Inovasi Kurikulum *- p-ISSN 1829-6750 & e-ISSN 2798-1363*
Volume 22 No 1 (2025) 161-180

JURNAL UPI

Inovasi Kurikulum

https://ejournal.upi.edu/index.php/JIK

HIPKIN

Improving assessment quality: Development of evaluation instruments for sixth-grade Mathematics learning

Eny Cahyaningsih[1], Wardani Rahayu[2], Riyan Arthur[3]
[1,2,3] Univesitas Negeri Jakarta, Jakarta, Indonesia
cahyaningsiheny@gmail.com[1]

ABSTRACT

Valid and reliable assessment instruments are crucial for accurately measuring student competencies, yet many fail. This study aims to develop a Mathematics learning outcome assessment instrument based on the Kurikulum Merdeka for sixth-grade elementary school students, focusing on validity and reliability using the Rasch model. This instrument is designed to encompass cognitive, affective, and psychomotor competencies following the needs of the Kurikulum Merdeka. Analysis was conducted on 213 students using a quantitative approach with Winsteps software to evaluate item quality, unidimensionality, reliability, difficulty level, and potential differential item functioning (DIF). The research results indicate that the instrument is highly reliable and meets the unidimensionality criteria. The distribution of question difficulty levels varies from very easy to very difficult, reflecting the instrument's ability to measure students with a range of abilities. All items meet the fit criteria based on fit statistics (Outfit MNSQ, ZSTD, and Point Measure Correlation). However, two items (Item 1 and Item 5) show significant DIF bias based on gender analysis. This study concludes that this assessment instrument based on the Kurikulum Merdeka is valid, reliable, and suitable for assessing students' mathematical abilities.

ARTICLE INFO

Article History:
Received: 10 Sep 2024
Revised: 23 Dec 2024
Accepted: 27 Dec 2024
Available online: 5 Jan 2025
Publish: 28 Feb 2025

Keyword:
educational assessment; instrument evaluation; kurikulum merdeka; Rasch model; reliability; validity

Open access
Inovasi Kurikulum is a peer-reviewed open-access journal.

ABSTRAK

Instrumen penilaian yang valid dan reliabel sangat penting untuk mengukur kemampuan peserta didik secara objektif, namun banyak instrumen yang ada belum merepresentasikan kompetensi peserta didik dengan akurat. Penelitian ini bertujuan untuk mengembangkan instrumen penilaian hasil belajar matematika berbasis Kurikulum Merdeka untuk peserta didik kelas VI Sekolah Dasar, dengan fokus pada validitas dan reliabilitas menggunakan model Rasch. Analisis dilakukan pada 213 peserta didik menggunakan pendekatan kuantitatif dengan software Winsteps untuk mengevaluasi kualitas item, unidimensionalitas, reliabilitas, tingkat kesulitan, dan potensi bias diferensial (Differential Item Functioning/ DIF). Hasil penelitian menunjukkan bahwa instrumen memiliki reliabilitas tinggi dan memenuhi kriteria unidimensionalitas. Sebaran tingkat kesulitan soal cukup bervariasi dari sangat mudah hingga sangat sulit, mencerminkan kemampuan instrumen dalam mengukur peserta didik dengan beragam tingkat kemampuan. Semua item memenuhi kriteria kesesuaian berdasarkan statistik kecocokan (Outfit MNSQ, ZSTD, dan Point Measure Correlation), namun terdapat dua item (Item 1 dan Item 5) yang menunjukkan bias DIF signifikan berdasarkan analisis gender. Kesimpulan dari penelitian ini adalah bahwa instrumen penilaian berbasis Kurikulum Merdeka ini valid, reliabel, dan cocok untuk digunakan dalam menilai kemampuan matematika peserta didik.

Kata Kunci: *evaluasi instrumen; Kurikulum Merdeka; Model Rasch; penilaian pendidikan; reliabilitas; validitas*

How to cite (APA 7)
Cahyaningsih, E., Rahayu, W., & Arthur, R. (2025). Improving assessment quality: Development of evaluation instruments for sixth-grade Mathematics learning. *Inovasi Kurikulum, 22*(1), 161-180.

Peer review
This article has been peer-reviewed through the journal's standard double-blind peer review, where both the reviewers and authors are anonymised during review.

Copyright
2025, Eny Cahyaningsih, Wardani Rahayu, Riyan Arthur. This an open-access is article distributed under the terms of the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) https://creativecommons.org/licenses/by-sa/4.0/, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author, and source are credited. *Corresponding author: cahyaningsiheny@gmail.com

https://doi.org/10.17509/jik.v22i1.77242

INTRODUCTION

Penerapan Kurikulum Merdeka telah membawa perubahan signifikan dalam sistem pendidikan di Indonesia dengan memberikan keleluasaan kepada pendidik untuk merancang pembelajaran yang relevan dengan kebutuhan peserta didik. Penilaian menjadi bagian penting dalam sistem pembelajaran, yang harus dilakukan guru untuk mengukur efektivitas proses pembelajaran mereka. Proses evaluasi ini sangat krusial karena dapat memberikan gambaran komprehensif tentang penguasaan peserta didik terhadap topik tertentu, tantangan yang mereka hadapi saat belajar, dan posisi mereka dibandingkan dengan teman-teman sekelasnya (Safitri & Widyanti, 2024; Saputri *et al*., 2024). Dalam konteks mata pelajaran matematika, peserta didik kelas VI diharapkan mampu mengembangkan keterampilan berpikir logis dan analitis yang dapat diterapkan dalam kehidupan sehari-hari. Penilaian yang efektif menjadi kunci untuk mengukur dan meningkatkan keterampilan tersebut.

Studi menemukan bahwa meskipun instrumen penilaian kemampuan berpikir kritis di sekolah menengah memiliki reliabilitas tinggi (r = 0,89), sekitar 25% butir soal tidak valid, sehingga tidak sepenuhnya mengukur konstruk yang dimaksud (Juliani & Erita, 2023). Studi lainnya menyatakan bahwa instrumen penilaian rasa percaya diri peserta didik memiliki validitas dan reliabilitas yang memadai, namun beberapa butir pernyataan perlu direvisi untuk meningkatkan akurasi pengukuran (Maulana, 2022). Hasil penelitian ini menegaskan pentingnya pengembangan dan analisis instrumen penilaian yang cermat agar valid dan reliabel dalam mencerminkan kemampuan peserta didik secara akurat. Oleh karena itu, pengembangan instrumen penilaian berbasis Kurikulum Merdeka menjadi kebutuhan mendesak untuk memastikan kualitas pembelajaran dan evaluasi peserta didik. Instrumen penilaian yang baik harus mencakup aspek berpikir kritis, kreatif, dan kemampuan pemecahan masalah, yang terbukti tidak hanya meningkatkan hasil belajar tetapi juga memotivasi peserta didik.

Pendekatan model Rasch dalam pengembangan instrumen penilaian berbasis psikometri menawarkan metode yang komprehensif untuk mengevaluasi kualitas butir soal. Model ini memastikan bahwa instrumen yang dihasilkan valid, reliabel, dan sesuai dengan kebutuhan peserta didik (Tarigan *et al*., 2022). Penilaian sendiri merupakan elemen krusial dalam proses pembelajaran yang berfungsi untuk mengukur pencapaian peserta didik sekaligus memberikan umpan balik bagi guru dan peserta didik. Evaluasi terhadap validitas dan reliabilitas instrumen menjadi langkah penting untuk menjamin bahwa alat ukur dapat mengukur pencapaian peserta didik secara akurat. Validitas mengukur sejauh mana instrumen sesuai dengan tujuan pengukurannya, sedangkan reliabilitas memastikan konsistensi hasil yang diperoleh dalam kondisi serupa. Dengan demikian, pengembangan instrumen penilaian yang berkualitas tinggi akan mendukung peningkatan efektivitas pembelajaran serta akurasi evaluasi hasil belajar peserta didik.

Penggunaan model Rasch juga memungkinkan evaluasi butir soal secara mendalam, mencakup validitas dan unidimensionalitas (Abdullaev *et al*., 2024; Latifah *et al*., 2024; Yusuf *et al*., 2021). Melalui pendekatan ini, instrumen penilaian dapat disesuaikan dengan kemampuan peserta didik, menciptakan alat ukur yang lebih adil dan objektif. Penelitian mengenai analisis kualitas butir soal Ujian Tengah Semester (UTS) Biologi Umum dengan menggunakan model Rasch menunjukkan bahwa model Rasch mampu mengidentifikasi butir soal yang tidak sesuai (misfit) dan memberikan rekomendasi untuk meningkatkan efektivitas soal dalam mengukur pencapaian mahasiswa (Novriyanti & Arthur, 2024).

Psikometri menyediakan dasar teoritis yang kuat untuk pengembangan instrumen penilaian *berbasis Item Response Theory* (IRT), yang lebih adaptif dibandingkan dengan teori tes klasik (CTT). Penelitian terbaru menunjukkan bahwa instrumen berbasis IRT memiliki keunggulan dalam mengidentifikasi item yang kurang valid serta memberikan panduan penyempurnaan. IRT memungkinkan analisis mendalam terhadap karakteristik butir soal, seperti validitas, reliabilitas, tingkat kesulitan, dan daya beda, sehingga instrumen penilaian dapat dioptimalkan untuk mengukur kemampuan peserta didik secara akurat.

https://doi.org/10.17509/jik.v22i1.77242

Misalnya, penelitian lainnya menerapkan IRT model 2PL untuk menganalisis parameter item dan kemampuan peserta didik, tetapi hanya berfokus pada tingkat kesulitan dan daya beda soal tanpa mengevaluasi aspek reliabilitas, unidimensionalitas, atau bias DIF (Jumini & Retnawati, 2022). Selain itu, penelitian tersebut tidak mengembangkan instrumen baru. Berbeda dengan penelitian sebelumnya, penelitian ini lebih komprehensif karena mengembangkan instrumen baru berbasis model Rasch dan mencakup evaluasi bias DIF, unidimensionalitas, serta reliabilitas, sehingga memberikan hasil analisis yang lebih menyeluruh.

Penilaian berbasis konteks, yang melibatkan soal-soal relevan dengan kehidupan sehari-hari, terbukti efektif dalam meningkatkan motivasi peserta didik dan kemampuan berpikir kritis (Smith *et al*., 2022). Hal ini selaras dengan prinsip Kurikulum Merdeka yang menekankan aplikasi praktis dalam pembelajaran. Pendekatan ini tidak hanya meningkatkan keterlibatan peserta didik tetapi juga membantu mereka mengaitkan konsep abstrak dengan situasi nyata. Pada fase C, peserta didik kelas VI diharapkan mampu menguasai operasi bilangan, konsep pecahan, perbandingan, dan pengukuran, serta menerapkannya dalam konteks kehidupan nyata. Studi terbaru menunjukkan bahwa penggunaan media pembelajaran inovatif, seperti kartu bilangan dan pendekatan kooperatif, efektif dalam meningkatkan pemahaman peserta didik tentang operasi pecahan dan konsep matematika lainnya (Nurdiana, 2023; Ruswan, 2020; Wibowo *et al.,* 2024). Selain itu, metode demonstrasi juga terbukti mampu meningkatkan hasil belajar peserta didik dengan memfasilitasi proses pembelajaran yang lebih aktif dan interaktif (Rustiati, 2023). Model pembelajaran berbasis masalah (PBL) juga berkontribusi secara signifikan dalam memperkuat keterampilan pemecahan masalah dan pemahaman konsep pecahan di kalangan peserta didik kelas VI (Adiyana, 2024).

Temuan ini menunjukkan bahwa penggunaan strategi pembelajaran yang tepat dapat meningkatkan kualitas pembelajaran dan hasil belajar peserta didik secara signifikan, terutama dalam memahami konsep pecahan yang sering menjadi tantangan bagi peserta didik Sekolah Dasar. Oleh karena itu, pengembangan instrumen penilaian berbasis Kurikulum Merdeka harus mampu mencerminkan kebutuhan ini untuk memastikan relevansi pembelajaran dengan tujuan pendidikan yang holistik. Analisis butir soal secara kuantitatif merupakan pendekatan yang didasarkan pada data empiris dari soal yang telah diuji. Data ini digunakan untuk mengevaluasi kualitas soal dan memastikan bahwa instrumen pengukuran memiliki validitas dan reliabilitas yang tinggi. Tes yang berkualitas mampu mengukur kemampuan peserta didik secara akurat dan hasil pengukurannya dapat dipercaya. Sebuah tes dianggap memiliki validitas tinggi jika mampu mengukur sesuai dengan tujuan yang telah ditetapkan. Penelitian ini bertujuan untuk mengembangkan instrumen penilaian hasil belajar matematika yang valid dan reliabel untuk kelas VI sesuai dengan Kurikulum Merdeka. Secara khusus, tujuan penelitian ini yaitu 1) Menganalisis kualitas instrumen penilaian yang saat ini digunakan; dan 2) Menghasilkan instrumen penilaian yang memenuhi standar validitas dan reliabilitas untuk menilai hasil belajar peserta didik secara objektif.

LITERATURE REVIEW

Dalam dunia pendidikan, penilaian memegang peranan penting sebagai alat untuk mengevaluasi keberhasilan proses pembelajaran sekaligus memberikan umpan balik yang konstruktif bagi peserta didik dan pendidik. Penilaian yang baik tidak hanya mengukur pencapaian akademik peserta didik, tetapi juga menjadi instrumen untuk mendorong pengembangan kompetensi dan kemampuan peserta didik secara holistik. Dalam konteks ini, validitas dan reliabilitas instrumen menjadi aspek krusial yang memastikan penilaian dapat dilakukan secara akurat dan konsisten. Berbagai pendekatan, seperti teori klasik tes (*Classical Test Theory*) dan teori respons butir (*Item Response Theory*), telah digunakan untuk meningkatkan kualitas instrumen penilaian, termasuk penerapan model analisis Rasch. Model ini tidak hanya membantu dalam mengidentifikasi validitas butir soal, tetapi juga memberikan kemampuan untuk

https://doi.org/10.17509/jik.v22i1.77242

mengevaluasi kecocokan data dengan model teoretis, yang menjadikannya alat yang sangat efektif dalam pengembangan penilaian berbasis kompetensi dengan tambahan elemen kontekstual dalam instrumen penilaian, seperti relevansi soal terhadap kehidupan sehari-hari, proses penilaian menjadi lebih bermakna dan aplikatif, sejalan dengan tuntutan pendidikan modern seperti Kurikulum Merdeka.

Konsep Penilaian dalam Pendidikan

Penilaian pendidikan merupakan salah satu aspek krusial dalam proses belajar mengajar yang tidak dapat diabaikan. Proses ini tidak hanya berfungsi untuk mengukur hasil belajar peserta didik, tetapi juga sebagai alat untuk memberikan umpan balik yang konstruktif bagi peserta didik dan guru. Asesmen perlu dikembangkan sebagai bagian dari proses evaluasi untuk mengukur kemajuan belajar peserta didik (Marwa *et al.*, 2024; Sholikhah & Hidayati, 2024). Evaluasi merupakan subsistem yang sangat penting dan dibutuhkan dalam sistem pendidikan. Melalui evaluasi, pencapaian proses pembelajaran dapat dianalisis, sehingga memungkinkan pengambilan keputusan terkait perbaikan yang diperlukan untuk meningkatkan kualitas pembelajaran di masa depan (Yektiana & Nursikin, 2023).

Salah satu elemen utama dalam penilaian pendidikan adalah validitas dan reliabilitas instrumen yang digunakan. Validitas mengacu pada sejauh mana instrumen penilaian mampu mengukur apa yang seharusnya diukur. Misalnya, jika sebuah tes dirancang untuk mengukur kemampuan matematika peserta didik, maka tes tersebut harus secara spesifik mencerminkan kemampuan matematika yang dimaksud, tanpa terpengaruh oleh aspek lain seperti kemampuan membaca atau menulis. Sementara itu, reliabilitas mengacu pada konsistensi hasil pengukuran. Instrumen penilaian yang reliabel akan memberikan hasil yang konsisten jika digunakan berulang kali dalam kondisi yang serupa.

Model Analisis Rasch

Penggunaan model analisis Rasch dalam pengembangan instrumen telah terbukti efektif untuk meningkatkan kualitas butir soal. Sebagai pendekatan yang berbasis *Item Response Theory* (IRT), analisis Rasch menyediakan metode matematis yang presisi untuk mengevaluasi kecocokan data empiris dengan model teoretis. Melalui *fit statistics* seperti Infit dan Outfit, analisis ini mampu mendeteksi butir soal yang tidak sesuai dengan konstruksi yang diukur, sehingga mempermudah proses penyempurnaan instrumen (Novriyanti & Arthur, 2024). Model ini juga memiliki keunggulan dalam mengatasi permasalahan seperti bias responden dan ketidakseragaman tingkat kesulitan butir soal. Penelitian menunjukkan bahwa analisis Rasch tidak hanya membantu merancang soal yang lebih efektif, tetapi juga memastikan setiap butir soal mampu membedakan kemampuan peserta secara akurat (Nudin & Hidayatullah, 2023). Selain itu, model ini mendukung pengujian *local independence*, yaitu asumsi penting dalam pengembangan instrumen yang mengharuskan setiap item bersifat independen satu sama lain (Latifah *et al.*, 2024; Abdullaev *et al.*, 2024).

Analisis Rasch juga mampu mengkonfirmasi bahwa instrumen bersifat unidimensional, di mana setiap item diharapkan dapat mengukur satu konstruk laten yang sama. Melalui *Principal Component Analysis of Residuals* (PCAR), model ini dapat mengidentifikasi dimensi tambahan yang tidak diinginkan dalam instrumen (Yusuf *et al.*, 2021). Dengan demikian, analisis Rasch tidak hanya menjamin reliabilitas instrumen, tetapi juga memastikan validitasnya dalam mengukur kompetensi tertentu secara konsisten. Lebih lanjut, analisis item soal melalui model Rasch menawarkan berbagai keuntungan. Model ini mampu mendeteksi jawaban yang tidak konsisten, menangani data yang hilang, dan menunjukkan bahwa kemampuan responden tidak hanya ditentukan oleh jawaban yang benar, tetapi juga oleh pola respon yang konsisten (Azizah & Wahyuningsih 2020; Eliza & Yusmaita, 2021).

https://doi.org/10.17509/jik.v22i1.77242

Penelitian terbaru juga mengonfirmasi bahwa model Rasch dapat diterapkan dalam berbagai konteks penilaian, mulai dari pendidikan hingga bidang medis dan psikologi. Dalam dunia pendidikan, instrumen berbasis Rasch terbukti meningkatkan validitas dan reliabilitas penilaian berbasis kompetensi, seperti yang diterapkan dalam Kurikulum Merdeka di Indonesia (Widodo, 2020). Model ini memungkinkan pengembangan instrumen yang lebih responsif dan adaptif terhadap kebutuhan peserta didik, mendukung asesmen yang lebih inklusif dan berkeadilan. Psikometri menyediakan kerangka kerja yang kokoh untuk mengembangkan instrumen penilaian berbasis Teori Klasik Tes (*Classical Test Theory*, CTT) maupun Teori Respons Butir (*Item Response Theory*, IRT). Penggunaan IRT dalam pengembangan instrumen penilaian menawarkan keunggulan signifikan dibandingkan CTT, terutama dalam hal independensi parameter item dari sampel dan kemampuannya untuk merancang tes adaptif. IRT memungkinkan evaluasi instrumen secara lebih mendalam, mendukung perancangan tes yang efisien dan relevan bagi peserta didik dengan berbagai tingkat kemampuan (Firdaus *et al*., 2022).

Penelitian terbaru semakin menegaskan pentingnya psikometri dalam pengembangan instrumen penilaian. Instrumen berbasis IRT memiliki keunggulan dalam mengidentifikasi item dengan validitas rendah serta memberikan panduan untuk penyempurnaan soal. Selain itu, pendekatan psikometri memungkinkan analisis reliabilitas instrumen melalui parameter seperti kesulitan item (*item difficulty*), daya beda item (*item discrimination*), dan parameter menebak (*guessing parameter*), yang tidak sepenuhnya tercakup dalam pendekatan CTT. Penelitian menyoroti peran psikometri dalam merancang instrumen yang mendukung penilaian berbasis kompetensi, yang semakin relevan dalam kurikulum modern seperti Kurikulum Merdeka (Jones *et al*., 2021). Lebih lanjut, psikometri juga berperan dalam memastikan bahwa instrumen penilaian memenuhi prinsip unidimensionalitas, yang penting untuk mengukur satu konstruk laten secara konsisten. Melalui analisis Rasch, sebagai salah satu aplikasi IRT, para peneliti dapat mengidentifikasi item yang tidak sesuai dengan model dan mengoptimalkan skala penilaian. Instrumen berbasis Rasch lebih unggul dalam mengevaluasi pencapaian peserta didik secara holistik, mencakup kemampuan numerasi dan pemecahan masalah (Kim & Kim, 2022).

Psikometri modern juga mencakup pengembangan teknologi penilaian berbasis komputer atau *Computerized Adaptive Testing* (CAT). Melalui CAT, tes dapat disesuaikan secara dinamis berdasarkan respons peserta, memungkinkan pengukuran yang lebih presisi dengan jumlah item yang lebih sedikit. CAT yang dirancang dengan prinsip IRT mampu meningkatkan pengalaman tes peserta didik sekaligus menghasilkan data yang lebih valid dan reliabel (Wang & Zheng, 2023). Dengan landasan teoritis yang kuat dan bukti empiris yang mendukung, psikometri terus menjadi pilar utama dalam pengembangan instrumen penilaian. Penerapannya tidak hanya memastikan validitas dan reliabilitas instrumen, tetapi juga meningkatkan relevansi dan efisiensi proses penilaian di berbagai ranah pendidikan.

Penilaian Kontekstual

Penilaian berbasis konteks, seperti soal-soal yang relevan dengan kehidupan sehari-hari, terbukti efektif dalam meningkatkan keterlibatan peserta didik dalam pembelajaran. Soal-soal tersebut, seperti menghitung kembalian uang atau mengukur luas bangun datar, memberikan pengalaman belajar yang lebih bermakna. Pendekatan ini selaras dengan Kurikulum Merdeka, yang menekankan pengembangan kompetensi numerasi dan kemampuan pemecahan masalah yang aplikatif dalam kehidupan nyata. Penilaian kontekstual tidak hanya meningkatkan motivasi peserta didik, tetapi juga mendukung pengembangan keterampilan berpikir kritis dan kemampuan pemecahan masalah. Peserta didik yang terpapar pada soal berbasis konteks cenderung lebih aktif dalam menyelesaikan masalah karena mereka merasa materi yang diajarkan relevan dengan kehidupan mereka (Widodo, 2020).

https://doi.org/10.17509/jik.v22i1.77242

Pendekatan ini membantu peserta didik mengaitkan konsep abstrak dengan aplikasi praktis, sehingga memperkuat pemahaman mereka terhadap materi pembelajaran (Smith *et al*., 2022). Lebih lanjut, penilaian kontekstual berkontribusi pada pengembangan literasi dan numerasi peserta didik. Penelitian menunjukkan bahwa penggunaan soal yang menggambarkan situasi sehari-hari, seperti perbandingan harga atau waktu perjalanan, tidak hanya meningkatkan keterlibatan peserta didik tetapi juga memberikan pemahaman yang lebih mendalam tentang konsep matematika. Temuan ini menyoroti pentingnya memasukkan elemen kontekstual ke dalam kurikulum untuk menciptakan pengalaman belajar yang relevan dan aplikatif (Jones *et al*., 2021).

Dalam konteks pendidikan berbasis kompetensi, penilaian kontekstual juga membantu mengidentifikasi kesenjangan pemahaman peserta didik. Penggunaan soal-soal berbasis situasi nyata memungkinkan guru mengevaluasi sejauh mana peserta didik mampu menerapkan pengetahuan mereka dalam skenario praktis. Sebagai contoh, penelitian mengungkapkan bahwa peserta didik yang dilatih menggunakan soal kontekstual menunjukkan peningkatan signifikan dalam kemampuan pemecahan masalah dibandingkan dengan mereka yang hanya menggunakan pendekatan tradisional (Nguyen *et al*., 2023). Selain itu, penilaian berbasis konteks mendukung pengembangan keterampilan abad ke-21, seperti kolaborasi, kreativitas, dan komunikasi. Peserta didik yang terlibat dalam diskusi berbasis konteks cenderung lebih mampu bekerja dalam tim dan mempresentasikan solusi mereka secara kreatif. Dengan demikian, pendekatan ini tidak hanya mendukung pencapaian akademik, tetapi juga mempersiapkan peserta didik menghadapi tantangan di dunia nyata.

Penerapan penilaian kontekstual juga selaras dengan pendekatan adaptif dan personalisasi dalam pembelajaran. Penilaian berbasis konteks yang dipersonalisasi berdasarkan pengalaman peserta didik menunjukkan hasil yang lebih efektif dalam meningkatkan keterlibatan dan pemahaman. Penilaian yang relevan secara personal membantu peserta didik merasa lebih termotivasi dan terlibat, menciptakan lingkungan pembelajaran yang lebih inklusif dan mendukung semua latar belakang peserta didik. Seiring dengan berkembangnya kebutuhan pendidikan yang relevan dan aplikatif, penilaian kontekstual memberikan landasan penting bagi pendidikan berbasis kompetensi. Pendekatan ini memastikan bahwa peserta didik tidak hanya belajar untuk tes, tetapi juga untuk kehidupan nyata, memperkuat prinsip-prinsip Kurikulum Merdeka. Dengan demikian, penilaian kontekstual berperan krusial dalam membentuk generasi yang memiliki keterampilan berpikir kritis, kolaboratif, dan adaptif untuk menghadapi tantangan masa depan.

METHODS

Penelitian ini menggunakan pendekatan kuantitatif dengan analisis model Rasch sebagai metode psikometrik untuk menilai kapasitas instrumen. Analisis model Rasch diterapkan untuk mengukur validitas dan reliabilitas instrumen, serta melakukan analisis *Differential Item Functioning* (DIF) guna mendeteksi kemungkinan adanya bias (Dwilesanti & Yudiarso, 2022; Latifah, *et al*. 2024). Penelitian ini melibatkan 213 peserta didik yang berasal dari Jakarta dan Tangerang Selatan. Responden dipilih menggunakan dua metode, yaitu purposive sampling dan snowball sampling. Melalui purposive sampling, responden yang dipilih adalah peserta didik Kelas VI di Sekolah Prestasi Global yang mengikuti ujian sumatif tengah semester. Selanjutnya, peneliti juga membagikan soal kepada peserta didik Kelas VI lainnya yang bersedia berpartisipasi yang merupakan bagian dari metode snowball sampling (Kennedy-Shaffer *et al*., 2021). Kriteria utama dalam penelitian ini adalah peserta didik yang sedang menempuh pendidikan di kelas VI sekolah dasar. Analisis data dilakukan dengan menggunakan perangkat lunak Winstep untuk memastikan keakuratan hasil pengukuran.

https://doi.org/10.17509/jik.v22i1.77242

RESULTS AND DISCUSSION

Unidimensionalitas

Dalam analisis menggunakan model Rasch, terdapat dua asumsi fundamental yang harus dipenuhi, yaitu unidimensionalitas dan lokal independensi. Unidimensionalitas mengacu pada asumsi bahwa setiap item dalam instrumen hanya mengukur satu konstruk laten utama. Untuk memastikan unidimensionalitas, dilakukan *Principal Component Analysis of Residuals* (PCAR) guna mengevaluasi pola residual yang mungkin menunjukkan dimensi tambahan.

Penelitian mengeksplorasi validitas dan reliabilitas skala pembelajaran mandiri menggunakan model Rasch. Analisis ini memberikan informasi mendalam mengenai unidimensionalitas dan local independence, yang sangat penting untuk memastikan bahwa instrumen hanya mengukur satu konstruk utama (unidimensi) dan butir-butir soal bersifat independen satu sama lain (Nizaruddin *et al*., 2024).

Pengujian unidimensionalitas digunakan untuk menilai validitas model Rasch. Model ini mensyaratkan bahwa variabel-variabel yang ada harus bersifat unidimensional. Peta dimensionalitas dianalisis menggunakan perangkat lunak Winstep melalui nilai "*raw variance explained by measure*". Pengukuran unidimensionalitas dianggap terpenuhi jika varians mentah yang dijelaskan oleh pengukuran sama dengan atau lebih besar dari 20% (Latifah *et al*., 2024). Dalam konteks model Rasch, batas unidimensionalitas idealnya mencapai minimal 40%, dan lebih baik jika melebihi angka tersebut (Hidayat *et al*., 2020). Kriteria Unidimensionalitas disajikan pada **Tabel 1** sebagai berikut:

Tabel 1. Kriteria Unidimensionalitas

Persentase Varians Kasar yang Dijelaskan oleh Pengukuran	Kriteria
> 60%	Sangat Baik
40% - 60%	Baik
20% - 40%	Cukup

Sumber: Latifah et al.., 2024.

Dari hasil yang disajikan pada **Gambar 1** total varians data adalah 24.4 (100%), dengan varians yang dijelaskan oleh model sebesar 9.4 (38.4%). Analisis ini menunjukkan bahwa model cukup mampu menjelaskan varians, dengan kontribusi varians antar person sebesar 4.1 (16.9%) dan antar item sebesar 5.2 (21.5%).

```
     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                          -- Empirical --    Modeled
Total raw variance in observations     =  24.4 100.0%          100.0%
  Raw variance explained by measures   =   9.4  38.4%           38.3%
    Raw variance explained by persons  =   4.1  16.9%           16.9%
    Raw Variance explained by items    =   5.2  21.5%           21.4%
  Raw unexplained variance (total)     =  15.0  61.6% 100.0%    61.7%
    Unexplned variance in 1st contrast =   1.5   6.2%  10.1%
    Unexplned variance in 2nd contrast =   1.4   5.8%   9.5%
    Unexplned variance in 3rd contrast =   1.3   5.2%   8.4%
    Unexplned variance in 4th contrast =   1.2   5.1%   8.3%
    Unexplned variance in 5th contrast =   1.2   4.8%   7.8%
```

Gambar 1. Variansi Residual Terstandarisasi (dalam satuan Eigenvalue)
Sumber: *Dokumentasi Penulis 2024*

Peta dimensionalitas juga menguji independensi lokal, yang dianalisis berdasarkan varians yang tidak dijelaskan dalam komponen residual (PCAR). Nilai ini mencerminkan independensi item dari konstruk tambahan yang tidak diukur oleh instrumen utama. Kriteria untuk varians yang tidak dijelaskan disajikan pada **Tabel 2**.

https://doi.org/10.17509/jik.v22i1.77242

Tabel 2. Kriteria Varians Yang tidak Dijelaskan

Varians yang Tidak Dijelaskan pada Komponen 1-5 PCA Residual	Kriteria
< 3%	Sangat Baik
3 - 5%	Baik Sekali
5 - 10%	Baik
10 - 15%	Cukup
> 15%	Buruk

Sumber: Ocy et al., 2023

Dari **Gambar 1** nilai eigenvalue < 2.0 pada kontras pertama (1.5) termasuk kategori baik, yang menunjukkan bahwa instrumen memenuhi kriteria unidimensionalitas. Meskipun *unexplained variance* total sebesar 61.6% tergolong cukup besar, penyebarannya mendukung pengukuran satu konstruk utama. Proporsi Varians Dijelaskan: Model menjelaskan 38.4% varians, yang meskipun moderat, cukup untuk mendukung penerapan model Rasch. Dengan *unexplained variance* yang rendah di kontras pertama, dapat disimpulkan bahwa item dalam instrumen ini mengukur satu konstruk utama secara unidimensional. Model ini menunjukkan fit yang memadai, sehingga dapat digunakan sebagai alat ukur yang valid dan reliabel untuk analisis psikometrik lebih lanjut.

Lokal Independensi

Lokal independensi mengacu pada asumsi bahwa respons peserta didik terhadap suatu item tidak dipengaruhi oleh respons mereka terhadap item lain setelah memperhitungkan kemampuan laten peserta didik. Pelanggaran terhadap asumsi ini dapat menyebabkan bias dalam estimasi parameter item, terutama jika terdapat kesamaan konten atau keterkaitan logis antar item (Latifah *et al*., 2024; Abdullaev *et al*., 2024).

SUMMARY OF 212 MEASURED (NON-EXTREME) Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	7.5	15.0	.00	.67	1.00	.0	1.01	.1
S.D.	2.9	.0	1.25	.09	.31	1.0	.76	.9
MAX.	14.0	15.0	3.38	1.08	2.15	3.0	8.43	3.2
MIN.	1.0	15.0	-3.34	.62	.44	-2.2	.27	-1.8

REAL RMSE	.71	TRUE SD	1.03	SEPARATION	1.44	Person RELIABILITY .68
MODEL RMSE	.67	TRUE SD	1.05	SEPARATION	1.56	Person RELIABILITY .71
S.E. OF Person MEAN = .09						

SUMMARY OF 213 MEASURED (EXTREME AND NON-EXTREME) Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	7.5	15.0	-.02	.67				
S.D.	2.9	.0	1.29	.12				
MAX.	14.0	15.0	3.38	1.86				
MIN.	.0	15.0	-4.66	.62	.44	-2.2	.27	-1.8

REAL RMSE	.72	TRUE SD	1.07	SEPARATION	1.48	Person RELIABILITY .69
MODEL RMSE	.68	TRUE SD	1.09	SEPARATION	1.59	Person RELIABILITY .72
S.E. OF Person MEAN = .09						

Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .70

Gambar 2. Reliabilitas Orang dan Butir
Sumber: *Dokumentasi Penulis 2024*

https://doi.org/10.17509/jik.v22i1.77242

Hasil uji lokal independensi berdasarkan korelasi antar-residual menunjukkan bahwa tidak ada nilai korelasi antar-item yang melampaui ambang batas kritis ±0,2. Sebagian besar nilai korelasi berada di sekitar nol atau mendekati nol, yang mengindikasikan hubungan antar-item relatif rendah atau tidak signifikan. Temuan ini mencerminkan bahwa item memiliki independensi yang baik, sesuai dengan salah satu asumsi penting dalam model Rasch.

Selain itu, tidak ditemukan korelasi tinggi (mendekati 1 atau -1) antar-item, yang menunjukkan bahwa setiap item mengukur aspek yang berbeda dari konstruk yang diukur. Ini juga menghilangkan potensi multikolinearitas yang dapat memengaruhi validitas pengukuran. Dengan demikian, semua item dalam instrumen memenuhi asumsi lokal independensi, yaitu respons terhadap suatu item tidak dipengaruhi oleh respons terhadap item lain setelah kemampuan laten diperhitungkan.

Hasil ini diperkuat dengan data visual pada **Gambar 2** yang menunjukkan bahwa tidak ada korelasi antar-item yang melampaui ambang batas kritis ±0,2. Korelasi residual yang rendah memastikan bahwa tidak terdapat keterkaitan sistematis antar-item, sehingga instrumen dapat dinyatakan bebas dari bias lokal dan memenuhi persyaratan analisis model Rasch.

Keandalan

Keandalan Rasch dapat dimanfaatkan untuk mengevaluasi kestabilan individu dan item dalam instrumen, yang merupakan informasi penting yang disediakan oleh pemodelan Rasch. Nilai reliabilitas Rasch, yang berkisar antara 0 hingga 1, sering disamakan dengan Alpha Cronbach dan mencerminkan konsistensi baik dari responden maupun item dalam instrumen. Nilai ini menunjukkan bahwa instrumen memiliki reliabilitas yang cukup hingga sangat baik. Kriteria reliabilitas individu dan item yang digunakan dalam analisis dirangkum dalam **Tabel 3**.

Tabel 3. Kriteria Reliabilitas Orang dan Butir

Reliabilitas Orang dan Item	Kriteria
> 0,94	Sangat Baik
0,91 - 0,94	Baik Sekali
0,81 - 0,90	Baik
0,67 - 0,80	Cukup
< 0,67	Buruk

Sumber: Erfan, et al., 2020

Hasil analisis pada **Tabel 3** memberikan ringkasan reliabilitas responden, baik dari kelompok ekstrem maupun non-ekstrem. Koefisien Cronbach Alpha yang diperoleh dalam penelitian ini adalah 0,70 untuk reliabilitas responden, yang tergolong dalam kategori cukup. Sementara itu, reliabilitas item mencapai 0,99, yang diklasifikasikan sebagai sangat baik.

Dengan demikian, hasil ini menunjukkan bahwa instrumen yang diuji memiliki konsistensi internal yang tinggi dan dapat diandalkan sebagai alat pengukuran dalam penelitian. Nilai-nilai ini juga memperkuat validitas dan reliabilitas instrumen, sejalan dengan standar psikometrik yang diacu dalam literatur (Erfan *et al*., 2020).

Kesesuaian/Kecocokan Item

Dalam konteks analisis model Rasch, evaluasi kecocokan item bertujuan untuk menilai sejauh mana item efektif dalam mengukur konstruk yang menjadi fokus penelitian. Proses ini melibatkan statistik kecocokan, seperti infit dan outfit mean square (MNSQ), yang memberikan gambaran mengenai kesesuaian respons

https://doi.org/10.17509/jik.v22i1.77242

peserta dengan model yang diharapkan. Selain itu, model Rasch menerapkan kriteria khusus untuk menilai kualitas item (Latifah *et al*., 2024), yang meliputi:

1. Outfit MNSQ (Mean Square): 0,5 < MNSQ < 1,5
2. Outfit ZSTD (Z-Standard): -2,0 < ZSTD < +2,0
3. Point Measure Correlation: 0,4 < Pt Measure Corr < 0,85

Item dalam instrumen yang tidak memenuhi ketiga kriteria tersebut dianggap sebagai "misfit" dan memerlukan revisi atau penggantian. Namun, jika suatu item memenuhi minimal dua dari tiga kriteria, maka item tersebut tetap dapat dinyatakan fit dengan model dan layak digunakan dalam instrumen (Azizah & Wahyuningsih, 2020).

Lebih lanjut, nilai Outfit MNSQ berfungsi sebagai indikator utama dalam menilai kecocokan item. Item dengan nilai di luar rentang yang telah ditetapkan memerlukan evaluasi tambahan untuk memastikan bahwa instrumen memiliki kualitas yang memadai dan bebas dari bias (Noben *et al*., 2021).

Tabel 4. Kesesuaian/Kecocokan Item

Item	Outfit		PT-MEASURE			Fit/misfit
	MNSQ	MNSQ kriteria	ZSTD	Korelasi	PM Kriteria	
Item10	1.34	Baik	1.3	0.33	Baik	Fit
Item5	1.3	Baik	1.0	0.34	Baik	Fit
Item9	1.17	Baik	0.8	0.36	Baik	Fit
Item8	1.11	Baik	1.0	0.43	Sangat Baik	Fit
Item6	1.07	Baik	0.6	0.42	Sangat Baik	Fit
Item12	1.07	Baik	0.5	0.42	Sangat Baik	Fit
Item15	0.85	Baik	-0.5	0.36	Baik	Fit
Item13	1.04	Baik	0.5	0.48	Sangat Baik	Fit
Item2	0.99	Baik	0.0	0.36	Baik	Fit
Item7	0.94	Baik	-0.4	0.49	Sangat Baik	Fit
Item3	0.83	Baik	-1.0	0.47	Sangat Baik	Fit
Item14	0.95	Baik	-0.3	0.48	Sangat Baik	Fit
Item11	0.80	Baik	0.8	0.33	Baik	Fit
Item1	0.88	Baik	-1.1	0.53	Sangat Baik	Fit
Item4	0.77	Baik	-1.7	0.55	Sangat Baik	Fit

Sumber: Data primer 2024

Tabel 4 menyajikan hasil evaluasi kesesuaian item berdasarkan tiga parameter utama, yaitu Outfit MNSQ, ZSTD (Z-Standard), dan Point Measure Correlation (PM).

1. **Parameter Outfit MNSQ**
   Outfit MNSQ mengukur sejauh mana data sesuai dengan model Rasch. Hasil analisis menunjukkan bahwa semua item memiliki nilai dalam rentang ideal (0,5-1,5). Ini menunjukkan bahwa semua item memenuhi kriteria baik berdasarkan kecocokan dengan model. Item dengan nilai Outfit MNSQ tertinggi adalah Item10 (1,34), sedangkan nilai terendah adalah Item4 (0,77). Tidak ada item yang menunjukkan misfit ekstrem, sehingga semua item dianggap sesuai dengan model.
2. **Parameter ZSTD (Z-Standard)**
   ZSTD mengevaluasi deviasi statistik dalam data. Semua item memiliki nilai ZSTD yang berada dalam rentang ideal -2 hingga +2, yang menunjukkan bahwa tidak ada item dengan deviasi signifikan dari model. Nilai ZSTD tertinggi ditemukan pada Item10 (1,3), yang menunjukkan sedikit deviasi tetapi masih dalam kategori wajar. Nilai ZSTD terendah adalah Item4 (-1,7), yang menunjukkan kesesuaian sangat baik dengan model.

https://doi.org/10.17509/jik.v22i1.77242

3. **Parameter Point Measure Correlation (PM)**
   Parameter PM mengukur hubungan antara pola respons item dengan kemampuan peserta. Nilai ideal untuk PM adalah 0,4-0,85. Sebagian besar item memiliki korelasi di atas 0,4, kecuali Item10 (0,33), Item5 (0,38), dan Item9 (0,39). Item dengan korelasi tertinggi adalah Item4 (0,55), yang menunjukkan hubungan kuat dengan kemampuan peserta. Meskipun nilai PM pada Item10 sedikit di bawah batas ideal, item ini masih dianggap fit dengan model.

Berdasarkan evaluasi ketiga parameter tersebut (Outfit MNSQ, ZSTD, dan PM), semua item dinyatakan fit dengan model Rasch. Tidak ada item yang perlu dihapus atau direvisi karena ketidaksesuaian model. Hasil ini menunjukkan bahwa instrumen memiliki kualitas yang sangat baik dan dapat diandalkan untuk mengukur konstruk yang dimaksud. Evaluasi ini juga menegaskan bahwa respons peserta terhadap item konsisten dengan asumsi model Rasch, memperkuat validitas dan reliabilitas instrumen yang digunakan.

Tingkat Kesulitan Butir

Tingkat kesulitan item dalam pemodelan Rasch dibagi ke dalam empat kategori berdasarkan ukuran logit dan deviasi standar (SD) dari logit item (Ocy *et al*., 2023). Nilai logit yang lebih tinggi menunjukkan tingkat kesulitan soal yang lebih besar. Sebelum menganalisis tingkat kesulitan butir soal, standar deviasi (SD) ditetapkan sebesar 1,38.

Tabel 5. Tingkat Kesulitan Item

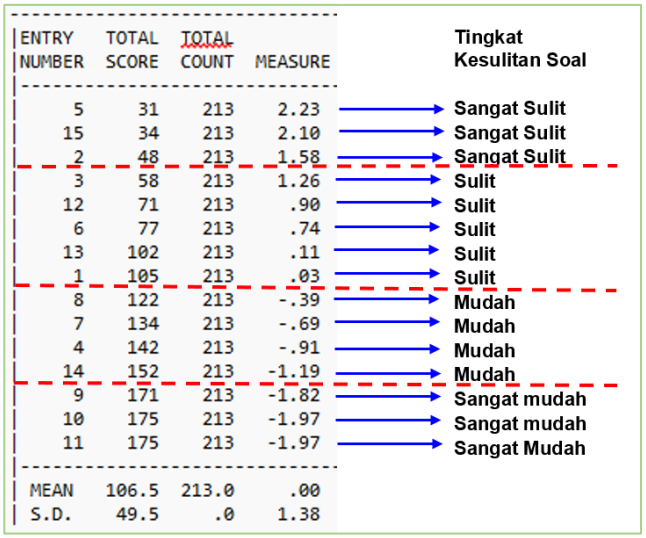
Nilai Logit	Kategori	Jumlah Item
Lebih besar dari +1,38	Sangat Sulit	2,5,15
Lebih dari 0,0 logit dan kurang dari +01,38 SD	Sulit	1,3,6,12,13
Kurang dari 0,0 logit - dan Lebih dari -1,38 SD	Mudah	4,7,8,14
Kurang dari -1,38	Sangat Mudah	9, 10,11

Sumber: Data primer 2024

Berdasarkan data yang disajikan pada **Tabel 5**, kategori kesulitan item ditentukan menggunakan standar deviasi (SD) sebesar 1,38. Terdapat tiga item yang tergolong sangat sulit, yaitu item 2, 5, dan 15. Soal-soal ini memiliki tingkat kesulitan yang tinggi dan cenderung hanya dapat dijawab benar oleh peserta dengan kemampuan sangat tinggi.

Soal dengan nilai logit antara 0 hingga +1,38 termasuk dalam kategori sulit. Item yang tergolong sulit adalah 1, 3, 6, 12, dan 13. Soal-soal ini menantang, tetapi masih dapat dijawab oleh peserta dengan kemampuan di atas rata-rata. Item dengan nilai logit antara 0 hingga -1,38 masuk dalam kategori mudah. Soal-soal ini meliputi 4, 7, 8, dan 14, yang relatif lebih sederhana dan cocok untuk peserta dengan kemampuan menengah hingga rendah. Sementara itu, item dengan nilai logit kurang dari -1,38 tergolong sangat mudah. Soal-soal ini, yaitu 9, 10, dan 11, dirancang untuk peserta dengan kemampuan sangat rendah, yang memungkinkan mereka meraih skor dengan lebih mudah. Lebih jelas dapat dilihat pada **Gambar 3**.

https://doi.org/10.17509/jik.v22i1.77242



Gambar 3. Tingkat Kesulitan Butir
Sumber: *Dokumentasi Penulis 2024*

Analisis Distribusi Kesulitan Item: Distribusi tingkat kesulitan soal dalam instrumen ini menunjukkan keberagaman tingkat kesulitan, mulai dari sangat mudah hingga sangat sulit. Hal ini mencerminkan kemampuan instrumen untuk mengukur peserta dengan berbagai tingkat kemampuan secara efektif. Namun, jumlah soal dalam kategori sangat sulit dan sangat mudah relatif lebih sedikit dibandingkan soal dalam kategori sulit dan mudah. Distribusi ini menunjukkan bahwa instrumen lebih berfokus pada pengukuran di sekitar kemampuan rata-rata peserta. Struktur ini dapat meningkatkan sensitivitas instrumen dalam mengukur kemampuan peserta pada tingkat yang paling umum dijumpai dalam populasi.

Tingkat Abilitas Peserta Didik dalam Menjawab Item

Sejalan dengan analisis tingkat kesulitan butir, tingkat kemampuan (abilitas) peserta didik dalam menjawab item pada pemodelan Rasch dikelompokkan ke dalam empat kategori berdasarkan ukuran logit dan nilai Deviasi Standar (SD) dari logit item (Ocy *et al*., 2023). Tingkat abilitas peserta didik ini digunakan untuk mengidentifikasi kemampuan mereka dalam menjawab pertanyaan, yang diurutkan dari nilai logit tertinggi hingga terendah.

Pada analisis ini, standar deviasi (SD) ditetapkan sebesar 1,29, dengan nilai rata-rata logit person sebesar -0,02. Nilai ini menjadi titik acuan dalam mengategorikan kemampuan peserta didik ke dalam beberapa kelompok.

Tabel 6. Abilitas Peserta Didik dalam Menjawab Item

Nilai Logit Abilitas Peserta Didik	Kriteria	Jumlah Peserta Didik
Logit > +1,29	Sangat Tinggi	33
-0.02 < logit ≤ 1,29	Tinggi	77
-1,29 ≤ logit ≤ -0,02	Rendah	67
Logit < -1,29	Sangat Rendah	36

Sumber: Data primer 2024

Analisis Distribusi Kemampuan Peserta Didik: Berdasarkan data yang disajikan pada **Tabel 6**, distribusi kemampuan peserta didik menunjukkan bahwa sebagian besar peserta didik, yaitu 110 dari 213 peserta didik (51,6%), memiliki kemampuan yang tergolong tinggi hingga sangat tinggi. Namun, masih terdapat 103 peserta didik (48,4%) yang masuk dalam kategori rendah hingga sangat rendah.

https://doi.org/10.17509/jik.v22i1.77242

Hasil ini menunjukkan bahwa instrumen pengukuran cukup efektif dalam mengidentifikasi variasi kemampuan peserta didik di berbagai tingkat. Namun, temuan ini juga mengindikasikan perlunya intervensi pembelajaran yang lebih terfokus pada kelompok peserta didik dengan kemampuan rendah dan sangat rendah. Langkah ini penting untuk meningkatkan kompetensi mereka dan memastikan pemerataan kualitas pembelajaran di seluruh kelompok kemampuan.

Tingkat Kesulitan Butir vs Tingkat Abilitas Peserta Didik dalam Menjawab Item

Tabel 7 menggambarkan hubungan antara tingkat kesulitan butir soal (logit item) dan kemampuan peserta didik (logit person) dalam menjawab item tertentu.

Tabel 7. Tingkat Kesulitan Butir vs Tingkat Abilitas Peserta Didik dalam Menjawab Item

Person Measure	Responden/ Peserta Didik	Nomor dan Tingkat Kesulitan Item													
		-1.97	-1.97	-1.82	-1.19	-.91	-.69	-.39	.03	.11	.74	.90	1.26	1.58	2.10	2.23
		10	11	9	14	4	7	8	1	13	6	12	3	2	15	5
3.38	140P	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
3.38	290L	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
3.38	207P	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
2.50	001P	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
2.50	055L	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1
2.50	079P	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1
2.50	093L	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
2.50	141L	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1
2.50	161L	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1
1.90	044L	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0
1.90	071L	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1
1.90	072L	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0
1.90	075P	1	1	1	1	1	1	0	1	1	1	1	0	1	0	1
1.90	109P	1	1	1	1	1	1	1	0	1	0	1	1	1	1	0
1.90	138L	1	1	1	0	1	0	1	1	1	1	1	0	1	1	1
1.41	026P	1	1	1	1	1	1	0	1	1	1	1	0	1	0	0
1.41	050L	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1
1.41	058P	1	1	0	1	1	1	1	1	1	1	0	1	0	1	0
1.41	060P	1	1	1	1	1	1	1	1	0	1	0	1	0	1	0
1.41	064P	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0

Sumber: Data primer 2024

Berdasarkan data yang ditampilkan pada **Tabel 7**, terdapat beberapa temuan menarik terkait hubungan antara **kemampuan peserta Didik** dan **tingkat kesulitan item**:

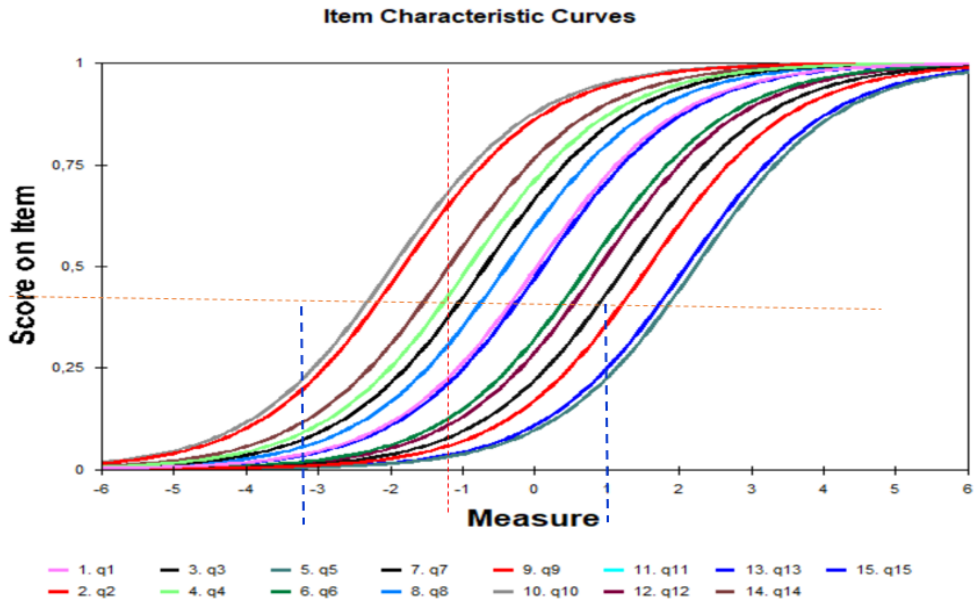
1. Peserta didik dengan Logit Tinggi yang tidak menjawab soal sulit. Beberapa peserta didik dengan logit tinggi (kemampuan di atas 2.5), seperti 140P, 290L, 207P, 001P, dan 055L, berada di atas tingkat kesulitan soal nomor 5 (logit 2.23). Namun, mereka tidak dapat menjawab soal ini dengan benar. Kegagalan ini dapat disebabkan oleh kurangnya kecermatan (*carelessness*) saat menjawab soal, meskipun kemampuan mereka seharusnya cukup untuk menyelesaikan soal tersebut dengan benar.
2. Peserta didik dengan Logit Rendah yang Berhasil Menjawab Soal Sulit: Sebaliknya, peserta didik dengan logit lebih rendah (1.90), seperti 071L, berhasil menjawab soal nomor 5 (logit 2.23) dengan benar. Keberhasilan ini mungkin disebabkan oleh tebakan yang benar (*lucky guess*), mengingat tingkat kesulitan soal tersebut berada di atas kemampuan peserta didik yang sebenarnya.
3. Polarisasi kemampuan dan tingkat kesulitan. Secara umum, peserta didik dengan kemampuan tinggi (logit besar) cenderung dapat menjawab soal dengan tingkat kesulitan rendah hingga sedang. Namun, terdapat beberapa anomali, seperti: Peserta didik berkemampuan tinggi yang gagal menjawab soal sulit, yang mungkin dipengaruhi oleh kecermatan, motivasi rendah, atau gangguan konsentrasi. Atau, peserta didik berkemampuan rendah yang berhasil menjawab soal sulit, yang dapat menunjukkan adanya strategi tebakan yang benar.

https://doi.org/10.17509/jik.v22i1.77242

Hubungan antara kemampuan peserta didik dan tingkat kesulitan item dalam instrumen ini menunjukkan pola yang umumnya sesuai dengan model Rasch. Namun, adanya anomali seperti yang disebutkan di atas mengindikasikan perlunya evaluasi lebih lanjut terhadap instrumen. Evaluasi ini bertujuan untuk: 1) Mengidentifikasi faktor-faktor non-kemampuan yang memengaruhi hasil, seperti motivasional dan strategi menjawab; 2) Mengurangi potensi jawaban yang didasarkan pada keberuntungan atau kurangnya kecermatan melalui pengujian ulang atau pelatihan tambahan; dan 3) Memperbaiki validitas dan reliabilitas instrumen dengan menyempurnakan item yang memiliki hasil tidak konsisten dengan model Rasch.

Analisis Kurva Karakteristik Butir (ICC) dalam Teori Respons Butir (IRT)

Kurva Karakteristik Butir (*Item Characteristic Curve*/ICC) dalam Teori Respons Butir (IRT) merupakan alat yang digunakan untuk memvisualisasikan hubungan antara kemampuan peserta dan probabilitas menjawab item dengan benar. ICC memberikan informasi penting mengenai tingkat kesulitan dan daya diskriminasi item, sehingga memungkinkan analisis kualitas instrumen tes secara mendalam. Alat ini efektif untuk mengevaluasi kualitas item, memantau konsistensi performa item, dan mengidentifikasi butir yang memerlukan revisi. Selain itu, ICC mendukung pengukuran yang valid dan reliabel dengan memastikan item dapat mengukur kemampuan peserta secara akurat dan efisien (Oktaviyanthi *et al*., 2024).



Gambar 4. Grafik ICC
Sumber: Dokumentasi Penulis 2024

Gambar 4 mengenai *Item Characteristic Curves* (ICC) menunjukkan bahwa item dengan kurva yang bergeser ke kanan memiliki tingkat kesulitan yang lebih tinggi, seperti pada Item15 dan Item5. Sebaliknya, item dengan kurva yang bergeser ke kiri memiliki tingkat kesulitan yang lebih rendah, seperti Item11 dan Item10. Tingkat kemiringan kurva mencerminkan daya diskriminasi setiap item, yaitu seberapa baik item dapat membedakan peserta didik berdasarkan kemampuan mereka.

Semua item dalam grafik menunjukkan kemiringan yang seragam, yang menandakan bahwa setiap item memiliki kemampuan diskriminatif yang memadai. Kurva pada ICC juga mendekati nilai 1 (100% probabilitas menjawab benar) untuk peserta didik dengan kemampuan tinggi dan mendekati nilai 0 (0% probabilitas menjawab benar) untuk peserta didik dengan kemampuan rendah. Hal ini menunjukkan bahwa model Rasch berhasil mencerminkan hubungan antara kemampuan peserta didik dan tingkat

https://doi.org/10.17509/jik.v22i1.77242

kesulitan item dengan baik. Sebaran tingkat kesulitan terlihat merata, mencakup item yang sangat mudah hingga sangat sulit, yang mengindikasikan bahwa tes ini memiliki distribusi item yang cukup beragam untuk mengukur peserta didik dengan berbagai tingkat kemampuan secara efektif.

Item-item dalam tes ini menunjukkan kinerja yang baik berdasarkan karakteristik kurva. Item dengan tingkat kesulitan tinggi (seperti **Item15** dan **Item5**) memberikan informasi yang bermanfaat untuk mengukur peserta didik dengan kemampuan tinggi. Sebaliknya, item dengan tingkat kesulitan rendah (seperti **Item11** dan **Item10**) berguna untuk mengukur peserta didik dengan kemampuan rendah. Secara keseluruhan, tes ini memiliki kualitas yang memadai dalam mengukur kemampuan peserta didik di berbagai tingkatan, sesuai dengan prinsip model Rasch.

Differential Item Functioning (DIF)

Differential Item Functioning (DIF) adalah fenomena di mana item dalam tes menunjukkan perilaku yang berbeda terhadap kelompok peserta dengan kemampuan yang sama, yang dapat menyebabkan bias dan menurunkan validitas instrumen pengukuran. DIF digunakan untuk menilai apakah suatu item berfungsi secara adil di berbagai kelompok berdasarkan karakteristik tertentu, seperti jenis kelamin, usia, budaya, atau latar belakang pendidikan (El Fahmi *et al*, 2021; Peng *et al*., 2024).

Beberapa penelitian menyoroti pentingnya deteksi dan penanganan DIF untuk memastikan keadilan dan validitas instrumen. Analisis Rasch pada *Autism Behavior Checklist* (ABC) menunjukkan adanya DIF yang memerlukan perhatian khusus terhadap karakteristik populasi (Peng *et al*., 2024). Selain itu, metode semi-otomatis untuk analisis Rasch yang mempertimbangkan kemungkinan DIF, menyederhanakan proses analisis yang kompleks dan meminimalkan subjektivitas dalam pengambilan keputusan (Wijayanto *et al*., 2023).

Person CLASSES	SUMMARY DIF CHI-SQUARE	D.F.	PROB.	BETWEEN-CLASS MEAN-SQUARE	t=ZSTD	Item Number	Name
2	3.9379	1	.0472	1.9670	1.0080	1	q1
2	.0975	1	.7549	.0476	-.8809	2	q2
2	2.8261	1	.0927	1.4052	.7261	3	q3
2	.1298	1	.7187	.0640	-.8015	4	q4
2	.0394	1	.8427	.0197	-1.0772	5	q5
2	.7629	1	.3824	.3774	-.1169	6	q6
2	1.1649	1	.2804	.5734	.1124	7	q7
2	.6848	1	.4079	.3377	-.1726	8	q8
2	.1906	1	.6624	.0931	-.6885	9	q9
2	.0869	1	.7682	.0425	-.9096	10	q10
2	.0869	1	.7682	.0425	-.9096	11	q11
2	.8195	1	.3653	.4040	-.0817	12	q12
2	.1088	1	.7415	.0535	-.8505	13	q13
2	.8966	1	.3437	.4383	-.0385	14	q14
2	1.7254	1	.1890	.8513	.3606	15	q15

Person CLASS	OBSERVATIONS COUNT	AVERAGE	BASELINE EXPECT	MEASURE	DIF SCORE	DIF MEASURE	DIF SIZE	DIF S.E.	DIF t	Prob.	Item Number	Name
L	124	.53	.48	.03	.05	-.24	-.27	.21	-1.31	.1930	1	q1
L	124	.23	.22	1.58	.01	1.53	-.05	.24	-.20	.8431	2	q2
L	124	.23	.27	1.26	-.04	1.53	.27	.24	1.12	.2667	3	q3
L	124	.66	.65	-.91	.01	-.96	-.05	.22	-.24	.8116	4	q4
L	124	.14	.14	2.23	.00	2.27	.04	.28	.14	.8904	5	q5
L	124	.33	.35	.74	-.02	.86	.13	.22	.58	.5639	6	q6
L	124	.59	.62	-.69	-.03	-.54	.15	.21	.70	.4838	7	q7
L	124	.58	.56	-.39	.02	-.50	-.11	.21	-.54	.5875	8	q8
L	124	.80	.79	-1.82	.01	-1.89	-.07	.25	-.29	.7755	9	q9
L	124	.81	.81	-1.97	.01	-2.02	-.05	.26	-.19	.8461	10	q10
L	124	.81	.81	-1.97	.01	-2.02	-.05	.26	-.19	.8461	11	q11
L	124	.35	.32	.90	.02	.77	-.13	.21	-.59	.5556	12	q12
L	124	.48	.47	.11	.01	.06	-.04	.21	-.22	.8297	13	q13
L	124	.68	.70	-1.19	-.02	-1.06	.13	.22	.61	.5437	14	q14
L	124	.13	.15	2.10	-.03	2.35	.25	.29	.87	.3844	15	q15
P	88	.44	.52	.03	-.07	.40	.36	.24	1.51	.1343	1	q1
P	88	.23	.24	1.58	-.01	1.65	.07	.28	.24	.8092	2	q2
P	88	.34	.29	1.26	.05	.94	-.32	.25	-1.27	.2070	3	q3
P	88	.68	.69	-.91	-.01	-.84	.07	.25	.27	.7871	4	q4
P	88	.16	.15	2.23	.01	2.18	-.05	.32	-.14	.8865	5	q5
P	88	.41	.38	.74	.03	.58	-.16	.24	-.66	.5120	6	q6
P	88	.69	.66	-.69	.04	-.90	-.21	.25	-.83	.4115	7	q7
P	88	.57	.60	-.39	-.03	-.23	.15	.24	.63	.5321	8	q8
P	88	.82	.83	-1.82	-.01	-1.72	.10	.30	.33	.7410	9	q9
P	88	.84	.85	-1.97	-.01	-1.90	.07	.31	.22	.8245	10	q10
P	88	.84	.85	-1.97	-.01	-1.90	.07	.31	.22	.8245	11	q11
P	88	.32	.35	.90	-.03	1.07	.18	.26	.69	.4920	12	q12
P	88	.49	.50	.11	-.01	.17	.06	.24	.25	.8026	13	q13
P	88	.77	.74	-1.19	.03	-1.39	-.20	.28	-.73	.4675	14	q14
P	88	.20	.17	2.10	.04	1.81	-.29	.29	-.99	.3250	15	q15

Gambar 5. Reliabilitas Orang dan Butir
Sumber: *Dokumentasi Penulis 2024*

Analisis data menggunakan model Rasch pada perangkat lunak Winsteps disajikan pada **Gambar 5**.

Sebagian besar item memiliki nilai probabilitas yang cukup tinggi (> 0.05) pada kolom Chi-Square, yang menunjukkan tidak adanya perbedaan signifikan dalam fungsi item di antara kelompok peserta. Namun, **Item1** memiliki nilai probabilitas 0.0472, di bawah ambang batas 0.05, yang mengindikasikan potensi bias DIF (Wahyuni, 2022).

Analisis lebih lanjut pada kolom Between-Class MEAN-SQUARE dan t-ZSTD menunjukkan bahwa Item1 memiliki nilai t-ZSTD sebesar 1.0800, mendekati ambang batas signifikan (±2). Ini memperkuat indikasi

https://doi.org/10.17509/jik.v22i1.77242

bias pada item tersebut. Selain itu, Item5 memiliki perbedaan DIF Score Measure yang signifikan antara kelompok L (-1.53) dan kelompok P (2.18), yang menunjukkan bahwa peserta perempuan memiliki peluang lebih besar untuk menjawab item ini dengan benar dibandingkan peserta laki-laki.

DIF digunakan untuk menilai apakah suatu item bekerja secara adil di berbagai kelompok peserta berdasarkan karakteristik tertentu. Sebagian besar item dalam tabel memiliki nilai probabilitas yang cukup tinggi (> 0.05) pada kolom Chi-Square, menunjukkan tidak ada perbedaan signifikan dalam fungsi item di antara kelompok peserta. Namun, Item1 memiliki nilai probabilitas 0.0472, yang berada di bawah 0.05, mengindikasikan potensi bias DIF (Wahyuni, 2022). Pada kolom Between-Class MEAN-SQUARE dan t-ZSTD, Item 1 menunjukkan nilai t-ZSTD sebesar 1.0800, mendekati ambang batas signifikan (±2), yang memperkuat indikasi bias. Selain itu, tabel detail menunjukkan bahwa Item5 memiliki perbedaan DIF Score Measure signifikan antara kelas L (-1.53) dan kelas P (2.18), yang menunjukkan bahwa peserta di kelas perempuan memiliki peluang lebih besar untuk menjawab item ini dengan benar dibandingkan kelas laki-laki.

Gambar 6. Plot DIF Bias Gender pada Signifikansi Item Jawaban
Sumber: *Dokumentasi Penulis 2024*

Gambar 6 memperlihatkan bahwa beberapa item, seperti **Item1** dan **Item5**, memiliki garis yang tidak berimpit, menunjukkan adanya perbedaan signifikan antar kelompok. Sebaliknya, item seperti **Item8** dan **Item10** menunjukkan garis yang berimpit, yang menandakan tidak adanya perbedaan signifikan dalam kesulitan item antar kelompok.

Sebagian besar item menunjukkan fungsi yang seragam, yang menandakan kualitas instrumen secara keseluruhan cukup baik. Namun, item seperti **Item1** dan **Item5** memerlukan evaluasi lebih lanjut untuk memastikan keadilan pengukuran. Bias DIF pada item tertentu mungkin disebabkan oleh faktor eksternal, seperti konteks budaya atau pengalaman spesifik peserta (Bialo & Li., 2024; Ray *et al*, 2024).

Penanganan *Differential Item Functioning* (DIF) bertujuan untuk memastikan keadilan dan validitas instrumen pengukuran melalui serangkaian langkah strategis. Proses ini diawali dengan identifikasi item bermasalah, di mana item yang menunjukkan bias DIF diidentifikasi menggunakan analisis statistik berbasis model Rasch. Setelah itu, langkah berikutnya adalah revisi atau penghapusan item. Item yang terdeteksi memiliki bias dapat direvisi untuk menghilangkan ketidakseimbangan atau, jika revisi terbukti tidak efektif, item tersebut dapat dihapus dari instrumen. Tahap selanjutnya adalah evaluasi ulang, di mana instrumen yang telah direvisi diuji kembali pada populasi yang lebih beragam untuk memastikan tidak ada bias tambahan yang terdeteksi.

https://doi.org/10.17509/jik.v22i1.77242

Selain itu, dalam konteks validasi instrumen, apabila penghapusan item dengan DIF memengaruhi validitas isi, maka item tersebut tidak lagi digunakan dalam perbandingan ukuran individu, seperti rata-rata skor antar kelompok sampel. Namun, jika item dengan DIF tetap dipertahankan, penting untuk mempertimbangkan faktor budaya, bahasa, dan latar belakang responden saat instrumen diaplikasikan. Dengan demikian, langkah-langkah ini diharapkan dapat menjaga keadilan dan akurasi instrumen dalam mengukur kemampuan peserta secara objektif dan konsisten.

Discussion

Penelitian ini menyoroti pengembangan instrumen penilaian hasil belajar matematika berbasis Kurikulum Merdeka dengan pendekatan model Rasch. Temuan utama menunjukkan bahwa instrumen ini memiliki validitas dan reliabilitas yang tinggi, didukung oleh analisis menggunakan perangkat lunak Winsteps. Dengan Alpha Cronbach sebesar 0,70 dan reliabilitas item sebesar 0,99, instrumen ini memenuhi kriteria unidimensionalitas serta mencerminkan sebaran tingkat kesulitan yang bervariasi, mulai dari sangat mudah hingga sangat sulit. Namun, analisis lebih lanjut mengungkapkan adanya bias DIF pada Item 1 dan Item 5 berdasarkan gender, yang mengindikasikan perlunya revisi atau penyesuaian untuk memastikan instrumen lebih adil dalam mengukur kemampuan peserta didik di berbagai kelompok demografi.

Penelitian ini memperkuat hasil studi sebelumnya mengenai validitas dan reliabilitas instrumen berbasis model Rasch. Kemampuan model Rasch dalam mendeteksi butir soal yang tidak sesuai dan memberikan wawasan tentang daya beda serta tingkat kesulitan soal. Penelitian lainnya menunjukkan efektivitas model ini dalam menilai unidimensionalitas, memastikan bahwa setiap item mengukur konstruk laten yang sama (Latifah *et al*., 2024; Tarigan *et al*., 2022). Selain itu, pentingnya deteksi bias DIF untuk meningkatkan keadilan evaluasi berdasarkan karakteristik demografi (Wahyuni, 2022). Studi ini melampaui penelitian sebelumnya dengan mengintegrasikan analisis DIF dan evaluasi unidimensionalitas secara komprehensif, mendukung penggunaan model Rasch untuk mengidentifikasi anomali respons sebagaimana disampaikan (Eliza & Yusmaita, 2021).

Penelitian ini juga melengkapi studi tambahan yang menggunakan Teori Respons Butir (IRT) dan Psikometri dalam mengevaluasi kualitas instrumen penilaian. Penelitian menyoroti pengembangan instrumen berbasis kompetensi untuk pembelajaran kontekstual yang menekankan validitas dan reliabilitas sebagai komponen utama evaluasi (Komisia *et al*., 2021). Penggabungan analisis reliabilitas, daya beda, dan bias DIF mendukung pendekatan yang diadopsi dalam penelitian ini (Natanael *et al*., 2022). Selain itu, penggunaan *machine learning* untuk mendeteksi bias DIF, yang memberikan inspirasi bagi pengembangan instrumen evaluasi yang lebih adaptif dan canggih di masa depan (Peng *et al*., 2024).

Kontribusi penting dari penelitian ini mencakup pengembangan instrumen baru yang valid dan reliabel berbasis Kurikulum Merdeka untuk menilai kemampuan matematika peserta didik kelas VI. Penelitian ini juga memperkenalkan pendekatan yang menggabungkan analisis DIF, unidimensionalitas, dan reliabilitas item, menghasilkan alat ukur yang komprehensif dan berkualitas. Selain itu, penelitian ini mengusulkan *framework* praktis untuk memperbaiki item yang bias berdasarkan gender, meningkatkan keadilan dan efektivitas instrumen evaluasi.

CONCLUSION

Penelitian ini berhasil mengembangkan instrumen penilaian hasil belajar matematika berbasis Kurikulum Merdeka untuk peserta didik kelas VI Sekolah Dasar dengan fokus pada validitas dan reliabilitas menggunakan model Rasch. Instrumen ini mencakup aspek kognitif, afektif, dan psikomotor, serta telah diuji pada 213 peserta didik dengan pendekatan kuantitatif menggunakan perangkat lunak Winsteps. Hasil

https://doi.org/10.17509/jik.v22i1.77242

analisis menunjukkan bahwa instrumen memiliki reliabilitas yang tinggi (Alpha Cronbach = 0,70; reliabilitas item = 0,99) dan memenuhi kriteria unidimensionalitas. Sebaran tingkat kesulitan butir soal yang bervariasi dari sangat mudah hingga sangat sulit mencerminkan kemampuan instrumen dalam mengukur peserta didik dengan tingkat kemampuan yang beragam. Seluruh item memenuhi kriteria kecocokan berdasarkan statistik fit (Outfit MNSQ, ZSTD, dan *Point Measure Correlation*). Namun, dua item (Item 1 dan Item 5) menunjukkan bias *Differential Item Functioning* (DIF) signifikan berdasarkan analisis gender, yang memerlukan evaluasi dan penyesuaian lebih lanjut. Dengan demikian, instrumen yang dikembangkan dalam penelitian ini dinyatakan valid, reliabel, dan layak digunakan sebagai alat evaluasi kemampuan matematika peserta didik kelas VI sesuai dengan prinsip Kurikulum Merdeka. Penggunaan model Rasch terbukti efektif dalam menganalisis kualitas item secara mendalam, memastikan instrumen yang dihasilkan mampu memberikan pengukuran yang adil, akurat, dan komprehensif. Hasil penelitian ini diharapkan dapat menjadi referensi bagi pengembangan instrumen penilaian lainnya yang berorientasi pada peningkatan kualitas pembelajaran dan evaluasi hasil belajar peserta didik. Peningkatan kualitas instrumen penilaian memerlukan pemeriksaan ulang terhadap item dengan bias DIF signifikan, terutama dalam konteks bahasa, budaya, dan pengalaman peserta. Modifikasi atau penggantian item dapat dilakukan untuk memastikan keadilan pengukuran di berbagai kelompok. Selain itu, penting untuk menambah jumlah item dalam kategori sangat mudah dan sangat sulit guna memperluas rentang kemampuan yang dapat diukur, sehingga instrumen lebih adaptif terhadap kelompok peserta didik dengan kemampuan ekstrem. Pelatihan bagi guru juga menjadi prioritas agar mereka dapat memahami penggunaan instrumen berbasis model Rasch, membaca hasil analisis dengan benar, dan memberikan umpan balik yang lebih efektif kepada peserta didik. Uji validasi lanjutan dengan melibatkan sampel yang lebih besar dan beragam dari berbagai daerah juga diperlukan untuk memvalidasi instrumen dan meningkatkan generalisasi hasil penelitian. Akhirnya, instrumen ini dapat diimplementasikan secara praktis sebagai alat penilaian dalam pembelajaran berbasis Kurikulum Merdeka, mendukung pengukuran yang lebih holistik dan adil bagi peserta didik.

AUTHOR'S NOTE

Penulis menyatakan bahwa tidak ada konflik kepentingan terkait publikasi artikel ini. Penulis menegaskan bahwa data dan isi artikel bebas dari plagiarisme.

REFERENCES

Abdullaev, D., Shukhratovna, D. L., Rasulovna, J. O., Umirzakovich, J. U., & Staroverova, O. V. (2024). Examining local item dependence in a cloze test with the Rasch Model. *International Journal of Language Testing*, *14*(1), 75-81.

Adiyana, S. (2024). Peningkatan kemampuan menghitung pecahan melalui Model Problem Based Learning pada siswa kelas VI SDN 01 Ngunut. *Jurnal Edukasi Indonesia, 12*(3), 120-136.

Azizah, & Wahyuningsih, S. (2020). Penggunaan model Rasch untuk analisis instrumen tes pada Mata kuliah Matematika Aktuaria**.** *Jurnal Pendidikan Matematika (Jupitek)*, *3*(1), 45-50.

Bialo, J. A., & Li, H. (2024). An analysis of DIF and sources of DIF in achievement motivation items using anchoring vignettes. *Educational Assessment*, *29*(4), 293-318.

Dwilesanti, W. G., & Yudiarso, A. (2022). Rasch analysis of the Indonesian version of INDIVIDUAL Work Performance Questionnaire (IWPQ). *JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)*, *11*(2), 153-167.

El Fahmi, E., Khoirot, U., & Astutik, F. (2021). Analisis psikometri aitem need of agression tes EPPS pada remaja akhir. *Psikoislamika: Jurnal Psikologi dan Psikologi Islam, 18*(2), 295-306.

https://doi.org/10.17509/jik.v22i1.77242

Eliza, W., & Yusmaita, E. (2021). Pengembangan butir soal literasi Kimia pada materi sistem koloid kelas XI IPA SMA/MA. *Jurnal Eksakta Pendidikan (JEP)*, *5*(2), 197-204.

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.

Firdaus, F., Huda, A., Irfan, D., & Hebdriyani, Y. (2022). Pengembangan sistem Computer Adaptive Test (CAT) dengan pendekatan Item Response Theory (IRT). *EduTech: Jurnal Teknologi Pendidikan*, *21*(3), 272-286.

Hidayat, R., Patras, Y. E., Harijanto, S., & Hasanah, L. (2020). Analisis instrumen dan prioritas tindakan untuk kepuasan kerja guru di Indonesia berdasarkan pemodelan Rasch. *Kelola: Jurnal Manajemen Pendidikan*, *7*(2), 110-130.

Jones, R. J., Brown, D. E., & Smith, T. L. (2021). Competency-based assessment in modern curriculum: A contextual approach. *Educational Measurement Quarterly, 45*(3), 150-168.

Juliani, R. P., & Erita, S. (2023). Analisis validitas dan reliabilitas instrumen penilaian kemampuan berpikir kritis dalam konteks sekolah menengah. *JEID: Journal of Educational Integration and Development*, *3*(3), 169-179.

Jumini, J., & Retnawati, H. (2022). Estimating item parameters and student abilities: An IRT 2PL analysis of mathematics examination. *Al-Ishlah: Jurnal Pendidikan*, *14*(1), 385-398.

Kennedy-Shaffer, L., Qiu, X., & Hanage, W. P. (2021). Snowball sampling study design for serosurveys early in disease outbreaks. *American Journal of Epidemiology*, *190*(9), 1918-1927.

Kim, S., & Kim, J. (2022). Advancing Rasch analysis for holistic student assessment. *Journal of Educational Measurement, 59*(1), 78-95.

Komisia, F., Tukan, M. I. B., & Leba, M. A. U. (2021). Pengembangan perangkat pembelajaran berbasis pendekatan kontekstual untuk siswa SMA. *Indonesian Journal of Educational Science (IJES)*, *3*(2), 98-104.

Latifah, M., Saripah, I., Suryana, D., & Sunarya, Y. (2024). Validity and reliability of self-concept instrument using Rasch Model. *Jurnal Kajian Bimbingan dan Konseling*, *9*(1), 26-35.

Marwa, N. W. S., Pitria, P. R., & Madani, F. (2024). Development of authentic assessment of 21st-century skills in kurikulum merdeka. *Inovasi Kurikulum, 21*(2), 635-646.

Maulana, A. (2022). Analisis validitas, reliabilitas, dan kelayakan instrumen penilaian rasa percaya diri siswa. *Jurnal Kualita Pendidikan*, *3*(3), 133-139.

Natanael, Y., Salsabilla, R., Aulia, D., Khoirunnisa, D., Munawar, H. N., Hidayat, N. S., & Firdaus, R. F. (2022). Rasch rating scale model: Bias detection and validation test of Indonesian-adolescent life satisfaction scale. *Psympathic: Jurnal Ilmiah Psikologi, 9*(1), 31-44.

Nguyen, T., Pham, L., & Tran, H. (2023). Context-based learning and its impact on problem-solving skills*. Educational Research Review, 58*(1), 45-67.

Nizaruddin, N., Muhtarom, M., Murtianto, Y. H., & Sutrisno, S. (2024). Examining the self-regulated learning scale using the Rasch model approach. *Indonesian Journal of Science and Mathematics Education*, *7*(3), 518-530.

Noben, I., Maulana, R., Deinum, J. F., & Hofman, W. A. (2021). Measuring university teachers' teaching quality: A Rasch modelling approach. *Learning Environments Research*, *24*(1), 87-107.

Novriyanti, E., & Arthur, R. (2024). Analisis kualitas butir soal ujian tengah semester Biologi umum menggunakan Model Rasch. *JagoMIPA: Jurnal Pendidikan Matematika dan IPA*, *4*(4), 718-733.

Nudin, I., & Hidayatullah, R. S. (2023). Analisis butir soal penilaian tengah semester menggunakan model Rasch di SMK Negeri 5 Surabaya. *JPTM*, *12*(2), 132-139.

https://doi.org/10.17509/jik.v22i1.77242

Nurdiana, N. (2023). Meningkatkan hasil belajar operasi hitung bilangan pecahan dengan kartu bilangan siswa kelas VI SD Negeri Krueng Baung. *Jurnal Bima: Pusat Publikasi Ilmu Pendidikan Bahasa dan Sastra*, *1*(3), 338-348.

Ocy, D. R., Rahayu, W., & Makmuri, M. (2023). Rasch model analysis: Development of hots-based mathematical abstraction ability instrument according to Riau Islands Culture. *Aksioma: Jurnal Program Studi Pendidikan Matematika*, *12*(4), 3542-3560.

Oktaviyanthi, R., Agus, R. N., Garcia, M. L. B., & Lertdechapat, K. (2024). Cognitive load scale in learning formal definition of limit: A Rasch model approach. *Infinity Journal of Mathematics Education, 13*(1), 99-118.

Peng, K., Chen, M., Zhou, L., & Weng, X. (2024). Differential item functioning in the autism behavior checklist in children with autism spectrum disorder based on a machine learning approach. *Frontiers in Psychiatry, 15*(1), 1-14.

Ray, J. V., Baker, T., & Peck, J. H. (2024). An examination of differential item functioning in a measure of self-reported offending across race and ethnicity among a sample of justice-involved youth. *Justice Quarterly, 1*(1), 1-25.

Rustiati, T. (2023). Upaya meningkatkan hasil belajar siswa kelas VI SD pada konsep operasi hitung pecahan pada mata pelajaran Matematika melalui metode demostrasi. *Jurnal Pendidikan Abad Ke-21*, *1*(1), 17-29.

Ruswan, R. (2020). Penggunaan pendekatan kooperatif dalam pembelajaran Matematika tentang operasi hitung pecahan untuk meningkatkan hasil belajar siswa sekolah dasar. *Pedadidaktika: Jurnal Ilmiah Pendidikan Guru Sekolah Dasar*, *7*(3), 58-67.

Safitri, E., & Widyanti, E. (2024). Analisis penilaian guru yang efektif pada pencapaian kompetensi pengetahuan siswa. *Ihsan: Jurnal Pendidikan Islam, 2*(2), 227-235.

Saputri, R. E., Firmansyah, R., & Silfiya, S. (2024). Pentingnya evaluasi pembelajaran untuk meningkatkan kompetensi peserta didik di sekolah dasar. *Sindoro: Cendikia Pendidikan*, *3*(8), 21-30.

Smith, J. K., Lee, M., & Davis, K. (2022). Integrating real-life scenarios into classroom assessments. *Journal of Educational Innovation, 20*(4), 101-120.

Sholikhah, M., & Hidayati, Y. M. (2024). Summative assessment planning in the kurikulum merdeka on two-dimensional figure materials. *Inovasi Kurikulum, 21*(1), 467-480.

Tarigan, E. F., Nilmarito, S., Islamiyah, K., Darmana, A., & Suyanti, R. D. (2022). Analisis instrumen tes menggunakan Rasch model dan Software SPSS 22.0. *Jurnal Inovasi Pendidikan Kimia*, *16*(2), 92-96.

Wahyuni, A. (2022). Detection of gender biased using DIF (Differential Item Functioning) analysis on item test of school examination Yogyakarta. *Jurnal Evaluasi Pendidikan*, *13*(1), 46-49.

Wang, X., & Zheng, Y. (2023). Improving adaptive testing through psychometric modeling. *International Journal of Educational Technology, 14*(2), 112-126.

Wibowo, S. A., Degeng, M. D. K., & Praherdhiono, H. (2024). Interactive video for learning Mathematics element of measurement in elementary school. *Inovasi Kurikulum, 21*(2), 723-736.

Widodo, H. (2020). Penilaian kontekstual untuk meningkatkan kompetensi numerasi. *Jurnal Pendidikan dan Kebudayaan, 26*(4), 127-140.

Wijayanto, F., Bucur, I. G., Mul, K., Groot, P., van Engelen, B. G., & Heskes, T. (2023). Semi-automated Rasch analysis with differential item functioning. *Behavior Research Methods*, *55*(6), 3129-3148.

Yektiana, N., & Nursikin, M. (2023). Konsep dasar pengukuran, penilaian, dan evaluasi hasil belajar pendidikan agama Islam. *J-Ceki: Jurnal Cendekia Ilmiah*, *2*(2), 263-266.

Yusuf, S., Budiman, N., Yudha, E. S., Suryana, D., & Yusof, S. M. J. B. (2021). Rasch analysis of the Indonesian mental health screening tooals. *The Open Psychology Journal*, *14*(1), 198-203.

https://doi.org/10.17509/jik.v22i1.77242