

---

# Academia Open



*By Universitas Muhammadiyah Sidoarjo*

---

## Table Of Contents

|                                       |   |
|---------------------------------------|---|
| <b>Journal Cover</b>                  | 1 |
| <b>Author[s] Statement</b>            | 3 |
| <b>Editorial Team</b>                 | 4 |
| <b>Article information</b>            | 5 |
| Check this article update (crossmark) | 5 |
| Check this article impact             | 5 |
| Cite this article                     | 5 |
| <b>Title page</b>                     | 6 |
| Article Title                         | 6 |
| Author information                    | 6 |
| Abstract                              | 6 |
| <b>Article content</b>                | 7 |

## **Originality Statement**

The author[s] declare that this article is their own work and to the best of their knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the published of any other published materials, except where due acknowledgement is made in the article. Any contribution made to the research by others, with whom author[s] have work, is explicitly acknowledged in the article.

## **Conflict of Interest Statement**

The author[s] declare that this article was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## **Copyright Statement**

Copyright © Author(s). This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at <http://creativecommons.org/licences/by/4.0/legalcode>

# Academia Open

Vol. 10 No. 1 (2025): June  
DOI: 10.21070/acopen.10.2025.12271

## EDITORIAL TEAM

### Editor in Chief

Mochammad Tanzil Multazam, Universitas Muhammadiyah Sidoarjo, Indonesia

### Managing Editor

Bobur Sobirov, Samarkand Institute of Economics and Service, Uzbekistan

### Editors

Fika Megawati, Universitas Muhammadiyah Sidoarjo, Indonesia

Mahardika Darmawan Kusuma Wardana, Universitas Muhammadiyah Sidoarjo, Indonesia

Wiwit Wahyu Wijayanti, Universitas Muhammadiyah Sidoarjo, Indonesia

Farkhod Abdurakhmonov, Silk Road International Tourism University, Uzbekistan

Dr. Hindarto, Universitas Muhammadiyah Sidoarjo, Indonesia

Evi Rinata, Universitas Muhammadiyah Sidoarjo, Indonesia

M Faisal Amir, Universitas Muhammadiyah Sidoarjo, Indonesia

Dr. Hana Catur Wahyuni, Universitas Muhammadiyah Sidoarjo, Indonesia

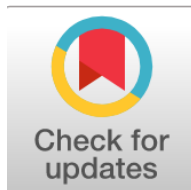
Complete list of editorial team ([link](#))

Complete list of indexing services for this journal ([link](#))

How to submit to this journal ([link](#))

## Article information

**Check this article update (crossmark)**



**Check this article impact <sup>(\*)</sup>**



**Save this article to Mendeley**



<sup>(\*)</sup> Time for indexing process is various, depends on indexing database platform

# Quality Analysis of Motion and Force Science Test Items Using TAP: Analisis Kualitas Item Uji Sains Gerak dan Gaya Menggunakan TAP

Salma Ramadhina Fatahillah, [salmaramadhina9@gmail.com](mailto:salmaramadhina9@gmail.com) (\*)

*Universitas Muhammadiyah Sidoarjo, Indonesia*

Nur Efendi, [nur.efendi@umsida.ac.id](mailto:nur.efendi@umsida.ac.id)

*Universitas Muhammadiyah Sidoarjo [<https://ror.org/017hvgd88>], Indonesia*

(\*) Corresponding author

## Abstract

**General Background:** In the era of Industry 4.0, educational assessment must align with 21st-century competencies, requiring high-quality test instruments capable of measuring students' cognitive abilities accurately. **Specific Background:** The quality of multiple-choice items is determined by parameters such as validity, reliability, difficulty level, discriminating power, distractor effectiveness, and cognitive level distribution according to Bloom's Taxonomy. **Knowledge Gap:** In practice, many school examinations are developed from existing question banks without systematic quantitative analysis, resulting in instruments that may not fully represent students' competencies or higher-order thinking skills. **Aims:** This study aimed to analyze the quality of Grade VII science test items on Motion and Force using the Test Analysis Program (TAP) based on standard psychometric criteria and cognitive domains. **Results:** Analysis of 40 items from 52 students indicated that 95% of the items were valid, with very high reliability ( $KR-20 = 0.929$ ; split-half = 0.925). The difficulty level was dominated by easy items (82.5%), most items showed high discriminating power (60%), and 90% of distractors functioned effectively. The cognitive distribution was concentrated at the analyzing level (C4), with very limited representation of higher levels. **Novelty:** The study integrates TAP-based statistical analysis with cognitive domain mapping to provide a comprehensive evaluation of science test quality. **Implications:** The findings suggest that while the instrument meets most quality standards, improvements are needed in difficulty balance and higher-order thinking coverage to support meaningful assessment in contemporary science education.

## Highlights:

- Most questions met statistical standards for psychometric soundness.
- Items were heavily concentrated in low-to-moderate challenge categories.
- Higher cognitive levels were minimally represented in the instrument.

## Keywords:

Test Analysis Program; Science Assessment; Item Analysis; Motion And Force; Junior High School

Published date: 2025-06-21

## 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]. Penilaian sebagai salah satu elemen penting dalam proses pendidikan dituntut mampu beradaptasi agar dapat mengukur kompetensi yang relevan dengan kebutuhan abad ke-21 [2], [3]. 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) [4]. 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 [5]. 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 [6], [7]. 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) [8]. 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 [9], [10]. 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) [11], [12]. 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 [13]. 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) [14].

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 sebagai 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, [15]. 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 [16].

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..

## 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 [17]. 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 [18], [19]. 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" [20]. 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 [21]. 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:

| Parameter         | Metode Analisis                                 | Kriteria Penilaian                                                                        |
|-------------------|-------------------------------------------------|-------------------------------------------------------------------------------------------|
| Validitas         | Korelasi Point Biserial                         | Valid jika $r \geq 0,30$ [22]                                                             |
| Reliabilitas      | KR-20 dan Split-Half                            | Tinggi jika $\geq 0,70$ [23]                                                              |
| Tingkat Kesukaran | Proporsi Jawaban Benar                          | Mudah (0,71–1,00), Sedang (0,31–0,70), Sukar (0,00–0,30) [24]                             |
| 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 ( $< 0,20$ ) [25] |

Table 1. Kriteria Penilaian Kualitas Butir Soal Berdasarkan Parameter Analisis  
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 [26], [27]. 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.

## Hasil 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:

| No. Soal | Butir | Tingkat Kesukaran | Kriteria Kesukaran | Daya Pembeda (D) | Kriteria Validitas | Kriteria Keterangan Validitas |
|----------|-------|-------------------|--------------------|------------------|--------------------|-------------------------------|
|----------|-------|-------------------|--------------------|------------------|--------------------|-------------------------------|

# Academia Open

Vol. 10 No. 1 (2025): June  
DOI: 10.21070/acopen.10.2025.12271

|    | (P)  |              |      | Pembeda     |      |               |                                                                                                           |
|----|------|--------------|------|-------------|------|---------------|-----------------------------------------------------------------------------------------------------------|
| 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,                                                                                          |

# Academia Open

Vol. 10 No. 1 (2025): June  
DOI: 10.21070/acopen.10.2025.12271

|    |      |              |      |             |      |             |                                                                        |
|----|------|--------------|------|-------------|------|-------------|------------------------------------------------------------------------|
| 13 | 0,90 | Sangat Mudah | 0,31 | Cukup       | 0,60 | Sangat Baik | pengecoh berfungsi 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                                         |

|    |      |              |      |             |      |             |                                                                   |
|----|------|--------------|------|-------------|------|-------------|-------------------------------------------------------------------|
| 31 | 0,90 | Sangat Mudah | 0,25 | Jelek       | 0,46 | Baik        | sedang. Diskriminasi rendah, perlu revisi besar.                  |
| 32 | 0,65 | Sedang       | 0,88 | Sangat Baik | 0,72 | Sangat Baik | Sangat baik dari semua parameter.                                 |
| 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.                          |

Table 2. Rekapitulasi Hasil Analisis Butir Soal

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[28].

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[29]. 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[30].

Validitas instrumen dianalisis dengan tiga cara. Pertama, validitas isi dicek berdasarkan keterkaitan soal dengan Kompetensi Dasar (KD) dan indikator pembelajaran[31]. 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) [32]. 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 mengidikasi 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.

| Kategori        | Jumlah Butir Soal | Instrumen |
|-----------------|-------------------|-----------|
| Layak           | 25                | 62,5 %    |
| Perlu Perbaikan | 10                | 25 %      |
| Tidak Layak     | 5                 | 12,5 %    |

Table 3. Rekapitulasi Kualitas Instrumen

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 instrument berikutnya, dengan meratakan tingkat kesukaran, optimalisasi daya beda, dan peningkatan validitas pada semua dimensi yang akan diukur.

## 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 instrument 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 pembanding 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

## References

- [1] I. Pada and P. Di, Pendidikan dan Keguruan dalam Menyongsong Era Revolusi 4.0, 2019.
- [2] I. A. Rahman, M. A. Viola, Masita, and F. A. Vilanti, "Uji Validitas dan Reliabilitas Kualitas Sarana dan Prasarana Akademik terhadap Prestasi Belajar Mahasiswa FKIP Universitas Jambi," *Jurnal Pendidikan 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, Banda Aceh: UIN Ar-Raniry Repository, 2021.
- [4] A. Doni, H. Baitika, L. Amalia, V. Y. Putri, and A. Saifudin, "Analisa dan Perancangan Sistem Aplikasi Ujian Sekolah Berbasis Komputer," *Jurnal Ilmu Komputer dan Pendidikan*, vol. 1, no. 1, pp. 59-65, 2022.
- [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," *Jurnal Pendidikan Tambusai*, vol. 6, no. 2, pp. 16834-16841, 2022.
- [6] S. Ono, "Uji Validitas dan Reliabilitas Alat Ukur SG Posture Evaluation," *Jurnal Keterampilan Fisik*, 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," *Jurnal Pendidikan Tambusai*, vol. 7, no. 3, pp. 21160-21171, 2023.

8. [8] B. Adibah and R. Vebrianto, "Pengembangan Instrumen untuk Mengukur Penggunaan E-Module sebagai Bahan Ajar Peserta Didik," *Bedelau Journal of Education and Learning*, vol. 3, no. 2, pp. 52-61, 2022.
9. [9] A. Widayanti, Z. R. Ridlo, and N. Ahmad, "Pengembangan Instrumen Penilaian Berbantuan Media," *Prosiding Seminar Nasional Pendidikan*, pp. 963-975, 2024.
10. [10] D. Wicaksana, "Uji Validitas dan Reliabilitas Alat Ukur Indonesia Implicit Self-Esteem Test," *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia*, vol. 1, no. 4, 2019.
11. [11] R. Suhartini, M. Cholik, T. Rijanto, and M. S. Sumbawati, "Upaya Peningkatan Kompetensi Guru Era 4.0 di Sekolah Indonesia Kuala Lumpur," *Jurnal Pengabdian Kepada Masyarakat*, vol. 4, no. 2, pp. 366-373, 2024, doi: 10.54259/pakmas.v4i2.3074.
12. [12] A. Setyaningtyas, B. A. Saputra, M. A. Yaqin, and M. R. Ridho, "Pendidikan Matematika," in *Prosiding Seminar Nasional Pendidikan FPMIPA IKIP PGRI Bojonegoro*, vol. 1, no. 1, pp. 573-576, 2023.
13. [13] Universitas Pendidikan Ganesha, *International Journal of Elementary Education*, vol. 6, no. 2, 2022.
14. [14] S. Pokhrel, "No Title," *Agane Journal*, vol. 15, no. 1, pp. 37-48, 2024.
15. [15] Risnawati and U. S. B. Pasaribu, "Mengkonstruksi Instrumen Penilaian Keterampilan (Psikomotor)," *Jurnal Pendidikan Tambusai*, vol. 8, no. 1, pp. 3107-3113, 2024.
16. [16] S. Ayuna, Analisis Kompetensi TI pada Guru Era 4.0 serta Perannya dalam Meningkatkan Prestasi Siswa SMKN 1 Al Mubarkeya, Banda Aceh: UIN Ar-Raniry Repository, 2021.
17. [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 Journal of Mathematics, Science and Technology Education*, vol. 18, no. 3, pp. 2-15, 2022, doi: 10.29333/ejmste/11794.
18. [18] A. Haleem, M. Javaid, M. A. Qadri, and R. Suman, "Understanding the Role of Digital Technologies in Education: A Review," *Sustainable Operations and Computers*, vol. 3, pp. 275-285, 2022, doi: 10.1016/j.susoc.2022.05.004.
19. [19] R. Andriani et al., "A Review of Digital Assessment in Education: Tools, Feature, and Effectiveness," *MICESHI Journal*, vol. 1, no. 1, pp. 1-11, 2023.
20. [20] N. M. S. Mertasari and I. M. Candiasa, "Formative Evaluation of Digital Learning Materials," *Journal of Education Technology*, vol. 6, no. 3, pp. 507-514, 2022, doi: 10.23887/jet.v6i3.44165.
21. [21] F. G. Nandini, R. S. Cahyawati, and C. Hasanudin, "Langkah-Langkah Uji Validitas dan Reliabilitas Instrumen Pemahaman Materi Teks Laporan Hasil Observasi," 2022.
22. [22] W. A. Fietri, Lufri, Syamzurizal, and Zulyusri, "Analisis Butir Soal Biologi Kelas VIII Madrasah Tsanawiyah Negeri 6 Kerinci," *Jurnal Pendidikan Biologi Undiksha*, vol. 8, no. 2, pp. 50-60, 2021.
23. [23] N. Z. Zuhri, A. Sopian, S. Sauri, and Y. Nurbayan, "Analisis Validitas dan Reliabilitas Soal Bahasa Arab melalui Website OpExams Pembuat Soal Berbasis AI," *Jurnal Pendidikan Modern*, vol. 9, no. 2, pp. 87-91, 2024, doi: 10.37471/jpm.v9i2.863.
24. [24] N. S. Azzahra, S. Sumarni, and H. Putranta, "Analisis Validitas dan Reliabilitas Kualitas Soal Pilihan Ganda AKM Mata Pelajaran Pendidikan Agama Islam," *QuranicEdu Journal*, vol. 4, no. 1, pp. 85-94, 2024.
25. [25] N. Z. Zuhri, S. Syihabuddin, and T. Tatang, "Analisis Validitas, Reliabilitas, dan Tingkat Kesukaran Soal Bahasa Arab Tingkat SMP Berbasis Artificial Intelligence melalui Platform QuestionWell," *Jurnal Pendidikan dan Pembelajaran Indonesia*, vol. 4, no. 2, pp. 693-704, 2024, doi: 10.53299/jppi.v4i2.576.
26. [26] V. Saadah, Uji Analisis Validitas dan Reliabilitas Butir Soal Pilihan Ganda Ulangan Akhir Semester Mata Pelajaran PAI dan Budi Pekerti Kelas IX di SMP Negeri 2 Sirampog Brebes, Semarang: UIN Walisongo Repository, 2020.
27. [27] S. A. Damayanti, N. Efendi, and S. B. Sartika, "Validitas Butir Soal Penilaian Akhir Semester Kelas VIII untuk Mata Pelajaran Ilmu Pengetahuan Alam," *Banua Science Education Journal*, vol. 2, no. 2, pp. 59-66, 2022, doi: 10.20527/jbse.v2i2.70.
28. [28] R. Damayanti, W. D. Halidjah, and S. Pranata, "Analisis Tingkat Kesukaran Butir Soal Pilihan Ganda pada Penilaian Tengah Semester Kelas IV," *Jurnal Pendidikan dan Pengajaran Khatulistiwa*, vol. 10, no. 11, pp. 1-10, 2021.
29. [29] K. M. Jaiswal, S. Dudhgaonkar, P. Gharade, and N. Sharma, "Post-Valuation Quality Check of Multiple-Choice Questions," *International Journal of Basic and Clinical Pharmacology*, vol. 12, no. 1, p. 43, 2022, doi: 10.18203/2319-2003.ijbcp20223353.
30. [30] I. Sholichah and S. B. Sartika, "Analisis Validitas Butir Soal IPA SMP pada Materi Pengukuran," 2022.
31. [31] N. Widari, "Kualitas Butir Soal Pilihan Ganda Ujian Akhir Semester Genap Bahasa Indonesia Kelas X Ditinjau dari Taraf Kesukaran, Daya Pembeda, dan Efektivitas Pengecoh," vol. 1, no. 2, pp. 1-11, 2023.
32. [32] N. Ngarifillaili, B. Kartowagiran, and U. Yvette, "Evaluation of the Implementation of Educational Assessment Standards at Madrasah Tsanawiyah Modern Islamic Boarding School," *Research and Evaluation in Education*, vol. 7, no. 2, pp. 106-117, 2021, doi: 10.21831/reid.v7i2.43672.
33. [33] P. Raya, "Analyzing the Validity and Reliability of an Assessment Tool for Senior High School Students," pp. 1-17, 2024.
34. [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 Medical Education*, vol. 24, no. 1, pp. 1-7, 2024, doi: 10.1186/s12909-024-05433-y.
35. [35] T. J. Wood, "Guidelines for Reviewing Multiple-Choice Questions on UGME Examinations," 2020.
36. [36] K. B. Yuksel and N. Dogan, "Investigation of Psychometric Properties of Multiple-Choice Items Developed by Turkish Teachers," *Sakarya University Journal of Education*, vol. 12, no. 1, pp. 130-149, 2022, doi: 10.19126/suje.1007897.
37. [37] A. D. Slepikov, 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. [38] E. H. Tyas and L. Naibaho, "HOTS Learning Model Improves the Quality of Education," *International Journal of*

Research – GRANTHAALAYAH, vol. 9, no. 1, pp. 176-182, 2021, doi: 10.29121/granthaalayah.v9.i1.2021.3100.

39. [39] T. Haladyna, "Creating Multiple-Choice Items for Testing Student Learning," *International Journal of Assessment Tools in Education*, vol. 9, Special Issue, pp. 6-18, 2022, doi: 10.21449/ijate.1196701.
40. [40] M. Ansari, R. Sadaf, A. Akbar, S. Rehman, Z. R. Chaudhry, and S. Shakir, "Assessment of Distractor Efficiency of MCQs in Item Analysis," *Professional Medical Journal*, vol. 29, no. 5, pp. 730-734, 2022, doi: 10.29309/tpmj/2022.29.05.6955.
41. [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. [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: Jurnal Ilmu Pendidikan*, vol. 6, no. 1, pp. 964-975, 2024, doi: 10.31004/edukatif.v6i1.5931.
43. [43] E. Novriyanti and R. Arthur, "Analisis Kualitas Butir Soal Ujian Tengah Semester Biologi Umum Menggunakan Model Rasch," *JagoMIPA: Jurnal Pendidikan Matematika dan IPA*, vol. 4, no. 4, pp. 718-733, 2024, doi: 10.53299/jagomipa.v4i4.927.
44. [44] I. Prabowo, "Analisis Kualitas Soal Asesmen Sumatif Pendidikan Agama Islam dan Budi Pekerti," *Jurnal Ilmiah Insan Mulia*, vol. 1, no. 1, pp. 1-7, 2024.
45. [45] N. Nadia, Analisis Butir Soal HOTS pada Soal Ulangan Kelas X Mata Pelajaran Fisika, 2023.