Teorema: *Teori dan Riset Matematika*, 09(02), 193–206 September 2024
p-ISSN 2541-0660, e-ISSN 2597-7237 © 2024
https://jurnal.unigal.ac.id/teorema/article/view/14867

Konstruksi Instrumen Tes Kemampuan Pemecahan Masalah Menggunakan Teori Respon Butir

Nur Aini Amalia Sari[1], Vindy Antia[2], Ummu Soim Daimah[3], Ihtiyatul Muhakimah[4], Sintha Sih Dewanti[5*]
[1, 2, 3, 4]Universitas Islam Negeri Sunan Kalijaga Yogyakarta, Jl. Laksda Adisucipto, Kabupaten Sleman, DIY, Indonesia
E-mail: sintha.dewanti@uin-suka.ac.id
**Corresponding Author*

ABSTRACT

Mathematical problem-solving skills equip students with critical and analytical thinking skills essential to face real-world challenges and prepare them to become adaptive and innovative thinkers in various fields. The right instrument for measuring mathematical problem-solving is the key to evaluating students' in-depth ability to analyze, formulate, and solve mathematical problems accurately. Therefore, this study aims to develop a test instrument to measure students' mathematical problem-solving skills. The method used in this study is development research with the ADDIE model. The subjects of this study were 121 students at one of the junior high schools in Bantul Regency, Yogyakarta. This study used an instrument developed by researchers and then validated by experts. After being validated, the instrument was tested on students who had studied the Two-Variable Linear Equation System material. The results of the test items were analyzed using Item Response Theory (IRT). Based on the results and discussion, it can be concluded that the problem-solving ability test instrument that has been developed is included in the valid category with a value of 0.97 and reliable with a value of 0.736. Based on the results of the exploratory factor analysis, it is known that the problem-solving ability test measures 2 factors. Of the 6 questions, 5 questions are suitable for use. Based on the results of the model fit test after modification, it shows a p-value of 0.08021., so the model can be accepted as a fit model. Based on the PCM results, it shows that items that have a threshold of 1 <threshold 2 are only in items 1 and 3, while the other two items have a threshold of 4. The problem-solving test instrument can accurately capture high, medium, and low abilities in each problem-solving indicator, namely identifying problems, developing strategies, implementing strategies, and re-checking solutions.

Keywords: *problem solving ability, item response theory, CFA, EFA*

ABSTRAK

Kemampuan pemecahan masalah matematika membekali siswa dengan keterampilan berpikir kritis dan analitis yang esensial untuk menghadapi tantangan dunia nyata dan mempersiapkan mereka menjadi pemikir yang adaptif dan inovatif di berbagai bidang. Instrumen yang tepat untuk pengukuran pemecahan masalah matematika adalah kunci untuk mengevaluasi kemampuan siswa secara mendalam dalam menganalisis, merumuskan, dan menyelesaikan masalah matematika dengan akurat. Oleh sebab itu, penelitian ini bertujuan untuk mengembangkan instrumen tes dalam mengukur kemampuan pemecahan masalah matematis peserta didik. Metode yang digunakan dalam penelitian ini yaitu penelitian pengembangan dengan model ADDIE. Subjek penelitian ini sebanyak 121 peserta didik pada salah satu madrasah tsanawiyah di Kabupaten Bantul Yogyakarta. Penelitian ini menggunakan instrumen yang telah dikembangkan oleh peneliti dan kemudian divalidasi oleh ahli. Setelah divalidasi, instrumen tersebut diujicobakan kepada peserta didik yang telah mempelajari materi Sistem Persamaan Linear Dua Variabel (SPLDV). Hasil uji coba butir soal dianalisis dengan menggunakan *Item Response Theory* (IRT). Berdasarkan hasil dan pembahasan dapat disimpulkan bahwa instrumen tes kemampuan pemecahan masalah yang telah dikembangkan termasuk dalam kategori valid dengan nilai sebesar 0,97 dan reliabel dengan nilai sebesar 0,736. Berdasarkan hasil analisis faktor eksploratori diketahui bahwa tes kemampuan pemecahan masalah mengukur 2 faktor. Dari 6 butir soal terdapat 5 butir soal yang layak untuk digunakan. Berdasarkan hasil uji kecocokan model setelah dimodifikasi menunjukkan nilai p-value sebesar 0, 08021., sehingga model dapat diterima sebagai model yang fit. Berdasarkan hasil PCM menunjukkan item yang memiliki threshold 1 < threshold 2 hanya pada item 1 dan 3, sedangkan kedua item lainnya yang memiliki threshold 4. Instrumen tes pemecahan masalahan dapat memotret kemampuan tinggi, sedang, rendah secara akurat pada setiap indikator pemecahan masalah, yaitu mengidentifikasi masalah, menyusun strategi, menerapkan strategi, dan memeriksa kembali solusi.

Kata kunci: kemampuan pemecahan masalah, teori respon butir, CFA, EFA

Dikirim: Juni 2024; Diterima: Agustus 2024; Dipublikasikan: September 2024
Cara sitasi: Sari Nur Aini, A., Antia, V., Daimah ummu, S., Muhakimah, I., Dewanti Sintha, S, (2024). Konstruksi Instrumen Tes Kemampuan Pemecahan Masalah Menggunakan Teori Respon Butir. *Teorema*: *Teori dan Riset Matematika, 09*(02), 193–206. DOI:http://dx.doi.org/10.25157/teorema.v9i2.14867

PENDAHULUAN

Kemampuan pemecahan masalah dalam pendidikan matematika merupakan salah satu kompetensi utama yang harus dikuasai peserta didik, karena berperan penting dalam pengembangan keterampilan berpikir kritis dan analitis yang dibutuhkan di berbagai aspek kehidupan sehari-hari dan pekerjaan . Menurut Polya (1957) pemecahan masalah bukan hanya tentang menemukan solusi, tetapi juga melibatkan proses berpikir logis yang mendalam yang membantu peserta didik mengembangkan strategi berpikir yang sistematis. Selain itu, penelitian Jonassen (2021) menunjukkan bahwa peserta didik yang terampil dalam pemecahan masalah matematika memiliki kemampuan adaptasi yang lebih baik terhadap tantangan yang kompleks dalam kehidupan nyata. Lebih lanjut, pemecahan masalah matematika juga membantu peserta didik membangun pemahaman yang lebih baik terhadap konsep-konsep matematika melalui aplikasi langsung, bukan hanya menghafal rumus atau prosedur tanpa konteks. Menurut studi yang dilakukan oleh Schoenfeld (2022), kemampuan ini memperkuat pemahaman konseptual peserta didik serta meningkatkan motivasi belajar mereka karena melihat relevansi langsung matematika dalam kehidupan. Oleh karena itu, pemecahan masalah dianggap sebagai fondasi penting yang mendukung penguasaan matematika dan pengembangan keterampilan berpikir tingkat tinggi di kalangan peserta didik.

Pengukuran kemampuan pemecahan masalah secara valid dan reliabel merupakan tantangan yang kompleks dalam pendidikan matematika. Salah satu kesulitan utama adalah perbedaan dalam definisi dan interpretasi pemecahan masalah itu sendiri, yang dapat mempengaruhi validitas instrumen pengukuran (Pape & Smith, 2019). Menurut studi yang dilakukan oleh Lester (2020), masalah validitas juga muncul karena karakteristik pemecahan masalah yang multidimensional, yang mencakup kemampuan analisis, penalaran logis, dan kreativitas, sehingga sulit untuk mengembangkan alat ukur yang mampu menangkap semua aspek tersebut secara komprehensif. Selain itu, reliabilitas instrumen sering dipertanyakan karena perbedaan individu dalam pendekatan penyelesaian masalah, di mana peserta didik dapat menggunakan strategi yang sangat bervariasi meskipun menghadapi masalah yang sama (Boesen et al., 2021). Penelitian Karadag (2022) menunjukkan bahwa variasi dalam tingkat kesulitan soal pemecahan masalah juga berpengaruh pada reliabilitas pengukuran, karena instrumen yang tidak diadaptasi dengan baik untuk berbagai tingkat kemampuan peserta didik cenderung menghasilkan hasil yang bias. Selain itu, faktor eksternal seperti lingkungan pengujian, tekanan waktu, dan kelelahan peserta didik dapat mempengaruhi performa mereka saat mengerjakan soal pemecahan masalah, yang pada akhirnya memengaruhi reliabilitas hasil tes (Greiff & Funke, 2021). Oleh karena itu, mengembangkan instrumen pengukuran yang valid dan reliabel untuk pemecahan masalah membutuhkan perencanaan yang teliti dan penggunaan metode pengukuran yang canggih.

Kebutuhan akan instrumen tes yang mampu mengevaluasi kemampuan pemecahan masalah secara objektif semakin mendesak dalam pendidikan matematika. Instrumen yang objektif diperlukan agar evaluasi tidak hanya mengandalkan intuisi atau penilaian subjektif dari guru, melainkan berdasarkan hasil pengukuran yang konsisten dan dapat diandalkan (Anderson & Shattuck, 2018). Menurut kajian oleh Park dan Leung (2019), instrumen tersebut harus dirancang untuk menangkap berbagai aspek pemecahan masalah, seperti identifikasi masalah, perencanaan solusi, dan evaluasi hasil, yang semuanya memerlukan pendekatan yang terstruktur dan sistematis. Selain itu, objektivitas dalam pengukuran juga berperan penting dalam memastikan bahwa hasil tes mencerminkan kemampuan peserta didik yang sebenarnya, tanpa dipengaruhi oleh faktor-faktor eksternal seperti bias pengajar atau situasi pengujian (Freeman & Higgins, 2020). Studi oleh He & Larkin (2021) menekankan bahwa penggunaan teknologi dan algoritma berbasis data dapat meningkatkan objektivitas tes dengan mengurangi ketergantungan pada penilaian manusia. Oleh karena itu, pengembangan instrumen yang objektif dan berbasis bukti menjadi prioritas untuk mengukur kemampuan pemecahan masalah secara lebih akurat dan adil di kalangan peserta didik (Wu & Adams, 2022).

Tes merupakan cara penilaian yang disiapkan dan diterapkan kepada peserta didik pada waktu dan lokasi yang spesifik serta dalam kondisi memenuhi syarat-syarat yang telah ditentukan (Angriani et al., 2018). Tes mencakup pertanyaan, tugas, atau rangkaian tugas yang dirancang secara strategis untuk

memperoleh informasi mengenai sifat atau atribut pendidikan maupun psikologis (Magdalena et al., 2023). Dalam penilaian hasil belajar, tes diharapkan mampu mencerminkan sampel perilaku dan menghasilkan nilai yang objektif dan akurat, sehingga tes yang digunakan oleh pendidik harus memiliki kualitas yang baik dari berbagai aspek, disusun berdasarkan prinsip dan prosedur penyusunan tes, serta setelah digunakan perlu dievaluasi untuk menentukan apakah tes tersebut berkualitas baik atau tidak (Ikawati et al., 2022). Dengan demikian, tes adalah salah satu hal yang perlu diperhatikan karena kualitas sebuah tes menentukan keakuratan dalam mengukur hasil belajar peserta didik dalam berbagai ranah, salah satunya pada ranah kognitif.

Matematika merupakan salah satu bidang studi yang menekankan pengembangan kemampuan pada ranah kognitif, yang mencakup proses berpikir dan pemahaman intelektual. Ranah kognitif berhubungan erat dengan tujuan belajar yang mengarah pada kemampuan berpikir, di mana peserta didik diharapkan dapat mencapai berbagai tingkat kemampuan yang terstruktur dalam hierarki kognitif (Oktaviana & Prihatin, 2018). Menurut taksonomi Bloom yang telah diperbarui, tingkatan kognitif tersebut mencakup kemampuan mengingat, memahami, menerapkan, menganalisis, mengevaluasi, dan menyintesis (Anderson & Krathwohl, 2001). Setiap tingkatan ini penting dalam pembelajaran matematika karena peserta didik diharapkan tidak hanya mampu menghafal rumus, tetapi juga memahami konsep, menerapkannya dalam berbagai konteks, serta menganalisis masalah secara kritis untuk menemukan solusi yang tepat (Angriani et al., 2018). Salah satu kemampuan kognitif yang menjadi fokus utama dalam mata pelajaran matematika adalah kemampuan pemecahan masalah. Kemampuan ini menuntut siswa untuk dapat mengidentifikasi masalah, merancang solusi, dan mengevaluasi hasilnya, yang merupakan inti dari keterampilan berpikir kritis dan analitis dalam matematika (Polya, 2020). Penguasaan kemampuan pemecahan masalah tidak hanya penting dalam konteks akademis, tetapi juga memiliki relevansi yang signifikan dalam kehidupan sehari-hari dan dunia kerja.

Kemampuan pemecahan masalah merupakan proses belajar yang melibatkan metode-metode ilmiah atau berpikir secara matematis, logis, sistematis, dan teliti (Aziza, 2019). Kemampuan ini adalah bagian yang paling pokok untuk dimiliki peserta didik, mengingat adanya perkembangan IPTEK yang semakin pesat (Situmorang & Bunawan, 2022). Dalam hal ini, peserta didik didorong untuk lebih giat dan aktif dalam proses belajar mengajar agar kemampuan mereka dalam memecahkan masalah matematis meningkat, sehingga mereka dapat menyusun dan menguji teori mereka sendiri, serta menguji teori teman sebaya, bahkan jika teori tersebut tidak sesuai dengan kenyataan, peserta didik harus mampu meninggalkannya dan mencoba teori lain (Situmorang & Bunawan, 2022). Peserta didik harus mampu mencari berbagai solusi dari suatu permasalahan yang kompleks melalui *point of view* yang berbeda.

Berdasarkan penjelasan di atas, kemampuan pemecahan masalah dapat ditentukan dengan melihat setiap tahapan yang digunakan dalam memecahkan suatu soal pemecahan masalah. Tahapan tersebut diantaranya mengidentifikasi masalah, membuat rencana, melaksanakan rencana, dan memeriksa kembali (Polya, 1957). Kemudian pendapat lain menyatakan bahwa tahapan pada kemampuan pemecahan masalah yaitu mengidentifikasi unsur-unsur yang diketahui, yang ditanyakan, dan kecukupan unsur yang diperlukan; merumuskan masalah matematik atau menyusun model matematika; menerapkan strategi untuk menyelesaikan berbagai masalah (sejenis dan masalah baru) dalam atau di luar matematika, menjelaskan atau menginterpretasikan hasil sesuai permasalahan asal, dan menggunakan matematika secara bermakna (NCTM, 2000). Lalu, terdapat juga pendapat lain yang berpendapat bahwa tahapan pada kemampuan pemecahan masalah yaitu mengidenfikasi masalah, menentukan tujuan; mengeksplorasi strategi yang mungkin; menerapkan strategi; dan meninjau kembali solusi yang ditemukan (Bransford & Stein, 1993). Selain itu, ada juga yang menyebutkan bahwa tahapan pada kemampuan pemecahan masalah adalah merumuskan masalah, menganalisis masalah; merumuskan hipotesis, hasil yang mungkin dan kemudian bertindak berdasarkan strategi itu, merumuskan rekomendasi pemecahan masalah; dan pengujian hipotesis (Dewey, 1933). Dapat disimpulkan bahwa tahapan yang dapat digunakan untuk memecahkan soal pemecahan masalah yaitu mengidentifikasi masalah, menyusun strategi, menerapkan strategi atau rencana, dan memeriksa kembali solusi yang didapat.

Satu di antara kemampuan matematis yang dibutuhkan oleh peserta didik yaitu pemecahan masalah. Pemecahan masalah matematis merupakan salah satu keterampilan utama yang harus dikuasai oleh peserta didik sebagai bagian dari proses pembelajaran untuk mengaplikasikan ilmu yang mereka pelajari (Lukman et al., 2023). Pemecahan masalah perlu dikembangkan dalam proses pembelajaran matematika dan peserta didik perlu dibiasakan untuk memecahkan berbagai masalah, baik yang bersifat matematis maupun yang terkait dengan kehidupan sehari-hari (Fitrianty et al., 2022). Kemampuan pemecahan masalah sangat dibutuhkan oleh peserta didik, hal ini karena seiring dengan perkembangan zaman menyebabkan ilmu pengetahuan dan teknologi berkembang sangat pesat dan memungkinkan siapapun untuk mendapatkan informasi secara cepat dan mudah dari berbagai sumber (Rahmani & Widyasari, 2018). Berdasarkan penjelasan di atas dapat diketahui bahwa pentingnya kemampuan pemecahan masalah dalam pembelajaran matematika, sehingga kemampuan tersebut perlu dimiliki oleh setiap peserta didik.

Teori Respon Butir (Item Response Theory, IRT) adalah model psikometrik yang digunakan untuk menganalisis hubungan antara kemampuan individu dan respons mereka terhadap butir soal, dengan mempertimbangkan karakteristik tiap butir, seperti tingkat kesulitan dan daya beda (Embretson & Reise, 2020). IRT memiliki keunggulan dibandingkan dengan Teori Tes Klasik (Classical Test Theory, CTT), terutama dalam hal keakuratan pengukuran kemampuan, karena IRT tidak bergantung pada distribusi sampel atau karakteristik tes secara keseluruhan, melainkan pada parameter individu dari tiap butir soal (De Ayala, 2018). Menurut Kim dan Bolt (2021), salah satu kekuatan utama IRT adalah kemampuannya untuk memberikan estimasi kemampuan yang lebih tepat pada berbagai tingkat kesulitan soal, sehingga lebih efektif dalam mengukur kemampuan peserta didik yang heterogen.

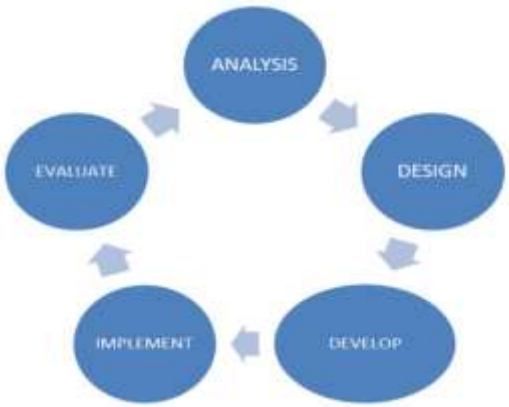
Relevansi IRT dalam pengembangan instrumen pengukuran yang lebih akurat semakin terlihat dalam domain pemecahan masalah matematika. Seperti yang dikemukakan oleh Wang dan Cheng (2019), IRT memungkinkan pengembang instrumen untuk mendesain soal-soal yang lebih sesuai dengan variasi strategi dan keterampilan pemecahan masalah yang dimiliki peserta didik. Dengan model ini, instrumen dapat menyesuaikan dengan perbedaan individual dalam kemampuan pemecahan masalah, serta memberikan hasil pengukuran yang lebih valid dan reliabel dibandingkan dengan CTT yang cenderung lebih terbatas dalam mengakomodasi variasi tersebut (Baker & Kim, 2022). Oleh karena itu, IRT sangat relevan dalam meningkatkan kualitas instrumen evaluasi matematika, terutama dalam mengukur kemampuan kompleks seperti pemecahan masalah.

Kurangnya instrumen yang dirancang menggunakan Teori Respon Butir (Item Response Theory, IRT) untuk mengukur kemampuan pemecahan masalah matematika masih menjadi tantangan besar dalam dunia pendidikan. Menurut studi yang dilakukan oleh Embretson dan Reise (2020), IRT memiliki potensi besar untuk memberikan estimasi kemampuan yang lebih akurat dan berbasis data pada berbagai level kemampuan peserta didik, namun penerapannya dalam konteks pemecahan masalah matematika masih terbatas. Hal ini diperkuat oleh temuan Wang dan Cheng (2019), yang menunjukkan bahwa sebagian besar instrumen matematika yang ada masih menggunakan pendekatan klasik seperti Classical Test Theory (CTT), yang sering kali tidak cukup sensitif untuk mengukur aspek-aspek kompleks dari pemecahan masalah.

Selain itu, Kim dan Bolt (2021) mengungkapkan bahwa salah satu kendala utama dalam penerapan IRT adalah kurangnya pengembangan butir soal yang benar-benar mampu mengukur keterampilan pemecahan masalah dengan berbagai tingkat kesulitan. Banyak instrumen yang ada tidak didesain secara spesifik untuk mengevaluasi variasi dalam strategi dan pendekatan pemecahan masalah peserta didik, yang pada akhirnya membatasi keandalan hasil pengukuran. Menurut penelitian yang dilakukan oleh Baker dan Kim (2022), instrumen berbasis IRT masih jarang digunakan dalam pendidikan matematika karena keterbatasan sumber daya dan keahlian teknis yang diperlukan untuk pengembangan dan kalibrasi butir soal yang sesuai. Oleh karena itu, dibutuhkan lebih banyak upaya dalam pengembangan instrumen yang dirancang menggunakan IRT untuk memberikan penilaian yang lebih valid dan reliabel terhadap kemampuan pemecahan masalah matematika.

METODE PENELITIAN

Metode penelitian ini adalah penelitian pengembangan dengan model ADDIE (*Analyze, Design, Development, Implementation, and Evaluation*). Penelitian dan pengembangan atau biasa dikenal dengan Research and Development (R&D) adalah metode penelitian yang digunakan untuk menghasilkan produk tertentu (Sugiyono, 2010). Prosedur pengembangan soal tes menggunakan model ADDIE yaitu *analyze, design, development, implementation*, dan *evaluation* (Dick dan Carey, 1996).



Gambar 1. Tahapan ADDIE

Pada tahap analyze peneliti melakukan analisis terhadap kebutuhan dan konteks di sekolah. Pada tahap design, peneliti melakukan sintesis indikator kemampuan yang akan digunakan. Kemudian, pada tahap development, butir tes disusun berdasarkan indikator yang telah disintesis. Setelah butir tes siap digunakan, butir tes tersebut diimplementasikan dengan diujikan pada subjek penelitian, yaitu kelas VIII di salah satu madrasah tsanawiyah di Kabupaten Bantul, Yogyakarta sebanyak 121 peserta didik. Instrumen yang digunakan dalam mengumpulkan data berupa instrumen tes yang telah dikembangkan dan lembar validasi ahli. Instrumen yang dikembangkan telah divalidasi oleh 4 validator. Tahap terakhir, yaitu evaluation, peneliti mengevaluasi hasil uji coba dengan menganalisis untuk menentukan validitas dan reliabilitas instrumen tes kemampuan pemecahan masalah yang telah dikembangkan.

Hasil uji coba butir soal dianalisis dengan menggunakan *Item Response Theory* (IRT). Sebelum di analisis dengan menggunakan IRT, dilakukan analisis faktor dengan menggunakan *Exploratory* Faktor Analisis (EFA). EFA atau analisis faktor adalah sebuah metode analisis yang mengevaluasi hubungan antara semua variabel atau faktor yang ada, sehingga menghasilkan pengelompokan dari banyak variabel menjadi beberapa variabel atau faktor baru (Hanim & Ragimun, 2010). Setelah dilakukan analisis EFA, kemudian dilanjutkan analisis CFA untuk uji validitas dan reliabilitas. Validitas yang digunakan dalam artikel ini adalah validitas *construct* yang dibuktikan melalui analisis CFA yang meliputi: (1) menghitung kecukupan sampel, (2) menguji multikoliniearitas, (3) melakukan analisis CFA order-2 (Retnawati, 2016). Setelah analisis data dinyatakan fit, dilanjutkan analisis PCM. PCM adalah pengembangan dari model Rasch butir dikotomi yang diterapkan pada butir politomi (Safaruddin et al., 2012).

HASIL DAN PEMBAHASAN

Pengembangan soal tes untuk mengukur kemampuan pemecahan masalah pada materi Sistem Persamaan Linear Dua Variabel (SPLDV). Berdasarkan hasil reliabilitas diperoleh nilai koefisien Cronbach's Alpha sebesar 0,736. Hasil ini termasuk dalam kategori reliabilitas yang tinggi. Reliabilitas instrumen dapat dikatakan reliabel apabila nilai koefisien ≥ 0,70 (Sudijono, 2004). Setiap butir soal pada penelitian ini dikatakan reliabel karena Corrected Item-Total Correlation yang dihasilkan lebih dari 0,3. Dapat disimpulkan bahwa setiap butir soal pada penelitian ini termasuk reliabel dan dapat dilanjutkan ke tahap pengembangan selanjutnya.

Pengembangan soal tes selanjutnya melalui tahapan ADDIE yaitu *analysis, design, development, implementation,* dan *evaluation.* Pada tahap analisis, tujuan pengembangan soal tes adalah untuk mengukur kemampuan peserta didik dalam memecahkan masalah yang berkaitan dengan Sistem Persamaan Linear Dua Variable (SPLDV). Analisis ini mencakup beberapa indikator yaitu mengidentifikasi masalah, menyusun strategi, menerapkan strategi atau rencana, dan memeriksa kembali solusi yang didapat. Kemudian, pada tahap *design,* peneliti merancang butir tes uraian yang akan digunakan. Selanjutnya, pada tahap *development,* butir tes disusun berdasarkan indikator yang telah disintesis. Berikut butir soal yang telah dikembangkan oleh peneliti. Butir soal nomor 1 digunakan untuk mengukur indikator mengidentifikasi masalah.

1. Perhatikan sistem persamaan di bawah ini!

   (i). $\begin{cases} 5x + 4y = 13 \\ 3p + 2q = 7 \end{cases}$ (iii). $\begin{cases} 2p + q = 8 \\ pq - 2q = -3 \end{cases}$

   (ii). $\begin{cases} \frac{1}{x} + \frac{1}{y} = 5 \\ 5x - y = 3 \end{cases}$ (iv). $\begin{cases} \frac{5x}{2} = \frac{2y}{3} - 10 \\ \frac{2x-y}{4} = 3 \end{cases}$

   Manakah yang termasuk Sistem Persamaan Linear Dua Variabel? Tuliskan beserta alasannya!

Gambar 2. Butir Soal Nomor 1

Butir soal nomor 2 dan 3 digunakan untuk mengukur indikator menyusun strategi.

2. Vinda dan Dewa membeli buku dan bolpoin di toko yang sama. Vinda membeli 5 buku dan 3 bolpoin dengan total yang harus dibayar Vinda yaitu Rp 21.000. Sedangkan Dewa membeli 2 buku dan 1 bolpoin dengan total harga yang harus dibayar Dewa yaitu Rp 8.000. Buatlah model SPLDV dari permasalahan tersebut!
3. Diketahui keliling suatu persegi panjang yaitu 116 cm. Apabila selisih dari ukuran panjang dan lebarnya yaitu 6 cm. Bagaimanakah pemodelan SPLDV dari permasalahan tersebut?

Gambar 3. Butir Soal Nomor 2 dan 3

Butir soal nomor 4 dan 5 digunakan untuk mengukur indikator menerapkan strategi atau rencana.

4. Selesaikan sistem persamaan berikut ini menggunakan metode eliminasi!

   $\begin{cases} x = y + 6 \\ x + y = 42 \end{cases}$

5. Selesaikan sistem persamaan berikut ini menggunakan metode eliminasi dan substitusi!

   $\begin{cases} 2x - 3y = 7 \\ 3(x + y) - y = 4 \end{cases}$

Gambar 4. Butir Soal Nomor 4 dan 5

Butir soal nomor 6 digunakan untuk mengukur indikator memeriksa kembali solusi.

6. Diketahui jumlah uang saku Bram dan Daim adalah Rp 38.000,00. Apakah benar jika 2 kali uang saku Bram ditambah 2 kali uang saku Daim adalah Rp 76.000,00? (misalkan banyaknya uang saku Bram = x dan banyaknya uang saku Daim = y)

Gambar 5. Butir Soal Nomor 6

Setelah instrumen di kembangkan, peneliti melakukan uji validitas kepada 4 validator yang terdiri dari satu dosen pendidikan matematika UIN Sunan Kalijaga dan 3 mahasiswa. Validator memilih jawaban dengan memperhatikan kesesuaian aspek indikator dan pernyataannya. Berdasarkan hasil validasi butir soal pada instrumen diperoleh nilai koefisien Aikens sebesar 0,97. Validitas suatu butir soal dapat

dikatakan valid apabila nilai koefisien Aikens V ≥ 0,75 (Aiken, 1985). Dengan demikian, semua butir soal yang dikembangkan dalam penelitian ini dapat dikatakan valid. Tetapi, terdapat kritik dan saran ahli terhadap butir soal nomor 2 dan 3 yang telah dikembangkan. Pada butir soal nomor 2 terdapat saran yaitu penulisan soal yang tadinya Rp 21.000 seharusnya menjadi Rp 21.000,00. Kemudian, penulisan yang tadinya eliminasi-substitusi seharusnya menjadi eliminasi dan substitusi. Selanjutnya, pada butir soal nomor 3 lebih diperjelas ukuran yang lebih panjang antara sisi panjang dan sisi lebar pada persegi panjang, karena dapat menimbulkan perbedaan pada saat memodelkan. Berdasarkan kritik dan saran yang diberikan, selanjutnya peneliti melakukan revisi terhadap butir soal nomor 2 dan 3 yang dikembangkan. Berikut hasil revisi butir soal yang dikembangkan.

2. Vinda dan Dewa membeli buku dan bolpoin di toko yang sama. Vinda membeli 5 buku dan 3 bolpoin dengan total yang harus dibayar Vinda yaitu Rp 21.000,00. Sedangkan Dewa membeli 2 buku dan 1 bolpoin dengan total harga yang harus dibayar Dewa yaitu Rp 8.000,00. Buatlah model SPLDV dari permasalahan tersebut!
3. Diketahui keliling suatu persegi panjang yaitu 116 cm. Apabila selisih dari ukuran panjang dan lebarnya yaitu 6 cm (dimana ukuran lebar lebih kecil daripada ukuran panjangnya). Bagaimanakah pemodelan SPLDV dari permasalahan tersebut?

Gambar 6. Revisi Butir Soal Nomor 2 dan 3

Tahap selanjutnya yang dilakukan oleh peneliti adalah tahap *implementation.* Pada tahap ini, butir soal tes yang telah direvisi diujicobakan kepada seluruh peserta didik yang dijadikan subjek penelitian. Uji coba dilakukan dengan tujuan untuk memeroleh data nilai peserta didik. Kemudian, tahap yang terakhir adalah tahap *evaluation.* Pada tahap ini, dilakukan analisis data dengan menggunakan teori respons butir yaitu menguji dimensi-dimensi dari data empiris. Proses pengujian dilakukan dengan analisis faktor eksploratori dengan menggunakan metode *Exploratory Faktor Analysis* (EFA). Asumsi awal yang harus dipenuhi dalam EFA adalah uji KMO dan Bartlett. Hasil output KMO dan Bartlett Test.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.705
Bartlett's Test of Sphericity	Approx. Chi-Square	194.807
	df	15
	Sig.	.000

Gambar 7. KMO and Bartlett's Test

Hasil output menunjukkan bahwa Kaiser-Meyer-Olkin Measure of Sampling Adequacy = 0.705 menunjukkan KMO > 0.5, artinya penelitian ini memiliki data yang cukup. Selanjutnya, signifikansi dari Bartlett's Test menunjukkan bahwa matriks korelasi bukan merupakan matriks identitas, sehingga membentuk matriks korelasi dengan hubungan yang erat antara variable. Kemudian, berdasarkan korelasi anti-image, MSA > 0,5 sehinga semua data memenuhi syarat untuk dilakukan analisis faktor. Berikut hasil output *Anti-image Matrices*.

Tabel 1. Anti-image Matrices

		butir1	butir2	butir3	butir4	butir5	butir6
Anti-image Correlation	butir1	.712[a]	-.535	-.206	.041	-.043	-.124
	butir2	-.535	.693[a]	-.133	-.198	-.195	.094
	butir3	-.206	-.133	.695[a]	.105	.091	-.423
	butir4	.041	-.198	.105	.700[a]	-.289	-.353
	butir5	-.043	-.195	.091	-.289	.777[a]	-.097
	butir6	-.124	.094	-.423	-.353	-.097	.680[a]

a. Measures of Sampling Adequacy(MSA)

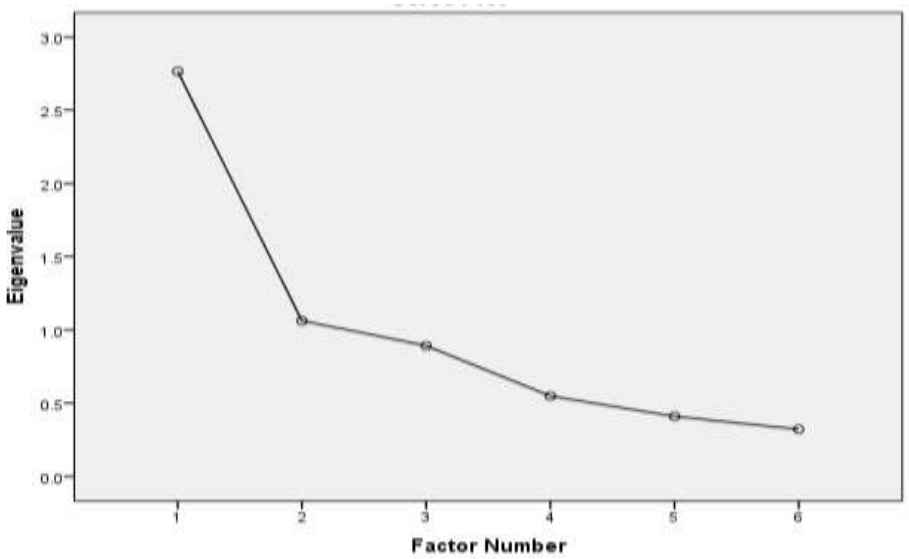
Kemudian untuk meminimalkan jumlah variabel yang diamati, sehingga sejumlah kecil komponen utama terbentuk dari sebagian besar varians variabel yang diamati menggunakan *Principal Component Analysis* (PCA). Jumlah faktor dapat ditentukan dengan memilih faktor yang memiliki nilai eigen lebih besar dari 1. Berikut nilai eigenvalues dari PCA.

Tabel 2. Total Variance Explained

Component	Intial Eigenvalues		
	Total	% of Variance	Cumulative %
1	2.765	46.081	46.081
2	1.063	17.708	63.790
3	.892	14.871	78.661
4	.549	9.147	87.809
5	.410	6.833	94.641
6	.322	5.359	100.000

Extraction Method: Principal Component Analysis.

Jumlah faktor dapat ditentukan dengan memilih faktor yang memiliki nilai Eigen lebih besar dari 1. Hasil analisis faktor menunjukkan bahwa terdapat 2 faktor yang memiliki nilai eigenvalue lebih dari 1. Hal ini juga dapat dilihat pada scree plot untuk menentukan banyak dimensi yang diukur oleh instrumen.



Gambar 8. Scree plot dari analisis komponen utama

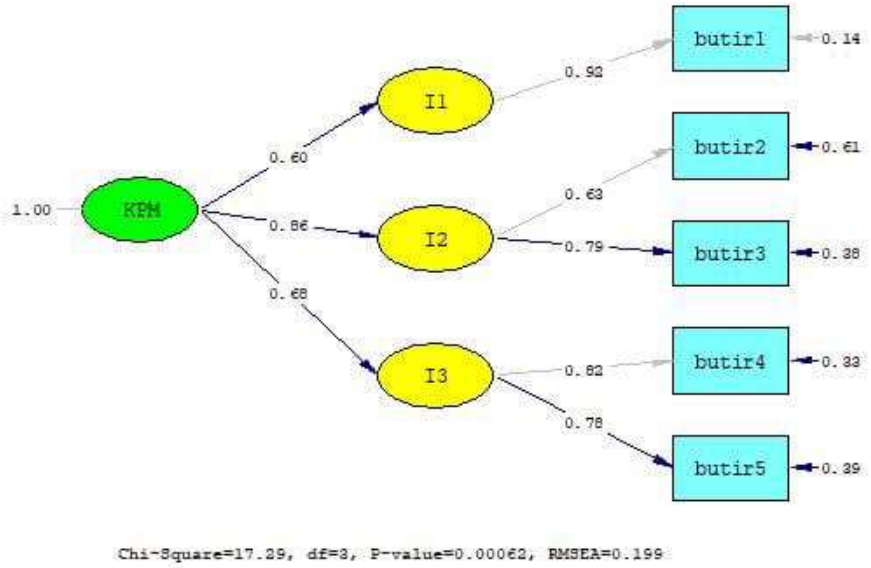
Kemudian dilakukan rotasi untuk mempermudah dalam interpretasi. Hasil dari rotasi faktor tersebut dapat dilihat terdapat 2 faktor. Faktor pertama diukur oleh butir 1, butir 2, butir 3 dan butir 6. Angka ini menunjukkan angka loading factor. Loading factor yang baik di atas 0,5 atau semakin tinggi semakin baik, artinya butir soal tersebut semakin berperan terhadap faktor yang mau diukur, sedangkan faktor kedua diukur melalui butir 4 dan butir 5. Berikut hasil rotated component matrix.

Tabel 3. Rotated component matrix

	Component	
	1	2
butir1	.768	
butir2	.631	
butir3	.851	
butir4		.797
butir5		.832
butir6	.617	

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

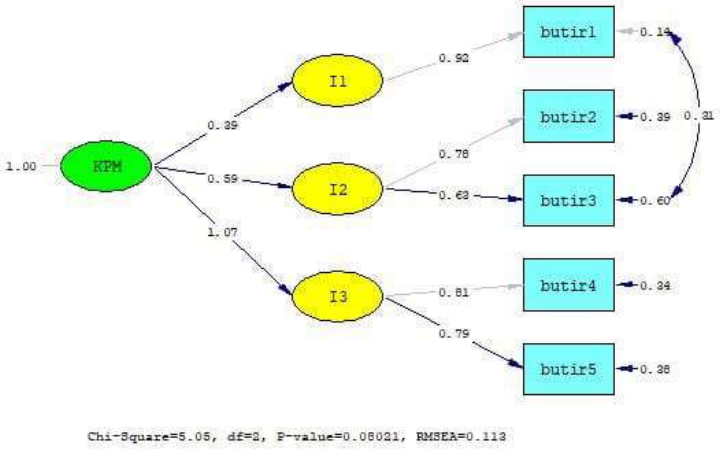
Second order CFA (*Confirmatory Factor Analysis*) merupakan model pengukuran yang terdiri dari dua tingkat. Tingkat kedua adalah sebuah CFA yang menunjukkan hubungan antara peubah-peubah laten pada tingkat pertama sebagai tahapan-tahapan dari sebuah peubah laten pada tingkat kedua. Langkah pertama yang dilakukan saat analisis CFA yaitu uji validitas. Validitas konstruk kemampuan pemecahan masalah matematis ditentukan oleh analisis faktor konfirmatori. Suatu peubah dikatakan mempunyai validitas yang baik terhadap konstruk atau peubah latennya, jika muatan faktor standar (standardized loading factors) sebesar ≥ 0,05. Output t-value disajikan pada gambar berikut.



Gambar 9. Diagram Path

Diagram path menunjukkan semua jalur valid, karena loading faktor menunjukkan di atas 0,50. Tetapi nilai p value = 0.00062, sehingga untuk uji kecocokan model tidak signifikan. Untuk itu perlu diperhatikan indeks modifikasi untuk menemukan saran perbaikan model.

Kemudian langkah yang selanjutnya dalam analisis CFA yaitu uji fit model. Ketentuan untuk uji kecocokan model nilai RMSEA ≤ 0,05. menandakan *close fit*, sedangkan jika nilai tersebut berada pada rentang 0,05 < RMSEA ≤ 0,08 model masih dapat diterima sebagai model yang fit. Artinya diharuskan lebih kecil dari 0, 05 dan paling minimal harus berada dibawah 0, 08. Karena pada output nilai RMSEA = 0.199, maka harus dilakukan modifikasi. Modifikasi dilakukan dengan melihat output diagram di paling terakhir, ternyata yang paling besar pertumbuhannya adalah butir 3 ke butir 1.



Gambar 10. Hasil Modifikasi Diagram Path

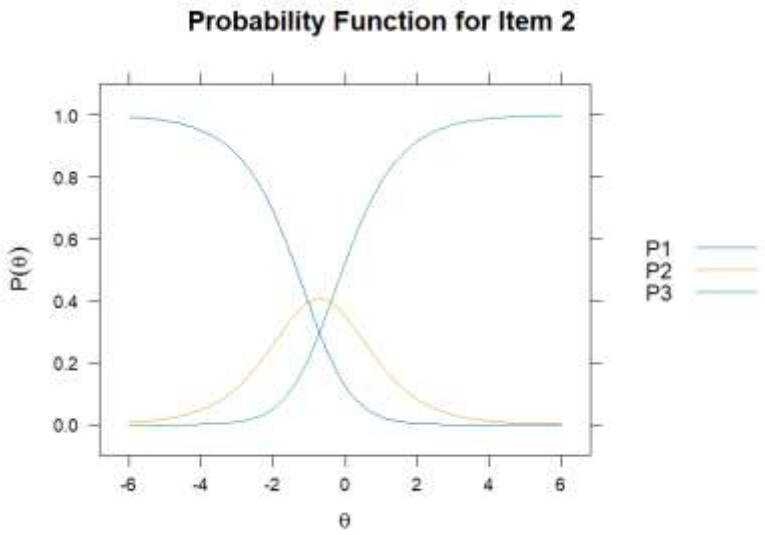
Setelah dilakukan modifikasi, nilai RMSEA mengalami penurunan namun tidak signifikan, yaitu menunjukkan RMSEA = 0, 113. Artinya nilai RMSEA masih di atas 0, 08, sedangkan untuk model yang dapat dikatakan fit seharusnya memiliki nilai RMSEA $\leq$ 0,08 (Tazkiyah et al., 2020). Output di atas menunjukkan nilai p-value mengalami peningkatan menjadi 0, 08021. Diagram di atas sudah fit dan tidak memerlukan modifikasi kembali. Diagram phat di atas adalah model fit yang didapat dari hasil uji dan dapat dilanjutkan untuk analisis PCM.

Dalam analisis PCM terdapat threshold yang merupakan titik temu dari setiap kategori untuk menunjukkan minimum poin tertentu (Dewanti et al., 2020). Kategori dalam pedoman penskoran pada penelitian ini berbeda-beda. Terdapat 4 butir soal yang memiliki kategori mulai dari 0 sampai 2. Oleh sebab itu, akan ada maksimal 2 threshold pada keempat butir soal tersebut di antaranya butir 1, butir 2, dan butir3. Kemudian, pada butir soal lainnya memiliki kategori mulai dari 0 sampai 4, sehingga akan ada maksimal 4 threshold pada butir 4 dan butir 5. Berikut threshold dari setiap butir soal disajikan pada tabel berikut.

Tabel 4. Output Analisis PCM

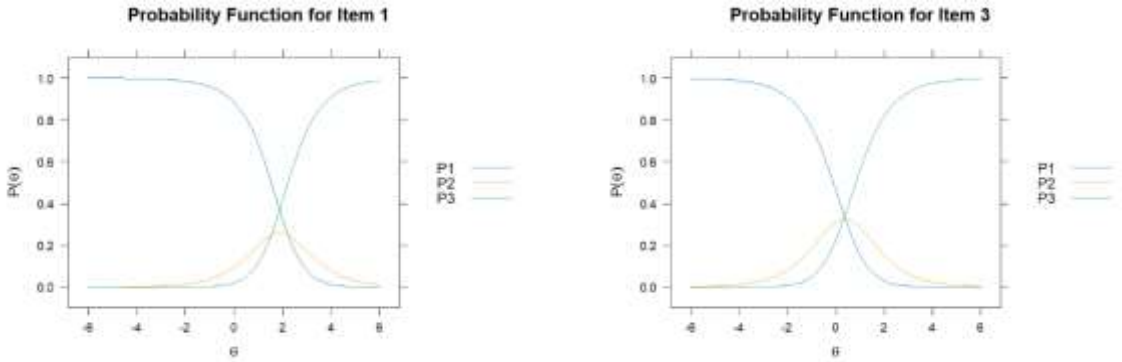
a	butir1	butir2	butir3	butir4	Location
butir1	2.2312480	1.52815483	NA	NA	1.8797014
butir2	-1.0219676	-0.39758279	NA	NA	-0.7097752
butir3	0.4095171	0.32465147	NA	NA	0.3670843
butir4	0.6498720	-0.03998354	1.569873	-0.4518987	0.4319656
butir5	1.2257592	-0.90233646	2.740074	0.1812801	0.8111941

Dari 5 butir soal, terdapat 2 butir soal yang memiliki 4 threshold dan 3 butir soal yang memiliki 2 threshold. Secara umum, threshold 1 < threshold 2 < threshold 3 < threshold 4, karena dalam menjawab menuliskan prosedur penyelesaian dengan lengkap dan tepat seharusnya memiliki kecenderungan jawaban yang lebih benar daripada menuliskan prosedur penyelesaian dengan lengkap namun kurang tepat. Akan tetapi, PCM tidak mengharuskan langkah-langkah dalam menyelesaikan setiap butir soal harus berurutan dan juga tidak mengharuskan setiap butir soal memiliki tingkat kesulitan yang sama (De Ayala, 1993). Hal tersebut mengakibatkan threshold penilaian PCM dari satu kategori ke kategori selanjutnya tidak selalu lebih besar. Dalam penelitian ini, butir soal yang memiliki threshold 1 < threshold 2 hanya pada butir soal 2 saja seperti yang ditunjukkan pada gambar berikut.

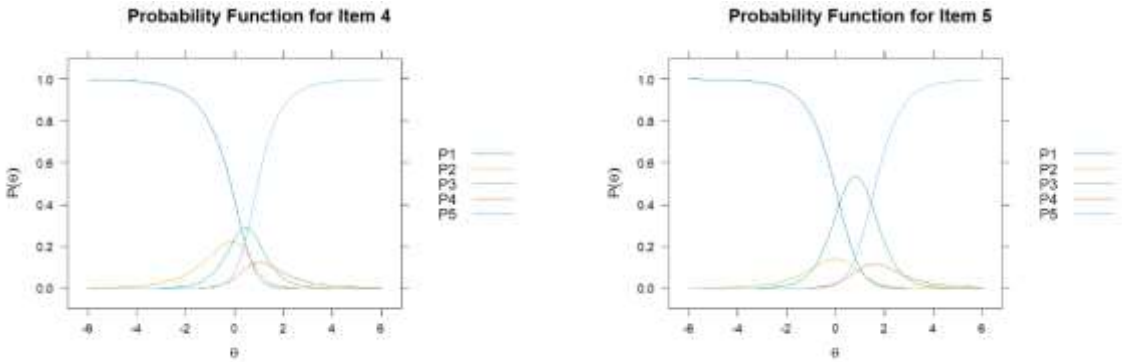


Gambar 11. Butir Soal dengan Threshold 1 < Threshold 2

Gambar 12 menunjukkan bahwa item 1 dan 3 memiliki threshold 2 < threshold 1. Kemudian, untuk kedua item lainnya yang memiliki threshold 4 nilai PCM dari satu kategori ke kategori berikutnya juga tidak lebih besar. Hal tersebut ditunjukkan pada Gambar 13 dan Gambar 14. Gambar 13 menunjukkan bahwa item 4 memiliki threshold 4 < threshold 2 < threshold 1 < threshold 3, sedangkan pada Gambar 14 menunjukkan bahwa item 5 memiliki threshold 2 < threshold 4 < threshold 1 < threshold 3.



Gambar 12. Butir Soal dengan Threshold 2 < Threshold 1



Gambar 13. Butir Soal dengan Threshold 4 < Threshold 2 < Threshold 1 < Threshold 3

Gambar 14. Butir Soal dengan Threshold 2 < Threshold 4 < Threshold 1 < Threshold 3.

Hasil analisis data menunjukkan bahwa instrumen yang dikembangkan telah menjalani serangkaian proses pengujian untuk memastikan validitas dan reliabilitasnya. Dengan menggunakan teori respons butir dan metode analisis faktor, struktur data dapat diidentifikasi dan dipahami lebih dalam. Analisis data dimulai dengan penggunaan teori respons butir untuk memahami karakteristik instrumen yang dikembangkan. Dimulai dengan kecukupan sampel sebagai syarat untuk uji asumsi. Sampel sebanyak 121 peserta didik dikatakan cukup jika nilai *Kaiser-Meyer-Olkin Measure of Sampling Adequacy* (KMO) > 0,5, sedangkan uji Bartlett harus < 0,05. Dalam penelitian ini keduanya memenuhi karena nilai KMO-MSA sebesar 0.705 menunjukkan bahwa data yang digunakan dalam penelitian sudah mencukupi untuk dilakukan analisis faktor. Kemudian, *Bartlett's Test of Sphericity* dengan nilai signifikansi 0.000 mengkonfirmasi bahwa matriks korelasi tidak identitas. Hal ini penting dilakukan, karena dalam teori respon butir sampel harus memenuhi ukuran untuk kelayakan data.

Kemudian, setelah sampel dinyatakan cukup dan layak, maka analisis faktor dilakukan menggunakan *Principal Component Analysis* (PCA). Hasilnya menunjukkan dua faktor utama dengan eigen value lebih dari 1, yaitu sebesar 2,765 dan 1,063 yang menjelaskan sebagian besar varians data. Artinya, jika peserta didik mampu menjawab soal yang sulit, maka soal yang mudah tentu akan dijawab dengan benar (Mardapi, 2012). Proses rotasi faktor dengan metode *Varimax* memperjelas hubungan antara variabel dan memudahkan interpretasi. Selanjutnya, uji validitas dan uji fit model dilakukan untuk menguji kecocokan model. Setelah modifikasi, model dinyatakan fit dengan nilai p-value yang memadai. Analisis PCM menentukan threshold untuk setiap kategori, menunjukkan tingkat kesulitan dan konsistensi penilaian. Keseluruhan analisis ini menghasilkan pemahaman yang mendalam tentang struktur, validitas, dan reliabilitas instrumen yang dikembangkan, serta memberikan dasar yang kuat untuk interpretasi dan penggunaan hasil penelitian.

KESIMPULAN

Penelitian ini berhasil mengembangkan instrumen tes kemampuan pemecahan masalah matematika menggunakan Teori Respon Butir (IRT). Instrumen ini terbukti lebih akurat dalam mengukur kemampuan peserta didik dibandingkan dengan pendekatan berbasis Teori Tes Klasik (CTT), karena mempertimbangkan karakteristik butir soal seperti tingkat kesulitan dan daya beda. Instrumen tes kemampuan pemecahan masalah yang telah dikembangkan termasuk dalam kategori valid dengan nilai sebesar 0,97 dan reliabel dengan nilai sebesar 0,736. Berdasarkan hasil analisis factor eksploratori diketahui bahwa tes kemampuan pemecahan masalah mengukur 2 faktor. Dari 6 butir soal terdapat 5 butir soal yang layak untuk digunakan. Berdasarkan hasil uji kecocokan model setelah dimodifikasi menunjukkan nilai p-value sebesar 0, 08021., sehingga model dapat diterima sebagai model yang fit. Berdasarkan hasil PCM menunjukkan item yang memiliki threshold 1 < threshold 2 hanya pad item 1 dan 3, sedangkan kedua item lainnya yang memiliki threshold 4. Hasil analisis menunjukkan bahwa instrumen ini mampu mengevaluasi kemampuan pemecahan masalah peserta didik secara lebih objektif dan adaptif terhadap berbagai tingkat kemampuan. Dengan demikian, IRT memberikan solusi yang lebih tepat dalam menangkap variasi strategi pemecahan masalah yang digunakan oleh peserta didik. Selain itu, penggunaan IRT dalam pengembangan instrumen ini memberikan keandalan dan validitas yang lebih tinggi dalam mengukur kemampuan peserta didik secara keseluruhan.

REKOMENDASI

Penelitian lanjutan direkomendasikan untuk memperluas cakupan soal yang dikembangkan berdasarkan IRT, khususnya dalam domain pemecahan masalah yang lebih kompleks dan di berbagai tingkatan pendidikan. Hal ini akan memperkaya validitas instrumen dan memastikan kemampuannya untuk mengevaluasi berbagai aspek pemecahan masalah. IRT sebaiknya digunakan lebih luas dalam pengembangan instrumen evaluasi di bidang lain, tidak hanya matematika. Penerapan IRT dapat meningkatkan kualitas pengukuran pada berbagai domain keterampilan yang membutuhkan evaluasi kemampuan kompleks.

UCAPAN TERIMAKASIH

Peneliti mengucapkan terima kasih kepada semua pihak salah satu madrasah tsanawiyah di Kabupaten Bantul Yogyakarta terutama Kepala Madrasah serta pendidik matematika yang telah berkenan mengizinkan peneliti untuk melakukan proses penelitian atau uji coba sehingga penyusunan artikel ini dapat diselesaikan dengan baik. Terakhir, peneliti mengucapkan terima kasih pada semua pihak yang telah membantu dalam menyusun artikel ini.

DAFTAR PUSTAKA

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Anderson, T., & Shattuck, J. (2018). Design-based research: A decade of progress in education research? *Educational Researcher*, 41(1), 16-25.

Angriani, A. D., Nursalam, N., Fuadah, N., & Baharuddin, B. (2018). Pengembangan instrumen tes untuk mengukur kemampuan pemecahan masalah matematika siswa. *AULADUNA: Jurnal Pendidikan Dasar Islam*, *5*(2), 212.

Aziza, M. (2019). Kemampuan pemecahan masalah siswa dalam menyelesaikan soal tertutup dan terbuka pada pokok bahasan lingkaran. *Pythagoras: Jurnal Pendidikan Matematika*, *14*(2), 126–138.

Baker, F. B., & Kim, S. H. (2022). *The basics of item response theory using R*. Springer.

Boesen, J., Lithner, J., & Palm, T. (2021). The relation between types of assessment tasks and the mathematical reasoning students use. *Educational Studies in Mathematics*, 81(3), 371-389.

Bransford, J. D., & Stein, B. S. (1993). *The Ideal Problem Solver: A Guide for Improving Thinking, Learning, and Creativity*.

De Ayala, R. J. (1993). An introduction to polytomous item response theory models. *Measurement and Evaluation in Counseling and Development*, *25*(4), 172–189.

De Ayala, R. J. (2018). *The theory and practice of item response theory*. Guilford Press.

Dewanti, S. S., Ayriza, Y., & Setiawati, F. A. (2020). The application of item response theory for development of a students' attitude scale toward mathematics. *New Educational Review*, *60*, 108–123. doi: 10.15804/tner.2020.60.2.09

Dewey, J. (1933). How we think. In *D.C. Heath and Co. Publishers*.

Embretson, S. E., & Reise, S. P. (2020). *Item response theory for psychologists*. Psychology Press.

Fitrianty, F., Yunita, A., & Juwita, R. (2022). Pengembangan instrumen tes kemampuan pemecahan masalah matematika di SMP Negeri 12 Padang. *Lattice Journal : Journal of Mathematics Education and Applied*, *2*(1), 91–102.

Freeman, S., & Higgins, J. (2020). Assessment in mathematical problem-solving: Validity and reliability challenges. *Journal of Educational Measurement*, 57(2), 168-189.

Greiff, S., & Funke, J. (2021). Measuring complex problem-solving skills: A review and recommendation for educational applications. *Journal of Educational Psychology*, 112(3), 442-456.

Hanim, A., & Ragimun. (2010). Analisis faktor-faktor yang mempengaruhi minat investasi di daerah: Study kasus di Kabupaten Jember Jawa Timur. *Kajian Ekonomi Dan Keuangan*, *14*(3), 5.

He, W., & Larkin, K. (2021). Technology-enhanced assessment for problem-solving in mathematics: Benefits and challenges. *Computers & Education*, 149, 103818.

Ikawati, H. D., Jayadi, A., & Hermansyah. (2022). Analisis kualitas tes dan butir soal sejarah di SMAN 1 Praya Timur. *Jurnal Pendidikan Dan Konseling*, *4*(4), 6263.

Jonassen, D. H. (2021). *Learning to solve problems: A handbook for designing problem-solving learning environments*. Routledge.

Karadag, E. (2022). Influence of problem difficulty on the reliability of problem-solving tests: A psychometric analysis. *Educational Assessment*, 27(1), 1-21.

Kim, S. H., & Bolt, D. M. (2021). Challenges in applying item response theory to educational assessment. *Educational Measurement: Issues and Practice*, 40(1), 12-21.

Lester, F. K. (2020). On the theoretical, conceptual, and philosophical foundations for research on mathematics teaching and learning. *Journal for Research in Mathematics Education*, 51(1), 1-17.

Lukman, H. S., Setiani, A., & Agustiani, N. (2023). Pengembangan instrumen tes kemampuan pemecahan masalah matematis berdasarkan teori krulik dan rudnick: Analisis validitas konten. *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *7*(1), 326–339.

Magdalena, I., Arwindi, S., & Hasan, S. N. (2023). Menyusun alat penilaian hasil belajar. *Sindoro: Cendikia Pendidikan*, *2*(4), 8.

Mardapi, D. (2012). *Pengukuran, penilaian, dan evaluasi pendidikan*. Yogyakarta: Nuha Medika.

NCTM. (2000). Principles and standars for school mathematics. In *Mathematics, The National Council of Teachers of*. United States of America: The National Council of Teachers of Mathematics, Inc.

Oktaviana, D., & Prihatin, I. (2018). Analisis hasil belajar siswa pada materi perbandingan berdasarkan ranah kognitif revisi taksonomi bloom. *Buana Matematika : Jurnal Ilmiah Matematika Dan Pendidikan Matematika*, *8*(2), 82.

Pape, S. J., & Smith, C. E. (2019). The complexity of measuring mathematical problem-solving skills. *Mathematical Thinking and Learning*, 21(2), 85-105.

Park, M., & Leung, F. K. S. (2019). Framework for assessing problem-solving skills in mathematics education. *International Journal of Educational Research*, 97, 110-123.

Polya, G. (1957). How to solve it: a new aspect of mathematical method second edition. In *Princeton University Press: United States of America*.

Rahmani, W., & Widyasari, N. (2018). Meningkatkan kemampuan pemecahan masalah matematis siswa melalui media tangram. *FIBONACCI: Jurnal Pendidikan Matematika Dan Matematika*, *4*(1), 17–24. doi: 10.24853/fbc.4.1.17-23

Retnawati, H. (2016). *Validitas dan reliabilitas dan karakteristik butir*. Parama Publising.

Safaruddin, Anisa, & AF, M. S. (2012). Partial Credit Model (PCM) dalam penskoran politomi pada teori respon butir. *Jurnal Matematika, Statistika, & Komputasi*, *9*(1), 41.

Schoenfeld, A. H. (2022). *Mathematical problem solving*. Academic Press.

Situmorang, H. F., & Bunawan, W. (2022). Pengembangan instrumen tes untuk mengukur kemampuan pemecahan masalah siswa pada materi gerak parabola. *Jurnal Inovasi Pembelajaran Fisika*, *10*(3), 45.

Sugiyono. (2010). *Metode penelitian pendidikan*. Alfabeta.

Tazkiyah, F., Ihsan, H., & Musthofa, M. A. (2020). Prophetic leadership scale's validation and the tendency of normative response. *Jurnal Psikologi Islam Dan Budaya*, *3*(2), 152.

Wang, W., & Cheng, Y. (2019). Classical test theory vs. item response theory: An empirical comparison in mathematics problem-solving assessment. *Journal of Educational Measurement*, 56(4), 591-608.

Wu, M. L., & Adams, R. J. (2022). Advances in objective testing: A practical approach to evaluating problem-solving skills. *Educational Measurement: Issues and Practice*, 41(2), 54-66.