

Analysis of the Quality of Summative Test Items in Junior High School Science Subjects Using the Rasch Model

[Analisis Butir Soal Sumatif Mata Pelajaran IPA SMP Menggunakan RASCH Model]

Nurul Zainiyah¹⁾, Septi Budi Sartika^{*,2)}

¹⁾Program Studi Pendidikan IPA, Universitas Muhammadiyah Sidoarjo, Indonesia

²⁾Program Studi Pendidikan IPA, Universitas Muhammadiyah Sidoarjo, Indonesia

*Email Penulis Korespondensi: septibudi1@umsida.ac.id

Abstract. - This study aims to analyze the quality of summative test items in science subjects for grade IX junior high school using the Rasch Model. The research was conducted using a descriptive quantitative approach through documentation study of 25 multiple-choice items covering the material on the reproductive system in plants and animals. Data analysis was carried out with the assistance of Ministep software to examine item validity, reliability, and difficulty characteristics. From the total of 25 items evaluated, most items (20 items) showed acceptable fit to the Rasch Model and were considered valid. However, five items did not meet the fit criteria and therefore need further revision. In addition, the reliability analysis revealed a Cronbach's Alpha value of 0.89, which suggests that the instrument consistently measures student's abilities. The distribution of item difficulty levels varies with 3 very difficult items, 8 difficult items, 10 easy items, and 4 very easy items. These findings indicate that the evaluation instrument overall has good quality, but revisions need to be made to misfit items to improve measurement accuracy. This research provides practical contributions for teachers in improving the quality of science learning evaluation through systematic and objective item analysis.

Keywords – Rasch Model; Item validity; reliability; science learning evaluation; item analysis

Abstrak. – Penelitian ini bertujuan untuk menganalisis kualitas butir soal sumatif mata pelajaran IPA kelas IX SMP dengan menggunakan Model Rasch. Penelitian ini dilakukan melalui pendekatan kuantitatif deskriptif dengan metode studi dokumentasi terhadap 25 butir soal pilihan ganda yang mencakup materi sistem reproduksi tumbuhan dan hewan. Analisis data dilakukan dengan bantuan perangkat lunak ministep untuk mengkaji validitas, reliabilitas, serta tingkat kesukaran butir soal. Dari 25 butir soal yang dianalisis, sebagian besar butir (20 butir) menunjukkan kecocokan dengan kriteria *fit statistics* Model Rasch dan dinyatakan valid. Sementara lima butir lainnya belum memenuhi kriteria tersebut sehingga memerlukan revisi. Hasil pengujian reliabilitas menunjukkan nilai *Cronbach's Alpha* sebesar 0,89 yang mencerminkan konsistensi internal instrumen yang sangat baik. Tingkat kesukaran butir soal bervariasi dari mudah hingga sukar, meskipun beberapa butir masih perlu diperbaiki agar dapat mengukur kemampuan siswa secara lebih akurat. Secara praktis, temuan ini dapat menjadi bahan pertimbangan bagi guru dalam meningkatkan kualitas evaluasi pembelajaran IPA melalui analisis butir soal yang lebih terencana.

Kata Kunci – Model Rasch; validitas butir; reliabilitas; tingkat kesukaran; evaluasi pembelajaran IPA; analisis butir soal

I. PENDAHULUAN

Evaluasi hasil belajar merupakan komponen penting dalam proses pembelajaran karena digunakan untuk menilai sejauh mana tujuan pembelajaran telah tercapai. Salah satu bentuknya adalah evaluasi sumatif, yaitu penilaian yang dilakukan pada akhir satuan pembelajaran untuk mengukur tingkat pencapaian kompetensi siswa[1]. Evaluasi sumatif umumnya dilaksanakan melalui tes tertulis yang terdiri atas butir-butir soal, yang berfungsi sebagai instrumen utama untuk memperoleh data objektif mengenai kemampuan peserta didik[2][3]. Kualitas butir soal sangat berpengaruh terhadap keberhasilan asesmen, karena soal yang tidak valid dan tidak reliabel dapat menghasilkan informasi yang keliru tentang kemampuan siswa. Oleh karena itu, analisis butir soal perlu dilakukan secara sistematis untuk memastikan bahwa soal yang digunakan benar-benar mampu mengukur kemampuan yang diharapkan[4]. Penelitian yang dilakukan oleh Ramadhani menyatakan bahwa analisis butir soal membantu guru dalam memperbaiki kualitas evaluasi, terutama dalam hal tingkat kesukaran dan daya pembeda soal[5]. Analisis butir soal dilakukan untuk semua mata pelajaran tak terkecuali pada mata pelajaran Ilmu Pengetahuan Alam (IPA).

Pada jenjang SMP, pembelajaran IPA dievaluasi melalui asesmen sumatif yang bertujuan untuk mengukur pencapaian kompetensi peserta didik pada akhir periode pembelajaran. Keakuratan hasil evaluasi tersebut sangat bergantung pada kualitas instrumen tes yang digunakan, termasuk validitas, reliabilitas, serta kesesuaian butir soal dengan tujuan pengukuran. Penelitian yang dilakukan oleh Ernawati dkk dan Wardhani dkk menyatakan bahwa

Copyright © Universitas Muhammadiyah Sidoarjo. This preprint is protected by copyright held by Universitas Muhammadiyah Sidoarjo and is distributed under the Creative Commons Attribution License (CC BY). Users may share, distribute, or reproduce the work as long as the original author(s) and copyright holder are credited, and the preprint server is cited per academic standards.

Authors retain the right to publish their work in academic journals where copyright remains with them. Any use, distribution, or reproduction that does not comply with these terms is not permitted.

analisis kualitas butir soal merupakan aspek penting dalam menentukan apakah instrumen tersebut benar-benar mampu mengukur kemampuan siswa secara akurat dan objektif dalam evaluasi pembelajaran[6][7]. Hasil analisis empiris seperti itu sangat diperlukan agar keputusan evaluasi pembelajaran yang diambil berdasarkan tes menjadi tepat sasaran dan dapat dipertanggungjawabkan secara ilmiah.

Namun demikian, dalam praktiknya analisis butir soal di sekolah masih sering dilakukan secara terbatas atau belum berbasis analisis statistik yang komprehensif, sehingga kualitas instrumen tes dalam evaluasi sumatif belum terjamin sepenuhnya. Hal ini berpotensi menyebabkan hasil evaluasi kurang akurat dan kurang konsisten dalam mengukur kompetensi siswa[7]. Untuk menjawab permasalahan ini, pendekatan analisis yang lebih objektif dan sistematis seperti pemodelan Rasch diperlukan karena mampu mengevaluasi karakteristik butir soal termasuk tingkat kesulitan, kesesuaian terhadap model, serta pemisahan antar butir secara empiris[6][7]. Dengan demikian, penerapan Model Rasch dapat memperkuat kualitas evaluasi instrumen soal dan memberikan dasar ilmiah yang lebih kuat dalam penilaian hasil belajar.

Analisis butir soal di tingkat sekolah merupakan salah satu langkah penting dalam upaya meningkatkan mutu evaluasi pembelajaran. Huda dan Wahyuni, melalui penelitiannya pada Try Out USBN IPA, menunjukkan bahwa sejumlah sekolah telah mulai melakukan analisis butir soal untuk menilai kualitas soal yang digunakan. Meskipun demikian, temuan tersebut juga mengindikasikan bahwa penerapan analisis butir soal belum dilakukan secara merata. Hal ini mengisyaratkan masih perlunya peningkatan kesadaran dan kemampuan guru agar analisis butir soal dapat dilaksanakan secara sistematis dan berkelanjutan[8]. Analisis butir soal sendiri berfungsi untuk mengetahui sejauh mana suatu butir soal mampu mengukur kompetensi yang diharapkan serta memberikan informasi mengenai kualitas instrumen penilaian[9][1]. Aspek yang dianalisis meliputi tingkat kesukaran, daya pembeda, efektivitas pengecoh, validitas, dan reliabilitas soal[3]. Penelitian oleh Nurani di SMP Negeri 8 Kota Malang menunjukkan bahwa analisis soal penilaian tengah semester IPA kelas IX yang dilakukan menggunakan Model Rasch mampu mengidentifikasi butir soal yang tidak *fit* serta soal dengan tingkat kesulitan yang tidak seimbang[10]. Studi serupa mengungkapkan bahwa analisis karakteristik butir soal pada penilaian IPA di tingkat sekolah menengah pertama membantu mengevaluasi kualitas item soal, termasuk sulit mudahnya soal, kemampuan soal membedakan peserta didik, dan reliabilitas instrumen, yang pada gilirannya mendukung kesesuaian instrumen penilaian dengan tujuan pembelajaran dan standar kompetensi yang ditetapkan[11]. Penelitian serupa oleh Widyaningsih dkk. menunjukkan bahwa analisis kualitas butir soal melalui pendekatan yang sistematis mampu memberikan gambaran yang komprehensif mengenai mutu instrumen penilaian serta tingkat pencapaian hasil belajar peserta didik[3]. Temuan-temuan ini menegaskan bahwa pelaksanaan analisis butir soal di sekolah-sekolah telah memberikan kontribusi positif dalam meningkatkan kualitas pembelajaran IPA di tingkat SMP.

Seiring dengan berkembangnya pendekatan pengukuran dalam evaluasi pendidikan, Model Rasch menjadi salah satu pendekatan yang semakin banyak digunakan dalam analisis instrumen. Dibandingkan dengan Teori Tes Klasik (*Classical Test Theory/CTT*), Model Rasch mampu mengubah data ordinal menjadi skala interval (logit) serta mengestimasi parameter butir dan kemampuan responden secara simultan[12]. Pendekatan CTT memiliki keterbatasan karena parameter tingkat kesukaran dan daya pembeda sangat dipengaruhi oleh karakteristik sampel yang digunakan. Berbeda dengan CTT, Model Rasch memiliki prinsip *specific objectivity*, yaitu estimasi parameter butir dan kemampuan responden relatif independen sepanjang data memenuhi asumsi model. Dengan demikian, Model Rasch mampu memberikan estimasi yang lebih stabil dan objektif dalam mengevaluasi kualitas Instrumen[13]. Penelitian oleh El Hasna menunjukkan bahwa Model Rasch menghasilkan nilai reliabilitas butir yang lebih tinggi serta pemetaan kemampuan siswa yang lebih representatif dibandingkan dengan pendekatan *Classical Test Theory* (CTT)[14]. Selain itu, Model Rasch juga memungkinkan pemetaan antara tingkat kesulitan soal dan kemampuan siswa dalam satu skala visual melalui *Wright Map*, serta mendeteksi item dan responden yang tidak sesuai (*misfit*)[15].

Dari sisi keadilan pengukuran, Model Rasch mendukung prinsip pengukuran yang objektif karena estimasi parameter butir dan kemampuan responden secara terpisah dalam satu skala pengukuran yang sama[16]. Selain itu, Model Rasch tetap mampu memberikan estimasi kemampuan yang stabil pada instrumen pilihan ganda tanpa memerlukan parameter tambahan yang kompleks[17]. Meskipun memiliki berbagai kelebihan, model ini juga memiliki sejumlah keterbatasan, seperti menuntut asumsi unidimensionalitas yang ketat, memerlukan jumlah sampel yang memadai, serta membutuhkan perangkat lunak khusus seperti *Winstep* atau *Ministep*[13]. Dengan melihat berbagai kelebihan dan keterbatasannya, Model Rasch dianggap sesuai untuk digunakan dalam penelitian ini karena dapat membantu peneliti melihat kualitas butir soal secara lebih terukur dan objektif. Melalui model ini, setiap butir soal dapat dianalisis berdasarkan tingkat kesulitan dan kesesuaiannya dengan kemampuan siswa, sehingga hasil evaluasi menjadi lebih jelas dan sistematis.

Dalam pembelajaran IPA di tingkat SMP, instrumen penilaian memiliki peran yang sangat penting. IPA tidak hanya menekankan pada penguasaan materi, tetapi juga pada kemampuan berpikir ilmiah, seperti

menganalisis, mengamati, dan menarik kesimpulan. Namun pada kenyataannya, masih banyak siswa kelas IX yang mengalami kesulitan dalam memahami konsep-konsep IPA, terutama yang bersifat abstrak. Kondisi ini menunjukkan bahwa diperlukan instrumen penilaian yang benar-benar mampu mengukur pemahaman siswa secara tepat. Penelitian lain menunjukkan bahwa analisis kesulitan belajar IPA pada siswa SMP memberikan wawasan penting tentang faktor-faktor penyebab hambatan belajar, seperti motivasi, minat, dan konsentrasi, yang berpengaruh terhadap hasil [18]. Kondisi tersebut menunjukkan bahwa evaluasi pembelajaran tidak hanya berfokus pada pencapaian hasil belajar, tetapi juga pada ketepatan instrumen yang digunakan untuk mengukur kemampuan siswa. Oleh karena itu, analisis butir soal diperlukan untuk memastikan kesesuaian antara soal, tujuan pembelajaran, serta karakteristik peserta didik.

Materi yang dianalisis dalam penelitian ini berfokus pada topik sistem reproduksi kelas IX. Pemilihan materi tersebut didasarkan pada karakteristiknya yang konseptual dan melibatkan proses biologis yang bersifat abstrak. Penelitian sebelumnya menunjukkan bahwa siswa SMP masih mengalami kesulitan dan miskonsepsi pada materi sistem reproduksi manusia [19]. Dengan demikian, analisis kualitas butir soal pada materi ini menjadi penting untuk memastikan bahwa instrumen yang digunakan benar-benar mampu merepresentasikan kompetensi yang diharapkan.

Penelitian ini dilakukan di salah satu SMP Negeri di Kota Pasuruan. Berdasarkan hasil wawancara dengan guru mata pelajaran, penyusunan soal ujian di sekolah tersebut disesuaikan dengan kondisi dan kemampuan siswa, termasuk siswa inklusi. Penyesuaian ini bertujuan agar seluruh siswa tetap dapat mengikuti evaluasi sesuai dengan kebutuhannya masing-masing. Namun, adanya perbedaan bentuk soal yang diberikan kepada siswa berdampak pada hasil analisis butir soal. Hal ini menyebabkan kualitas instrumen yang dihasilkan belum sepenuhnya menunjukkan kondisi yang ideal. Situasi tersebut menunjukkan bahwa praktik penilaian di sekolah masih perlu ditinjau kembali secara lebih terarah. Instrumen evaluasi seharusnya mampu menggambarkan kemampuan siswa secara adil dan objektif, tanpa dipengaruhi oleh perbedaan perlakuan yang terlalu jauh dalam bentuk soal. Jika analisis butir soal tidak dilakukan secara sistematis, maka hasil evaluasi berpotensi kurang akurat dalam mencerminkan kemampuan sebenarnya. Selain itu, variasi dalam penyusunan soal yang belum melalui analisis mendalam juga dapat memengaruhi tingkat kevalidan dan keandalan instrumen. Oleh karena itu, diperlukan kajian yang lebih menyeluruh untuk memastikan bahwa setiap butir soal benar-benar sesuai dengan kompetensi yang hendak diukur.

Meskipun beberapa penelitian sebelumnya telah mengkaji kualitas butir soal IPA menggunakan Model Rasch, sebagian besar penelitian tersebut menggunakan instrumen yang dikembangkan sendiri oleh peneliti atau dilakukan pada sekolah dengan karakteristik yang berbeda. Hal ini menunjukkan bahwa analisis pada konteks sekolah yang menjadi lokasi penelitian ini masih relevan dan perlu dilakukan. Penelitian ini memiliki kebaruan pada penerapan Model Rasch untuk menganalisis instrumen sumatif berbasis *Computer Based Test* (CBT) yang benar-benar digunakan dalam praktik evaluasi di salah satu SMP Negeri di Kota Pasuruan.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menganalisis kualitas butir soal sumatif mata pelajaran IPA kelas IX SMP menggunakan Model Rasch. Secara khusus, penelitian ini bertujuan untuk mengetahui tingkat kesukaran butir soal, kesesuaian (*fit*) butir dan responden, validitas, serta reliabilitas instrumen berdasarkan parameter dalam Model Rasch. Hasil analisis dalam penelitian ini diharapkan dapat digunakan sebagai bahan pertimbangan dalam memperbaiki dan mengembangkan instrumen evaluasi pembelajaran agar kualitasnya semakin baik. Dengan adanya analisis yang lebih terarah, soal-soal yang digunakan diharapkan dapat mengukur kemampuan siswa secara lebih tepat dan konsisten.

Penelitian ini juga diharapkan memberi manfaat secara praktis, khususnya bagi guru dan pihak sekolah, sebagai bahan refleksi dalam menyusun dan menelaah soal secara lebih sistematis. Selain itu, temuan penelitian ini dapat dimanfaatkan oleh komunitas guru atau MGMP sebagai referensi dalam merancang kegiatan pelatihan maupun pendampingan yang berkaitan dengan penyusunan dan analisis butir soal, sehingga kualitas penilaian pembelajaran dapat terus ditingkatkan. Secara akademis, penelitian ini juga dapat menjadi rujukan bagi kajian lanjutan untuk memperluas analisis pada mata pelajaran atau konteks sekolah yang berbeda termasuk pengembangan perangkat analisis butir soal berbasis teknologi yang lebih aplikatif.

II. METODE

Penelitian ini menggunakan pendekatan kuantitatif deskriptif dengan metode studi dokumentasi yang bertujuan untuk mendeskripsikan kualitas butir soal Sumatif mata pelajaran IPA kelas IX berdasarkan analisis Model Rasch dengan bantuan *software Ministep* versi 4.8.2.9 (*Evaluation/Student Version*).

Data penelitian berupa dokumen soal formatif, kisi-kisi, dan hasil jawaban siswa kelas IX yang diperoleh secara *purposive sampling*, dengan kategori telah mengikuti ujian sumatif dan memiliki data hasil belajar lengkap. Jumlah responden dalam penelitian ini sebanyak 70 siswa. Jumlah tersebut dinilai memadai untuk analisis menggunakan Model Rasch karena pendekatan ini tidak mensyaratkan ukuran sampel yang sangat besar seperti pada analisis klasik. Estimasi parameter butir dan kemampuan responden dalam Model Rasch dilakukan secara simultan melalui

transformasi logit sehingga tetap dapat menghasilkan estimasi yang stabil pada ukuran sampel sedang. Boone, et.al. menjelaskan bahwa analisis Rasch masih dapat memberikan estimasi parameter yang akurat pada rentang sampel 30-100 responden, khususnya dalam penelitian instrumen pendidikan skala kelas[20]. Oleh karena itu, jumlah responden 70 siswa dalam penelitian ini dinilai telah memenuhi kecukupan analisis untuk memperoleh estimasi parameter yang stabil dan dapat diinterpretasikan secara ilmiah. Analisis butir soal dilakukan terhadap 25 butir soal pilihan ganda dengan skor dikotomis (1 untuk jawaban benar dan 0 untuk jawaban salah). Pelaksanaan ujian Sumatif tersebut dilakukan melalui *computer based test* atau tes berbasis komputer dengan jumlah butir soal yang diujikan berjumlah 25 soal. Distribusi materi dari soal tersebut tercantum dalam Tabel 1.

Tabel 1. Sebaran Materi

Materi	Submateri	Item (No Soal)
Perkembangbiakan hewan dan tumbuhan	Tujuan reproduksi	1, 3
	Organ reproduksi	2, 4, 5
	Proses penyerbukan tumbuhan	6, 7, 8, 9, 10
	Penyebaran biji	11, 12, 13, 14, 15
	Metagenesis lumut	16, 17, 18
	Metagenesis paku	19, 20
	Fertilisasi hewan (internal & eksternal)	21, 22
	Cara reproduksi pada hewan (ovovivipar, tunas, metamorfosis)	23, 24, 25

Data yang diperoleh dari hasil jawaban siswa dianalisis menggunakan pendekatan Model Rasch. Model ini pertama kali diperkenalkan oleh Georg Rasch dan menggunakan fungsi logaritmik untuk mengubah data ordinal menjadi skala interval, sehingga hasil pengukurannya bersifat lebih objektif dan akurat[21].

Analisis validitas butir soal pada penelitian ini dilakukan dengan menggunakan tiga indikator dalam Model Rasch, yaitu nilai *Outfit Mean Square (MNSQ)*, *Outfit Z-Standard (ZSTD)*, dan *Point Measure Correlation (Pt Measure Corr)*. Ketiga indikator tersebut digunakan untuk melihat sejauh mana setiap butir soal sesuai dengan model pengukuran yang digunakan. Penelitian ini memusatkan perhatian pada tiga aspek utama, yakni validitas, reliabilitas, dan tingkat kesukaran soal. Ketiga aspek tersebut dipilih karena dianggap mewakili kualitas instrumen secara menyeluruh dalam pendekatan Model Rasch. Berbeda dengan teori tes klasik (*Classical Test Theory/CTT*), Model Rasch memiliki karakteristik tersendiri dalam menganalisis hubungan antara kemampuan peserta didik dan tingkat kesulitan butir soal, sehingga aspek-aspek tersebut menjadi penting untuk dikaji. Dalam Model Rasch, daya pembeda tidak dianalisis secara terpisah karena informasi mengenai daya pembeda sudah terintegrasi dalam analisis validitas melalui statistik *fit* (*Outfit MNSQ*, *Outfit ZSTD*, dan *Pt Measure Corr*). Model Rasch mengasumsikan bahwa semua butir soal yang *fit* dengan model memiliki daya pembeda yang ekuivalen, sehingga parameter diskriminasi dianggap konstan untuk seluruh butir[22][23]. Ketika suatu butir memiliki daya pembeda yang berbeda secara signifikan, hal tersebut akan terdeteksi melalui analisis *misfit* dan akan dikategorikan sebagai butir yang tidak valid[24][25]. Dengan kriteria kelayakan butir mengacu pada Sumintono dan Widhiarso, yaitu[21]:

Tabel 2. Indikator Validitas Butir Soal

No	Indikator	Rentang Ideal	Keterangan
1	<i>Outfit Mean Square (MNSQ)</i>	$0,5 < \text{MNSQ} < 1,5$	Menunjukkan kesesuaian antara respons dan model.
2	<i>Outfit Z-Standard (ZSTD)</i>	$-2,0 < \text{ZSTD} < +2,0$	Menggambarkan standar deviasi dari kesesuaian model.
3	<i>Point Measure Correlation (Pt Measure Corr)</i>	$0,4 < \text{Pt Measure Corr} < 0,85$	Menunjukkan korelasi antara butir dan kemampuan responden.

Soal yang tidak memenuhi salah satu kriteria tersebut dikategorikan sebagai misfit, sehingga tidak dianggap mewakili konstruk pengukuran yang dimaksud dan perlu dikaji ulang atau direvisi. Model Rasch dinilai lebih objektif dan unggul dibandingkan teori klasik karena tidak bergantung pada karakteristik kelompok responden tertentu[26]

Reliabilitas instrumen dianalisis menggunakan nilai *Cronbach's Alpha* (KR-20) yang dihasilkan dari *output Summary Statistic* pada perangkat lunak *Ministep*. Nilai reliabilitas mengukur konsistensi internal antarbutir soal dalam instrumen. Interpretasi nilai α mengacu pada kriteria Ernanda et al., pada tabel berikut[27]:

Tabel 3. Indikator Reliabilitas Instrumen

Nilai <i>Cronbach's Alpha</i>	Kategori
$\alpha \geq 0,80$	Istimewa
$0,70 \leq \alpha < 0,80$	Bagus Sekali
$0,60 \leq \alpha < 0,70$	Bagus
$0,50 \leq \alpha < 0,60$	Cukup
$\alpha < 0,50$	Lemah

Tingkat kesukaran butir soal ditentukan berdasarkan nilai *Measure (logit)* yang diperoleh dari analisis *Ministep* dan divisualisasikan melalui *Wright Map*. Nilai logit menunjukkan posisi relatif butir soal pada skala kesukaran: semakin tinggi nilainya, semakin sukar butir tersebut dijawab dengan benar. Kategori tingkat kesukaran mengacu pada Sumintono & Widhiarso sebagai berikut[21]:

Tabel 4. Indikator Tingkat Kesukaran Butir Soal

Nilai <i>Measure (Logit)</i>	Kategori
$Measure < -0,5 \text{ logit}$	Sangat Mudah
$-0,5 \text{ s.d} + 0,5 \text{ logit}$	Mudah
$+0,5 \text{ s.d} + 1,0 \text{ logit}$	Sukar
$Measure > 1,0 \text{ logit}$	Sangat Sukar

Dalam Model Rasch, semakin tinggi nilai logit suatu butir soal, maka akan semakin besar pula tingkat kesulitannya. Setiap butir memiliki nilai *measure* yang menunjukkan tingkat kesukaran tersebut, di mana nilai ini diperoleh dari hasil transformasi data ordinal (skor jawaban) menjadi skala logit dengan satuan interval yang setara. Berdasarkan kriteria yang telah ditetapkan, butir soal dikatakan valid apabila memenuhi *ketiga* kriteria berikut secara simultan: (1) nilai *Outfit MNSQ* berada pada rentang $0,5 < \text{MNSQ} < 1,5$; (2) nilai *Outfit ZSTD* berada pada rentang $-2,0 < \text{ZSTD} < +2,0$; dan (3) nilai *Pt Measure Corr* berada pada rentang $0,4 < x < 0,85$. Butir soal yang tidak memenuhi salah satu atau lebih dari kriteria tersebut dikategorikan sebagai *misfit* dan perlu direvisi atau dihapus dari instrumen pengukuran[21][28].

Instrumen dikatakan reliabel apabila nilai *Cronbach's Alpha* (KR-20) $\geq 0,70$ yang mengindikasikan konsistensi internal yang baik dan kemampuan instrumen untuk menghasilkan pengukuran yang stabil dan dapat dipercaya[29][30]. Nilai reliabilitas yang tinggi menunjukkan bahwa instrumen dapat membedakan kemampuan responden secara konsisten dan butir-butir soal saling berkorelasi dalam mengukur konstruk yang sama.

Distribusi tingkat kesukaran butir soal dikatakan seimbang dan berkualitas apabila butir-butir soal tersebar secara proporsional pada kategori mudah, sedang, dan sukar berdasarkan nilai *measure* (logit). Distribusi yang proporsional memastikan instrumen mampu mengukur kemampuan siswa pada berbagai tingkat kompetensi, sehingga menghasilkan pengukuran yang lebih presisi dan informatif[21][31].

III. HASIL DAN PEMBAHASAN

Penelitian ini difokuskan pada penilaian kualitas butir soal sumatif mata pelajaran IPA di salah satu SMP di Kota Pasuruan. Pengujian ini dilakukan terhadap 25 butir soal pilihan ganda menggunakan Rasch Model dengan bantuan *software Ministep*. Soal-soal tersebut mencakup materi sistem reproduksi pada tumbuhan dan hewan yang meliputi topik: (1) Reproduksi vegetatif dan generatif pada tumbuhan, (2) Penyerbukan dan

penyebaran biji, (3) Metagenesis tumbuhan lumut dan paku, serta (4) Fertilisasi dan metamorfosis hewan. Distribusi soal disajikan pada Tabel 1 di bagian Metode.

Berdasarkan *output Ministep* (lihat Lampiran 1), terdapat 5 item yang terindikasi tidak valid atau *misfit* yaitu item 5, item 10, item 12, item 20, dan item 21. Kelima item tersebut dinyatakan tidak sesuai dengan Model Rasch karena nilai *Outfit* MNSQ yang diperoleh berada di luar kriteria yang ditetapkan. Nilai *Outfit* item 5 dan item 12 juga tidak memenuhi kriteria ZSTD yang sudah ditentukan. Selain item 5, item 10, item 12, item 20, dan item 21, 20 item lainnya merupakan item yang *fit* sesuai dengan *Rasch Model*. Berdasarkan *Output* tersebut, maka hasil analisis validitas butir soal menggunakan *Rasch Model* dapat dibagi sebagai berikut:

Tabel 6. Hasil Analisis Validitas Butir Soal

Kriteria	Item
<i>Valid</i>	1, 2, 3, 4, 6, 7, 8, 9, 11, 13, 14, 15, 16, 17, 18, 19, 22, 23, 24, 25
<i>Invalid/Revised</i>	5, 10, 12, 20, 21

Berdasarkan Tabel 6, dari 25 butir soal yang dianalisis, sebanyak 20 butir dinyatakan *fit* atau *valid*, sedangkan 5 butir terindikasi *misfit*, yaitu butir nomor 5, 10, 12, 20, dan 21. Kelima butir tersebut dinyatakan *misfit* karena memiliki nilai *Outfit* MNSQ yang melebihi batas ideal 0,5–1,5. Secara spesifik, butir nomor 5 memiliki nilai *Outfit* MNSQ sebesar 4,92 dan butir nomor 12 sebesar 5,49, yang menunjukkan *underfit* berat atau pola respons yang sangat tidak konsisten dengan ekspektasi model Rasch. Sementara itu, butir nomor 10 (1,54), butir nomor 20 (1,59), dan butir nomor 21 (1,62) juga melebihi batas toleransi, sehingga termasuk kategori *underfit*. Kondisi ini mengindikasikan bahwa respons siswa terhadap butir-butir tersebut tidak sepenuhnya sejalan dengan pola probabilistik yang diharapkan berdasarkan perbedaan kemampuan siswa dan tingkat kesukaran soal. Reliabilitas instrumen soal dalam *software Ministep* diuji menggunakan *Cronbach's Alpha* (lihat Lampiran 2). Berdasarkan *output* tersebut, nilai *Cronbach's Alpha* (KR-20) yang diperoleh adalah 0,89. Menurut kriteria Ernanda et al., nilai tersebut termasuk dalam kategori istimewa ($0,8 < \alpha$), yang menunjukkan reliabilitas antarbutir instrumen sangat tinggi [27]. Tingkat kesukaran butir soal dianalisis berdasarkan nilai *measure* (*logit*) pada *Wright Map* yang dihasilkan oleh *ministep* (lihat Lampiran 3).

Berdasarkan nilai *measure* *logit* yang diperoleh dari *output* analisis (lihat Lampiran 4), hasil pengelempokkan tingkat kesukaran butir soal dirangkum dan disajikan pada Tabel 7.

Tabel 7. Rekap Hasil Analisis Tingkat Kesukaran Butir Soal

Kriteria	No Soal
Sangat Sukar	2, 3, 5, 21
Sukar	1, 8, 10, 11, 14, 15, 17, 19, 20, 24, 25
Mudah	4, 9, 12, 16, 22, 23
Sangat Mudah	6, 7, 13, 18

Berdasarkan Tabel 7, distribusi tingkat kesukaran butir soal menunjukkan variasi yang cukup luas. Dari 25 butir soal, terdapat 4 butir yang tergolong sangat sukar (2, 3, 5, 21), 11 butir tergolong sukar (1, 8, 10, 11, 14, 15, 17, 19, 20, 24, 25), 6 butir tergolong mudah (4, 9, 12, 16, 22, 23), dan 4 butir tergolong sangat mudah (13, 14, 18). Kategori tingkat kesukaran ini mengacu pada kriteria Sumintono dan Widhiarso yang membagi tingkat kesukaran menjadi empat kategori berdasarkan nilai *measure* *logit*: sangat sukar ($> +1,0$ *logit*), sukar ($+0,5$ s.d. $+1,0$ *logit*), mudah ($-0,5$ s.d. $+0,5$ *logit*), dan sangat mudah ($< -0,5$ *logit*). Distribusi ini menunjukkan bahwa instrumen memiliki variasi yang baik dalam mengukur kemampuan siswa pada berbagai tingkat kompetensi.

Tabel 8. Justifikasi Analisis Kualitas Butir Soal

Aspek	Item/Butir Soal	Kesimpulan
Validitas	Valid: 1, 2, 3, 4, 6, 7, 8, 9, 11, 13, 14, 15, 16, 17, 18, 19, 22, 23, 24, 25 Tidak Valid: 5, 10, 12, 20, 21	20 butir valid dan layak digunakan 5 butir perlu direvisi atau dihapus

Reliabilitas	Semua 25 butir soal	<i>Cronbach's Alpha</i> = 0,89 (Sangat Tinggi). Instrumen reliabel dan konsisten
Tingkat Kesukaran	Sangat Sukar: 2, 3, 5, 21 (4 butir) Sukar: 1, 8, 10, 11, 14, 15, 17, 19, 20, 24, 25 (11 butir) Mudah: 4, 9, 12, 16, 22, 23 (6 butir) Sangat Mudah: 6, 7, 13, 18 (4 butir)	Distribusi bervariasi dengan baik dan dapat mengukur berbagai tingkat kemampuan siswa

Berdasarkan Tabel 8, dapat disimpulkan bahwa dari hasil analisis kualitas butir soal, terdapat 20 butir soal yang layak digunakan, yaitu butir nomor 1, 2, 3, 4, 6, 7, 8, 9, 11, 13, 14, 15, 16, 17, 18, 19, 22, 23, 24, dan 25. Butir-butir soal ini telah memenuhi kriteria validitas dengan nilai *fit statistics* yang berada dalam rentang yang diterima, memiliki distribusi tingkat kesukaran bervariasi, dan didukung oleh nilai reliabilitas instrumen yang sangat tinggi ($\alpha = 0,89$). Dengan demikian, ke-20 butir soal ini dapat digunakan untuk mengukur pemahaman siswa mengenai materi sistem reproduksi pada tumbuhan dan hewan secara akurat dan konsisten.

Terdapat 5 butir soal yang perlu direvisi, yaitu butir nomor 5, 10, 12, 20, dan 21, karena tidak memenuhi kriteria *fit statistics* yang telah ditentukan. Nilai *Outfit* MNSQ pada kelima butir tersebut berada di atas batas toleransi, yang menunjukkan adanya ketidakteraturan pola jawaban siswa (*underfit*). Kondisi ini dapat dipengaruhi oleh beberapa faktor, seperti redaksi soal yang kurang jelas, penggunaan istilah teknis yang belum dipahami siswa, pengecoh yang kurang berfungsi optimal, maupun tingkat kesukaran yang tidak proporsional dengan kemampuan siswa.

Berdasarkan hasil analisis menggunakan Model Rasch, diperoleh hasil dari 25 butir soal yang dianalisis, sebanyak 20 butir soal dinyatakan valid. Sedangkan 5 butir soal lainnya (soal nomor 5, 10, 12, 20, dan 21) teridentifikasi tidak valid atau *misfit*. Hasil analisis menunjukkan bahwa sebagian besar butir soal telah sesuai dengan model pengukuran Rasch, yang berarti pola jawaban siswa selaras dengan tingkat kemampuan yang diukur oleh masing-masing butir soal. Namun demikian, terdapat lima butir soal yang dinyatakan tidak valid karena menunjukkan ketidaksesuaian antara data empiris dengan model teoritis, sehingga butir-butir tersebut memerlukan revisi. Penentuan validitas butir soal dalam penelitian ini didasarkan pada nilai *Outfit* MNSQ, *ZSTD*, dan *Point Measure Correlation* (*Pt Measure Corr*) sebagai indikator utama. Butir 5 dan 12 menunjukkan nilai *Outfit* MNSQ yang melebihi batas kriteria, mengindikasikan ketidaksesuaian antara respons siswa dengan pola kemampuan yang diharapkan. Kondisi ini diduga disebabkan oleh redaksi soal yang kurang jelas, pengecoh yang belum berfungsi optimal, atau tingkat kesukaran yang terlalu tinggi. Kedua butir ini memerlukan revisi agar dapat memberikan hasil pengukuran yang lebih akurat.

Kemudian, hasil analisis reliabilitas menunjukkan bahwa nilai *Cronbach's Alpha* sebesar 0,89 yang mana angka ini termasuk dalam kategori sangat tinggi. Nilai ini mengindikasikan bahwa antarbutir dalam instrumen memiliki konsistensi yang kuat dan mampu mengukur kemampuan siswa secara stabil. Dengan demikian, secara keseluruhan, instrumen soal ini sudah cukup reliabel dan dapat dipercaya dalam menilai kemampuan siswa dalam mata pelajaran IPA.

Berdasarkan *output Wright Map* (lihat Lampiran 3), kemampuan peserta tes tersebar pada rentang logit -3 hingga +6, dengan nilai rata-rata kemampuan berada pada logit +1. Mayoritas peserta berada pada kategori kemampuan sedang, yaitu pada rentang logit 0 hingga 2. Selain itu, terdapat beberapa peserta dengan kemampuan sangat tinggi hingga mencapai logit +6, serta sebagian kecil peserta dengan kemampuan rendah pada kisaran logit -2 hingga -3. Di sisi lain, tingkat kesulitan butir soal tersebar pada rentang logit -2 hingga +5. Butir soal dengan tingkat kesulitan tertinggi adalah S5 yang berada pada logit +5, sedangkan butir termudah adalah S13 dengan nilai logit -2. Rata-rata kesulitan butir berada pada logit 0. Distribusi ini menunjukkan bahwa tes dapat mencakup rentang kemampuan siswa dengan cukup baik. Namun, terdapat sejumlah siswa dengan kemampuan tinggi yang melebihi tingkat kesulitan butir tertinggi. Sehingga, tes ini dapat ditingkatkan dengan menambahkan beberapa butir yang lebih sulit untuk mengukur kemampuan kelompok tersebut secara lebih akurat.

Selanjutnya, hasil analisis tingkat kesukaran menunjukkan bahwa butir-butir soal memiliki variasi yang cukup merata dengan mempertimbangkan karakteristik materi sistem reproduksi pada tumbuhan dan hewan. Materi reproduksi tumbuhan cenderung dipersepsikan lebih sulit oleh siswa karena memuat konsep-konsep yang bersifat abstrak, seperti metagenesis, berbagai bentuk penyerbukan, serta proses penyebaran biji yang menuntut pemahaman menyeluruh mengenai siklus hidup tumbuhan. Di sisi lain, materi reproduksi hewan cenderung lebih mudah dipahami oleh siswa karena contohnya lebih nyata dan dekat dengan kehidupan sehari-hari. Misalnya, pembahasan tentang metamorfosis pada serangga atau proses perkembangbiakan pada hewan vertebrata yang sering mereka temui dalam pembelajaran maupun lingkungan sekitar. Hal ini membuat siswa

lebih mudah membayangkan dan memahami konsep yang disampaikan. Secara keseluruhan, tingkat kesukaran butir soal dalam penelitian ini cukup beragam, mulai dari soal yang sangat mudah hingga yang tergolong sangat sulit. Adanya variasi tersebut menunjukkan bahwa instrumen evaluasi tidak hanya mengukur siswa dengan kemampuan tertentu saja, tetapi telah mencakup berbagai tingkat kemampuan yang dimiliki siswa. Meskipun demikian, masih terdapat beberapa butir soal dengan tingkat kesukaran yang berada pada kategori ekstrem, baik terlalu mudah maupun terlalu sukar. Kondisi tersebut perlu diperhatikan dengan baik agar tingkat kesukaran soal menjadi lebih seimbang dan mampu merepresentasikan kemampuan siswa secara lebih proporsional. Untuk memahami lebih dalam tentang kualitas butir soal yang diteliti, perlu dikaji lebih lanjut mengenai butir-butir yang terindikasi *misfit* dan faktor-faktor yang menyebabkan ketidaksesuaian tersebut. Butir soal nomor 5 yang membahas cara perkembangbiakan tumbuhan memiliki nilai *Outfit* MNSQ sebesar 4,92. Nilai tersebut menunjukkan adanya ketidaksesuaian yang cukup tinggi dengan model Rasch, sehingga respons siswa pada butir ini dinilai sangat tidak konsisten. Kondisi ini bisa jadi dipengaruhi oleh pilihan jawaban yang kurang jelas atau adanya konsep yang serupa sehingga membingungkan siswa. Mengingat materi reproduksi tumbuhan bersifat cukup konseptual dan tidak selalu dapat diamati secara langsung, tidak menutup kemungkinan siswa mengalami miskonsepsi atau perbedaan penafsiran terhadap istilah yang digunakan dalam soal.

Pada butir nomor 10 yang membahas simbiosis mutualisme antara lebah dan bunga, diperoleh nilai *Outfit* MNSQ sebesar 1,54 sedikit melewati batas ideal. Hal ini menunjukkan adanya ketidaksesuaian ringan antara pola jawaban siswa dengan prediksi model. Walaupun konsep simbiosis tergolong konkret dan dekat dengan kehidupan sehari-hari, kemungkinan terdapat redaksi soal atau pilihan pengecoh yang kurang optimal, sehingga belum sepenuhnya mampu membedakan siswa dengan kemampuan tinggi dan rendah. Sementara itu, butir nomor 12 mengenai penyebaran biji secara mamokori memiliki nilai *Outfit* MNSQ sebesar 5,49 yang termasuk dalam kategori ketidaksesuaian berat. Nilai ini mengindikasikan bahwa respons siswa pada butir tersebut cukup menyimpang dari pola yang diharapkan oleh model, sehingga butir ini perlu ditelaah kembali, baik dari segi kejelasan materi maupun konstruksi soalnya.

Nilai ini menunjukkan respons yang sangat tidak terprediksi. Konsep mamokori merupakan istilah teknis yang cukup spesifik dan mungkin kurang familiar bagi siswa. Selain itu, materi penyebaran biji menuntut pemahaman klasifikasi jenis-jenis dispersal secara tepat, sehingga jika redaksi kurang memberikan konteks yang jelas, siswa dapat menafsirkannya secara berbeda-beda.

Butir nomor 20 tentang metagenesis tumbuhan paku memiliki nilai *Outfit* MNSQ sebesar 1,59. Konsep metagenesis tergolong cukup abstrak karena menjelaskan adanya pergiliran keturunan antara fase gametofit dan sporofit. Proses tersebut tidak dapat diamati secara langsung oleh siswa, sehingga pemahamannya sangat bergantung pada penjelasan dan kemampuan membayangkan alur siklusnya. Tingkat kompleksitas ini berpotensi membuat jawaban siswa menjadi kurang konsisten, terutama jika mereka belum benar-benar memahami hubungan antara tahap dalam siklus tersebut. Sementara itu, butir nomor 21 yang membahas alat reproduksi vegetatif pada tumbuhan paku memperoleh nilai *Outfit* MNSQ sebesar 1,62. Nilai ini menunjukkan adanya ketidaksesuaian yang perlu diperhatikan. Hal tersebut kemungkinan dipengaruhi oleh tingkat kesulitan materi yang cukup tinggi, serta adanya pilihan jawaban yang belum sepenuhnya efektif dalam membedakan siswa berdasarkan tingkat kemampuannya. Secara keseluruhan, kelima butir tersebut menunjukkan pola respons yang tidak konsisten dengan model Rasch sehingga perlu direvisi dari aspek redaksi, kejelasan istilah, serta efektivitas pengecoh agar lebih representatif dalam mengukur pemahaman siswa terhadap materi reproduksi tumbuhan dan hewan.

Keberadaan butir-butir soal yang tergolong misfit memberikan masukan yang bernilai dalam upaya pengembangan instrumen evaluasi. Beberapa aspek yang perlu mendapat perhatian meliputi ketepatan penyusunan pengecoh, kesesuaian tingkat kesukaran soal dengan kemampuan siswa, serta kejelasan penggunaan istilah teknis dalam soal. Perbaikan pada aspek-aspek tersebut diharapkan mampu meningkatkan ketepatan pengukuran pemahaman siswa terhadap materi sistem reproduksi tumbuhan dan hewan. Secara umum, instrumen evaluasi sumatif mata pelajaran IPA di SMP X Kota Pasuruan dapat dikatakan telah memiliki kualitas yang baik, yang ditunjukkan oleh dominasi butir soal yang valid, reliabel, serta memiliki variasi tingkat kesukaran yang memadai. Meskipun demikian, beberapa butir soal yang tidak valid dan berada pada tingkat kesukaran ekstrem tetap memerlukan revisi agar instrumen evaluasi dapat memberikan hasil pengukuran yang lebih akurat, adil, dan konsisten dalam menilai hasil belajar siswa.

VII. KESIMPULAN

Berdasarkan hasil analisis menggunakan Model Rasch terhadap 25 butir soal sumatif mata pelajaran IPA kelas IX di salah satu SMP Negeri Kota Pasuruan, instrumen evaluasi yang digunakan secara keseluruhan

menunjukkan kualitas yang baik. Ditinjau dari aspek validitas, sebagian besar butir soal yang dianalisis telah memenuhi kriteria kesesuaian dengan mode dan dinyatakan layak digunakan, meskipun masih terdapat beberapa butir yang perlu diperbaiki karena menunjukkan ketidaksesuaian. Instrumen evaluasi yang digunakan juga menunjukkan tingkat reliabilitas yang sangat baik, yang mencerminkan konsistensi internal dalam mengukur kemampuan siswa. Selain itu, tingkat kesukaran butir soal memperlihatkan sebaran yang cukup beragam sehingga mampu menjangkau kemampuan siswa pada berbagai tingkat, walaupun masih ditemukan sejumlah butir dengan tingkat kesukaran yang berada pada kategori ekstrem dan perlu diseimbangkan kembali. Penerapan Model Rasch dalam penelitian ini menunjukkan kemampuannya dalam memberikan informasi yang bersifat objektif mengenai kualitas instrumen evaluasi, tanpa bergantung pada karakteristik populasi responden. Temuan dalam penelitian ini menunjukkan bahwa analisis butir soal secara sistematis memang perlu dilakukan sebagai bagian dari upaya memperbaiki kualitas evaluasi pembelajaran. Hal ini menjadi semakin penting pada mata pelajaran IPA, karena materi yang dipelajari banyak menuntut pemahaman konsep secara mendalam, bukan sekedar hafalan.

Hasil penelitian ini juga memberikan gambaran yang bisa dimanfaatkan oleh pihak sekolah dalam meningkatkan kualitas instrumen evaluasi. Butir soal yang teridentifikasi *misfit* serta penyebaran tingkat kesukaran yang belum sepenuhnya seimbang dapat dijadikan bahan pertimbangan untuk melakukan perbaikan atau penyusunan ulang soal agar lebih sesuai dengan kompetensi yang ingin diukur. Selain itu, informasi mengenai reliabilitas instrumen dapat digunakan sebagai dasar dalam menyusun bank soal yang lebih terarah dan memiliki standar kualitas yang lebih baik. Dengan langkah tersebut, kualitas asesmen IPA di sekolah diharapkan dapat terus ditingkatkan secara bertahap dan berkelanjutan.

Berdasarkan hasil penelitian ini, penelitian selanjutnya disarankan melibatkan jumlah sampel yang lebih luas, misalnya dengan menganalisis butir soal dari beberapa sekolah atau wilayah yang berbeda. Hal ini bertujuan agar diperoleh gambaran kualitas instrumen revaluasi IPA yang lebih mewakili kondisi pada jenjang SMP. Selain itu, penelitian berikutnya juga dapat mempertimbangkan penggunaan pendekatan kualitatif, seperti wawancara dengan siswa maupun tenaga pendidik, untuk memperdalam pemahaman terhadap butir soal yang teridentifikasi *misfit* dalam analisis Model Rasch. Melalui pendekatan tersebut, faktor-faktor yang memengaruhi ketidaksesuaian butir, baik dari segi materi, bahasa, maupun fungsi pengecoh, dapat dikaji secara lebih mendalam sehingga hasil analisis menjadi lebih menyeluruh. Pendekatan lanjutan juga dapat mengeksplorasi potensi bias butir soal terhadap kelompok tertentu melalui analisis *Differential Item Functioning* (DIF) untuk mengidentifikasi potensi bias butir soal terhadap kelompok tertentu serta memastikan keadilan dalam pengukuran. Di samping itu, kajian serupa pada mata pelajaran lain dan pengembangan bank soal berbasis Model Rasch yang terstandarisasi dapat dilakukan untuk mendukung peningkatan mutu evaluasi pembelajaran secara berkelanjutan.

UCAPAN TERIMA KASIH

Penulis menyampaikan apresiasi kepada berbagai pihak yang telah memberikan dukungan selama pelaksanaan penelitian ini. Terima kasih kepada Kepala Sekolah dan guru mata pelajaran IPA di SMP Negeri Kota Pasuruan yang telah memberikan izin dan kerja sama dalam pengumpulan data penelitian. Ucapan terima kasih juga disampaikan kepada Program Studi Pendidikan IPA Universitas Muhammadiyah Sidoarjo atas dukungan akademik dan fasilitas yang diberikan. Penghargaan khusus penulis sampaikan kepada dosen pembimbing yang telah memberikan arahan, bimbingan, dan masukan konstruktif selama proses penelitian. Penulis juga mengucapkan terima kasih kepada rekan-rekan mahasiswa yang telah memberikan dukungan moral dan diskusi yang bermanfaat. Semoga hasil penelitian ini dapat memberikan kontribusi positif bagi peningkatan kualitas evaluasi pembelajaran IPA di Indonesia.

REFERENSI

- [1] Fitriani, D. (2024). Validitas dan Reliabilitas Soal Ujian Sumatif Tengah Semester Kelas X SMA Negeri 8 Palembang. Alfabeta: Jurnal Pendidikan dan Pembelajaran, Universitas Islam Batanghari.
- [2] Hartono, N., Ningsih, A. W., & Tenriawaru, A. (2024). Analisis Butir Soal Sumatif IPA Kelas VII SMP Negeri 3 Pontianak dengan Rasch Model. Bioscientist: Jurnal Ilmiah Pendidikan Biologi, e-Journal Undikma.
- [3] Widyaningsih, E., dkk. (2023). Analisis Kualitas Butir Soal Fisika pada Penilaian Akhir Semester Gasal Kelas X IPA SMA YPK Oikoumene Manokwari. Al-Khazini: Jurnal Pendidikan Fisika, UIN Alauddin Makassar.
- [4] Purwati, H., Retnawati, H., Jailani, J., & Retnowati, T. H. (2021). Analisis karakteristik butir soal Ujian Nasional Matematika SMP/MTs ditinjau dari teori tes klasik. Jurnal Penelitian Pendidikan Matematika dan Sains, 7(1), 15–26. <https://ejournal.uksw.edu/juses/article/view/3820>

- [5] Ramadhani, N. (2020). Analisis Butir Soal Ulangan Harian Pada Mata Pelajaran Bahasa Indonesia Semester Genap Tahun Ajaran 2018/2019 Di Sd Negeri 7 Jagong Jeget (*Doctoral dissertation*, Universitas Bina Bangsa Getsempena).
- [6] Ernawati, E., Habibah, R. Y., Syarifah, N., Firmansyah, F., & Attamimi, H. A. R. (2024). Item analysis test of science, Indonesian language, and mathematics using the rasch model in elementary schools. *Jurnal Penelitian dan Evaluasi Pendidikan*, 28(2), 195-209.
- [7] Wardhani, D. F., Suratno, A. P. P., & Putra, A. P. (2019). Analisis Model Fit Butir Soal pada Instrumen Penilaian Kognitif Ilmu Pendidikan Alam (IPA) Sekolah Menengah Pertama. *JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)*, 8(2), 105-117.
- [8] Huda, N., & Wahyuni, T. S. (2019). Analisis butir soal IPA Try Out USBN Tahun Ajaran 2018/2019 dalam kaitannya dengan level kognitif. *MADRASAH: Jurnal Pendidikan dan Pembelajaran Islam*, 4(2), 1–10. <https://doi.org/10.30829/madrasah.v4i2.1234>
- [9] Kusumadani, A. I., & Andira, D. M. (2019). Pelaksanaan praktikum IPA berdasarkan evaluasi guru dan siswa di SMP Kabupaten Sragen Tahun Ajaran 2018/2019. *Jurnal Pendidikan IPA Indonesia*, 8(1), 45–53. <https://doi.org/10.15294/jpii.v8i1.17890>
- [10] Nurani, N. S. (2024). Analisis Soal Penilaian Tengah Semester Ipa Kelas IX Dengan Rasch Model. *Jurnal Pembelajaran, Bimbingan, dan Pengelolaan Pendidikan*, 4(4), 17-17.
- [11] Sari, I. N., Anggraini, M. P., Maryoni, R., & Walid, A. (2023). Item Analysis of Mid-Term Science Examination Questions in Junior Secondary Education: Evidence from Indonesia. *ISEJ: Indonesian Science Education Journal*, 4(2), 46-53.
- [12] Nisfatulsanah, A., & Sugiharto, B. (2023). Comparing Classical Test Theory and Item Response Theory for Respiratory System Instruments Menggunakan Rasch. *PIONIR*.
- [13] Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.
- [14] El Hasna, A. (2022). Analisis Pengukuran Hasil Ulangan Menggunakan Teori Tes Klasik dan Model Rasch pada Siswa SMA. *Bio-Pedagogi*.
- [15] Tutz, G. (2023). *Unidimensionality in Rasch Models: Efficient Item Selection and Hierarchical Clustering*. arXiv.
- [16] Zeileis, A. (2024). *Examining Exams Using Rasch Models and Assessment of Measurement Invariance*. arXiv.
- [17] Hayat, B. (2022). Adjustment for guessing in a basic statistics test for Indonesian undergraduate psychology students using the Rasch model. *Cogent Education*, 9(1), 2059044.
- [18] Amaliyah, M., Suardana, I. N., & Selamat, K. (2021). Analisis kesulitan belajar dan faktor-faktor penyebab kesulitan belajar IPA siswa SMP Negeri 4 Singaraja. *Jurnal Pendidikan Dan Pembelajaran Sains Indonesia (JPPSI)*, 4(1), 90-101
- [19] Tyas, R. A. (2023). Analisis Miskonsepsi Siswa SMP Materi Sistem Reproduksi Manusia berdasarkan Jenis Kelamin. *Jurnal Ilmiah WUNY*, 5(2).
- [20] Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences* (Vol. 10, pp. 978-94). Dordrecht: Springer.
- [21] Sumintono, B., & Widhiarso, W. (2015). Aplikasi pemodelan rasch pada assessment pendidikan. *Trim komunikasi*
- [22] Linacre, J. M. (2005). A user's guide to WINSTEPS & MINISTEP Rasch-model computer programs. *Program manual*. Chicago, IL: Winsteps. com..
- [23] Debelak, R., Strobl, C., & Zeigenfuse, M. D. (2022). *An introduction to the Rasch model with examples in R*. Chapman and Hall/CRC..
- [24] Wu, M., & Adams, R. J. (2013). *Properties of Rasch residual fit statistics*. *Journal of Applied Measurement*, 14(4), 339-355..
- [25] Tesio, L., Caronni, A., Kumbhare, D., & Scarano, S. (2024). *Interpreting results from Rasch analysis 1. The “most likely” measures coming from the model*. *Disability and rehabilitation*, 46(3), 591-603.. <https://doi.org/10.1080/09638288.2023.2169771>
- [26] Fauziana, A., Wulansari, A. D., & Kurniawati, S. D. N. (2021). Analisis Kualitas Butir Soal Ulangan Harian di Sekolah Dasar dengan Model Rasch. *Jurnal Evaluasi Pendidikan*, 6(1), 10–19
- [27] Ernanda, M. R. D., Rizqiah, S., Binta, S. A., Oktarisa, Y., & Sukma Suryana, T. G. (2024). Pengembangan instrumen tes diagnostik tingkat konsep pada materi dinamika. *Jurnal Nebula*, 1(2), 28-30. <https://jurnal.untirta.ac.id/index.php/nebula/>
- [28] Rosli, R., Abdullah, M., Siregar, N. C., Hamid, N. S. A., Abdullah, S., Beng, G. K., Halim, L., Daud, N. M., Bahari, S. A., Majid, R. A., & Bais, B. (2021). Student awareness of space science: Rasch model analysis for validity and reliability. *Journal of Turkish Science Education*, 18(4), 757-770.

- [29] Dunbar-Wallis, A. K., Katcher, J., Moore, W., & Corwin, L. A. (2024). Bee The CURE: Increasing Student Science Self-Efficacy, Science Identity, and Predictors of Scientific Civic Engagement in a Community College CURE. *CBE—Life Sciences Education*, 23(4), ar46.. <https://doi.org/10.1187/cbe.24-01-0015>
- [30] Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). Analisis kualitas soal kemampuan membedakan rangkaian seri dan paralel melalui teori tes klasik dan model rasch. *Indonesian Journal of Educational Research and Review*, 3(1), 11-19.
- [31] Dwi, M. R. (2024). Pengembangan Instrumen Tes Diagnostik Tingkat Konsepsi pada Materi Dinamika. *Nebula: Jurnal Penelitian dan Pembelajaran Fisika*, 1(2).

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.

