Jurnal Pendidikan FISIKA

ISSN: 3025-9215

Diterbitkan oleh
Universitas Negeri Yogyakarta

Volume 13 Edisi 2, Oktober, 2026, 52-63
DOI: *10.21831/jpf.v13i2.29980*

PENGEMBANGAN INSTRUMEN PENILAIAN UNTUK MENGUKUR PEMAHAMAN KONSEP PESERTA DIDIK MATERI MOMENTUM DAN IMPULS

Lilik Muyassaroh*, Universitas Negeri Yogyakarta, Indonesia
Supahar, Universitas Negeri Yogyakarta, Indonesia
*e-mail: lilikmuyassaroh.2021@student.uny.ac.id

Abstrak. Penelitian ini bertujuan untuk: (1) Mengetahui konstruk instrumen penilaian yang dapat mengukur kemampuan pemahaman konsep peserta didik; (2) Menguji kualitas instrumen penilaian untuk mengukur kemampuan pemahaman konsep peserta didik; (3) Mengetahui kepraktisan instrumen penilaian untuk mengukur kemampuan pemahaman konsep peserta didik; (4) Mengetahui profil kemampuan pemahaman konsep perserta didik berdasarkan hasil pengukuran. Penelitian ini merupakan penelitian pengembangan (R&D) dengan menggunakan pengembangan hasil modifikasi model Wilson serta Oriondo dan Antonio melalui tiga tahapan, yaitu perancangan tes, uji coba tes, dan pengukuran. Produk yang dihasilkan berupa instrumen penilaian berbasis *four-tier multiple choice*. Validasi isi butis soal melibatkan tujuh *expert judgement* yang dianalisis menggunakan persamaan *Aiken's V*. Subjek penelitian melibatkan 10 peserta didik pada uji coba terbatas, 261 peserta didik pada uji coba luas, dan 93 peserta didik pada pengukuran. Subjek penelitian merupakan peserta didik kelas XI di SMA Yogyakarta. Analisis karakteristik butir menggunakan PCM 1-PL (Rasch Model) berbantuan program Winsteps. Hasil penelitian ini berupa instrumen tes *four-tier multiple choice* yang terdiri dari paket A dan paket B, dengan masing-masing paket berjumlah 20 butir soal dengan 4 butir bersama (*anchor item*). Instrumen dinyatakan valid dengan rata-rata *Aiken's V* senilai 0,89; 30 butir soal dinyatakan *fit* dengan model PCM; memiliki reliabilitas senilai 0,92 yang tergolong kategori sangat baik; tingkat kesukaran seluruh butir berkategori sedang; 34 butir soal tidak bias gender; serta dinyatakan praktis untuk digunakan berdasarkan respons guru dan peserta didik. Hasil pengukuran menunjukkan bahwa sebagian besar peserta didik memiliki kemampuan pemahaman konsep pada tingkat sedang dengan persentase 62,36%. Instrumen ini dinyatakan efektif dan layak digunakan untuk mengukur kemampuan pemahaman konsep peserta didik kelas XI.
Kata Kunci: instrumen penilaian, pemahaman konsep, four-tier multiple choice, momentum dan impuls

Abstract. *This study aims to: (1) determine the construct of an assessment instrument capable of measuring students' conceptual understanding; (2) test the quality of the assessment instrument in measuring students' conceptual understanding; (3) evaluate the practicality of the assessment instrument in measuring students' conceptual understanding; and (4) profile students' conceptual understanding based on the measurement results. This research and development (R&D) study utilized a modified development model combining Wilson with Oriondo and Antonio, which progressed through three stages: test design, test trial, and measurement. The resulting product is a four-tier multiple choice assessment instrument.*

Item content validation involved seven expert judges and was analyzed using the Aiken's V equation. The research subjects included 10 students for the limited trial, 261 students for the expanded trial, and 93 students for measurement phase. All subjects were eleventh-grade students at a senior high school in Yogyakarta. Item characteristics were analyzed using the PCM 1-PL assisted by the Winsteps program. The results of this stud developed a four-tier multiple choice test instrument consisting of Form A and Form B, each containing 20 items with 4 anchor items. The instrument was declared valid with an average Aiken's V value of 0,89; 30 items were found to fit the PCM model; it demonstrated a reliability coefficient of 0,92 which is categorized as excellent; the difficulty level of all items fell into the moderate category; 34 items showed no gender bias; and it was deemed practical for use based on teacher and student responses. Furthermore, the measurement results indicated that the majority of students possessed a moderate level of conceptual understanding, accounting for 62,36%. This instrument is effective and feasible for measuring the conceptual understanding of eleventh-grade students.
Keywords: assessment instrument, conceptual understanding, four-tier multiple choice, momentum and impulse

PENDAHULUAN

Asesmen internasional *Programme for International Student Assessment* (PISA) mengidentifikasi bahwa kualitas hasil belajar peserta didik Indonesia relatif rendah dibandingkan dengan negara-negara lain (OECD, 2023). Hal tersebut tidak terlepas dari proses pembelajaran dan sistem evaluasi yang digunakan di sekolah. Dalam upaya mencapai tujuan pendidikan tersebut, pembelajaran memegang peran penting sebagai proses utama dalam mengembangkan kompetensi peserta didik. Dalam konsteks pendidikan sains termasuk fisika, pembelajaran bertujuan agar peserta didik tidak sekadar menghafal rumus, melainkan mampu mendalami pengetahuan secara bermakna. Keberhasilan pembelajaran juga ditentukan oleh sistem penilaian yang berlaku. Penilaian merupakan komponen esensial dari pembelajaran guna memperoleh informasi mengenai perkembangan serta capaian belajar peserta didik (Imania & Bariah, 2019).

Pemahaman konsep merupakan aspek yang penting dan mendasar. Pemahaman konsep merupakan suatu kemampuan menerima, menyerap, serta mengerti suatu materi maupun informasi yang diperoleh melalui serangkaian kejadian atau peristiwa yang dapat diaplikasikan dalam kehidupan sehari-hari (Susanti et al., 2021). Peserta didik yang memiliki pemahaman konsep berkemampuan untuk menginterpretasi, menerjemahkan, mengekstrapolasi, dan menginterpolasi (Hidayat et al., 2019). Kemampuan tersebut menggiring pada pemahaman sesuai dengan konsep yang benar serta memicu peserta didik untuk menghubungkan dengan pengetahuan awal. Berbagai penelitian menunjukkan bahwa peserta disih masih mengalami kesulitan dalam memahami konsep-konsep fisika dengan benar sehingga memunculkan miskonsepsi yang menghambat tujuan kompetensi yang diharapkan. Miskonsepsi merupakan pola pikir yang salah yang terus terjadi dalam permasalahan (Didik et al., 2020). Miskonsepsi hampir terjadi di semua materi pembelajaran fisika (Asmin & Rosdianti, 2021; Azizah et al., 2020; Alawiyah et al., 2017; Sakinah et al., 2023; Paramita & Jauhariyah, 2024) mengidentifikasi pemahaman konsep yang cenderung kurang dan tingkat miskonsepsi yang tinggi.

Hal tersebut menjadi pemicu pentingnya alat ukur yang akurat dalam bentuk instrumen evaluasi untuk mengidentifikasi kemampuan pemahaman konsep peserta didik sebagai langkah awal dalam menciptakan peserta didik yang dapat bersaing di era pendidikan abad ke-21. Salah satu bentuk instrumen tes yang sering dan efektif digunakan untuk mengukur kemampuan pemahaman konsep peserta didik adalah tes objektif pilihan ganda bertingkat. *Four-tier multiple choice* atau pilihan ganda empat tingkat menjadi salah satu

diantaranya yang menjadi instrumen paling akurat, valid, dan reliabel untuk mengidentifikasi miskonsepsi dan mengukur pemahaman konsep peserta didik (Pratama, 2023).

Berbagai penelitian telah menggunakan instrumen tes ini dengan tujuan yang sama, diantaranya adalah Yunita & Nasution (2025), Roistiya et al. (2019), Yasthophi & Ritongga (2019), serta Ananda et al. (2026). Melalui penerapan instrumen *four-tier multiple choice*, kemampuan pemahaman konsep peserta didik dapat diukur dengan baik, sehingga dapat dijadikan sebagai bahan evaluasi untuk memperbaiki kualitas pembelajaran selanjutnya. Hal tersebut yang menjadi acuan dalam mengembangkan instrumen penilaian yang dapat digunakan untuk mengukur kemampuan pemahaman konsep peserta didik menggunakan instrumen *four-tier multiple choice*.

METODE

Penelitian ini menggunakan metode pengembangan modifikasi model Wilson serta Oriondo dan Antonio (Istiyono, Mardapi, & Suparno, 2014). Pengembangan model ini melalui tiga tahap, yaitu perancangan tes, uji coba tes, dan pengukuran. Subjek yang digunakan dalam penelitian adalah peserta didik kelas XI SMA di Yogyakarta yang telah mempelajari materi momentum dan impuls. Instrumen tes yang digunakan pada penelitian ini berupa *four-tier multiple choice* yang dikemas dalam dua paket soal yang berisi 20 butir soal pada masing-masing paket dengan 4 butir bersama.

Pengumpulan data dimulai dengan melakukan uji coba terbatas, uji coba luas, dan dilanjutkan dengan pengukuran. Pengumpulan data menggunakan instrumen tes dan non tes. Instrumen tes berupa instrumen penilaian pemahaman konsep digunakan pada saat uji coba luas dan pengukuran. Instrumen non tes berupa lembar validasi isi dan angket respon digunakan pada tahap validasi isi, uji coba terbatas, dan kepraktisan instrumen. Pengembangan instrumen dimulai dengan menetapkan tujuan tes, kompetensi dan materi, penyusunan matriks dan kisi-kisi tes, penulisan butir soal dan pedoman penskoran, yang kemudian dilakukan validasi isi butir soal oleh *expert judgment*. Hasil validasi isi butir soal dianalisis menggunakan acuan koefisien validitas *Aiken's V* dengan persamaan *Aiken's V* (Aiken, 1985).

$$V = \frac{\sum s}{[n(c-1)]}$$

Keterangan:
s = $r - l_0$
V = koefisien validitas isi
l_0 = skor penilaian validitas terendah
c = skor penilaian validitas tertinggi
r = skor dari penilaian ahli
n = jumlah *expert judgement*

Nilai validitas berada pada rentang 0,0 hingga 1,0. Kriteria validitas isi butir soal dapat dilihat pada Tabel 1.

Tabel 1. Kriteria validitas

Rentang Nilai Koefisien Validitas	Interpretasi Validitas
0,80 < V ≤ 1,00	Valid
0,40 < V ≤ 0,80	Sedang
V ≤ 0,40	Kurang

(Retnawati, 2016)

Butir soal dirakit kembali sesuai dengan saran perbaikan dari validator. Butir soal yang telah diperbaiki akan digunakan pada tahap uji coba produk yang meliputi uji coba terbatas dan uji coba luas. Uji coba terbatas dilakukan untuk mengetahui kualitas produk berdasarkan penulisan, ejaan, tanda baca, serta penggunaan kata yang tepat. Uji coba luas menghasilkan data empiris yang dianalisis berdasarkan jawaban peserta didik dengan menggunakan pedoman penskoran menurut Gurel et al. (2015) dan Istiyono et al. (2014) pada Tabel 2.

Tabel 2. Penskoran *four-tier test*

Jawaban	Tingkat Kepercayaan	Alasan	Tingkat Kepercayaan	Kategori	Skor
Benar	Yakin	Benar	Yakin	*Scientific Correct*	7
Benar	Yakin	Benar	Tidak Yakin	*Lack of Knowledge type 1*	6
Benar	Tidak Yakin	Benar	Yakin	*Lack of Knowledge type 1*	6
Benar	Tidak Yakin	Benar	Tidak Yakin	*Lack of Knowledge type 1*	6
Benar	Yakin	Salah	Yakin	*False Positive*	5
Benar	Yakin	Salah	Tidak Yakin	*Lack of Knowledge type 2*	4
Benar	Tidak Yakin	Salah	Yakin	*Lack of Knowledge type 2*	4
Benar	Tidak Yakin	Salah	Tidak Yakin	*Lack of Knowledge type 2*	4
Salah	Yakin	Benar	Yakin	*False Negative*	3
Salah	Yakin	Benar	Tidak Yakin	*Lack of Knowledge type 3*	2
Salah	Tidak Yakin	Benar	Yakin	*Lack of Knowledge type 3*	2
Salah	Tidak Yakin	Benar	Tidak Yakin	*Lack of Knowledge type 3*	2
Salah	Yakin	Salah	Yakin	*Misconception*	1
Salah	Yakin	Salah	Tidak Yakin	*Lack of Knowledge type 4*	0
Salah	Tidak Yakin	Salah	Yakin	*Lack of Knowledge type 4*	0
Salah	Tidak Yakin	Salah	Tidak Yakin	*Lack of Knowledge type 4*	0

(Istiyono et al., 2023)

Hasil analisis uji produk berupa validitas, reliabilitas, dan tingkat kesukaran butir yang menjadi kriteria kualitas instrumen yang dikembangkan. Analisis uji coba produk menggunakan model PCM 1-PL (Rasch) berbantuan program Winsteps. Validitas konstruk berupa kecocokan butir soal (*item fit*) dengan model Rasch berdasarkan nilai *Infit* MNSQ pada rentang 0,77 hingga 1,30 (Adams & Kho., 1996). Reliabilitas digunakan untuk mengukur konsistensi instrumen tes dalam mengukur suatu variabel yang sama. Kategori reliabilitas instrumen ditetapkan berdasarkan kriteria pengujian reliabilitas pada Tabel 3.

Tabel 3. Kriteria reliabilitas

Koefisien Reliabilitas	Kategori
0,80 - 1,00	Sangat Tinggi
0,60 – 0,80	Tinggi
0,40 – 0,60	Cukup
0,20 – 0,40	Rendah
-1,00 – 0,20	Sangat Rendah

(Suparman, 2020)

Tingkat kesulitan butir diukur untuk menunjukkan indeks kesukaran pada tiap butir soal berdasarkan jumlah peserta didik yang menjawab butir soal dengan benar. Suatu butir dianggap baik jika berada pada rentang $-2{,}00 < b < +2{,}00$ (Hambleton & Swaminathan, 1985). Kategori tingkat kesukaran butir soal ditetapkan berdasarkan kriteria tingkat kesukaran pada Tabel 4.

Tabel 4. Kriteria tingkat kesukaran

Nilai *Threshold* (b)	Kategori
$b > 2$	Sangat Sukar
$1 < b \leq 2$	Sukar
$-1 \leq b \leq 1$	Sedang
$-2 \leq b \leq -1$	Mudah
$b > -2$	Sangat Mudah

(Setyawarno, 2017)

Butir soal yang kurang memenuhi standar kriteria yang telah ditetapkan akan direvisi dan disusun ulang menjadi satu instrumen yang utuh untuk digunakan dalam tahap pengukuran. Tahap pengukuran dilakukan untuk mengetahui profil kemampuan pemahaman konsep peserta didik melalui nilai *person measure* pada *output* Winsteps. Berdasarkan panduan penggunaan Winsteps oleh Linacre (2026), nilai θ perlu dikonversi menjadi T-*score* dengan menggunakan persamaan.

$$T = 10\ (measure) + 50$$

Keterangan:
T = *T-score*
Measure = nilai *person measure* (*theta*) dari *output* analisis IRT
Nilai T yang diperoleh kemudian dikategorikan berdasarkan kriteria kemampuan yang melibatkan nilai rata-rata ideal dan standar deviasi ideal. Kategori kemampuan pemahaman konsep ditetapkan berdasarkan kriteria pengukuran pada Tabel 5.

Tabel 5. Kriteria pengukuran kemampuan

Rentang Nilai	Kategori Kemampuan
skor ≥ Mean + SDi	Tinggi
Mean - SDi ≤ skor < Mean + SDi	Sedang
skor ≤ Mean - SDi	Rendah

(Rahman, 2018)

Dalam tahap pengukuran, penilaian kepraktisan instrumen oleh guru dan peserta didik dilakukan untuk meninjau kemudahan penggunaan serta kebermanfaatan produk. Hasil angket kepraktisan produk oleh guru dan peserta didik dianalisis menjadi bentuk persentase.

Hasil analisis diinterpretasikan berdasarkan kriteia kepraktisan pada Tabel 6.

Tabel 6. Kriteria kepraktisan

Interval Kepraktisan	Kategori
81% - 100%	Sangat Praktis
61% - 80%	Praktis
41% - 60%	Sedang
21% - 40%	Kurang Praktis
0% - 20%	Tidak Praktis

(Akbar, 2013)

Interpretasi dilakukan berdasarkan hasil pengukuran untuk mengetahui profil kemampuan pemahaman konsep peserta didik. Kepraktisan produk dinilai dari segi penggunaan, penyajian, dan kebermanfaatan berdasarkan hasil angket respons guru dan peserta didik.

HASIL DAN PEMBAHASAN

Hasil

Instrumen dirancang berdasarkan sub materi momentum dan impuls, yaitu konsep momentum dan impuls, hukum kekekalan momentum, dan jenis tumbukan dengan persebaran level kognitif C2 dan C3. Instrumen juga dirancang berdasarkan matriks dan kisi-kisi yang telah disesuaikan dengan tujuan, kompetensi, dan materi tes. Matriks memuat aspek pemahaman konsep, indikator pemahaman konsep, serta nomor item. Aspek dan indikator yang digunakan merupakan hasil sintesa aspek pemahaman konsep dari peneliti terdahulu. Matriks yang menjadi acuan dalam penyusunan butir soal tercantum pada Tabel 7.

Tabel 7. Matriks instrumen penilaian

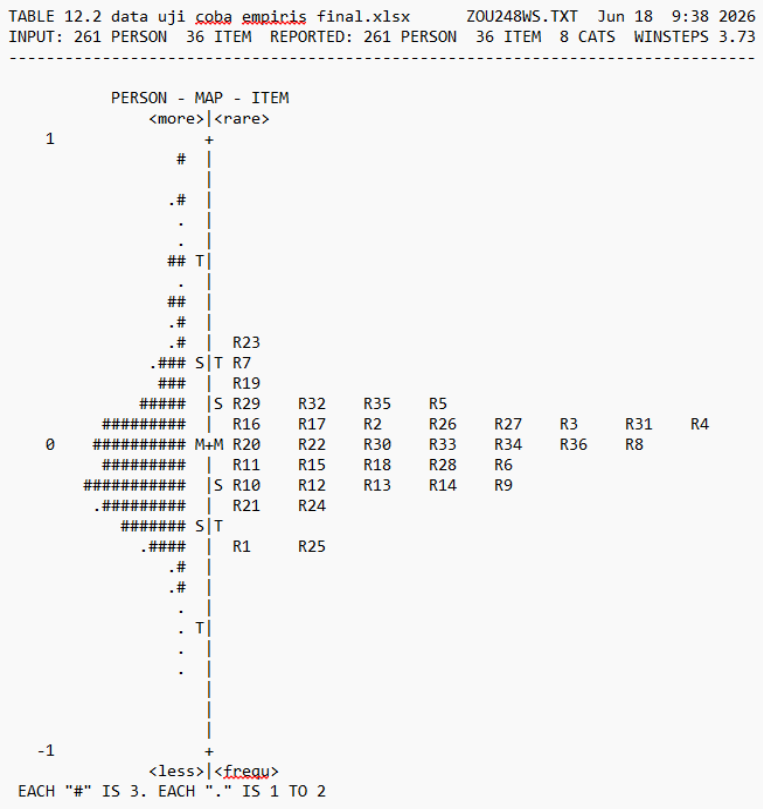
No.	Aspek Pemahaman Konsep	Indikator Pemahaman Konsep	Nomor Item
1.	Menafsirkan	Peserta didik manpu menafsirkan permasalahan dengan tepat	1a, 1b 2a, 2b 3a, 3b 4a, 4b
2.	Mengklasifikasikan	Peserta didik mampu mengklasifikasikan permasalahan sesuai dengan konsep fisika	5a, 5b 6a, 6b 7a, 7b 8c*
3.	Membedakan	Peserta didik mampu membedakan konsep yang serupa tetapi berbeda	9a, 9b 10c* 11a, 11b 12a, 12b
4.	Menarik inferensi	Peserta didik mampu menginterpretasikan konsep fisika dengan benar	13a, 13b 14a, 14b 15c* 16a, 16b
5.	Menerapkan	Peserta didik mampu mempraktikkan konsep fisika dalam peristiwa dengan benar	17a, 17b 18a, 18b 19a, 19b 20c*

Bentuk instrumen yang dihasilkan berupa *four-tier multiple choice* yang dikemas dalam dua paket soal dengan masing-masing paket berisi 20 butir soal dengan 4 butir bersama. Instrumen dilakukan validasi isi butir soal oleh 7 *expert judgement* yang menghasilkan rata-rata indeks *V* Aiken senilai 0,89 yang termasuk kategori validitas valid dengan beberapa saran perbaikan dari validator. Hasil revisi kemudian digunakan untuk uji coba luas pada 261 peserta didik kelas XI di lima SMA Sleman yang memiliki karakteristik mendekati subjek pengukuran. Hasil analisis uji coba luas menunjukkan kualitas instrumen dalam mengukur kemampuan pemahaman konsep peserta didik. Berdasarkan nilai *infit* MNSQ pada kecocokan butir dengan model, terdapat 6 butir soal tidak cocok (*misfit*) yang tercantum pada Tabel 8.

Tabel 8. Hasil analisis kecocokan butir

Butir Soal	Nomor soal	*Infit* MNSQ	Keterangan
1	1a	1,47	*Misfit*
8	8c	0,72	*Misfit*
17	17a	0,74	*Misfit*
23	3b	1,47	*Misfit*
26	6b	1,47	*Misfit*
32	14b	0,76	*Misfit*

Kualitas instrumen ditunjukkan pada hasil analisis reliabilitas dengan nilai 0,92 yang berada pada kategori sangat baik. Tingkat kesulitan seluruh butir pada instrumen yang dikembangkan termasuk dalam kategori sedang dengan peta distribusi pada Gambar 1.



Gambar 1. Peta distribusi kesukaran butir

Namun, terdapat butir soal bias (DIF) gender dengan nilai probabilitas < 0,05 yang menunjukkan terdapat perbedaan peluang untuk menjawab butir soal dengan benar antara

peserta didik laki-laki dan peserta didik perempuan. Butir soal tersebut tercantum pada Tabel 9.

Tabel 9. Hasil analisis bias butir

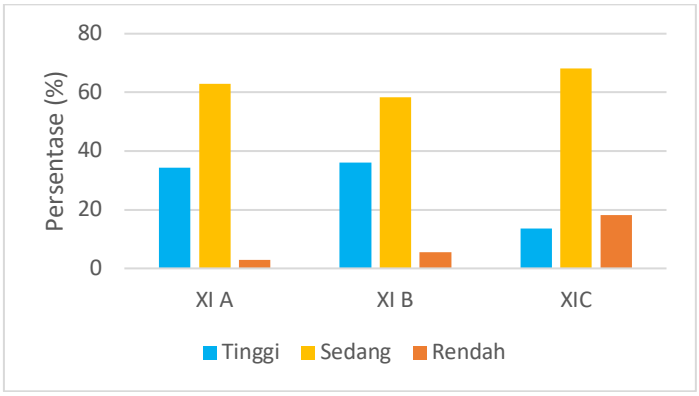
Butir Soal	Probabilitas	Keterangan
16a	0,0205	DIF
19b	0,0481	DIF

Tahap pengukuran dilakukan untuk menguji kepraktisan produk dan untuk mengetahui profil kemampuan pemahaman konsep peserta didik dilakukan dengan menggunakan instrumen tes yang telah disusun kembali berdasarkan perbaikan butir soal yang belum memenuhi kriteria kualitas instrumen tes. Pengukuran dilakukan pada 93 peserta didik kelas XI SMA Wates yang menghasilkan kemampuan pemahaman konsep yang beragam. Sebagian besar kemampuan pemahaman konsep termasuk dalam kategori sedang dengan rincian pada Tabel 10.

Tabel 10. Persentase kemampuan pemahaman konsep

Kategori	Skor	Jumlah Peserta Didik	Persentase
Tinggi	$X \geq 51{,}93$	28	30,11%
Sedang	$48{,}16 \leq X < 51{,}93$	58	62,36%
Rendah	$X \leq 48{,}16$	9	7,53%

Kemampuan pemahaman konsep pada Tabel 9 merupakan persebaran secara keseluruhan. Kemampuan pemahaman konsep pada setiap kelas juga ditunjukkan pada Gambar 2.



Gambar 2. Diagram persebaran kemampuan tiap kelas

Kepraktisan produk dilakukan dengan menggunakan angket respons oleh 93 peserta didik dan 3 guru setelah produk digunakan. Aspek kepraktisan yang dinilai meliputi kemudahan penggunaan, penyajian, dan kebermanfaatan produk. Hasil analisis kepraktisan respons peserta didik dan guru menunjukkan persentase dalam kategori sangat praktis yang dirinci pada Tabel 11.

Tabel 11. Hasil analisis kepraktisan produk

Aspek	Peserta Didik		Guru	
	Persentase	Kategori	Persentase	Kategori
Kemudahan penggunaan	86,56%	Sangat Praktis	91,67%	Sangat Praktis
Penyajian	82,69%	Sangat Praktis	95,00%	Sangat Praktis
Kebermanfaatan	87,85%	Sangat Praktis	98,61%	Sangat Praktis

Pembahasan

Perancangan instrumen tes untuk mengukur kemampuan peserta didik kelas XI SMA materi momentum impuls menghasilkan instrumen penilaian berbentuk pilihan ganda empat tingkat atau *four-tier multiple choice* dengan 36 butir soal. Butir soal dikemas dalam dua paket soal, yaitu paket A dan paket B dengan 20 butir soal termasuk 4 butir bersama pada masing-masing paket. Butir-butir tersebut disusun berdasarkan aspek dan indikator pemahaman konsep peserta didik yang kemudian divalidasi oleh 5 guru dan 2 teman sejawat untuk mengetahui validitas isi dari butir soal. Hasil validasi pada aspek materi, konstruksi, dan bahasa menyatakan bahwa 35 butir soal valid dan 1 butir soal termasuk kategori sedang menurut Retnawati (2016) pada Tabel 1. Butir soal yang memiliki validitas sedang akan direvisi berdasarkan saran dan komentar dari validator. Hampir seluruh validator memberi masukan mengenai penulisan kata yang salah, keutuhan soal, serta perbaikan kunci jawaban pada butir soal nomor 19. Namun demikian, rata-rata indeks *Aiken's V* menunjukkan nilai yang termasuk dalam kategori valid, yaitu di atas 0,80.

Setelah instrumen diperbaiki dan dirakit kembali, instrumen tes digunakan pada tahap uji coba luas untuk mendapatkan data empiris yang dijadikan bahan untuk mengetahui kualitas instrumen tes. Sebelum dilakukan analisis data, jawaban peserta didik akan diberi skor sesuai dengan pedoman penskoran pada Tabel 2. Kualitas instrumen tes diuji dengan menggunakan program Winsteps untuk mengukur validitas, reliabilitas, serta tingkat kesukaran pada tiap butir soal. Berdasarkan Tabel 8, nilai *infit* MNSQ butir soal yang dihasilkan berada pada rentang 0,72 hingga 1,47. Hal tersebut tidak sesuai dengan kriteria menurut Adams & Kho (1996) yang mengatakan bahwa kategori butir termasuk *fit* apabila nilai *infit* MNSQ berada pada rentang 0,77 hingga 1,30. Rata-rata *infit* MNSQ senilai 1,02 ± 0,19 dengan 36 butir soal termasuk 4 butir bersama. Kualitas instrumen tes diuji berdasarkan kestabilan produk dalam mengukur kemampuan pemahaman konsep ketika dilakukan berulang kali. Instrumen tes *four-tier multiple choice* digunakan kepada 261 peserta didik dari lima Sekolah Menengah Atas yang ada di Kabupaten Sleman. Perolehan angka 0,92 pada nilai *item reliability* dari *output* Winsteps menunjukkan instrumen tes memiliki reliabilitas yang sangat baik berdasarkan Tabel 3.

Butir-butir soal pada instrumen tes memiliki tingkat kesukaran sedang berdasarkan Tabel 4. Tingkat kesukaran sedang dianggap paling ideal karena menunjukkan efektivitas daya pembeda yang tinggi mengenai kemampuan subjek. Gambar 1 merupakan visualisasi yang mengindikasikan adanya kesesuaian antara tingkat kesukaran butir soal dengan kemampuan subjek. Pada Gambar 1, butir soal yang berada paling atas merupakan item paling sulit, yaitu butir soal nomor 3 paket B dengan nilai *measure* 0,31. Namun, pada peta tersebut terdapat beberapa subjek yang memiliki kapabilitas lebih sehingga dapat menjawab butir soal tersulit dengan benar. Subjek tersebut adalah peserta didik dengan kemampuan pemahaman konsep tinggi. Butir soal paling mudah, yaitu nomor 5 paket B dengan nilai *measure* -0,34. Pada peta distribusi terdapat pula beberapa subjek dengan kemampuan dibawah item termudah, sehingga instrumen tes kurang sesuai digunakan untuk mengukur kemampuan peserta didik pada level tersebut.

Analisis DIF dilakukan untuk mendeteksi bias pengukuran secara objektif sehingga dapat mengidentifikasi kelompok responden yang mendapatkan keuntungan. Dalam penelitian ini, bias yang diukur berupa gender subjek penelitian. Hal tersebut dianalisis berdasarkan nilai probabilitas < 0,05 menunjukkan terdapat peluang perbedaan antara repsonden dalam menjawab butir soal dengan butir soal. Berdasarkan Tabel 9, terdapat dua butir soal yang memiliki nilai probabilitas < 0,05 yang berarti butir soal tersebut menguntungkan salah satu kelompok. Hasil analisis menunjukkan bahwa kedua butir soal tersebut menguntungkan kelompok laki-laki dengan DIF *contrast* yang cukup tinggi dibandingkan dengan butir soal lain. Bagi peserta didik laki-laki, butir soal nomor 16 paket A dan 19 paket B mudah untuk dikerjakan sehingga berpeluang tinggi untuk menjawab dengan benar.

Hasil analisis kualitas butir soal yang kurang memenuhi standar perlu direvisi untuk digunakan sebagai alat ukur dalam tahap pengukuran. Jawaban peserta didik pada pengukuran dikoreksi dan diberi skor sesuai dengan pedoman penskoran pada Tabel 2. Data diolah menggunakan program Winsteps untuk mengetahui kemampuan pemahaman konsep peserta didik melalui nilai *person measure*. Analisis pemahaman konsep menggunakan nilai theta yang dikonversi menjadi T *score*. Penentuan kategori kemampuan pemahaman konsep peserta didik dilakukan berdasarkan nilai rata-rata ideal dan simpangan baku ideal yang menghasilkan nilai mean ideal 50,05 dan SDi 1,88. Secara keseluruhan, rata-rata kemampuan pemahaman konsep subjek pengukuran adalah 50,64 ± 0,20 yang dikategorikan berkemampuan sedang dengan persentase 62,36% seperti pada Tabel 10. Kelompok kelas yang berkemampuan tinggi paling banyak adalah kela XI B, sedangkan kelas XI C merupakan kelompok kelas dengan persentase paling tinggi untuk pemahaman konsep kategori sedang dan kategori rendah. Peserta didik yang berkemampuan tinggi telah menguasai seluruh indikator pemahaman konsep yang digunakan dalam instrumen tes. Sedangkan peserta didik yang berkemampuan sedang hanya menguasai dua atau tiga indikator saja, dan yang berkemampuan rendah kemungkinan hanya menguasai sebagian kecil dari indikator pemahaman konsep.

SIMPULAN

Instrumen tes yang dihasilkan memiliki indeks validitas yang baik sehingga instrumen dapat digunakan untuk uji coba tes setelah direvisi. Kualitas instrumen diidentifikasi menggunakan data uji coba tes. Terdapat enam butir soal yang tidak cocok dengan model Rasch, sehingga mempengaruhi kualitas instrumen. Namun demikian, terdapat 30 butir soal yang *fit* dengan model sehingga dapat digunakan untuk mengukur kemampuan pemahaman konsep. Secara keseluruhan, instrumen tes dapat digunakan berulang kali untuk mengukur kemampuan pemahaman konsep karena memiliki nilai reliabilitas yang tinggi. Seluruh butir soal memiliki kesukaran butir dalam kategori sedang. Namun, terdapat dua butir soal yang bias gender kepada laki-laki. Berdasarkan kualitas instrumen tersebut, butir soal yang kurang sesuai dengan kriteria diupayakan untuk diperbaiki dan disusun ulang untuk menjadi alat ukur dalam pengukuran. Profil kemampuan pemahaman konsep peserta didik sebagian besar tergolong dalam kategori sedang dengan persentase 62,36%. Hasil tersebut dapat dijadikan sebagai acuan untuk memperbaiki pembelajaran selanjutnya. Selain itu, peserta didik diharapkan dapat melatih kemampuan pemahaman konsep melalui instrumen tes ini sehingga dapat meningkatkan peringkat Indonesia pada PISA. Instrumen tes ini dapat dikembangkan untuk mengidentifikasi miskonsepsi secara mendalam. Instrumen tes ini juga dapat dikembangkan menggunakan aspek dan indikator pemahaman konsep pada level kognitif yang lebih tinggi.

UCAPAN TERIMA KASIH

Terima kasih diucapkan kepada validator yang telah berkenan menilai dan mengarahkan demi kualitas instrumen tes yang dikembangkan, kepada subjek penelitian yang bersedia menjadi responden penelitian secara sungguh-sunggu, kepada pembimbing yang selalu memberikan dukungan dan arahan dalam penelitian, serta pihak lainnya yang telah membantu dalam pelaksanaan penelitian.

DAFTAR PUSTAKA

Adams, R. J., & Kho, S. T. (1996). *Acer Quest Version 2.1.* Camberwell, Victoria: The Australian Council for Educational Research.

Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement,* 45(1), 131-142. https://doi.org/10.1177/0013164485451012.

Akbar, S. (2013). *Instrumen Perangkat Pembelajaran.* Remaja Posakarya.

Alawiyah, N. S., Ngadimin, & Hamid, A. (2017). Identifikasi Miskonsepsi Siswa dengan Menggunakan Metode Indeks Respon Kepastian (IRK) pada Materi Impuls dan Momentum Linear di SMA Negeri 2 Banda Aceh. *Jurnal Ilmiah Mahasiswa (JIM) Pendidikan Fisika,* 2(2), 272-276.

Ananda, M. F. R., Maison, & Furqon, M. (2026). Analisis Miskonsepsi Siswa pada Materi Momentum dan Impuls Menggunakan Instrumen Four-Tier Diagnostic Test Berbasis Multirepresentasi. *OPTIKA: Jurnal Pendidikan Fisika,* 10(1), 156-169. https://doi.org/10.37478/optika.v10i1.7809.

Asmin, L. O., & Rosdianti. (2021). Analisis Miskonsepsi Siswa SMA Negeri 04 Bombana dengan Menggunakan CRI pada Konsep Suhu dan Kalor. *Jurnal Fisika dan Pendidikan Fisika,* 6(2), 80-87. https://doi.org/10.20414/konstan.v6i2.100.

Azizah, Z., Taqwa, M. R. A., & Assalam, I. T. (2020). Analisis Pemahaman Konsep Fisika Peserta Didik Menggunakan Instrumen Berbantuan Quizizz. *EduSains: Jurnal Pendidikan Sains & Matematika,* 8(2), 1-11.

Didik, L. A., Wahyudi, M., & Kafrawi, M. (2020). Identifikasi Miskonsepsi dan Tingkat Pemahaman Mahasiswa Tadris Fisika pada Materi Listrik Dinamis Menggunakan 3-Tier Diagnostic Test. *Journal of Natural Science and Integration,* 3(2), 128-137.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Kluwer-Nijhoff Publishing.

Hidayat, R., Hakim, L., & Lia, L. (2019). Pengaruh Model Guided Discovery Learning Berbantuan Media Simulasi PhET Terhadap Pemahaman Konsep Fisika Siswa. *Berkala Ilmiah Pendidikan Fisika,* 7(2), 97-104.

Imania, K. A. N., & Bariah, S. K. (2019). Rancangan Pengembangan Instrumen Penilaian Pembelajaran Berbasis Daring. *Jurnal PETIK,* 5(1), 31-47

Istiyono, E., Dwandaru, W. S. B., Fenditasari, K., Ayub, M. R. S. S. N., & Saepuzaman, D. (2023). The Development of a Four-Tier Diagnostic Test Based on Modern Test Theory in Physics Education. *European Journal of Educational Research,* 12(1), 371-285. https://doi.org/10.12973/eu-jer.12.1.371.

Istiyono, E., Mardapi, D., & Suparno. (2014). Pengembangan Tes Kemampuan Berpikir Tingkat Tinggi Fisika (PysTHOTS) Peserta Didik SMA. *Jurnal Penelitian dan Evaluasi Pendidikan,* 18(1), 1-12. https://doi.org/10.21831/pep.v18i1.2120.

Linacre, J. M. (2026). *Winsteps Rasch measurement computer program User's Guide.* From https://www.winsteps.com/a/Winsteps-Manual.pdf

OECD (Organization for Economic Co-operation and Development). (2023). *PISA 2022 Result (Volume I and II) – Country Notes: Indonesia.* https://www.oecd.org/en/publications/pisa-2022-results-volume-i-and-ii-country-notes_ed6fbcc5-en/indonesia_c2e1ae0e-en.html

Paramita, U. V., & Jauhariyah, M. N. R. (2024). Identifying Student Misconceptions on Momentum and Impulse Using Four-Tier Diagnostic Test Instrument with CRI. *Jurnal Pijar MIPA,* 19(1), 75-82. https://doi.org/10.29303/jpm.v19i1.6086.

Pratama, R. (2023). Pengembangan Tes Diagnostik Four-Tier Multiple Choice Pendeteksi Pemahaman Konsep Materi Sistem Koordinasi. *Journal on Education,* 5(3), 7012-7019. http://jonedu.org/index.php/joe.

Rahman, A. (2018). Hubungan Motivasi dan Sikap Siswa terhadap Prestasi Belajar Matematika Kelas VIII SMP Negeri 1 Wera. *Jurnal Pendidikan MIPA LPPM STKIP Taman Siswa Bima,* 8(1), 60-66.

Retnawati, H. (2016). *Analisis Kuantitatif Instrumen Penelitian.* Parama Publishing.

Roistiya, H., Putra, I. A., & Pertiwi, N. A. S. (2019). Pengembangan Instrumen MW4T (Mechanic Wave Four Tier) Diagnostic Test untuk Mengukur Pemahaman Konsep Gelombang Mekanik. *DIFFRACTION: Journal of Physics Education and Applied Physics,* 1(2), 14-21.

Sakinah, N., Amalia, I. F., & Adimayuda, R. (2023). Identifying Senior Higgh School Students' Misconception on The Momentum and Impulse Concepts Uses Three-Tier Diagnostic Test. *Research in Physics Education,* 2(1), 18-25. https://doi.org/10.31980/ripe.v2i1.26.

Setyawarno, D. (2017). *Analisis Data Pengukuran Menggunakan Program Quest.* FMIPA UNY.

Suparman. (2020). Menemukan Karakteristik Butir Menggunakan Quest. *Jurnal Komunikasi dan Pendidikan Islam,* 9(1), 83-104.

Susanti, N. K. E., Asrin, & Khair, B. N. (2021). Analisis Tingkat Pemahaman Konsep IPA Siswa Kelas V SDN Gugus V Kecamatan Cakranegara. *Jurnal Ilmiah Profesi Pendidikan,* 6(4), 686-609. https://doi.org/10.29303/jipp.v6i4.317.

Yasthophi, A., & Ritongga, P. S. (2019). Pengembangan Instrumen Test Diagnostik Multiple Choice Four Tier pada Materi Ikatan Kimia. *Jurnal Konfigurasi,* 3(1), 23-31.

Yunita, A., & Nasution, H. I. (2025). Pengembangan Instrumen Diagnostik Four-Tier untuk Mengetahui Tingkat Pemahaman Konsep Siswa pada Pokok Bahasan Laju Reaksi. *Jurnal Pendidikan Kimia Universitas Riau,* 10(1), 1-7. http://dx.doi.org/10.33578/jpk-unri.v7i1.xx.