

Analysis of the Quality of Junior High School Science Question Items Using the Test Analysis Program (TAP) Application for Motion and Force Material

[Analisis Kualitas Butir Soal IPA SMP dengan Aplikasi Test Analysis Program (TAP) Materi Gerak dan Gaya]

Salma Ramadhina Fatahillah¹⁾, Nur Efendi²⁾

Program Studi Pendidikan IPA, Universitas Muhammadiyah Sidoarjo, Indonesia

Program Studi Pendidikan IPA, Universitas Muhammadiyah Sidoarjo, Indonesia

Penulis Korespondensi: nur.efendi@umsida.ac.id

Abstract. *This study aims to analyze the quality of multiple-choice test items in Grade VII science subject on “Motion and Force” based on parameters of validity, reliability, difficulty level, discriminating power, distractor effectiveness, and their relationship with Bloom's Taxonomy cognitive domains. This research employed a quantitative descriptive-analytic approach. A total of 40 test items and student responses from 52 participants were analyzed using Test Analysis Program (TAP) version 14.7.4. The results showed that 95% of the items were valid ($r \geq 0.30$) with very high reliability ($KR-20 = 0.929$ and $split-half = 0.925$). The difficulty level was dominated by easy questions (82.5%), and most items had high discriminating power (60%). About 90% of the distractors were effective. The cognitive level distribution was dominated by level C4 (analyzing), with very few items at level C5 (evaluating) and none at level C6 (creating). Inferential analysis indicated a positive correlation between discriminating power and validity. These findings suggest that the analyzed items meet most of the quality standards. However, improvements are needed in terms of difficulty level variation and integration of higher-order thinking skills (HOTS). The use of technology such as TAP should be complemented with pedagogical development aligned with 21st-century educational needs.*

Keywords - TAP Application, Motion and Force, Science, Question Item Quality, Junior High School

Abstrak. *Penelitian ini bertujuan menganalisis kualitas butir soal pilihan ganda IPA kelas VII materi “Gerak dan Gaya” berdasarkan validitas, reliabilitas, tingkat kesukaran, daya pembeda, efektivitas pengecoh, dan ranah kognitif Taksonomi Bloom. Penelitian menggunakan pendekatan kuantitatif deskriptif-analitik dengan 40 soal dan jawaban 52 siswa yang dianalisis melalui aplikasi Test Analysis Program (TAP) versi 14.7.4. Hasil menunjukkan 95% soal valid ($r \geq 0,30$), reliabilitas sangat tinggi ($KR-20 = 0,929$; $split-half = 0,925$), tingkat kesukaran didominasi mudah (82,5%), daya pembeda mayoritas tinggi (60%), dan 90% pengecoh efektif. Distribusi kognitif lebih banyak pada C4 (analisis), sedikit pada C5 (evaluasi), dan tidak ada pada C6 (mencipta). Analisis inferensial mengonfirmasi adanya korelasi positif antara daya pembeda dan validitas. Secara umum, soal memenuhi kriteria baik, meski perlu peningkatan pada variasi kesukaran serta integrasi keterampilan berpikir tingkat tinggi (HOTS) untuk mendukung kebutuhan pendidikan abad ke-21.*

Kata Kunci - Aplikasi TAP, Gerak dan Gaya, IPA, Kualitas Butir Soal, SMP

I. PENDAHULUAN

Revolusi Industri 4.0 telah memberikan pengaruh signifikan pada berbagai aspek kehidupan manusia, termasuk dalam sektor pendidikan. Dalam era ini ditandai oleh munculnya teknologi mutakhir seperti kecerdasan buatan (AI), *big data*, *Internet of Things* (IoT), dan otomatisasi, yang menuntut dunia pendidikan untuk melakukan transformasi secara drastis [1], [2]. Penilaian sebagai salah satu elemen penting dalam proses pendidikan dituntut mampu beradaptasi agar dapat mengukur kompetensi yang relevan dengan kebutuhan abad ke-21 [3], [4]. Pendidikan abad ini menekankan penguasaan keterampilan berpikir kritis, kreativitas, kolaborasi, dan komunikasi yang terintegrasi dalam konsep 4C, serta kemampuan berpikir tingkat tinggi atau *Higher Order Thinking Skills* (HOTS) [5]. Oleh karena itu, dibutuhkan instrumen evaluasi misalnya, soal yang berkualitas agar penilaian benar-benar dapat merefleksikan kemampuan peserta didik secara objektif.

Kualitas butir soal menjadi faktor utama dalam mewujudkan penilaian yang bermakna. Butir soal yang berkualitas ditandai oleh terpenuhinya beberapa kriteria, yaitu validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh [6]. Validitas memastikan soal mengukur aspek yang sesuai dengan tujuan pembelajaran, reliabilitas menunjukkan konsistensi hasil tes, tingkat kesukaran menunjukkan proporsi peserta didik yang dapat menjawab dengan benar, daya pembeda menunjukkan kemampuan soal dalam membedakan peserta didik yang

berkemampuan tinggi dan rendah, serta efektivitas pengecoh menunjukkan apakah pilihan jawaban salah berfungsi sebagaimana mestinya [7], [8]. Selain itu, kualitas soal juga perlu memperhatikan tingkat ranah kognitif dalam Taksonomi Bloom (revisi) yang dimana mencakup enam level, mulai dari C1 (mengingat), C2 (memahami), C3 (menerapkan), C4 (menganalisis), C5 (mengevaluasi), dan C6 (mencipta) [9], [10]. Namun demikian, praktik di lapangan menunjukkan banyak soal masih terfokus pada tingkat kognitif rendah (C1 dan C2), sehingga belum sepenuhnya mendukung pengembangan HOTS peserta didik [11], [12]. Soal-soal diduga belum dianalisis kualitasnya sebagai bentuk penjaminan mutu pembelajaran belum optimal.

Dalam rangka menjamin mutu instrumen penilaian, guru memerlukan alat bantu yang dapat melakukan analisis soal secara objektif, efisien, dan komprehensif. Salah satu perangkat lunak yang dapat digunakan adalah Test Analysis Program (TAP) [13], [14]. TAP merupakan aplikasi yang dirancang untuk menganalisis butir soal pilihan ganda berdasarkan kriteria validitas (point-biserial), reliabilitas (KR-20, split-half), tingkat kesukaran, daya pembeda, dan efektivitas pengecoh [15]. Hasil analisis TAP lebih cepat, terukur, dan praktis dibandingkan analisis manual, serta mampu menampilkan data yang mudah dipahami guru. Dengan demikian, penggunaan TAP sangat relevan untuk meningkatkan kualitas evaluasi pembelajaran, terutama di tingkat Sekolah Menengah Pertama (SMP) [16], [17].

Selain itu, TAP memiliki beberapa kelebihan penting dibandingkan metode analisis tradisional. Pertama, TAP dapat memproses data secara otomatis dengan tingkat akurasi tinggi sehingga mengurangi potensi kesalahan manusia dalam perhitungan statistik. Kedua, hasil analisis yang dihasilkan TAP bersifat komprehensif, karena menyajikan berbagai parameter kualitas butir soal sekaligus dalam bentuk tabel maupun grafik yang mudah dipahami. Ketiga, TAP membantu guru dalam menghemat waktu dan tenaga, sehingga guru dapat lebih fokus pada tindak lanjut perbaikan soal maupun strategi pembelajaran. Keempat, aplikasi ini mendukung prinsip asesmen modern karena selaras dengan tuntutan efisiensi dan berbasis data dalam pendidikan abad ke-21. Dengan berbagai keunggulan tersebut, TAP layak dijadikan instrumen standar dalam evaluasi kualitas soal. Namun, pemanfaatan TAP oleh guru masih belum optimal, sehingga kualitas soal sering kali tidak dianalisis secara mendalam.

Berdasarkan hasil observasi awal di salah satu SMP Swasta di Kabupaten Sidoarjo, guru IPA belum melakukan analisis butir soal secara kuantitatif menggunakan aplikasi *Test Analysis Program* (TAP). Setiap akhir materi, guru menyusun soal ujian berbasis kisi-kisi kurikulum, namun soal tersebut jarang dianalisis secara kuantitatif. Dari wawancara dengan guru, diketahui bahwa soal-soal yang digunakan sebagian besar berasal dari bank soal lama dan soal dari internet, tanpa pengujian mendalam terkait validitas maupun reliabilitas. Fakta menunjukkan bahwa **soal** ujian akhir bab masih berada pada level kognitif C1–C2, sehingga kurang menantang peserta didik untuk berpikir kritis. Selain itu, hasil rekap nilai ulangan harian dan ujian semester ganjil tahun ajaran 2024/2025 menunjukkan bahwa rata-rata ketuntasan belajar peserta didik di bawah Kriteria Ketuntasan Minimal (KKM) 75. Hal ini memperkuat bukti bahwa kualitas instrumen evaluasi masih perlu ditingkatkan [18], [19]. Salah satu materi IPA yang dianggap sulit peserta didik ialah materi Gerak dan Gaya.

Materi Gerak dan Gaya dipilih sebagai fokus penelitian ini karena memiliki peran fundamental dalam pembelajaran IPA Fisika kelas VII. Materi ini merupakan kompetensi dasar yang menjadi landasan bagi pemahaman konsep-konsep fisika lanjutan, seperti Hukum Newton, energi, dan momentum. Namun, hasil evaluasi awal menunjukkan bahwa banyak peserta didik mengalami kesulitan dalam memahami materi ini, yang tercermin dari rendahnya ketuntasan belajar masih dibawah Kriteria Ketuntasan Minimal (KKM). Kondisi tersebut menunjukkan bahwa instrumen evaluasi yang digunakan belum sepenuhnya mampu mengukur capaian kompetensi dengan akurat. Oleh karena itu, dibutuhkan instrumen evaluasi yang berkualitas untuk memastikan bahwa capaian kompetensi peserta didik dalam materi “Gerak dan Gaya” dapat diukur dengan tepat [20].

Pentingnya kualitas butir soal yang dianalisis terletak pada kemampuannya menjamin mutu pembelajaran, sebab setiap soal harus benar-benar mengukur apa yang semestinya diukur sesuai indikator pembelajaran. Analisis butir soal yang sistematis memungkinkan guru memperoleh gambaran objektif mengenai validitas, reliabilitas, tingkat kesukaran, daya pembeda, serta efektivitas pengecoh dari setiap item yang digunakan. Hasil analisis tersebut dapat menjadi dasar bagi pengembangan instrumen evaluasi yang lebih akurat, sehingga mampu memberikan informasi yang dapat ditindaklanjuti dalam proses perbaikan pembelajaran. Dengan demikian, penelitian ini diharapkan berkontribusi terhadap peningkatan kualitas asesmen di sekolah, mendorong terciptanya praktik evaluasi yang lebih bermakna, serta mendukung pengembangan keterampilan berpikir tingkat tinggi (HOTS) yang esensial bagi peserta didik di abad ke-21..

II. METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan jenis deskriptif-analitik. Pendekatan ini dipilih karena sesuai untuk menganalisis data numerik dari hasil tes serta mendeskripsikan kualitas butir soal berdasarkan analisis statistik [21]. Analisis deskriptif diterapkan untuk mengevaluasi karakteristik soal seperti validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh, sedangkan analisis inferensial diterapkan untuk menguji

hubungan antar variabel. Melalui pendekatan ini, peneliti bisa memahami secara utuh, baik dari segi deskripsi maupun hubungan antar indikator kualitas soal, sehingga hasil analisis menjadi lebih mendalam dan menyeluruh.

Data dalam penelitian ini adalah data primer berupa dokumen soal pilihan ganda IPA kelas VII dan lembar jawaban peserta didik yang diperoleh dari salah satu SMP Swasta di Kabupaten Sidoarjo. Data dikumpulkan pada tanggal 27 Mei 2025 dan terdiri atas 40 butir soal pada materi Gerak dan Gaya beserta hasil jawaban dari 52 siswa. Sumber data dipilih secara purposive berdasarkan pertimbangan bahwa sekolah tersebut telah menerapkan evaluasi berbasis komputer dan memungkinkan penerapan aplikasi Test Analysis Program (TAP) versi 14.7.4 dalam pengolahan data, serta mengevaluasi hubungannya dengan ranah kognitif menurut Taksonomi Bloom. Pemilihan teknik purposive juga memastikan bahwa responden benar-benar relevan dengan tujuan penelitian, sehingga data yang diperoleh sesuai dengan kebutuhan analisis [22], [23]. Hal ini penting agar instrument yang dianalisis benar-benar cocok dengan konteks penelitian dan tujuan pengukuran yang diinginkan.

Populasi dalam penelitian ini adalah seluruh siswa kelas VII, sedangkan sampel penelitian adalah siswa kelas VII-ICP 1 dan VII-SCP 1 yang dipilih berdasarkan teknik purposive sampling. Pertimbangan pemilihan sampel adalah keterjangkauan dan kesesuaian dengan materi yang diujikan, yaitu "Gerak dan Gaya" [24]. Teknik pengumpulan data dilakukan melalui dokumentasi berupa naskah soal pilihan ganda dan lembar jawaban peserta didik, serta observasi terhadap proses evaluasi di sekolah. Selain itu, digunakan pula teknik tes untuk memperoleh data hasil belajar peserta didik sebagai dasar analisis kualitas soal [25]. Dengan prosedur ini, diperoleh data yang lengkap untuk dianalisis baik dari sisi isi soal maupun respon siswa.

Validitas soal dianalisis berdasarkan tiga jenis, yakni validitas isi yang digunakan untuk memastikan bahwa butir soal sesuai dengan kompetensi dasar dan indikator pembelajaran. Validitas konstruk digunakan untuk memastikan bahwa butir soal benar-benar mengukur kemampuan berpikir dan keterampilan yang menjadi sasaran, seperti berpikir kritis dan pemecahan masalah. Validitas empiris digunakan untuk mengevaluasi kesesuaian hasil tes dengan kriteria statistik, dalam hal ini korelasi point biserial, untuk melihat sejauh mana soal dapat membedakan peserta didik berdasarkan kemampuan yang diharapkan. Kombinasi ketiga jenis validitas ini membantu menilai lebih tepat tentang kualitas soal yang ada.

Data yang telah dikumpulkan kemudian dianalisis secara bertahap dengan bantuan aplikasi Test Analysis Program (TAP) versi 14.7.4. Analisis dilakukan secara bertahap terhadap beberapa parameter pokok sebagaimana disajikan pada tabel berikut:

Tabel 1. Kriteria Penilaian Kualitas Butir Soal Berdasarkan Parameter Analisis

Parameter	Metode Analisis	Kriteria Penilaian
Validitas	Korelasi Point Biserial	Valid jika $r \geq 0,30$ [26]
Reliabilitas	KR-20 dan Split-Half	Tinggi jika $\geq 0,70$ [27]
Tingkat Kesukaran	Proporsi Jawaban Benar	Mudah (0,71–1,00), Sedang (0,31–0,70), Sukar (0,00–0,30) [28]
Daya Pembeda	Selisih proporsi antara kelompok atas dan bawah	Sangat Baik ($\geq 0,40$), Baik (0,30–0,39), Cukup (0,20–0,29), Jelek ($< 0,20$) [29]

Sumber : (Data penelitian yang diadaptasi dari Arikunto, 2020)

Untuk mendukung pelengkap analisis, dilakukan pula uji inferensial berupa analisis korelasi Pearson antara tingkat kesukaran dan daya pembeda serta regresi linear sederhana untuk mengetahui pengaruh daya beda dan tingkat kesukaran terhadap validitas soal [32], [33]. Melalui penerapan metode ini, diharapkan hasil dari analisis dapat memberikan penjelasan yang menyeluruh dan akurat tentang kualitas pada butir soal serta hubungannya dengan ranah kognitif peserta didik.

III. HASIL DAN PEMBAHASAN

Hasil Penelitian Dan Pembahasan

Hasil dan pembahasan dari analisis butir soal IPA SMP Kelas VII pada materi Gerak dan Gaya yang terdiri atas 40 butir soal pilihan ganda. Secara umum, nilai rata-rata tingkat kesukaran adalah 0,780 yang termasuk kategori mudah, rata-rata daya pembeda sebesar 0,452 yang termasuk kategori baik, dan rata-rata point biserial sebesar 0,525 yang juga termasuk kategori baik. Nilai reliabilitas tes (KR-20) mencapai 0,929, yang menandakan tingkat konsistensi internal sangat tinggi, sementara standar error pengukuran (SEM) adalah 2,164. Ini menunjukkan bahwa secara umum tes memiliki konsistensi internal yang sangat baik, namun dominasi soal dengan kategori mudah cukup tinggi sehingga variasi soal pada kategori sedang dan sukar perlu ditingkatkan. Oleh karena itu, meskipun instrument ini memiliki

reliabilitas yang tinggi, penyeimbangan tingkat kesukaran perlu dilakukan agar evaluasi dapat mencerminkan kemampuan peserta didik dengan lebih akurat. Dari 40 butir soal didapat hasil analisis sebagai berikut:

Tabel 2. Rekapitulasi Hasil Analisis Butir Soal

No. Butir Soal	Tingkat Kesukaran (P)	Kriteria Kesukaran	Daya Pembeda (D)	Kriteria Daya Pembeda	Validitas	Kriteria Validitas	Keterangan
1	0,48	Sedang	0,62	Baik	0,30	Cukup	Kesukaran sedang, daya pembeda baik, PB cukup; masih dapat digunakan dengan sedikit perbaikan distraktor.
2	0,48	Sedang	0,68	Sangat Baik	0,47	Baik	Proporsi jawaban tepat seimbang, mampu membedakan kelompok tinggi dan rendah.
3	0,79	Mudah	0,31	Cukup	0,13	Rendah	Terlalu mudah dan daya pembeda rendah, kemungkinan pengecoh kurang efektif.
4	0,85	Mudah	0,12	Jelek	0,07	Sangat Rendah	Hampir semua siswa menjawab benar, tidak membedakan kemampuan.
5	0,62	Sedang	0,81	Sangat Baik	0,57	Baik	Daya pembeda tinggi, distraktor bekerja baik.
6	0,92	Sangat Mudah	0,19	Jelek	0,18	Rendah	Soal terlalu mudah, PB rendah, fungsi diskriminasi lemah.
7	0,52	Sedang	0,62	Baik	0,47	Baik	Tingkat kesukaran tepat, membedakan kelompok tinggi-rendah dengan baik.
8	0,87	Mudah	0,38	Cukup	0,42	Baik	Terlalu mudah, walau daya pembeda cukup.
9	0,87	Mudah	0,44	Baik	0,63	Sangat Baik	Walau mudah, tetap membedakan dengan baik; pengecoh efektif.
10	0,79	Mudah	0,31	Cukup	0,13	Rendah	PB rendah, beberapa pengecoh kurang menarik.
11	0,88	Mudah	0,31	Cukup	0,62	Sangat Baik	Daya pembeda cukup, PB tinggi; masih relevan digunakan.
12	0,63	Sedang	0,50	Baik	0,43	Baik	Parameter ideal, pengecoh berfungsi baik.

No. Butir Soal	Tingkat Kesukaran (P)	Kriteria Kesukaran	Daya Pembeda (D)	Kriteria Daya Pembeda	Validitas	Kriteria Validitas	Keterangan
13	0,90	Sangat Mudah	0,31	Cukup	0,60	Sangat Baik	Terlalu mudah, perlu tingkatkan kompleksitas.
14	0,75	Mudah	0,50	Baik	0,51	Baik	Daya pembeda baik dan PB tinggi.
15	0,81	Mudah	0,50	Baik	0,55	Baik	Soal mudah namun masih membedakan dengan efektif.
16	0,85	Mudah	0,44	Baik	0,67	Sangat Baik	PB sangat tinggi, daya pembeda baik.
17	0,79	Mudah	0,44	Baik	0,43	Baik	Parameter memadai meskipun tingkat kesukaran tinggi.
18	0,62	Sedang	0,63	Baik	0,41	Baik	Ideal, mampu membedakan dengan baik.
19	0,94	Sangat Mudah	0,19	Jelek	0,47	Baik	Sangat mudah, daya pembeda lemah.
20	0,87	Mudah	0,31	Cukup	0,55	Baik	PB baik, meski daya pembeda hanya cukup.
21	0,58	Sedang	0,62	Baik	0,54	Baik	Seimbang, pembeda efektif, pengecoh berfungsi baik.
22	0,75	Mudah	0,69	Sangat Baik	0,61	Baik	Walau mudah, mampu membedakan dengan sangat baik.
23	0,67	Sedang	0,62	Baik	0,54	Baik	Kriteria ideal, pembeda baik.
24	0,77	Mudah	0,63	Baik	0,59	Baik	Mudah tetapi tetap memiliki pembeda yang memadai.
25	0,88	Mudah	0,38	Cukup	0,61	Sangat Baik	Tingkat kesukaran tinggi, daya pembeda sedang.
26	0,81	Mudah	0,38	Cukup	0,39	Cukup	Daya pembeda dan PB sedang, pengecoh perlu perbaikan.
27	0,83	Mudah	0,44	Baik	0,54	Baik	Mampu membedakan cukup efektif.
28	0,73	Mudah	0,56	Baik	0,48	Baik	Parameter baik meski kesukaran agak tinggi.
29	0,90	Sangat Mudah	0,31	Cukup	0,56	Baik	Terlalu mudah, meskipun PB baik.
30	0,87	Mudah	0,38	Cukup	0,58	Baik	Kesukaran tinggi, daya pembeda sedang.
31	0,90	Sangat Mudah	0,25	Jelek	0,46	Baik	Diskriminasi rendah, perlu revisi besar.
32	0,65	Sedang	0,88	Sangat Baik	0,72	Sangat Baik	Sangat baik dari semua parameter.

No. Butir Soal	Tingkat Kesukaran (P)	Kriteria Kesukaran	Daya Pembeda (D)	Kriteria Daya Pembeda	Validitas	Kriteria Validitas	Keterangan
33	0,83	Mudah	0,50	Baik	0,74	Sangat Baik	Mudah tapi tetap membedakan dengan sangat baik.
34	0,87	Mudah	0,44	Baik	0,75	Sangat Baik	Walau mudah, indikator pembeda sangat baik.
35	0,83	Mudah	0,31	Cukup	0,41	Baik	Daya pembeda sedang, butuh revisi distraktor.
36	0,62	Sedang	0,81	Sangat Baik	0,51	Baik	Ideal, pembeda sangat baik.
37	0,87	Mudah	0,25	Jelek	0,35	Cukup	Tidak mampu membedakan kemampuan peserta didik secara signifikan.
38	0,88	Mudah	0,38	Cukup	0,46	Baik	Kesukaran tinggi, pembeda sedang.
39	0,88	Mudah	0,38	Cukup	0,57	Baik	PB baik, namun daya pembeda perlu ditingkatkan.
40	0,88	Mudah	0,31	Cukup	0,61	Baik	Soal terlalu mudah, diskriminasi sedang.

Sumber : (Data Penelitian yang diolah dengan Aplikasi *Test Analysis Program* (TAP), 2025)

Hasil analisis butir soal menunjukkan bahwa dari 40 soal, 25 butir soal (62,5%) dinyatakan layak digunakan tanpa adanya perbaikan, 10 butir soal (25%) butir soal perlu adanya perbaikan pada aspek tertentu seperti peningkatan kualitas pengecoh atau penyesuaian tingkat kesukaran, 5 butir soal (12,5%) dikategorikan tidak layak karena memiliki daya pembeda $< 0,20$ atau $PB < 0,20$, sehingga tidak efektif dalam membedakan kemampuan peserta didik[41].

Butir soal yang dinilai layak umumnya memiliki tingkat kesukaran sedang hingga mudah dengan daya pembeda $\geq 0,40$ dan $PB \geq 0,40$. Soal nomor 5, 12, 14, 15, dan 18 juga menunjukkan parameter ideal dan layak digunakan. Soal yang memerlukan perbaikan, seperti nomor 3, 8, 25, 26, 30, 35, 38, 39, dan 40, umumnya memiliki PB cukup atau buruk meskipun tingkat kesukaran dan daya pembeda memadai. Perbaikan diperlukan pada aspek konstruksi butir dan pengecoh agar validitas empiris meningkat soal tidak layak, seperti nomor 4, 6, 19, 31, dan 37, menunjukkan kelemahan signifikan pada daya pembeda dan PB, sehingga sama sekali tidak mampu membedakan peserta didik berkemampuan tinggi dan rendah[42]. Oleh karena itu, pemetaan kategori kelayakan butir soal memberikan dasar yang jelas untuk menetapkan prioritas perbaikan, mulai dari perbaikan distraktor atau pengecoh sampai perbaikan jumlah butir soal yang tidak efektif digunakan[43].

Validitas instrumen dianalisis dengan tiga cara. Pertama, validitas isi dicek berdasarkan keterkaitan soal dengan Kompetensi Dasar (KD) dan indikator pembelajaran[45]. Hasilnya, semua soal sudah sesuai dengan topik dan indikator yang diukur, meskipun ada 7 butir soal yang perlu direvisi pada konstruksi opsi jawabannya. Kedua, validitas konstruk dilihat dari cakupan ranah kognitif. Instrumen ini mencakup C1 hingga C4, namun proporsi C5 dan C6 masih rendah, sekitar 5%, sehingga perlu ditingkatkan agar mampu mengakomodir keterampilan berpikir tingkat tinggi (HOTS) [44]. Ketiga, validitas empiris dihitung menggunakan point biserial (PB). Hasilnya menunjukkan 28 butir soal (70%) berada dalam kategori baik sampai sangat baik ($PB \geq 0,40$), 5 butir (12,5%) dalam kategori cukup ($PB 0,30-0,39$), dan 7 butir (17,5%) dalam kategori buruk ($PB < 0,20$). Butir soal dengan PB rendah menunjukkan kurangnya kemampuan instrumen dalam membedakan peserta didik yang berkemampuan tinggi dengan peserta didik yang berkemampuan rendah. Hal tersebut mengindikasikan bahwasannya upaya untuk meningkatkan kualitas instrumen tidak hanya memerlukan revisi teknis pada butir soal, tetapi juga penyesuaian konten agar mencakup variasi kognitif yang lebih luas, khususnya pada level HOTS.

Tabel 3. Rekapitulasi Kualitas Instrumen

Kategori	Jumlah Butir Soal	Instrumen
Layak	25	62,5 %

Perlu Perbaikan	10	25 %
Tidak Layak	5	12,5 %

Sumber : (Data penelitian yang diolah dengan Aplikasi *Test Analysis Program* (TAP), 2025)

Data tersebut menunjukkan bahwa sebagian besar instrumen tes sudah berada pada kategori layak digunakan, dengan reliabilitas keseluruhan yang tinggi. Namun, sepertiga dari butir soal memerlukan revisi atau penggantian untuk meningkatkan kualitas pengukuran secara keseluruhan. Hasil ini menjadi pijakan penting untuk penyusunan instrumen berikutnya, dengan meratakan tingkat kesukaran, optimalisasi daya beda, dan peningkatan validitas pada semua dimensi yang akan diukur.

IV. SIMPULAN

Setelah dilakukannya penelitian tentang analisis butir soal IPA SMP pada materi Gerak dan Gaya di salah satu SMP Swasta di Kabupaten Sidoarjo, dapat disimpulkan bahwa instrumen memiliki reliabilitas tinggi sehingga konsisten dalam mengukur kemampuan peserta didik. Sebagian besar butir soal memenuhi kriteria kualitas, namun masih terdapat kelemahan berupa daya pembeda dan validitas yang rendah pada beberapa soal. Dominasi soal kategori mudah juga menunjukkan bahwa instrumen belum seimbang dalam distribusi tingkat kesukaran, serta kurang optimal dalam mengukur keterampilan berpikir tingkat tinggi. Oleh karena itu, penyusunan instrumen evaluasi ke depan perlu memperhatikan keseimbangan tingkat kesukaran, efektivitas pengecoh, dan cakupan ranah kognitif yang lebih luas agar penilaian dapat mencerminkan kompetensi peserta didik secara menyeluruh.

Sejalan dengan temuan tersebut, untuk penelitian berikutnya disarankan memperluas cakupan pada materi maupun jenjang yang berbeda sehingga hasil analisis lebih representatif. Pengembangan soal hendaknya menambah variasi butir yang menuntut keterampilan berpikir tingkat tinggi serta melibatkan soal uraian untuk mengeksplorasi kemampuan analisis dan pemecahan masalah peserta didik. Selain itu, penggunaan perangkat analisis lain seperti *Iteman* atau *Anates* dapat dipertimbangkan sebagai pembandingan terhadap hasil analisis *Test Analysis Program* (TAP), sehingga diperoleh evaluasi instrumen yang lebih komprehensif. Dengan kombinasi tersebut, kualitas asesmen diharapkan semakin meningkat dan mampu mendukung pembelajaran yang bermakna sesuai tuntutan abad ke-21.

UCAPAN TERIMA KASIH

Penulis mengungkapkan rasa syukur yang besar kepada Allah SWT atas segala berkat dan anugerah-Nya sehingga artikel ini yang berjudul “Analisis Validitas & Reliabilitas Butir Soal Berbasis *Test Analysis Program* (TAP) untuk Kualitas Soal di Era Industri 4.0” dapat diselesaikan dengan baik. Penulis mengucapkan terima kasih yang mendalam kepada dosen pembimbing yang telah memberikan bimbingan dengan kesabaran, ketelitian, dan arahan yang konstruktif selama penulisan artikel ini. Penulis mengucapkan terima kasih kepada dosen validator instrumen yang telah memvalidasi dan memberikan saran berharga untuk instrumen penelitian yang dipakai. Rasa hormat dan penghargaan penulis ucapkan kepada semua dosen Program Studi Pendidikan IPA Universitas Muhammadiyah Sidoarjo yang telah memberikan pengetahuan dan inspirasi selama masa kuliah. Penulis menyampaikan apresiasi kepada Bapak/Ibu guru IPA dan kepala sekolah di SMP lokasi penelitian yang telah memberikan izin dan dukungan dalam proses pengambilan data, serta kepada siswa kelas VII yang telah berkontribusi aktif dalam pengisian soal yang dianalisis. Penulis juga berterima kasih kepada rekan-rekan mahasiswa angkatan 2021 Pendidikan IPA atas dukungan moral, kebersamaan, serta diskusi yang memperkuat selama proses penelitian dilakukan. Ucapan terima kasih yang tulus penulis sampaikan kepada keluarga tercinta yang selalu memberikan doa, motivasi, serta dukungan moral dan materi. Akhirnya, penulis mengucapkan terima kasih kepada semua pihak yang telah memberikan bantuan dan kontribusi dalam berbagai bentuk yang tidak bisa disebutkan satu per satu. Semoga semua kebaikan yang diberikan menjadi amal jariyah dan semoga tulisan ini memberi manfaat bagi kemajuan dunia pendidikan di zaman Industri 4.0.

REFERENSI

- [1] I. Pada and P. Di, dan Keguruan Dalam Menyongsong Era Revolusi.
- [2] I. Afidah Rahman, M. Adinda Viola, Masita, and F. Aqilah Vilanti, “Uji Validitas dan Reliabilitas Kualitas Sarana dan Prasarana Akademik Terhadap Prestasi Belajar Mahasiswa FKIP Universitas Jambi,” *J. Pendidik. Tambusai*, vol. 7, no. 3, pp. 4–5, 2023.
- [3] Z. Raudhatul, “Analisis Tingkat Kesulitan Soal Ujian Semester Ganjil Mata Pelajaran Matematika Kelas V MIN 1 Banda Aceh,” p. 91, 2021, [Online]. Available: <https://repository.ar-raniry.ac.id/id/eprint/14251/>
- [4] A. Doni, H. Baitika, L. Amalia, V. Y. Putri, and A. Saifudin, “Analisa dan Perancangan Sistem Aplikasi Ujian Sekolah Berbasis Komputer,” *J. Ilmu Komput. dan Pendidik.*, vol. 1, no. 1, pp. 59–65, 2022, [Online]. Available: <https://journal.mediapublikasi.id/index.php/logic>

- [5] T. Septiana, D. Kurniawan, J. Juliati, I. Sunandi, and S. Z. Nurbaya, "Adopsi Teknologi dalam Pendidikan Hibrida: Tantangan dan Peluang bagi Institusi Pendidikan Tinggi," *J. Pendidik. Tambusai*, vol. 6, no. 2, pp. 16834–16841, 2022.
- [6] S. Ono, "Uji Validitas dan Reliabilitas Alat Ukur SG Posture Evaluation," *J. Keterapian Fis.*, vol. 5, no. 1, pp. 55–61, 2020, doi: 10.37341/jkf.v5i1.167.
- [7] Y. F. Basri and E. M. Karima, "Kualitas Butir Soal Penilaian Tengah Semester Ganjil Mata Pelajaran Sejarah Terhadap Hasil Belajar Kognitif Siswa di Fase E SMK Negeri 1 Painan," *J. Pendidik. Tambusai*, vol. 7, no. 3, pp. 21160–21171, 2023, [Online]. Available: <https://doi.org/10.31004/jptam.v7i3.9639>
- [8] B. Adibah and R. Vebrianto, "Pengembangan Instrumen untuk Mengukur Penggunaan E-Module sebagai Bahan Ajar Peserta Didik," *Bedelau J. Educ. Learn.*, vol. 3, no. 2, pp. 52–61, 2022.
- [9] A. Widayanti, Z. R. Ridlo, N. Ahmad, U. Jember, and J. Timur, "Pengembangan Instrumen Penilaian Berbantuan Media," pp. 963–975, 2024.
- [10] D. Wicaksana, "Uji Validitas Dan Reliabilitas Alat Ukur Indonesia Implicit Self-Esteem Test," *J. Pengukuran Psikol. dan Pendidik. Indones.*, vol. 1, no. 4, 2019.
- [11] R. Suhartini, M. Cholik, T. Rijanto, and M. S. Sumbawati, "Upaya Peningkatan Kompetensi Guru Era 4.0 di Sekolah Indonesia Kuala Lumpur," vol. 4, no. 2, pp. 366–373, 2024, doi: 10.54259/pakmas.v4i2.3074.
- [12] A. Setyaningtyas, B. A. Saputra, M. A. Yaqin, M. R. Ridho, and K. P. Email, "Pendidikan Matematika , FPMIPA , IKIP PGRI Bojonegoro," *Pros. Semin. Nas. Pendidik. FPMIPA*, vol. 1, no. 1, pp. 573–576, 2023.
- [13] U. P. Ganesha, *International Education of Elementary*, vol. 6 (2), 2022.
- [14] S. Pokhrel, "No TitleEΛENH," *Ayan*, vol. 15, no. 1, pp. 37–48, 2024.
- [15] Risnawati and U. S. B. Pasaribu, "Mengkonstruksi Instrumen Penilaian Keterampilan (Psikomotor)," *J. Pendidik. Tambusai*, vol. 8, no. 1, pp. 3107–3113, 2024.
- [16] S. Ayuna, *Analisis Kompetensi TI Pada Guru Era 4.0 Serta Perannya Dalam Meningkatkan Prestasi Siswa SMKN 1 Al Mubarkeya*. 2021. [Online]. Available: [https://repository.ar-raniry.ac.id/id/eprint/19416/%0Ahttps://repository.ar-raniry.ac.id/id/eprint/19416/1/Seri Ayuna%2C170212021%2CFTK%2CPTI%2C082273691608.pdf](https://repository.ar-raniry.ac.id/id/eprint/19416/%0Ahttps://repository.ar-raniry.ac.id/id/eprint/19416/1/Seri_Ayuna%2C170212021%2CFTK%2CPTI%2C082273691608.pdf)
- [17] S. Sovey, K. Osman, and M. E. E. M. Matore, "Rasch Analysis for Disposition Levels of Computational Thinking Instrument Among Secondary School Students," *Eurasia J. Math. Sci. Technol. Educ.*, vol. 18, no. 3, pp. 2–15, 2022, doi: 10.29333/ejmste/11794.
- [18] A. Haleem, M. Javaid, M. A. Qadri, and R. Suman, "Understanding the role of digital technologies in education: A review," *Sustain. Oper. Comput.*, vol. 3, no. May, pp. 275–285, 2022, doi: 10.1016/j.susoc.2022.05.004.
- [19] R. Andriani, Mellyzar, I. R. Lukman, Muttakin, A. I. Pasaribu, and M. R. Fadli, "A review of Digital Assessment in Education: Tools, Feature, and Effectiveness," *Miceshi*, vol. 1, no. 1, pp. 1–11, 2023, [Online]. Available: <https://ojs.unimal.ac.id/mijeshi/>
- [20] N. M. Sri Mertasari and I. M. Candiasa, "Formative Evaluation of Digital Learning Materials," *J. Educ. Technol.*, vol. 6, no. 3, pp. 507–514, 2022, doi: 10.23887/jet.v6i3.44165.
- [21] F. G. Nandini, R. S. Cahyawati, and C. Hasanudin, "Langkah-Langkah Uji Validitas dan Reliabilitas Instrumen Pemahaman Materi Teks Laporan Hasil Observasi".
- [22] W. A. Fietri, Lufri, Syamzurizal, and Zulyusri, "Analisis Butir Soal Biologi Kelas VIII Madrasah Tsanawiyah Negeri 6 Kerinci," *J. Pendidik. Biol. Undiksha*, vol. 8, no. 2, pp. 50–60, 2021, [Online]. Available: <https://ejournal.undiksha.ac.id/index.php/JJPB/index%0AAnalisis>
- [23] N. Zalfa Zuhri, A. Sopian, Sofyan Sauri, and Y. Nurbayan, "Analisis Validitas dan Reliabilitas Soal Bahasa Arab Melalui Website OpExams Pembuat Soal Berbasis AI," *J. Pendidik. Mod.*, vol. 9, no. 2, pp. 87–91, 2024, doi: 10.37471/jpm.v9i2.863.
- [24] N. S. Azzahra, S. Sumarni, and H. Putranta, "Analisis Validitas dan Realibilitas Kualitas Soal Pilihan Ganda Asesmen Kompetensi Minimum (AKM) Mata Pelajaran Pendidikan Agama Islam Menggunakan ...," *QuranicEdu J. ...*, vol. 4, no. 1, pp. 85–94, 2024, [Online]. Available: <https://jurnalannur.ac.id/index.php/QuranicEdu/article/view/681>
- [25] N. Z. Zuhri, S. Syihabuddin, and T. Tatang, "Analisis Validitas, Reliabilitas, dan Tingkat Kesukaran Soal Bahasa Arab Tingkat SMP Berbasis Artificial Intelligence (AI) melalui Platform QuestionWell," *J. Pendidik. dan Pembelajaran Indones.*, vol. 4, no. 2, pp. 693–704, 2024, doi: 10.53299/jppi.v4i2.576.
- [26] V. Saadah, *Uji Analisis Validitas dan Reabilitas Butir Soal Pilihan Ganda Ulangan Akhir Semester Mata Pelajaran PAI dan Budi Perkerti Kelas IX di SMO Negeri 2 Sirampog Brebes*. 2020. [Online]. Available: <http://eprints.walisongo.ac.id/id/eprint/6667/>
- [27] S. A. Damayanti, N. Efendi, and S. B. Sartika, "Validitas Butir Soal Penilaian Akhir Semester (Pas) Kelas Viii Untuk Mata Pelajaran Ilmu Pengetahuan Alam," *J. Banua Sci. Educ.*, vol. 2, no. 2, pp. 59–66, 2022, doi: 10.20527/jbse.v2i2.70.

- [28] R. Damayanti, Wi. D., Halidjah, S., & Pranata, "Analisis tingkat kesukaran butir soal pilihan ganda pada penilaian tengah semester kelas iv," *J. Pendidik. dan Pengajaran Khatulistiwa*, vol. 10, no. 11, pp. 1–10, 2021, [Online]. Available: <https://jurnal.untan.ac.id/index.php/jdpdp/article/view/50458/75676591120>
- [29] K. M. Jaiswal, S. Dudhgaonkar, P. Gharade, and N. Sharma, "Post-valuation quality check of multiple-choice questions," *Int. J. Basic Clin. Pharmacol.*, vol. 12, no. 1, p. 43, 2022, doi: 10.18203/2319-2003.ijbcp20223353.
- [30] I. Sholichah and S. B. Sartika, "The Analysis Item Validity of Natural Science Subject in Secondary School on The Measurement Material [Analisis Validitas Butir Soal IPA SMP pada Materi Pengukuran]," pp. 1–7.
- [31] N. A. Widari, "Kualitas Butir Soal Pilihan Ganda Ujian Akhir Semester Genap Bahasa Indonesia Kelas X Buatan Mahasiswa Ditinjau Dari Segi Taraf Kesukaran, Daya Pembeda, Dan Efektivitas Pengecoh," *Kualitas Butir Soal Pilihan Ganda Ujian Akhir Semester Gen. Bhs. Indones. Kelas X Buatan Mhs. Ditinjau Dari Segi Taraf Kesukaran, Daya Pembeda, Dan Ef. Pengecoh*, vol. 01, no. 02, pp. 1–11, 2023.
- [32] N. Ngarifillaili, B. Kartowagiran, and U. Yvette, "Evaluation of the implementation of educational assessment standards at Madrasah Tsanawiyah Modern Islamic Boarding School," *REID (Research Eval. Educ.)*, vol. 7, no. 2, pp. 106–117, 2021, doi: 10.21831/reid.v7i2.43672.
- [33] P. Raya, "Analyzing the validity and reliability of an assessment tool for senior high school students," no. 1951, pp. 1–17, 2024.
- [34] A. A. Rezigalla *et al.*, "Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items," *BMC Med. Educ.*, vol. 24, no. 1, pp. 1–7, 2024, doi: 10.1186/s12909-024-05433-y.
- [35] T. J. Wood, "Guidelines for reviewing multiple-choice questions on UGME examinations," pp. 1–6, 2020.
- [36] K. B. YÜKSEL and N. DOĞAN, "Investigation of Psychometric Properties of Multiple-Choice Items Developed By Turkish Teachers," *Sak. Univ. J. Educ.*, vol. 12, no. 1, pp. 130–149, 2022, doi: 10.19126/suje.1007897.
- [37] A. D. Slepko, M. L. Van Bussel, K. M. Fitze, and W. S. Burr, "A Baseline for Multiple-Choice Testing in the University Classroom," *SAGE Open*, vol. 11, no. 2, 2021, doi: 10.1177/21582440211016838.
- [38] E. H. Tyas and L. Naibaho, "Hots Learning Model Improves the Quality of Education," *Int. J. Res. - GRANTHAALAYAH*, vol. 9, no. 1, pp. 176–182, 2021, doi: 10.29121/granthaalayah.v9.i1.2021.3100.
- [39] T. HALADYNA, "Creating multiple-choice items for testing student learning," *Int. J. Assess. Tools Educ.*, vol. 9, no. Special Issue, pp. 6–18, 2022, doi: 10.21449/ijate.1196701.
- [40] Maliha Ansari, Rabia Sadaf, Aisha Akbar, Sabahat Rehman, Zunnera Rashid Chaudhry, and Sabeen Shakir, "Assessment of distractor efficiency of MCQS in item analysis.," *Prof. Med. J.*, vol. 29, no. 05, pp. 730–734, 2022, doi: 10.29309/tpmj/2022.29.05.6955.
- [41] M. Syafitri, "Analisis Butir Soal Penilaian Akhir Semester I Mata Pelajaran Matematika Tahun Ajaran 2019/2020 Kelas IV SD Negeri Se-Dabin Panggung Kecamatan Tegal Timur Kota Tegal," 2020.
- [42] M. I. Anshari, R. Nasution, M. Irsyad, A. Z. Alifa, and I. A. Zuhriyah, "Analisis Validitas dan Reliabilitas Butir Soal Sumatif Akhir Semester Ganjil Mata Pelajaran PAI," *Edukatif J. Ilmu Pendidik.*, vol. 6, no. 1, pp. 964–975, 2024, doi: 10.31004/edukatif.v6i1.5931.
- [43] E. Novriyanti and R. Arthur, "Analisis Kualitas Butir Soal Ujian Tengah Semester Biologi Umum Menggunakan Model Rasch," *JagoMIPA J. Pendidik. Mat. dan IPA*, vol. 4, no. 4, pp. 718–733, 2024, doi: 10.53299/jagomipa.v4i4.927.
- [44] I. Prabowo, "Analisis Kualitas Soal Asesmen Sumatif Pendidikan Agama Islam dan Budi Pekerti," *J. Ilm. Insa. Mulia*, vol. 1, no. 1, pp. 1–7, 2024, [Online]. Available: <https://doi.org/10.59923/jiim.v1i1.155>
- [45] N. Nadia, "Analisis Butir Soal HOTS (High Order Thingking Skill) Pada Soal Ulangan Kelas X Mata Pelajaran Fisika. Skripsi," 2023.

Conflict of Interest Statement:

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.