

**ANALISIS VALIDITAS UKUR DAN KONSISTENSI INTERNAL TES
PEMAHAMAN MEMBACA NARASI BERBASIS TEKS DIGITAL
MENGGUNAKAN MODEL PARTIAL CREDIT RASCH**

(Luh Putu Adhi Laksmini Susila¹), (Ida Bagus Putrayasa²), (Anjelina Septiana
Harianja³), (Ni Putu Eva Fransiska Dewi⁴), (Yunita Ratna Alfin⁵), (I Nengah
Martha⁶), (Kadek Wirahyuni⁷)

¹²³⁴⁵⁶⁷Universitas Pendidikan Ganesha

Alamat e-mail : ([1adhi.laksmini@student.undiksha.ac.id](mailto:adhi.laksmini@student.undiksha.ac.id)),

([2ib.putrayasa@student.undiksha.ac.id](mailto:ib.putrayasa@student.undiksha.ac.id)), ([3anjelina.septiana@student.undiksha.ac.id](mailto:anjelina.septiana@student.undiksha.ac.id)),

([4eva.fransiska@student.undiksha.ac.id](mailto:eva.fransiska@student.undiksha.ac.id)), ([5yunita.ratna@student.undiksha.ac.id](mailto:yunita.ratna@student.undiksha.ac.id)),

([6nengah.martha@student.undiksha.ac.id](mailto:nengah.martha@student.undiksha.ac.id)), ([7kadek.wirahyuni@student.undiksha.ac.id](mailto:kadek.wirahyuni@student.undiksha.ac.id))

ABSTRACT

Paradigma pendidikan global telah beralih secara masif menuju literasi digital, menuntut evaluasi yang mampu menangkap navigasi kognitif kompleks dalam membaca narasi digital. Penelitian ini bertujuan untuk mengkalibrasi kualitas psikometrik instrumen "Tes Pemahaman Membaca Narasi Digital" (TPMND) menggunakan pendekatan Item Response Theory (IRT), khususnya Model Partial Credit (PCM) Rasch, yang dianggap lebih unggul dari Teori Tes Klasik (CTT) yang terbatas dalam mengukur skor bertingkat. Penelitian dengan desain deskriptif kuantitatif ini melibatkan 450 siswa yang dipilih melalui multistage random sampling. Instrumen, yang terdiri dari 25 butir soal esai pendek dengan rubrik analitik skor 0-3, dianalisis menggunakan perangkat lunak Winsteps untuk menguji unidimensionalitas, kesesuaian butir, reliabilitas, dan keberfungsi kategori skor. Hasil analisis menunjukkan instrumen memiliki validitas konstruk sangat kuat dengan raw variance explained by measures sebesar 46,2% dan unexplained variance pada kontras pertama hanya 5,8%, membuktikan sifat unidimensionalnya. Dari 25 butir, 23 butir sesuai model, dengan dua butir (Item 14 dan 22) perlu revisi karena terminologi yang membingungkan. Indeks reliabilitas item (0,96) dan person (0,89) tergolong sangat baik, menunjukkan konsistensi internal tinggi dan kemampuan membedakan siswa ke dalam beberapa strata kemampuan. Analisis PCM membuktikan struktur penskoran 0-3 berfungsi optimal dengan ambang batas yang terurut. Temuan ini menegaskan bahwa model Rasch tidak hanya memberikan presisi pengukuran yang lebih tinggi, tetapi juga menawarkan fungsi diagnostik yang krusial, memungkinkan pendidik mengidentifikasi hambatan kognitif spesifik siswa dan memberikan umpan balik yang lebih targetted dalam evaluasi literasi digital.

Keywords: Validitas ukur, evaluasi, konsistensi

ABSTRAK

The global educational paradigm has shifted dramatically toward digital literacy, demanding assessments capable of capturing the complex cognitive navigation of reading digital narratives. This study aims to calibrate the psychometric quality of the "Digital Narrative Reading Comprehension Test" (TPMND) instrument using the Item Response Theory (IRT) approach, specifically the Rasch Partial Credit Model (PCM), which is considered superior to the Classical Test Theory (CTT) which is limited in measuring graded scores. This quantitative descriptive study involved 450 students selected through multistage random sampling. The instrument, consisting of 25 short essay items with an analytical rubric scoring 0-3, was analyzed using Winsteps software to test unidimensionality, item suitability, reliability, and the functionality of the score categories. The analysis results showed that the instrument has very strong construct validity with raw variance explained by measures of 46.2% and unexplained variance in the first contrast of only 5.8%, proving its unidimensional nature. Of the 25 items, 23 fit the model, with two items (Items 14 and 22) requiring revision due to confusing terminology. The item (0.96) and person (0.89) reliability indices were excellent, indicating high internal consistency and the ability to differentiate students across multiple ability strata. PCM analysis demonstrated that the 0-3 scoring structure functioned optimally with ordered thresholds. These findings confirm that the Rasch model not only provides higher measurement precision but also offers crucial diagnostic functions, enabling educators to identify students' specific cognitive barriers and provide more targeted feedback in digital literacy assessments.

Kata Kunci: Validity of measurement, evaluation, consistency

A. Pendahuluan

Lanskap pendidikan global saat ini tengah mengalami sebuah transformasi fundamental, sebuah pergeseran paradigma yang melampaui sekadar perubahan dalam kurikulum atau metode pengajaran. Pergeseran ini bergerak dari dominasi literasi konvensional, yang selama ini menjadi landasan pendidikan, menuju era literasi digital yang masif dan tidak terelakkan. Fenomena ini bukanlah

sebuah tren semata, melainkan sebuah urgensi global yang oleh Leu et al. (2017) secara cermat disebut sebagai kemunculan "literasi baru" (new literacies). Konsep ini menggarisbawahi bahwa kemampuan membaca, menulis, dan berkomunikasi di abad ke-21 tidak lagi terikat pada media cetak, melainkan secara inheren terkait dengan kemampuan untuk menavigasi, mengevaluasi, dan menciptakan

informasi dalam lingkungan digital yang dinamis dan sering kali tidak terstruktur. Dalam konteks ini, kemampuan membaca teks narasi, yang secara tradisional dianggap sebagai aktivitas linear dan prediktif, mengalami redefinisi yang drastis. Membaca tidak lagi sekadar memindai mata dari baris ke baris, tetapi telah bertransformasi menjadi proses navigasi kognitif yang sangat kompleks di atas layar digital, yang dipenuhi dengan hyperlink, media tertanam, dan elemen interaktif.

Mayer (2021) dalam teori Multimedia Learning-nya memberikan lensa teoretis yang tajam untuk memahami perubahan ini. Ia menekankan bahwa teks digital bukanlah sekadar pemindahan medium dari kertas ke layar. Sebaliknya, ia merupakan sebuah lingkungan kognitif yang sama sekali baru, yang secara aktif melibatkan pemrosesan kanal ganda—visual dan verbal—secara simultan. Pembaca tidak hanya memproses kata-kata, tetapi juga gambar, video, dan tata letak antarmuka, yang semuanya bersaing memperoleh perhatian kognitif. Hal ini menciptakan tantangan unik yang tidak dihadapi saat membaca teks cetak. Oleh

karena itu, dalam konteks mata kuliah Evaluasi Pembelajaran, fenomena ini menuntut adanya evolusi dalam instrumen penilaian. Para pendidik dan pengembang tes tidak lagi bisa mengandalkan alat ukur lama yang dirancang untuk literasi linear. Kini, dibutuhkan instrumen yang tidak hanya mampu mengukur aspek pemahaman konten secara mendalam, tetapi juga harus valid secara psikometrik dalam memotret kemampuan siswa untuk berinteraksi, menafsirkan, dan mensintesis informasi dari teks digital yang kompleks.

Tantangan utama dalam pengukuran pendidikan saat ini, sebagaimana diidentifikasi oleh Reckase (2015), bersumber dari keterbatasan inheren dalam Teori Tes Klasik (Classical Test Theory/CTT). CTT, yang telah menjadi dasar pengukuran selama beberapa dekade, sering kali gagal memberikan parameter yang independen terhadap sampel. Artinya, tingkat kesulitan sebuah soal dapat berubah secara drastis tergantung pada kemampuan rata-rata siswa yang mengerjakannya, sehingga menghilangkan sifat objektif yang seharusnya dimiliki oleh alat ukur. Lebih buruk lagi, CTT kesulitan

dalam mengelola data dari soal-soal dengan jawaban terbuka atau esai pendek yang memerlukan penskoran bertingkat (polytomous). Tanpa adanya kalibrasi yang tepat dan model pengukuran yang canggih, evaluasi terhadap kemampuan membaca siswa di era digital berisiko tinggi menghasilkan kesimpulan yang bias, tidak akurat, dan tidak representatif terhadap kemampuan sebenarnya.

Permasalahan ini semakin kompleks ketika disinggung dengan karakteristik intrinsik dari teks narasi digital itu sendiri. Menurut Smith (2019), teks digital sering kali memadukan elemen multimodal—seperti teks, gambar, audio, dan video—yang dapat secara signifikan memengaruhi beban kognitif (cognitive load) pembaca. Keterlibatan kanal ganda ini, meskipun berpotensi memperkaya pemahaman, juga dapat menyebabkan pembelahan perhatian dan menghambat pembentukan model mental yang koheren. Kintsch (2019) melalui model Construction-Integration (CI) mengingatkan bahwa keberhasilan pemahaman teks sangat bergantung pada kemampuan pembaca untuk membangun model situasi yang utuh. Namun, pada media

digital, proses ini sering kali terfragmentasi oleh berbagai gangguan dan jalur navigasi non-linier, sehingga siswa mungkin mengumpulkan potongan-potongan informasi tanpa pernah berhasil menyatukannya menjadi pemahaman yang utuh.

Di sinilah penggunaan model Partial Credit (PCM) dalam kerangka kerja Teori Respon Butir (Item Response Theory/IRT) muncul sebagai solusi metodologis yang sangat relevan dan mendesak. Sebagaimana dikemukakan oleh Bond dan Fox (2020), Model Rasch, yang menjadi dasar PCM, memberikan keunggulan krusial berupa objektivitas pengukuran. Hal ini dicapai melalui transformasi data kategorikal menjadi skala interval yang logaritmik (logit), yang memisahkan parameter kemampuan siswa (person) dengan parameter kesulitan soal (item). Penggunaan PCM secara spesifik, menurut Masters (2016), menjadi kunci untuk mengevaluasi respons yang kompleks. Model ini memungkinkan setiap butir soal memiliki struktur ambang batas (threshold) yang unik, sehingga memberikan apresiasi terhadap "pemahaman parsial" siswa

yang tidak dapat ditangkap oleh model dikotomus (benar-salah) sederhana. Seorang siswa mungkin tidak mampu memberikan jawaban sempurna (skor 3), tetapi ia telah menunjukkan pemahaman pada level tertentu (skor 1 atau 2). PCM mampu membedakan nilai kualitatif dari setiap lompatan kognitif ini. Dengan mengintegrasikan pendapat para ahli ini, penelitian ini berupaya mengisi celah metodologis yang krusial dalam evaluasi literasi digital, memastikan bahwa instrumen yang digunakan memiliki presisi tinggi dalam membedakan berbagai tingkatan kompetensi kognitif siswa secara adil dan akurat, yang pada akhirnya akan menjamin keadilan dalam penilaian dan meningkatkan kualitas mutu pendidikan di era digital.

B. Metode Penelitian

Penelitian ini mengadopsi desain deskriptif kuantitatif dengan pendekatan survei untuk secara sistematis menggali dan mendeskripsikan kualitas psikometrik dari instrumen "Tes Pemahaman Membaca Narasi Digital" (TPMND). Populasi target adalah siswa sekolah menengah pertama di wilayah urban yang telah akrab dengan pembelajaran berbasis gawai. Dipilih

sebanyak 450 siswa sebagai sampel melalui teknik multistage random sampling. Metode ini bertujuan untuk menjamin representasi yang akurat dari beragam tingkat kemampuan siswa, sehingga hasil analisis dapat dianggap mencerminkan populasi yang lebih luas.

Instrumen utama yang dikalibrasi adalah TPMND, yang terdiri dari 25 butir soal esai pendek respons polikotomus. Setiap butir dirancang untuk mengukur aspek berpikir tingkat tinggi dan dinilai menggunakan skala 0-3 berdasarkan rubrik analitik yang mencakup tiga kriteria utama: akurasi informasi, kedalaman analisis, dan penggunaan bukti tekstual. Sebelum pelaksanaan, instrumen ini telah divalidasi isinya melalui telaah oleh pakar untuk memastikan relevansi dan kecukupan konten.

Proses analisis data dilakukan secara komprehensif menggunakan perangkat lunak Winsteps. Tahap awal meliputi pembersihan data dari nilai yang hilang. Selanjutnya, penelitian menguji asumsi fundamental model Rasch, yaitu unidimensionalitas dan independensi lokal. Unidimensionalitas, yang menunjukkan bahwa tes hanya mengukur satu konstruk utama

(pemahaman membaca), diverifikasi melalui Principal Component Analysis (PCA) dari residu dengan kriteria varians yang dijelaskan oleh ukuran minimal 20%. Setelah asumsi terpenuhi, dilakukan analisis kesesuaian butir (Item Fit) dengan memeriksa nilai Infit/Outfit Mean Square (MNSQ) dan Z-Standard (ZSTD) untuk mengidentifikasi butir yang menyimpang dari model. Langkah terakhir adalah analisis keberfungsian skala rating, yang krusial untuk model Partial Credit, guna memastikan bahwa kategori skor 0, 1, 2, dan 3 dipahami dan digunakan secara konsisten oleh responden.

C. Hasil Penelitian dan Pembahasan

1. Analisis Validitas Konstruk (Unidimensionalitas)

Validitas konstruk dalam kerangka kerja Rasch diuji melalui Principal Component Analysis (PCA) dari residu. Hasil analisis menunjukkan bahwa raw variance explained by measures mencapai angka 46.2%. Merujuk pada kriteria yang ditetapkan oleh Linacre (2021), nilai varians di atas 40% mengindikasikan bahwa instrumen tersebut memiliki struktur dimensi yang sangat kuat. Hal ini

membuktikan bahwa meskipun teks narasi digital melibatkan berbagai elemen seperti navigasi dan interaksi antarmuka, instrumen ini tetap fokus mengukur satu dimensi tunggal, yaitu kemampuan pemahaman membaca. Lebih lanjut, analisis terhadap unexplained variance pada kontras pertama menunjukkan angka 5.8%. Menurut Bond dan Fox (2020), jika varians pada kontras pertama di bawah 15%, maka instrumen tersebut bebas dari gangguan dimensi sekunder yang signifikan. Dalam konteks evaluasi pembelajaran digital, hal ini krusial untuk memastikan bahwa skor siswa tidak dipengaruhi secara dominan oleh keterampilan teknis (misalnya kemahiran menggunakan mouse atau kecepatan mengetik), melainkan murni dari kapasitas kognitif mereka dalam mengolah teks narasi.

2. Kualitas Butir Soal: Analisis Item Fit dan Difficulty Measure

Berdasarkan parameter Infit dan Outfit Mean Square (MNSQ), ditemukan bahwa dari 25 butir soal yang diujikan, 23 butir berada pada rentang nilai 0.6 hingga 1.4. Hal ini menunjukkan bahwa butir-butir tersebut sangat produktif untuk pengukuran. Namun, terdapat dua

butir soal (Item 14 dan Item 22) yang menunjukkan nilai Outfit MNSQ di atas 1.5. Setelah dilakukan penelusuran kualitatif, Item 14 yang berkaitan dengan "Interpretasi Metafora Digital" cenderung membingungkan responden dengan tingkat kemampuan menengah, karena penggunaan istilah yang terlalu teknis.

Tingkat kesulitan soal (difficulty measure) tersebar dari -2.45 logit hingga +3.10 logit. Butir soal nomor 12, yang menuntut siswa untuk mengevaluasi sudut pandang penulis dalam narasi non-linear, menjadi butir tersulit. Sebaliknya, butir nomor 2 yang hanya meminta identifikasi latar tempat menjadi yang termudah. Sebaran ini menunjukkan bahwa instrumen ini mampu membedakan spektrum kemampuan siswa dari yang paling rendah hingga yang paling mahir. Reckase (2015) menekankan bahwa instrumen yang baik harus memiliki sebaran kesulitan yang setara dengan sebaran kemampuan responden (person ability) guna menghindari kesalahan standar pengukuran yang besar pada ekstremitas skala.

3. Analisis Keberfungsian Kategori Skor (Partial Credit Model)

Penggunaan Model Partial Credit (PCM) memberikan keunggulan dalam memvalidasi rubrik penskoran polikotomus (0-1-2-3). Melalui analisis Category Sky dan Threshold Calibration, peneliti menemukan bahwa setiap kategori skor berfungsi secara optimal. Nilai Average Measure meningkat secara monoton seiring dengan peningkatan kategori skor: skor 0 (-2.10), skor 1 (-0.45), skor 2 (+1.12), dan skor 3 (+2.88). Keberfungsian kategori ini dibuktikan dengan kurva probabilitas kategori yang menunjukkan puncak-puncak yang jelas dan tidak saling tumpang tindih secara ekstrem. Menurut Masters (2016), jika threshold atau ambang batas kategori tidak berurutan (disordered thresholds), hal itu menandakan bahwa definisi dalam rubrik penilaian membingungkan atau kriteria antar level (misalnya antara skor 'Cukup' dan 'Baik') terlalu tipis sehingga sulit dibedakan oleh penilai. Dalam penelitian ini, konsistensi ambang batas membuktikan bahwa rubrik yang dikembangkan dalam mata

kuliah Evaluasi Pembelajaran ini telah memiliki objektivitas yang tinggi.

4. Konsistensi Internal dan Reliabilitas

Indeks reliabilitas item sebesar 0.96 dan reliabilitas person sebesar 0.89 memberikan bukti kuat mengenai konsistensi internal instrumen. Nilai reliabilitas item yang mendekati sempurna menunjukkan bahwa sampel yang digunakan sudah cukup besar untuk mengonfirmasi hierarki kesulitan soal. Dengan kata lain, jika instrumen ini diberikan kepada kelompok siswa lain di wilayah berbeda, urutan soal dari yang termudah hingga tersulit akan tetap konsisten. Reliabilitas person sebesar 0.89 juga menunjukkan bahwa tes ini mampu membedakan siswa ke dalam 3 hingga 4 kelompok kemampuan yang berbeda (strata kemampuan). Hal ini sejalan dengan pendapat Sumintono dan Widhiarso (2015) bahwa model Rasch memberikan estimasi reliabilitas yang lebih jujur dibandingkan Cronbach's Alpha pada teori klasik, karena Rasch memperhitungkan adanya data yang hilang (missing data) dan tidak hanya bergantung pada korelasi antar-item.

5. Diskusi: Implikasi terhadap Evaluasi Pembelajaran Digital

Temuan ini memberikan implikasi serius pada bagaimana pemahaman membaca narasi digital harus dievaluasi. Tingginya tingkat kesulitan pada soal-soal evaluatif (Level C5 dan C6) mengonfirmasi teori Schoor et al. (2021) bahwa lingkungan digital menuntut "kewaspadaan kognitif" yang lebih tinggi. Siswa sering kali terjebak pada pemahaman literal namun gagal melakukan sintesis ketika teks narasi disajikan dalam format hyperlink atau multimodal.

Integrasi model PCM dalam evaluasi pembelajaran memungkinkan dosen dan praktisi pendidikan untuk memberikan umpan balik yang lebih spesifik. Bukan hanya sekadar skor total, tetapi guru dapat mengetahui pada "langkah kognitif" mana siswa mulai mengalami kegagalan. Misalnya, jika seorang siswa secara konsisten berhenti di skor 2 pada soal-soal integrasi, guru dapat menyimpulkan bahwa siswa tersebut sudah mampu menghubungkan informasi namun belum mampu menarik konklusi yang didukung oleh bukti textual yang kuat. Pendekatan diagnostik ini

adalah keunggulan utama dari penggunaan analisis Rasch dalam meningkatkan mutu pembelajaran secara berkelanjutan.

E. Kesimpulan

Penelitian ini menyimpulkan bahwa instrumen Tes Pemahaman Membaca Narasi Digital memiliki kualitas psikometrik yang sangat kuat dan valid untuk digunakan dalam evaluasi pembelajaran. Melalui pendekatan Model Partial Credit (PCM) Rasch, instrumen ini terbukti memenuhi kriteria unidimensionalitas, yang berarti tes tersebut murni mengukur kemampuan kognitif membaca tanpa terdistraksi oleh variabel kemampuan teknis digital lainnya. Kualitas butir soal dan reliabilitas instrumen berada pada kategori sangat bagus, dengan sebaran tingkat kesulitan yang mampu membedakan berbagai spektrum kemampuan siswa, dari tingkat dasar hingga mahir. Selain itu, struktur penskoran bertingkat (0-3) yang diterapkan berfungsi secara optimal dan objektif, di mana setiap kategori skor mampu menggambarkan tingkatan pemahaman siswa secara konsisten dan tidak saling tumpang tindih.

Secara keseluruhan, penggunaan model Rasch memberikan presisi pengukuran yang lebih tinggi dan jujur dibandingkan teori tes klasik, serta mampu memberikan fungsi diagnostik yang membantu pendidik memahami hambatan kognitif spesifik yang dialami siswa dalam literasi digital

DAFTAR PUSTAKA

- Bond, T. G., & Fox, C. M. (2020). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Routledge.
- Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE—Life Sciences Education*.
- Chen, H. J. (2021). Digital reading performance and its relationship with cognitive strategies. *Journal of Educational Computing Research*.
- Delgado, P., et al. (2018). Don't throw away your printed books: A meta-analysis on the advantage of reading narrative texts on paper. *Educational Research Review*.
- Englehard, G., & Wind, S. A. (2018). *Invariant Measurement with R*. Routledge.
- Hahne, C. (2020). *Rasch Analysis in Educational Research*. Springer.
- Kintsch, W. (2019). Revisiting the Construction-Integration Model of Text Comprehension. *Educational Psychologist*.
- Leu, D. J., et al. (2017). The new literacies of online research and

- comprehension. *The Reading Teacher*.
- Lin, S. W. (2018). Validating a reading comprehension test using the Rasch model. *Language Testing*.
- Masters, G. N. (2016). A partial credit model for objective measurement. *Psychometrika*.
- Mayer, R. E. (2021). *Multimedia Learning* (3rd ed.). Cambridge University Press.
- OECD. (2023). PISA 2022 Results: Learning During Disruptions. OECD Publishing.
- Panfalov, A. (2022). Psychometric evaluation of reading tests in the digital age. *Assessment in Education*.
- Prieto, G., & Nieto, C. (2019). Analysis of polytomous items in educational assessment. *Journal of Applied Measurement*.
- Putnam, S. M. (2020). Narrative comprehension in digital vs. print media. *Reading Research Quarterly*.
- Reckase, M. D. (2015). *Multidimensional Item Response Theory*. Springer.
- Salmerón, L., et al. (2018). Comprehension of digital texts: A review. *Psicothema*.
- Schoor, C., et al. (2021). Multiple document comprehension in digital environments. *Frontiers in Psychology*.
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Model Rasch untuk Penelitian Ilmu-Ilmu Sosial*. Trim Komunikata.
- Toste, J. R., et al. (2020). Reading comprehension and digital texts: A psychometric study. *Journal of Learning Disabilities*.
- Van der Vlies, R. (2020). Digital Reading Habits of Students. *OECD Education Working Papers*.
- Wang, W. C. (2017). Rasch measurement and its applications in education. *International Journal of Quantitative Research*.
- Wolf, A. D., et al. (2019). The impact of digital formats on reading comprehension. *Journal of Literacy Research*.
- Winsteps. (2024). *Winsteps Rasch Measurement Software User Guide*. Winsteps.com.
- Zhu, M., et al. (2022). Measuring digital literacy with IRT models: A systematic review. *Educational Psychology Review*.