responden 29P memiliki nilai *logit* paling tinggi yaitu 1,89, artinya responden 29P memiliki tingkat abilitas paling tinggi dibandingkan yang lain. Sedangkan, responden 52P memiliki nilai *logit* paling rendah sebesar -1,88, ini berarti bahwa responden 52P memiliki tingkat abilitas yang rendah.

Skala Peringkat (*Rating Scale*)

Analisis validitas skala peringkat dilakukan untuk memverifikasi apakah peringkat (*rating*) pilihan yang digunakan membingungkan bagi responden atau tidak, secara rinci dapat dilihat pada Tabel 5.

Tabel 5
Skala Peringkat (*Rating Scale*)

CATEGORY		OBSERVED		OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	UTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
LABEL	SCORE	COUNT	%							
0	0	526	31	-1.31	-1.24	.88	.91	NONE	(-2.12)	0
1	1	460	27	-.43	-.53	1.08	1.00	-.75	-.58	1
2	2	384	22	.23	.21	1.03	1.03	.02	-.59	2
3	3	340	20	.95	1.02	1.12	1.05	.73	(2.11)	3

Tabel 5 menunjukkan bahwa opsi yang diberikan sudah valid bagi responden karena nilai *logit* pada kolom OBSVD AVRGE menunjukkan bahwa nilai *logit* yang meningkat dari rendah sampai tinggi dan nilai *Andrich Threshold* bergerak dari NONE kemudian negatif dan terus mengarah ke positif secara berurutan (Andrich, 1988). Hal ini berarti siswa mampu memahami pilihan yang diberikan pada setiap tingkat kesulitan butir soal.

Berdasarkan uraian hasil dan pembahasan, disimpulkan bahwa soal tes *two-tier multiple choice* dapat dijadikan alternatif untuk mengukur kemampuan berpikir tingkat tinggi siswa terhadap materi fluida (fluida statis dan dinamis). Hal ini diperkuat oleh peneliti lain yang menghasilkan bahwa instrumen *two-tier multiple choice* dapat digunakan untuk mengukur kemampuan berpikir tingkat tinggi siswa (Shidiq, 2014; Treagust, 2006).

SIMPULAN

Berdasarkan hasil penelitian dan pembahasan dapat disimpulkan bahwa (1) soal tes berpikir tingkat tinggi materi fluida yang dikembangkan telah memenuhi standar kelayakan instrumen yaitu valid dan reliabel. Hasil analisis data menggunakan *Winstep* 3.73 semua soal telah memenuhi tiga kriteria kesesuaian butir soal menurut Boone *et al.* (2014), (2) soal tes HOTS memiliki reliabilitas sangat tinggi dengan *alpha chonbrach* sebesar 0,94. Dengan demikian soal tes dapat digunakan sebagai alternatif untuk mengukur kemampuan berpikir tingkat tinggi siswa, (3) soal tes berpikir tingkat tinggi yang dikembangkan berdasarkan nilai *logit* terdapat 5 soal sangat sulit, 10 soal sulit, 11 soal mudah, dan 4 soal sangat mudah, (4) responden memiliki konsistensi jawaban yang baik pada setiap butir soal yang diberikan, (5) opsi pengecoh pada semua soal yang diberikan kepada responden sudah valid (responden memahami pilihan yang diberikan pada setiap tingkat kesulitan butir soal), dan (6) semua butir soal memiliki daya beda yang sangat bagus karena nilai *Pt Measure Corr* positif dan lebih besar dari 0,40.

Penulis menyarankan agar guru dapat mengembangkan instrumen tes berpikir tingkat tinggi berdasarkan indikator berpikir tingkat tinggi Krathworl & Anderson pada setiap materi fisika, dan peneliti lain dapat mengembangkan soal berpikir tingkat tinggi dengan materi lainnya.

UCAPAN TERIMA KASIH

Para penulis sangat menghargai Direktorat Riset dan Pengabdian Masyarakat, Kementrian Riset, Teknologi, dan Pendidikan Tinggi untuk bantuan hibah keuangan melalui Hibah Lembaga Riset Strategis Nasional 2017-2018.

DAFTAR RUJUKAN

Adams, W. K., dan Wieman, C. E. (2011). Development and Validation of Instruments to Measure Learning of Expert Like Thinking. *Inter-national Journal of Science Education*, 33(9), 1289–1312.

Alagumalai, S., Curtis, D. D., dan Hungi, N. (2005). *Applied Rasch Measurement: A Book of Exemplars.* Dordrecht: Springer.

Anderson, L.W., dan Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching, and Assesing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Addison Wesley Longman, In.

Andrich, D. (1988). *Rasch Model for Measurement.* (Series: Quantitative Application in the Socials Sciences). Newburry Park, California: Sage Publication.

Arsad, N., Kamal, N., Ayob, A., Sarbani, N., Tsuey, C. S., Misran, N., dan Husein, H. (2013). Rasch Model Analysis on the Effectiveness or early Evaluation Questions as a Benchmark for New Student Ability. *Inter-national Education Studies*, 6(6), 185-190.

Barnett, J. E. & Francis, A. L. (2011). Using higher order thinking questions to foster Critical Thinking: a classroom study. *Educational Psychology: An International Journal of Experimental Educational Psychology*, 32(2), 201-211.

Bond, T. G., dan Fox, C. M. (2015). *Applying the rasch Model, Fundamental Measurement in the Human Sciences* (3rd edition). New York: Routledge.

Boone, W. J., Staver, J. R., dan Yale, M. S. (2014). *Rasch Analysis in the Human Science*. Dordrecht: Springer.

Conklin, W. 2012. *Higher Order Thinking Skills To Develop 21st Century Learners.* Huntington Beach: Shell Educational Publishing, Inc.

Cullinane, A., dan Liston, M. (2011). *Two-tier Multiple Choice Question: An Alternative Methode of Formatif Assesment for first Year Undergraduade Biology Students.* Linmark: National Center for Excellence In Mathematics and Education Science Teaching and Learning (NCE-MSTL).

Michael O. Martin, Ina V.S. Mullis, Pierre Foy, and Gabrielle M. Stanco. (2013). *TIMSS 2013 International Results in Science*. USA: TIMSS & PIRLS *International Study Center Lynch School of Education Boston College*.

Osman, S. A., Badaruzzaman, W. H. W., Hamid, R., Taib, K., Khalim, A. R., Hamzah R., dan Jaafar, O. (2011). Assesment On Students Performance Using Rasch Model In Reinforced Concrete Design Course Examination. *WSEAS Recent Researches in Education Proceeding,* 193-198.

Ramirez, R. P. B., dan Ganaden, M. S. (2006). Creative Activities and Students' Higher Order Thinking Skills. *Journal of Education Quarterly*, 66(1), 22-23.

Saidfudin, M., dan Ghulman, H. A. (2009). Modern Measurment Paradigm In Engineering Education: Easir To Read And Better Analysis Using Rasch-based Approach. *International Conference on Engineering Education,* 8(5), 591-602.

Shidiq, A.S., Masykuri, M., dan Susanti V. H., E. (2014). Pengembangan Penilaian Instrumen Two-Tier Multiple Choice Untuk Mengukur Kemampuan Berpikir Tingkat Tinggi (Higher Order Thinking Skills) Pada Materi Kelarutan dan Hasil Kali Kelarutan Untuk Siswa SMA/MA Kelas XI. *Jurnal Pendidikan Kimia*, 3(4), 83-92.

Smiley, J. (2015). Classical Test Theory or Rasch: A Personal Account From A Novice User. *SHIKEN,* 19(1), 16-29.

Sumintono, B., dan Widhiarso, W. (2014). *Aplikasi Model Rasch Untuk Penelitian Ilmu-Ilmu Sosial.* Cimahi: Trimkomunikata.

Sumintono, B., dan Widhiarso, W. (2015). *Aplikasi Permodelan Rasch Pada Assessment Pendidikan*. Cimahi: Trimkomunikata.

Tabatabaee-Yadzi, M., Motallebzadeh, K., Ashraf, H., dan Baghaei, P. (2018). Development and Validation of a Teacher Success Questionnaire Using the Rasch Model. *International Journal of Instruction*, 11(2), 129-144.

Treagust, D. F. (2006). Diagnostic Assessment In Science as A Means to Improving teaching Learning and Retention.*Uniserve Scince Assesments Symposium Proceedings*, 1-9.

Trilling, B., dan Fadel, C. (2009). *21st Century Skills, Learning for Life in Our Times*. US America: Jossey-Bass.

Yee. (2015). The Effectiveness of Higher Order Thinking Skills for Generating Idea among Technical Students. *Recent Advances in Educational Technologies*, 1(1), 223-241.