

Pengembangan Instrumen Diagnostik *Two-Tier* Literasi Sains SMP: Analisis Psikometrik dengan *Andrich Rating Scale Model*

Diterima:
19 November 2025
Disetujui:
05 Januari 2026
Diterbitkan:
14 Januari 2026

¹Nandang Purworatri, ^{2*}Rizki Nor Amelia
^{1,2}Program Studi Pendidikan IPA,
Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Negeri Semarang
^{1,2}Sekaran, Gunungpati, Semarang, Indonesia
E-mail: ¹purworatrinandang@students.unnes.ac.id,
^{2*}rizkinoramelia@mail.unnes.ac.id

*Corresponding Author

Abstrak—Literasi sains merupakan kompetensi penting dalam pembelajaran IPA. Pengukuran literasi masih menghadapi tantangan akibat keterbatasan instrumen yang mampu merefleksikan kemampuan dan penalaran peserta didik secara komprehensif. Penelitian ini bertujuan mengembangkan instrumen diagnostik *two-tier* untuk mengukur literasi sains peserta didik SMP kelas VIII serta menganalisis karakteristik psikometriknya menggunakan *Andrich Rating Scale Model*. Pengembangan instrumen mengikuti langkah Djemari Mardapi yang meliputi penyusunan spesifikasi tes, penulisan item, validasi isi, uji coba, dan analisis instrumen. Sampel penelitian terdiri atas 188 peserta didik yang dipilih melalui *cluster random sampling*. Validitas isi dibuktikan menggunakan Aiken's V dengan rerata koefisien sebesar 0,98. Hasil analisis menunjukkan reliabilitas item yang tinggi, kesesuaian seluruh item dengan model, variasi tingkat kesukaran antara -0,48 hingga 0,44 *logit*, serta struktur kategori respons 0–4 yang berfungsi secara optimal. Temuan ini menunjukkan bahwa instrumen memiliki kualitas psikometrik yang baik dan bersifat diagnostik dalam mengungkap literasi sains peserta didik.

Kata Kunci: *Andrich Rating Scale Model*; Instrumen Diagnostik; Literasi Sains.

Abstract—Science literacy is a crucial competency in science education; however, its assessment remains challenging due to the limited availability of instruments capable of comprehensively capturing students' abilities and reasoning. This study aimed to develop a two-tier diagnostic instrument to measure the science literacy of Grade 8 junior secondary school students and to examine its psychometric characteristics using the *Andrich Rating Scale Model*. The instrument development followed Djemari Mardapi's framework, including test specification, item writing, content validation, trial testing, and instrument analysis. The research sample consisted of 188 students selected using cluster random sampling. Content validity was established using Aiken's V, with a mean coefficient of 0.98. The analysis results indicate high item reliability, acceptable model fit for all items, a range of item difficulty estimates between -0.48 and 0.44 logits, and a well-functioning five-category rating scale (0–4). These findings demonstrate that the instrument possesses sound psychometric quality and provides diagnostic information for assessing students' science literacy.

Keywords: *Andrich Rating Scale Model*; Diagnostic Instrument; Science Literacy.

I. PENDAHULUAN

Pada era globalisasi dan Revolusi Industri 4.0, literasi sains menjadi kompetensi esensial bagi peserta didik untuk memahami fenomena ilmiah, berpikir kritis, dan mengambil keputusan berbasis bukti. Hasil *Programme for International Student Assessment* (PISA) menunjukkan bahwa skor literasi sains Indonesia berada pada angka 383, dibandingkan rata-rata global 485, dan menempati peringkat ke-67 dari 81 negara [1]. Kondisi ini menegaskan perlunya evaluasi literasi sains yang tidak hanya bersifat administratif, tetapi juga mampu memetakan kemampuan ilmiah peserta didik secara lebih autentik, khususnya pada konteks pembelajaran IPA di jenjang Sekolah Menengah Pertama (SMP). Dengan demikian, upaya peningkatan literasi sains perlu didukung oleh instrumen asesmen yang dirancang secara tepat dan berbasis bukti empiris.

Salah satu tantangan dalam pengukuran literasi sains terletak pada keterbatasan instrumen yang mampu menilai kemampuan secara komprehensif dan kontekstual. Penelitian di beberapa SMP menunjukkan bahwa peserta didik masih menghadapi kesulitan dalam menjelaskan fenomena ilmiah, merancang penyelidikan, dan menafsirkan bukti empiris [2]. Studi lain juga melaporkan bahwa peningkatan literasi sains peserta didik paling rendah tampak pada kemampuan mengevaluasi dan merancang penyelidikan ilmiah serta pada keterampilan menafsirkan data dan bukti ilmiah [3]. Penelitian pada peserta didik SMP terkait materi zat adiktif turut mengindikasikan adanya kesulitan dalam literasi sains maupun kemampuan pemecahan masalah pada konteks IPA [4]. Selain itu, penelitian lainnya menunjukkan bahwa kompetensi peserta didik yang paling lemah terdapat pada aspek penyusunan dan evaluasi inkuiri ilmiah, sedangkan kemampuan menafsirkan data dan fakta ilmiah juga masih berada pada kategori rendah hingga menengah [5]. Temuan-temuan tersebut menunjukkan bahwa kelemahan literasi sains peserta didik SMP tidak hanya bersifat konseptual, tetapi juga berkaitan dengan keterbatasan instrumen evaluasi yang mampu mengungkap proses penalaran ilmiah secara mendalam.

Tinjauan literatur di Indonesia juga mengungkap bahwa sebagian besar penelitian berfokus pada strategi pembelajaran, sementara pengembangan instrumen diagnostik yang kuat secara metodologis masih terbatas [6]. Studi lain juga mengungkap bahwa kualitas instrumen penilaian literasi sains yang ada sering kali belum memenuhi kriteria validitas dan reliabilitas empiris sehingga belum mampu memberikan informasi diagnostik yang mendalam [7]. Selain itu, penelitian mengenai asesmen kinerja literasi sains berbasis pendekatan STEM juga menegaskan bahwa instrumen yang dikembangkan masih memerlukan penguatan dari sisi kualitas psikometrik [8]. Temuan serupa disampaikan oleh penelitian yang menyoroti perlunya pengembangan

instrumen literasi sains yang tidak hanya valid dan reliabel, tetapi juga kontekstual dan sesuai dengan karakteristik peserta didik Indonesia [9]. Dengan demikian, masih terdapat kesenjangan penelitian pada pengembangan instrumen diagnostik literasi sains yang dikonstruksi secara sistematis dan dianalisis menggunakan pendekatan teori tes modern.

Berbagai pendekatan pembelajaran seperti *Problem-Based Learning* berbasis *socio-scientific issues* dan *structured inquiry learning* telah diterapkan dan terbukti meningkatkan literasi sains peserta didik [10], [11]. Namun, efektivitas pendekatan tersebut belum dapat dievaluasi secara optimal tanpa instrumen asesmen yang mampu memberikan gambaran mendalam mengenai kecakapan ilmiah peserta didik. Instrumen yang ada, seperti pada pengukuran literasi sains materi keanekaragaman hayati, masih menggunakan pendekatan klasik dan belum dianalisis secara mendalam menggunakan pendekatan teori tes modern [12]. Kondisi ini memperkuat urgensi pengembangan instrumen literasi sains yang tidak hanya berorientasi pada hasil akhir, tetapi juga mampu mengungkap kualitas penalaran dan konsistensi respon peserta didik.

Sejumlah studi menunjukkan bahwa Rasch Model merupakan salah satu pendekatan teori tes modern yang unggul dalam menganalisis kualitas instrumen literasi sains karena mampu menguji reliabilitas, kesesuaian item, tingkat kesulitan, dan kestabilan parameter pengukuran [13], [14], [15], [16]. Namun, penerapan Rasch Model dalam pengembangan instrumen diagnostik berbentuk *two-tier* untuk peserta didik SMP, khususnya pada materi IPA kelas VIII, masih belum banyak dilakukan. Selain itu, justifikasi penggunaan indikator literasi sains PISA sebagai dasar konstruksi tes pada konteks SMP belum banyak dielaborasi secara eksplisit dalam penelitian sebelumnya. Padahal, indikator PISA merepresentasikan kompetensi inti literasi sains yang mencakup kemampuan menjelaskan fenomena ilmiah, mengevaluasi dan merancang penyelidikan ilmiah, serta menafsirkan data dan bukti ilmiah, yang sejalan dengan tuntutan kurikulum IPA dan penguatan keterampilan abad ke-21.

Kebaruan penelitian ini dinyatakan secara eksplisit pada integrasi tiga aspek utama, yaitu pengembangan instrumen diagnostik literasi sains berbentuk *two-tier*, penggunaan indikator literasi sains PISA sebagai dasar konstruksi item soal, serta penerapan Rasch Model untuk menghasilkan informasi psikometrik yang komprehensif dan diagnostik. Kombinasi ketiga aspek tersebut masih jarang dilaporkan secara simultan dalam penelitian pendidikan sains pada jenjang SMP, khususnya pada konteks pembelajaran IPA kelas VIII. Dengan demikian, penelitian ini tidak hanya berkontribusi pada pengembangan instrumen, tetapi juga pada penguatan pendekatan evaluasi literasi sains berbasis teori tes modern.

Berdasarkan urgensi tersebut, penelitian ini mengembangkan instrumen diagnostik *two-tier* berpendekatan Rasch Model untuk mengukur literasi sains peserta didik pada konteks sekolah menengah pertama dengan karakteristik yang sebanding dengan lokasi penelitian. Tujuan penelitian meliputi: (1) menyusun spesifikasi tes berdasarkan indikator literasi sains PISA; (2) menguji validitas isi melalui penilaian ahli; (3) menganalisis reliabilitas instrumen; dan (4) mengevaluasi tingkat kesulitan item. Klaim penggunaan instrumen dalam penelitian ini dibatasi pada konteks sekolah dengan karakteristik serupa, sehingga hasil penelitian tidak dimaksudkan untuk digeneralisasikan secara luas ke seluruh SMP. Meskipun demikian, temuan penelitian ini diharapkan dapat menjadi rujukan awal dalam pengembangan instrumen literasi sains yang valid, reliabel, dan diagnostik pada jenjang SMP.

II. METODE PENELITIAN

Penelitian ini merupakan penelitian pengembangan dengan produk berupa tes diagnostik *two-tier* pada mata pelajaran IPA SMP, yang dirancang untuk mengukur literasi sains peserta didik secara komprehensif berdasarkan indikator PISA Tahun 2022. Tahapan pengembangan instrumen meliputi penyusunan spesifikasi tes sebagai dasar perumusan indikator dan cakupan materi, penulisan item soal sesuai kaidah konstruksi tes, penelaahan item melalui validasi isi oleh ahli, uji coba instrumen untuk memperoleh data empiris, analisis kualitas item dengan Rasch Model, revisi instrumen berdasarkan hasil analisis, dan penyusunan instrumen akhir yang siap digunakan sebagai alat ukur [17].

Penelitian dilaksanakan di SMP Negeri 1 Boja pada bulan Mei Tahun Ajaran 2024/2025. Sampel diambil menggunakan teknik *cluster random sampling*. Berdasarkan tabel Krejcie & Morgan, dari total populasi 252 peserta didik diperoleh sampel sebanyak 188 peserta didik kelas VIII. Instrumen dikembangkan berdasarkan ruang lingkup materi IPA kelas VIII yang meliputi enam topik utama, yaitu pengenalan sel; struktur dan fungsi makhluk hidup; usaha, energi dan pesawat sederhana; getaran, gelombang dan cahaya; unsur, senyawa dan campuran; serta struktur bumi dan perkembangannya. Pemilihan materi tersebut didasarkan pada keterwakilan tiga kompetensi utama literasi sains, yaitu menjelaskan fenomena secara ilmiah, mengevaluasi dan merancang penyelidikan ilmiah, serta menafsirkan data dan bukti ilmiah. Instrumen *two-tier* dalam penelitian ini menggunakan pedoman penskoran dengan rentang skor 0–4 sebagaimana yang disajikan pada Tabel 1, dan dirancang untuk menilai tidak hanya ketepatan jawaban pilihan ganda, tetapi juga kualitas alasan atau penjelasan ilmiah yang diberikan peserta didik.

TABEL 1. Pedoman Penskoran

| Skor | Pilihan Ganda | Alasan |
|------|--------------------------------------|---|
| 4 | Jawaban pilihan ganda benar (skor 1) | Peserta didik dapat menjawab 3 kata kunci dari kunci jawaban alasan |
| 3 | Jawaban pilihan ganda salah (skor 0) | Peserta didik dapat menjawab 3 kata kunci dari kunci jawaban alasan |
| | Jawaban pilihan ganda benar (skor 1) | Peserta didik dapat menjawab 2 kata kunci dari kunci jawaban alasan |
| 2 | Jawaban pilihan ganda salah (skor 0) | Peserta didik dapat menjawab 2 kata kunci dari kunci jawaban alasan |
| | Jawaban pilihan ganda benar (skor 1) | Peserta didik dapat menjawab 1 kata kunci dari kunci jawaban alasan |
| 1 | Jawaban pilihan ganda benar (skor 1) | Peserta didik tidak menjawab kata kunci dari kunci jawaban alasan |
| | Jawaban pilihan ganda salah (skor 0) | Peserta didik dapat menjawab 1 kata kunci dari kunci jawaban alasan |
| 0 | Jawaban pilihan ganda salah (skor 0) | Peserta didik tidak memberikan alasan kata dari jawaban alasan |

Skor ordinal yang disajikan pada Tabel 1 memiliki struktur kategori yang berurutan, sehingga memerlukan penerapan model Rasch yang sesuai untuk analisis data polytomous. Dalam kerangka Rasch Model, terdapat beberapa teknik yang dapat digunakan untuk menganalisis data bertingkat, salah satunya adalah *Andrich Rating Scale Model* yang dikembangkan oleh Andrich dan secara khusus dirancang untuk data ordinal dengan skala penilaian yang beraturan dan konsisten pada setiap item [18]. Model ini mengasumsikan bahwa seluruh item dalam instrumen menggunakan set kategori skor yang sama serta memiliki batas ambang (thresholds) antar kategori yang seragam di seluruh item. Dengan demikian, skala penilaian tertentu (misalnya 0–4) diperlakukan secara konsisten pada setiap item, dan setiap transisi antar kategori mencerminkan peningkatan kemampuan yang teratur sesuai dengan asumsi model [18]. Penggunaan *Andrich Rating Scale Model* memungkinkan diperolehnya informasi psikometrik yang komprehensif, khususnya dalam mengevaluasi fungsi kategori skor ordinal, termasuk identifikasi ambang batas (thresholds) dan kinerja kategori dalam instrumen asesmen pendidikan [19].

Teknik pengumpulan data dalam penelitian ini meliputi wawancara awal dengan guru IPA untuk mengidentifikasi kebutuhan asesmen dan konteks pembelajaran, validasi isi instrumen oleh para ahli yang selanjutnya dianalisis menggunakan formula Aiken, serta uji coba instrumen diagnostik *two-tier* kepada peserta didik sebagai sumber data utama. Dokumentasi digunakan sebagai data pendukung untuk melengkapi informasi administratif. Data yang diperoleh kemudian dianalisis menggunakan teknik *Andrich Rating Scale Model* yang mencakup estimasi reliabilitas instrumen untuk menilai konsistensi internal, analisis kesesuaian item (*item fit*)

terhadap model, analisis tingkat kesukaran item (*item difficulty*), dan struktur kategorisasi *rating scale*. Seluruh rangkaian prosedur analisis tersebut bertujuan untuk memastikan bahwa instrumen yang dikembangkan memiliki kualitas psikometrik yang kuat, akurat, dan layak digunakan dalam pengukuran literasi sains peserta didik SMP.

III. HASIL DAN PEMBAHASAN

Instrumen *two-tier* dikembangkan untuk menilai jawaban konseptual pada *tier* pertama dan alasan ilmiah pada *tier* kedua, sehingga memungkinkan analisis yang lebih komprehensif terhadap pemahaman peserta didik [20]. Melalui kombinasi kedua *tier*, instrumen dapat membedakan jawaban yang benar karena pemahaman dari jawaban yang tidak didukung alasan ilmiah yang tepat, sehingga memberikan gambaran lebih jelas mengenai kualitas penalaran peserta didik. Pendekatan ini selaras dengan prinsip literasi sains dalam Kurikulum Merdeka, yang menekankan kemampuan menjelaskan fenomena, mengevaluasi bukti, dan merancang penyelidikan ilmiah [21]. *Two-tier test* menjadi instrumen yang relevan untuk mengukur literasi sains secara lebih mendalam dan bermakna.

Studi pendahuluan menunjukkan bahwa asesmen IPA di sekolah mitra masih didominasi soal *one-tier* yang menekankan aspek mengingat. Format ini memang mudah disusun dan dinilai, namun kurang optimal untuk menilai kemampuan bernalar ilmiah atau memahami konsep secara mendalam. Temuan ini sejalan dengan berbagai penelitian yang menunjukkan bahwa praktik evaluasi di sekolah belum sepenuhnya mendukung pengukuran literasi sains secara kontekstual [22-23]. Instrumen *two-tier* yang menyertakan alasan konseptual memiliki keunggulan diagnostik karena mampu menelusuri kejelasan dan ketepatan penalaran peserta didik, bukan hanya hasil akhir jawaban [24-25]. Hal ini menunjukkan bahwa penerapan asesmen dua tingkat diperlukan untuk meningkatkan kualitas evaluasi pembelajaran sains.

Instrumen dikembangkan melalui penyusunan spesifikasi tes dan kisi-kisi yang memuat indikator literasi sains, cakupan materi, dan level kognitif. Sebanyak 20 item soal *two-tier* disusun lalu divalidasi secara isi oleh pakar menggunakan Aiken's V, dan menghasilkan indeks rerata sebesar 0,98. Nilai tersebut menunjukkan konsistensi penilaian terhadap kesesuaian *item* dengan indikator pembelajaran. Proses validasi melibatkan ahli materi dan ahli asesmen, sehingga *item* yang dihasilkan dapat dipastikan relevan secara konseptual dan teknis [26-27]. Instrumen yang telah dinyatakan layak selanjutnya diujicobakan kepada 188 peserta didik, dan data hasil uji coba dianalisis menggunakan *Andrich Rating Scale Model*. Ringkasan hasil analisis reliabilitas dan indeks pemisahan disajikan pada Tabel 2. Hasil estimasi menunjukkan nilai *item reliability*

sebesar 0,92 *logit*, yang berada di atas batas minimum 0,80 *logit* dan mendekati kategori ideal. Temuan ini mengindikasikan bahwa item-item dalam instrumen memiliki tingkat stabilitas dan konsistensi yang sangat baik dalam mengukur kemampuan literasi sains peserta didik [18, 28]. Nilai *person reliability* sebesar 0,55 *logit* yang berada pada kategori rendah. Nilai tersebut menunjukkan bahwa konsistensi respons peserta didik dalam menjawab instrumen masih belum stabil. Rendahnya *person reliability* mengindikasikan bahwa variasi kemampuan peserta didik relatif sempit atau karakteristik kemampuan responden cenderung homogen, sehingga instrumen belum sepenuhnya mampu membedakan tingkat kemampuan peserta didik secara optimal. Hal ini diperkuat oleh nilai *person separation index* sebesar 1,11 *logit*, yang menunjukkan bahwa instrumen hanya mampu mengelompokkan peserta didik ke dalam sekitar dua tingkat kemampuan.

TABEL 2. RINGKASAN RELIABILITAS ANDRICH RATING SCALE MODEL

| Indikator | Nilai (<i>logit</i>) | Kategori | Interpretasi |
|--------------------------------|------------------------|---------------|--|
| <i>Person Reliability</i> | 0,55 | Rendah | Konsistensi respons peserta didik masih rendah |
| <i>Item Reliability</i> | 0,92 | Sangat Tinggi | Item soal stabil dan berkualitas |
| <i>Person Separation Index</i> | 1,11 | Rendah | Membedakan ± 2 kelompok kemampuan |
| <i>Item Separation Index</i> | 3,45 | Tinggi | Membedakan ≥ 4 tingkat kesukaran item |

Kualitas item pada nilai *item reliability* sebesar 0,92 *logit* berada pada kategori sangat tinggi, yang menandakan bahwa item-item soal memiliki kestabilan dan konsistensi parameter kesukaran yang sangat baik. Nilai *item separation index* sebesar 3,45 *logit* menunjukkan bahwa hierarki tingkat kesukaran item tersusun secara jelas dan stabil, serta memiliki potensi replikasi yang baik apabila instrumen diterapkan pada kelompok responden lain dengan karakteristik yang sebanding. Dengan demikian, meskipun kemampuan instrumen dalam memisahkan tingkat kemampuan responden masih terbatas, kualitas item instrumen telah memenuhi kriteria pengukuran yang baik menurut *Andrich Rating Scale Model*. Selain itu, hasil analisis kesesuaian item pada Tabel 3 menunjukkan bahwa seluruh item berada dalam rentang nilai INFIT dan OUTFIT MNSQ sebesar 0,81–1,16 *logit*, yang masih berada dalam batas kelayakan model (0,5–1,5 *logit*). Tidak ditemukannya *item misfit*, menyimpulkan bahwa pola respons peserta didik terhadap setiap *item* telah sesuai dengan ekspektasi model [29-30]. Temuan ini menunjukkan bahwa instrumen bekerja secara konsisten dan mendukung validitas konstruksi literasi sains yang diukur.

TABEL 3. RINGKASAN ANALISIS *ITEM* BERDASARKAN *ANDRICH RATING SCALE MODEL*

| Nomor <i>Item</i> | <i>Item Fit</i> | | <i>Item Difficulty (logit)</i> |
|----------------------|--------------------|---------------------|--------------------------------|
| | INFIT MNSQ (logit) | OUTFIT MNSQ (logit) | |
| 1. | 0,93 | 0,93 | 0,13 |
| 2. | 1,04 | 1,03 | - 0,02 |
| 3. | 1,06 | 1,05 | - 0,20 |
| 4. | 1,03 | 1,02 | - 0,27 |
| 5. | 1,06 | 1,06 | - 0,29 |
| 6. | 1,00 | 1,01 | - 0,48 |
| 7. | 1,11 | 1,10 | - 0,17 |
| 8. | 0,98 | 0,98 | 0,04 |
| 9. | 0,99 | 0,99 | - 0,03 |
| 10. | 1,13 | 1,16 | 0,39 |
| 11. | 1,06 | 1,09 | 0,13 |
| 12. | 0,81 | 0,81 | 0,44 |
| 13. | 0,84 | 0,84 | 0,42 |
| 14. | 1,05 | 1,05 | 0,08 |
| 15. | 1,04 | 1,06 | 0,30 |
| 16. | 1,02 | 1,01 | 0,01 |
| 17. | 0,94 | 0,95 | - 0,05 |
| 18. | 0,91 | 0,90 | - 0,09 |
| 19. | 0,98 | 0,98 | - 0,30 |
| 20. | 1,02 | 1,01 | - 0,40 |

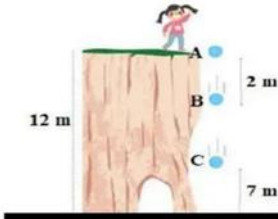
Analisis tingkat kesukaran item yang dirangkum dalam Tabel 3 menunjukkan bahwa nilai parameter kesukaran item berada pada rentang $-0,48$ hingga $+0,44$ logit. Berdasarkan kriteria pengelompokan tingkat kesukaran menggunakan standar deviasi sebesar 0,25 logit dengan mengacu pada nilai rata-rata 0,00 logit, item dapat diklasifikasikan ke dalam kategori mudah, sedang, sukar, dan sangat sukar [28]. Hasil analisis menunjukkan bahwa distribusi item terdiri atas kategori mudah (20%), sedang (35%), sukar (25%), dan sangat sukar (20%), yang mengindikasikan bahwa hierarki kesukaran item tersusun secara beragam dan mencakup spektrum kemampuan literasi sains yang diukur secara relatif seimbang. Meskipun demikian, terdapat empat item (nomor 10, 12, 13, dan 15) yang termasuk dalam kategori sangat sukar, yang menuntut penalaran tingkat tinggi serta pemahaman konteks yang lebih kompleks dibandingkan item lainnya. Oleh karena itu, keberadaan item dengan tingkat kesukaran ekstrem perlu ditinjau kembali untuk menjaga keseimbangan distribusi kesukaran item tanpa mengurangi kedalaman konsep yang diukur model [30-31].

Revisi terhadap item-item “sangat sukar” diharapkan dapat meningkatkan kejelasan stimulus, keterbacaan, dan aksesibilitas kognitif, sehingga respons peserta didik lebih merefleksikan kemampuan literasi sains yang diukur daripada kesulitan memahami redaksi soal. Salah satu contoh yang menunjukkan proses penyempurnaan ditampilkan pada Tabel 4 untuk

item nomor 12 mengenai energi mekanik. Pada versi awal, struktur item dinilai terlalu kompleks karena memuat uraian teks yang panjang, empat pernyataan dengan konsep berlapis, serta informasi numerik yang tidak seluruhnya diperlukan untuk menjawab pertanyaan. Kompleksitas tersebut berpotensi menambah beban kognitif, sehingga peserta didik kesulitan mengidentifikasi konsep inti yang ingin diukur. Selain itu, beberapa pernyataan dalam opsi mengandung ambiguitas, misalnya penggunaan istilah “menurun” yang tidak disertai rujukan visual yang jelas, sehingga membuka peluang interpretasi yang tidak konsisten antarresponden.

Melalui revisi, konteks dan stimulus dipadatkan, sementara pernyataan opsi diformulasikan ulang untuk menilai satu kompetensi utama, yakni pemahaman hubungan antara energi potensial, energi kinetik, dan energi mekanik pada gerak jatuh bebas. Penyederhanaan teks memastikan bahwa seluruh informasi yang diperlukan tersedia secara langsung, sehingga respon peserta didik lebih merefleksikan pemahaman konsep daripada kesulitan menafsirkan stimulus. Revisi ini juga mengurangi kemungkinan bias linguistik dan kesalahan pemahaman yang tidak terkait dengan tujuan pengukuran. Dengan demikian, versi baru *item* nomor 12 lebih fokus, lebih terstruktur, dan secara psikometrik lebih berpeluang menghasilkan respons yang stabil serta sesuai dengan prediksi model.

TABEL 4. ITEM SOAL KATEGORI SANGAT SUKAR

| Item 12 (Sebelum Revisi) |
|---|
| <p>Cermati teks berikut!</p> <p>Energi mekanik adalah energi yang dimiliki suatu benda karena sifat geraknya. Energi mekanik terdiri dari energi potensial dan energi kinetik. Energi potensial adalah energi yang dimiliki suatu benda karena kedudukan terhadap suatu acuan. Semakin rendah suatu benda mempunyai energi potensial semakin kecil. Seorang anak sedang bermain di dekat sebuah tebing. Ia menjatuhkan bola pejal bermassa 2 kg dari atas tebing tersebut seperti pada gambar berikut ini:</p>  <p>Berikut adalah hasil analisis berkaitan dengan infografis:</p> <ol style="list-style-type: none">(1) Energi potensial bola pejal bertambah 40 J ketika berubah dari titik A ke titik B(2) Energi mekanik bola pejal bernilai sama ketika berada diposisi A dan posisi C(3) Energi potensial bola pejal berkurang 60 J ketika berubah dari titik B ke titik C(4) Energi mekanik bola menurun dari titik A ke B karena sebagian berubah menjadi energi kinetik |

Hasil analisa yang benar ditunjukkan oleh nomor....

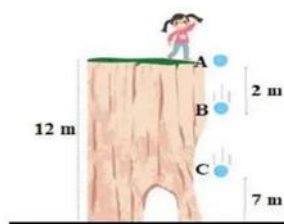
- A. (1) dan (2)
- B. (2) dan (3)
- C. (3) dan (4)
- D. (1) dan (4)

Alasan:

.....

Item 12 (Setelah Revisi)

Cermati gambar berikut yang menunjukkan sebuah bola pejal bermassa 2 kg yang dijatuhkan dari titik A, kemudian melewati titik B, dan mencapai titik C di dasar tebing.



Pernyataan berikut terkait perubahan energi ketika bola bergerak antara titik-titik tersebut:

1. Energi potensial bola berkurang 40 J saat berpindah dari A ke B.
2. Energi mekanik bola tetap sama pada posisi A dan posisi C.
3. Energi potensial bola berkurang 60 J saat berpindah dari B ke C.
4. Energi mekanik bola berkurang dari A ke B karena sebagian berubah menjadi energi kinetik.

Pernyataan yang benar adalah nomor ...

- A. (1) dan (2)
- B. (2) dan (3)
- C. (3) dan (4)
- D. (1) dan (4)

Alasan:

.....

Setelah kualitas *item* ditinjau dan direvisi, analisis selanjutnya difokuskan pada evaluasi struktur kategori penskoran untuk memastikan bahwa sistem respons ordinal yang digunakan berfungsi secara optimal. Evaluasi ini penting karena instrumen menggunakan skala bertingkat (0–4) yang dianalisis dengan *Andrich Rating Scale Model*. Ringkasan hasil analisis pada Tabel 5 menunjukkan bahwa instrumen terdiri atas lima kategori respons yang digunakan secara efektif. Distribusi frekuensi respons memperlihatkan bahwa seluruh kategori dipilih oleh responden, dengan kategori 2 memiliki frekuensi tertinggi (30%) dan kategori 4 terendah (9%). Tidak

ditemukan kategori yang jarang digunakan atau tidak berfungsi, sehingga setiap kategori respons dapat dikatakan berkontribusi secara bermakna dalam mengakomodasi variasi respons peserta didik.

TABEL 5. Struktur Kategorisasi *Andrich Rating Scale Model*

| Category | Observed Count (%) | Observed Average | INFIT MNSQ | OUTFIT MNSQ | Andrich Threshold | Thresholds between categories (Width) |
|----------|--------------------|------------------|------------|-------------|-------------------|---------------------------------------|
| 0 | 618 (16) | -0,42 | 1,03 | 1,03 | NONE | (-2,25) |
| 1 | 961 (26) | -0,30 | 0,96 | 0,95 | -0,81 | -0,87 |
| 2 | 1.129 (30) | -0,16 | 1,06 | 1,08 | -0,38 | 0,00 |
| 3 | 720 (19) | -0,01 | 0,96 | 0,95 | 0,37 | 0,87 |
| 4 | 332 (9) | 0,11 | 1,00 | 1,00 | 0,83 | (2,26) |

Nilai *Observed Average* pada setiap kategori menunjukkan pola peningkatan yang konsisten dari kategori terendah ke kategori tertinggi. Pola ini mengindikasikan bahwa semakin tinggi kategori yang dipilih, semakin tinggi pula estimasi kemampuan responden. Temuan tersebut menunjukkan bahwa responden mampu membedakan makna antar kategori secara konsisten dan bahwa urutan kategori telah dipahami dengan baik sesuai dengan asumsi *Andrich Rating Scale Model*. Selain itu, evaluasi kecocokan kategori berdasarkan nilai INFIT dan OUTFIT MNSQ menunjukkan bahwa seluruh kategori berada dalam rentang kelayakan model, dengan nilai yang mendekati 1,00 *logit*. Hal ini menandakan bahwa variasi respons pada setiap kategori sesuai dengan yang diharapkan oleh model, sehingga tidak terdapat kategori yang menunjukkan perilaku *overfit* maupun *underfit*.

Aspek struktur pada skala nilai *Andrich threshold* tersusun secara berurutan dan meningkat secara monotonik dari kategori rendah ke kategori tinggi. Pola ini menunjukkan bahwa setiap kategori memiliki batas transisi kemampuan yang jelas dan tidak terjadi tumpang tindih maupun pembalikan kategori. Dengan demikian, responden cenderung berpindah dari kategori yang lebih rendah ke kategori yang lebih tinggi seiring dengan meningkatnya tingkat kemampuan literasi sains yang diukur. Berdasarkan keseluruhan temuan tersebut, struktur kategori instrumen telah memenuhi kriteria kelayakan dalam analisis *Andrich Rating Scale Model* dan tidak memerlukan penggabungan atau perbaikan kategori. Instrumen ini dinilai layak digunakan untuk analisis lanjutan, seperti pemetaan *person-item* dan evaluasi kualitas item secara lebih mendalam.

Secara keseluruhan, instrumen diagnostik *two-tier* yang dikembangkan menunjukkan karakteristik psikometrik yang baik, meliputi validitas isi yang memadai, kestabilan item yang tinggi, kesesuaian item, serta variasi tingkat kesukaran item yang representatif. Instrumen ini tidak hanya menghasilkan skor pengukuran, tetapi juga menyediakan informasi diagnostik

mengenai tingkat pemahaman dan pola penalaran ilmiah peserta didik. Informasi tersebut berpotensi dimanfaatkan oleh guru sebagai dasar dalam merancang intervensi pembelajaran yang lebih tepat sasaran. Temuan ini sekaligus menegaskan efektivitas *Andrich Rating Scale Model* dalam mendukung pengukuran berbasis respons *item* yang objektif dan konsisten.

Penelitian ini memiliki beberapa keterbatasan ruang lingkup yaitu mencakup materi IPA kelas VIII dan dilaksanakan pada satu sekolah, sehingga generalisasi temuan masih terbatas. Selain itu, meskipun beberapa *item* dengan tingkat kesukaran sangat tinggi telah direvisi, pengujian ulang pada sampel yang berbeda tetap diperlukan untuk memastikan kestabilan parameter item. Penelitian ini juga belum mengkaji profil literasi sains peserta didik secara lebih mendalam, sehingga potensi instrumen dalam menggambarkan variasi capaian literasi antar kelompok responden belum sepenuhnya dieksplorasi.

Penelitian selanjutnya disarankan untuk mengembangkan instrumen serupa pada jenjang pendidikan yang berbeda, memperluas cakupan materi literasi sains, serta mengombinasikan analisis Rasch dengan kajian kualitatif terhadap alasan peserta didik. Selain itu, studi longitudinal dapat dilakukan untuk memantau perkembangan literasi sains dari waktu ke waktu dan mengevaluasi sensitivitas instrumen dalam mendeteksi perubahan kemampuan. Dengan pengembangan berkelanjutan, instrumen diagnostik *two-tier* berpendekatan teori tes modern berpotensi menjadi rujukan asesmen yang akurat, informatif, dan selaras dengan tuntutan pembelajaran IPA berbasis literasi sains.

IV. KESIMPULAN

Instrumen diagnostik *two-tier* literasi sains yang dikembangkan telah memenuhi kriteria kelayakan sebagai alat ukur berbasis psikometrik modern. Validitas isi instrumen didukung oleh penilaian pakar, sementara analisis Rasch dengan teknik *Andrich Rating Scale Model* menunjukkan kestabilan item yang tinggi, kesesuaian seluruh item dengan model, variasi tingkat kesukaran yang representatif, serta struktur kategori respons yang berfungsi secara optimal dan tersusun secara berurutan. Temuan tersebut mengindikasikan bahwa instrumen tidak hanya mampu mengukur tingkat literasi sains peserta didik secara kuantitatif, tetapi juga memberikan informasi diagnostik mengenai kualitas penalaran ilmiah melalui kombinasi jawaban dan alasan. Dengan karakteristik tersebut, instrumen ini berpotensi digunakan sebagai asesmen diagnostik literasi sains pada konteks pembelajaran IPA SMP dengan karakteristik sekolah yang sebanding, serta menjadi rujukan awal dalam pengembangan instrumen sejenis yang berbasis indikator PISA dan analisis Rasch.

DAFTAR PUSTAKA

- [1] N. F. Tillah and H. Subekti, "Analisis Kemampuan Literasi Sains Siswa SMP Berdasarkan Indikator dan Level Literasi Sains," *Edusaintek J. Pendidikan, Sains dan Teknol.*, vol. 12, no. 1, pp. 137–154, 2025.
- [2] L. Sholikah and F. Novika Pertiwi, "Analysis of Science Literacy Ability of Junior High School Students Based on Programme for International Student Assessment (PISA)," *Integr. Sci. Educ. Teach. Act. J.*, vol. 2, no. 1, pp. 95–104, 2021.
- [3] I. F. N. Ramdaniyah and K. Dwiningsih, "Penerapan LKS Berbasis Literasi Sains melalui Model Inkuiri Terbimbing untuk Meningkatkan Literasi Sains Siswa pada Submateri Faktor-Faktor yang Mempengaruhi Laju Reaksi," *UNESA J. Chem. Educ.*, vol. 6, no. 1, pp. 59–65, 2017.
- [4] I. A. Amala and Y. Yushardi, "Analisis Literasi Sains dan Kemampuan Pemecahan Masalah Siswa SMP pada Materi Zat Aditif dan Zat Adiktif," *J. Pendidik. Mipa*, vol. 12, no. 2, pp. 373–378, 2022, doi: 10.37630/jpm.v12i2.610.
- [5] S. R. Zahara, S. Alvina, and Mellyzar, "Literasi Sains dalam Pembelajaran Sains Siswa SMP," *Pendek. J. Pendidik. Berkarakter*, vol. 5, no. 2, pp. 119–124, 2022.
- [6] N. Mery, A. Rusilowati, S. Susilo, and P. Marwoto, "Meta-Analysis of Student Science Literacy in Indonesia," *Unnes Phys. Educ. J.*, vol. 3, no. 10, pp. 209–215, 2021.
- [7] J. Muftianah and Wahono, "Kualitas Instrumen Penilaian Literasi Sains Mata Pelajaran IPA untuk Peserta Didik Kelas VII," *E-Journal-Pensa*, vol. 6, no. 3, pp. 458–464, 2018.
- [8] K. Bashooir and Supahar, "Validitas dan Reliabilitas Instrumen Asesmen Kinerja Literasi Sains Pelajaran Fisika Berbasis STEM," *J. Penelit. dan Eval. Pendidik.*, vol. 22, no. 2, pp. 219–230, 2018.
- [9] A. Asyhari, "Pengembangan Instrumen Asesmen Literasi Sains Berbasis Nilai-Nilai Islam dan Budaya Indonesia dengan Pendekatan Kontekstual," *Lentera Pendidik.*, vol. 22, no. 1, pp. 1–13, 2019.
- [10] A. N. Khairrunisa, I. R. Yusup, and P. Epa, "Improving Student's Scientific Literacy Through a Problem-Based Learning Model Integrated With Socio-Scientific Issues on Ecosystem," *Biosf. Pendidik. Biol.*, vol. 18, no. 2, pp. 276–289, 2025.
- [11] A. R. Sarah and D. Astriani, "Improving Students' Science Literacy Competencies on Earth and its Satellites Through Structured Inquiry Learning," *J. Pendidik. Sains*, vol. 12, no. 1, pp. 32–37, 2020.
- [12] R. P. Alti, L. Lufri, H. Helendra, and R. Yogica, "Instrumen Asesmen Berbasis Literasi Sains Tentang Materi Keanekaragaman Hayati," *J. Lesson Learn. Stud.*, vol. 4, no. 1, pp. 53–58, 2021, doi: 10.23887/jlls.v4i1.34270.
- [13] R. Ashshaddiqah and A. Adiana, "Analisis Tingkat Literasi Sains Siswa pada Aspek Konteks, Konten, dan Kompetensi dengan Rasch," vol. 10, no. 1, pp. 28–39, 2021.
- [14] A. Riyadi, P. Susongko, and M. Munadi, "Model Asesmen Literasi Sains pada Peserta Didik Sekolah Dasar dengan Aplikasi Model Rasch," *J. Educ. Res.*, vol. 5, no. 3, pp. 3044–3054, 2024, doi: 10.37985/jer.v5i3.1411.
- [15] R. U. Sihombing, D. S. Naga, and W. Rahayu, "A Rasch Model Measurement Analysis on Indonesian Science Literacy Test: Smart Way to Improve the Learning Assessment," *Indones. J. Educ. Rev.*, vol. 6, no. 2, pp. 42–53, 2019.
- [16] W. T. Saputra, N. Y. Rustaman, and L. Rusyati, "Evaluating Validity and Reliability of Scientific Literacy Among Indonesian Secondary Students in PISA 2015 Context: Rasch Model Analysis," *J. Res. Educ. Res. Eval.*, vol. 13, no. 2, pp. 204–169, 2024.
- [17] D. Mardapi, *Penyusunan Tes dan Non Tes*. Mitra Cendekia Press, 2008.
- [18] C. M. Bond, T. G. Fox, *Applying the Rasch Model: Fundamental Measurement in the*

- Human Sciences* (3rd.). Routledge, 2015.
- [19] R. N. Amelia, A. Ristiyana, P. Sari, S. Rejeki, and D. Astuti, "Aplikasi Andrich Rating Scale Model Pada Analisis Psikometrik Tes Uraian Kimia Dasar I," vol. 14, no. 1, pp. 92–101, 2023.
 - [20] Z. A. Fadhilah *et al.*, "Socio-Scientific Issues using Two-Tier Instruments Based on Literacy Analysis," *J. Penelit. dan Pengemb. Pendidik.*, vol. 7, no. 3, pp. 448–455, 2023, doi: 10.23887/jppp.v7i3.63763.
 - [21] A. R. Firdaus and E. Erman, "Analisis Pemahaman Peserta Didik terhadap Konsep Zat dan Perubahan Zat," *BIOCHEPHY J. Sci. Educ.*, vol. 5, no. 1, pp. 499–505, 2025, doi: 10.52562/biochephy.v5i1.1567.
 - [22] D. K. Gurel, A. Eryilmaz, and L. C. McDermott, "A Review and Comparison of Diagnostic Instruments to Identify Students' Misconceptions in Science," *Eurasia J. Math. Sci. Technol. Educ.*, vol. 11, no. 5, pp. 989–1008, 2015, doi: 10.12973/eurasia.2015.1369a.
 - [23] I. W. E. Mahendra, N. N. Parmithi, E. Hermawan, D. P. Juwana, and I. W. Gunartha, "Teachers' Formative Assessment: Accessing Students' High Order Thinking Skills (HOTS)," *Int. J. Innov. Creat. Chang.*, vol. 12, no. 12, pp. 180–202, 2020.
 - [24] N. Rohmantika, E. S. Kurniawan, and S. Sriyono, "Effectiveness of Two-Tier Multiple Choice Diagnostic Test for Analyzing Students' Misconceptions in High School Physics Learning," *Radiasi J. Berk. Pendidik. Fis.*, vol. 15, no. 2, pp. 79–90, 2022, doi: 10.37729/radiasi.v15i2.2280.
 - [25] H. F. Firman, J. Ratnasari, and S. Windyariani, "Identifikasi Miskonsepsi Peserta Didik Menggunakan Two-Tier Test Berbantuan Certainty Of Response Index," *Biodik*, vol. 7, no. 2, pp. 33–44, 2021.
 - [26] H. Dumayanti, R. Putri Indriani, E. N. Hariyanti, R. H. Ristanto, and M. Miarsyah, "Two-Tier Test: Development of Critical Thinking Instruments for Ecosystem Concepts," *J. Phenom.*, vol. 12, no. 1, pp. 48–62, 2022.
 - [27] S. Nurhalimah, Y. Hidayati, I. Rosidi, and W. P. Hadi, "Hubungan Antara Validitas Item dengan Daya Pembeda dan Tingkat Kesukaran Soal Pilihan Ganda," *Nat. Sci. Educ. Res.*, vol. 4, no. 3, pp. 249–257, 2022.
 - [28] W. Sumintono, B., & Widhiarso, *Aplikasi Model Rasch untuk Penelitian Ilmu-Ilmu Sosial (Edisi Revisi)*. Bandung: Trim Komunikata Publishing House, 2014.
 - [29] S. Kassiavera, A. Suparmi, C. Cari, and S. Sukarmin, "Application of Rasch Model in Two-Tier Test for Assessing Critical Thinking in Physics Education," *J. Balt. Sci. Educ.*, vol. 23, no. 6, pp. 1227–1242, 2024, doi: 10.33225/jbse/24.23.1227.
 - [30] N. A. Putri, W. Liliawati, and R. Efendi, "Development Two Tier Multiple Choice Test of Mastery Concept on Heat Material using Rasch Model," *J. Pendidik. dan Ilmu Fis.*, vol. 04, no. 02, pp. 119–137, 2024.