PANGAN, Vol. 34, No. 2, Agustus 2025
ISSN: 0852-0607 DOI: 10.33964/jp.v34i2.894

Validasi Psikometrik Instrumen Faktor Penentu Kualitas Beras Premium Berbasis Teori Respons Butir

Psychometric Validation of an Instrument for Determinant Factors of Premium Rice Quality Based on Item Response Theory

Eny Cahyaningsih, Nurul Qomariyah Ahmad, Wardani Rahayu dan Achmad Ridwan

Program Studi Penelitian dan Evaluasi Pendidikan, Pasca Sarjana Universitas Negeri Jakarta, Jakarta, Indonesia
Jl. Rawamangun Muka Raya, RT. 11/RW.14, Rawamangun, Jakarta Timur, 13220
E-mail: cahyaningsiheny@gmail.com

Diterima: 27 Agustus 2025 Revisi: 11 September 2025 Disetujui : 12 September 2025

ABSTRAK

Pemahaman terhadap preferensi konsumen memerlukan alat ukur yang akurat, andal, serta mampu merepresentasikan persepsi terhadap kualitas produk secara menyeluruh. Pada komoditas strategis seperti beras premium, atribut sensorik dan fisik, meliputi aroma, warna, tekstur, kebersihan, dan masa simpan, memiliki peranan penting dalam membentuk persepsi konsumen. Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi instrumen pengukuran persepsi konsumen terhadap kualitas beras premium dengan pendekatan Teori Respons Butir (*Item Response Theory*/IRT), khususnya melalui Model Respons Terklasifikasi (*Graded Response Model*/GRM). Fokus penelitian diarahkan pada penyusunan instrumen yang valid dan reliabel untuk mengukur atribut-atribut utama yang relevan dalam konteks preferensi konsumen. Pengujian dilakukan terhadap sejumlah asumsi dasar GRM, yaitu unidimensionalitas, independensi lokal, dan monotonisitas, yang hasilnya menunjukkan bahwa ketiga asumsi terpenuhi. Sebagian besar butir dalam instrumen memiliki parameter diskriminasi yang tinggi (a > 1,0) serta ambang kesulitan (*threshold*) yang logis dan tersebar dengan baik di sepanjang spektrum tingkat kemampuan. Evaluasi kurva karakteristik butir (*Item Characteristic Curve*/ICC) dan indeks penerimaan menunjukkan bahwa skala ini mampu mengukur preferensi konsumen dari tingkat kesulitan sangat rendah hingga sangat tinggi secara efektif. Hasil uji efektivitas jumlah kategori respons memperlihatkan bahwa empat pilihan jawaban memberikan prediksi terbaik, dengan nilai *Root Mean Square Error* (RMSE) pada kisaran 0,1867–0,2138, lebih rendah dibandingkan dengan versi tiga maupun lima kategori. Dengan demikian, GRM terbukti efektif dalam membangun skala pengukuran preferensi konsumen terhadap beras premium, sehingga dapat dimanfaatkan untuk penyusunan strategi dan kebijakan berbasis data.

Kata Kunci: beras premium, *Graded Response Model*, persepsi kualitas, teori respons butir, validitas psikometrik

ABSTRACT

Understanding consumer preferences requires measurement tools that are accurate, reliable, and capable of representing perceptions of product quality comprehensively. For strategic commodities such as premium rice, sensory and physical attributes, such as aroma, color, texture, cleanliness, and shelf life, play a crucial role in shaping consumer perceptions. This study aimed to develop and evaluate a measurement instrument for consumer perceptions of premium rice quality using the Item Response Theory (IRT) framework, specifically the Graded Response Model (GRM). The focus of this research was on constructing a valid and reliable tool to assess key attributes that were relevant in the context of consumer preferences. The analysis tested several fundamental assumptions of the GRM, namely unidimensionality, local independence, and monotonicity, all of which were satisfied. Most items in the instrument demonstrated high discrimination parameters (a > 1.0) as well as logical and well-distributed threshold values across the ability spectrum. Evaluation of the Item Characteristic Curves (ICC) and acceptance indices indicated that the scale effectively measures consumer preferences across a wide range of difficulty levels, from very low to very high. Furthermore, the test of response category effectiveness showed that a four-option format provided the best prediction, with Root Mean Square Error (RMSE) ranging between 0.1867 and 0.2138, which was lower than the three- and five-option versions. Thus, the GRM is proven to be effective in constructing a measurement scale for consumer preferences toward premium rice, offering valuable insights for data-driven strategies and policies.

Keywords: *premium rice, Graded Response Model, quality perception, item response theory, psychometric*

I. PENDAHULUAN

Beras merupakan bahan pangan pokok utama di Indonesia, dan dalam beberapa tahun terakhir terjadi pergeseran pola konsumsi menuju beras premium. Konsumen kini makin memperhatikan atribut kualitas, seperti tekstur, aroma, kebersihan, dan daya simpan, dalam pengambilan keputusan pembelian (Zhou et al., 2020; Natasya et al., 2024). Tren ini menunjukkan pentingnya pengukuran persepsi kualitas beras secara sistematis dan berbasis bukti. Namun, instrumen yang digunakan dalam penelitian sebelumnya umumnya masih bersifat deskriptif dan hanya mengandalkan indikator konvensional dengan skala sederhana, tanpa disertai uji psikometrik yang mendalam (Chamhuri & Batt, 2013). Kelemahan utama terletak pada tidak dilakukannya analisis berbasis *Item Response Theory* (IRT) untuk menguji validitas dan reliabilitas pada tingkat butir, serta kurangnya pertimbangan mengenai efektivitas jumlah kategori respons, sehingga berpotensi menimbulkan bias pengukuran.

Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi instrumen baru berbasis *Graded Response Model* (GRM) untuk mengukur persepsi konsumen terhadap kualitas beras premium. Instrumen ini dirancang untuk menguji struktur faktor laten, menilai parameter diskriminasi dan ambang kategori, serta menentukan jumlah kategori optimal. Dengan demikian, penelitian ini menawarkan kontribusi orisinal berupa instrumen pengukuran yang lebih akurat, efisien, dan berbasis bukti psikometrik.

II. KERANGKA TEORI *GRADED RESPONSE MODEL* (GRM)

Studi ini didasarkan pada teori persepsi konsumen tentang kualitas produk dengan perhatian khusus pada produk makanan seperti beras premium. Persepsi konsumen dibentuk oleh atribut sensori dan fisik seperti warna, bau, tekstur, kebersihan, dan masa simpan produk. Atribut-atribut ini subjektif dan secara signifikan memengaruhi keputusan pembelian. Untuk menangkap persepsi tersebut secara andal dan akurat, Teori Respons Butir (IRT) digunakan, yang menawarkan keunggulan dibandingkan teori tes klasik (CTT) karena memodelkan hubungan antara kemampuan laten (θ) dan probabilitas.

Untuk mengukur persepsi dengan akurat dan dapat diandalkan, kita menggunakan pendekatan Teori Respons Butir (*Item Response Theory*/IRT). Pendekatan ini memiliki keunggulan dibandingkan teori tes klasik (*Classical Test Theory*/CTT) karena dapat memodelkan hubungan antara kemampuan laten (θ) dan probabilitas respon terhadap item dengan lebih tepat, serta menghasilkan estimasi parameter yang tidak tergantung pada sampel (Embretson & Reise, 2000). Dalam instrumen yang menggunakan skala ordinal seperti skala Likert, model IRT yang paling tepat adalah *Graded Response Model* (GRM), yang dikembangkan oleh Samejima pada tahun 1969. GRM dirancang untuk menganalisis item dengan kategori bertingkat dan telah banyak diterapkan di bidang psikologi, pendidikan, kesehatan, dan riset sosial (Samejima, 1969; Hambelton et al., 2011). Model ini menghitung probabilitas kumulatif bahwa seseorang dengan tingkat atribut tertentu akan memilih kategori tertentu atau yang lebih tinggi. Probabilitas untuk memilih kategori spesifik diperoleh dari selisih antara dua probabilitas kumulatif. Parameter diskriminasi (a_i) menunjukkan seberapa baik item dapat membedakan individu pada berbagai tingkat atribut, sementara ambang batas (b_k) menunjukkan titik transisi antar kategori. Agar penerapan GRM valid, terdapat tiga asumsi yang harus dipenuhi: monotonisitas, unidimensionalitas, dan independensi lokal.

Agar penerapan GRM valid, terdapat tiga asumsi yang harus dipenuhi: monotonisitas, unidimensionalitas, dan independensi lokal. Monotonisitas berarti makin tinggi kemampuan laten seseorang, makin besar peluangnya memilih kategori lebih tinggi. Unidimensionalitas menekankan bahwa semua item harus mengukur satu konstruk utama, yang dapat diuji dengan analisis paralel, Analisis Faktor Eksploratori *(Exploratory Factor Analysis*/EFA) atau Skala Multidimensional (*Multidimensional Scaling*/ MDS) (De Ayala & Hertzog, 1991; Brentari & Golia, 2007). Sementara itu, independensi lokal berarti bahwa setelah pengaruh θ diperhitungkan, jawaban terhadap suatu item tidak memengaruhi item lain, yang dapat diuji melalui korelasi residual (Chen & Thissen, 1997).

GRM merupakan perluasan dari model logistik dua parameter (2PL) dari data dikotomis menjadi politomus. Keunggulan GRM meliputi fleksibilitas untuk berbagai jenis instrumen, kemampuan memberikan estimasi lebih informatif dibandingkan metode skoring klasik, serta kemampuannya memetakan proses berpindah kategori dalam bentuk fungsi tangga respons. Oleh karena itu, GRM merupakan pendekatan yang ideal dalam pengembangan dan validasi instrumen berbasis skala Likert, seperti dalam pengukuran persepsi konsumen terhadap kualitas beras premium.

III. METODOLOGI

3.1 Data Kuesioner

Instrumen penelitian ini disusun berdasarkan data primer dari studi preferensi konsumen terhadap kualitas dan harga beras premium yang dilakukan oleh Perum BULOG pada tahun 2016. Data dikumpulkan menggunakan teknik *snowball sampling*, di mana responden dijaring melalui rujukan dari responden sebelumnya, khususnya di sejumlah pasar di kota-kota besar seperti Medan, DKI Jakarta, Semarang, Yogyakarta, Surabaya, Balikpapan, dan Makassar. Meskipun data tersebut tergolong lama, pada saat pengumpulannya analisis yang dilakukan masih terbatas pada pendekatan statistik konvensional seperti analisis deskriptif dan regresi linear sederhana. Belum dilakukan pengujian instrumen secara psikometris yang komprehensif, khususnya dengan pendekatan *Item Response Theory* (IRT). Oleh karena itu, penelitian ini memanfaatkan kembali data tersebut untuk dianalisis menggunakan pendekatan IRT, khususnya *Graded Response Model* (GRM), guna mengevaluasi validitas konstruk, estimasi parameter item, serta kelayakan model pengukuran. Dengan demikian, penelitian ini memberikan kontribusi baru berupa validasi psikometrik terhadap instrumen yang sebelumnya belum diuji secara mendalam, dan meningkatkan akurasi serta kredibilitas pengukuran persepsi konsumen terhadap beras premium.

Instrumen yang digunakan adalah kuesioner dengan skala Likert 5 poin yang terdiri dari 15 pernyataan tentang atribut kualitas beras premium. Item dalam kuesioner mencakup aspek tekstur, aroma, warna, kebersihan, ketahanan, dan kepuasan konsumen terhadap nasi hasil olahan beras premium. Butir-butir kuesioner tersebut disajikan dalam Tabel 1.

3.2 Metode

Penelitian ini dirancang untuk menganalisis kualitas instrumen pengukuran persepsi konsumen terhadap beras premium dengan mempertimbangkan variasi jumlah opsi respons dalam skala Likert. Fokus utama kajian ini adalah mengevaluasi *Root Mean Square Error* (RMSE) sebagai indikator ketepatan model dalam konteks *Item Response Theory* (IRT). Proses analisis dimulai dengan pengujian prasyarat utama IRT, yaitu: (1) unidimensionalitas, yang diperiksa melalui analisis faktor eksploratori (EFA) menggunakan paket *psych* dalam perangkat lunak R (Revelle, 2023); (2) independensi lokal, yang diuji melalui analisis korelasi residual (*Q3 statistics*) menggunakan fungsi dari paket *Mirt* (Chalmers, 2012); serta (3) monotonisitas, yang dinilai dengan analisis Mokken melalui koefisien Loevinger's H menggunakan paket *mokken* (van der Ark, 2007).

Setelah ketiga asumsi tersebut terpenuhi, tahap selanjutnya adalah pemilihan model IRT politomus yang sesuai dengan sifat data dan struktur skala. Model yang dibandingkan meliputi *Partial Credit Model* (PCM), *Generalized Partial Credit Model* (GPCM), *Graded Response Model* (GRM), dan *Nominal Response Model* (NRM), yang seluruhnya diestimasi menggunakan fungsi dari paket *Mirt* dalam R (Chalmers, 2012). Pemilihan model dilakukan berdasarkan evaluasi kecocokan model (*fit indices*), kemampuan model menangkap sifat ordinal data, serta efisiensi pengukuran.

Setelah GRM terpilih sebagai model terbaik, dilakukan analisis parameter psikometrik untuk mengevaluasi kinerja masing-masing item. Parameter yang dianalisis meliputi: (1) parameter diskriminasi (a), yang menunjukkan seberapa sensitif item membedakan responden berdasarkan tingkat kemampuan laten; (2) ambang kategori (*threshold* b1–b4), yang merepresentasikan batas transisi antar kategori respons; (3) *Item Characteristic Curve* (ICC), yang digunakan untuk memvisualisasikan fungsi item; serta (4) informasi tes (TIF) dan standar galat (SE) sebagai indikator akurasi estimasi pada berbagai tingkat kemampuan (θ) (Cai &

Tabel 1. Kuesioner Atribut Kualitas Beras Premium

No Item	Pernyataan tentang Atribut Kualitas Beras Premium
Item 1	Beras Premium menghasilkan nasi yang lembut dan tidak keras.
Item 2	Beras Premium menghasilkan nasi dengan aroma yang harum dan segar.
Item 3	Beras Premium memiliki warna putih alami yang bersih tanpa pewarna tambahan.
Item 4	Beras Premium bersih dari benda asing seperti gabah atau kotoran lainnya.
Item 5	Beras Premium memiliki butiran beras yang utuh dengan sedikit butir patah.
Item 6	Beras Premium memiliki ukuran butiran yang seragam dan tidak bercampur dengan jenis beras lain.
Item 7	Beras Premium menggunakan varietas beras yang terkenal dan berkualitas tinggi.
Item 8	Beras Premium memiliki daya tahan yang baik untuk disimpan dalam waktu lama.
Item 9	Nasi dari Beras Premium tidak mudah basi saat disimpan pada suhu ruang.
Item 10	Nasi dari Beras Premium tetap pulen meskipun sudah dingin.
Item 11	Beras Premium tidak mudah rusak atau berkutu selama penyimpanan.
Item 12	Beras Premium memiliki tekstur nasi yang konsisten setiap kali dimasak.
Item 13	Beras Premium tidak memerlukan pencucian berkali-kali untuk menjadi bersih.
Item 14	Beras Premium tidak mengandung bau apek atau bau kimia saat belum dimasak.
Item 15	Nasi dari Beras Premium memberikan rasa enak meskipun dimasak tanpa lauk

Monroe, 2014). Selanjutnya, dilakukan simulasi untuk mengevaluasi efektivitas pengukuran berdasarkan jumlah opsi respons yang berbeda (3, 4, dan 5 poin). Simulasi dilakukan dalam tiga kondisi eksperimental, masing-masing diuji dalam 12 replikasi dengan jumlah responden acak sebanyak n = 150 per kondisi. Akurasi model dalam setiap kondisi dibandingkan menggunakan nilai RMSE, dan untuk menguji signifikansi perbedaannya digunakan analisis ANOVA satu arah dan uji lanjut Tukey HSD.

Seluruh analisis dilakukan menggunakan perangkat lunak R versi 4.5.0 dengan dukungan paket psych (Revelle, 2023), *Mirt* (Chalmers, 2012), dan mokken (van der Ark, 2007). Visualisasi data dilakukan menggunakan paket *ggplot2* dan *ggpubr* untuk mendukung interpretasi hasil secara lebih komprehensif dan replikatif.

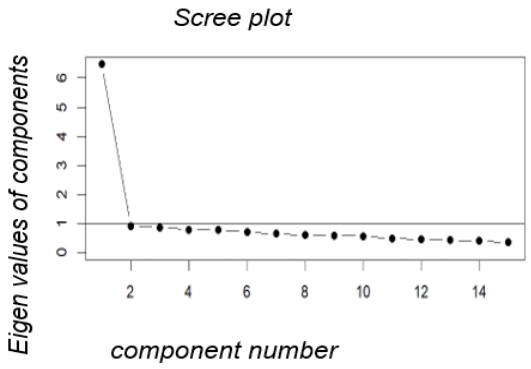
IV. HASIL DAN PEMBAHASAN

4.1 Hasil

4.1.1 Unidimensi

Unidimensionalitas merupakan asumsi mendasar dalam Teori Respons Butir (IRT) yang menyatakan bahwa sekumpulan butir hanya mengukur satu konstruk laten utama. Pemenuhan asumsi ini penting untuk menjamin validitas estimasi parameter, termasuk pada *Graded Response Model (GRM)*. Evaluasi unidimensionalitas biasanya dilakukan melalui analisis faktor eksploratori (EFA), konfirmatori (CFA), atau analisis *eigenvalue* dan *scree plot*, di mana satu faktor dominan dengan nilai *eigen* >1 dan tampilan “tekukan tajam” pada faktor pertama menjadi indikator utama. Meskipun unidimensionalitas murni sulit dicapai, yang terpenting adalah adanya dominasi satu dimensi laten terhadap respons butir (Reckase, 1979; Embretson & Reise, 2000). Hingga kini, analisis *eigenvalue* dan *scree plot* tetap banyak digunakan karena sederhana dan informatif, serta sering dipadukan dengan CFA untuk memperkuat bukti struktur dimensi data (Fokkema & Greiff, 2017).

Gambar 1 menyajikan hasil analisis *eigenvalue* dalam bentuk *scree plot* dan tabel nilai *eigen* untuk 15 faktor yang diekstraksi. Visualisasi tersebut digunakan untuk mengevaluasi struktur dimensi dari data



Faktor	Eigenvalues	Proportion of Variance	Cumulative Prop. Variance
Factor1	6.49	0.43	0.43
Factor2	0.91	0.06	0.49
Factor3	0.86	0.06	0.55
Factor4	0.79	0.05	0.60
Factor5	0.78	0.05	0.65
Factor6	0.69	0.05	0.70
Factor7	0.65	0.04	0.74
Factor8	0.60	0.04	0.78
Factor9	0.58	0.04	0.82
Factor10	0.56	0.04	0.86
Factor11	0.48	0.03	0.89
Factor12	0.45	0.03	0.92
Factor13	0.42	0.03	0.95
Factor14	0.39	0.03	0.98
Factor15	0.34	0.02	1.00

Gambar 1. Visualisasi Nilai *Eigen* dan Proporsi Varians dari 15 Faktor yang Diuji

dan menentukan jumlah faktor yang layak dipertahankan dalam analisis lebih lanjut. Hasil analisis *eigenvalue* menunjukkan bahwa Faktor 1 merupakan satu-satunya komponen dengan nilai *eigen* lebih dari 1, yaitu sebesar 6,49, yang secara umum memenuhi kriteria Kaiser untuk dipertahankan dalam analisis faktor. Sebaliknya, Faktor 2 hanya memiliki nilai *eigen* sebesar 0,91, sehingga tidak memenuhi ambang batas signifikansi. Rasio antara nilai *eigen* Faktor 1 terhadap Faktor 2 adalah 7,2, yang tergolong sangat tinggi dan merupakan indikator kuat bahwa struktur data bersifat unidimensional. Hal ini diperkuat oleh tampilan *scree plot*, yang menunjukkan adanya tekukan tajam (*elbow*) pada komponen pertama, menandakan bahwa kontribusi varians menurun drastis setelah faktor pertama.

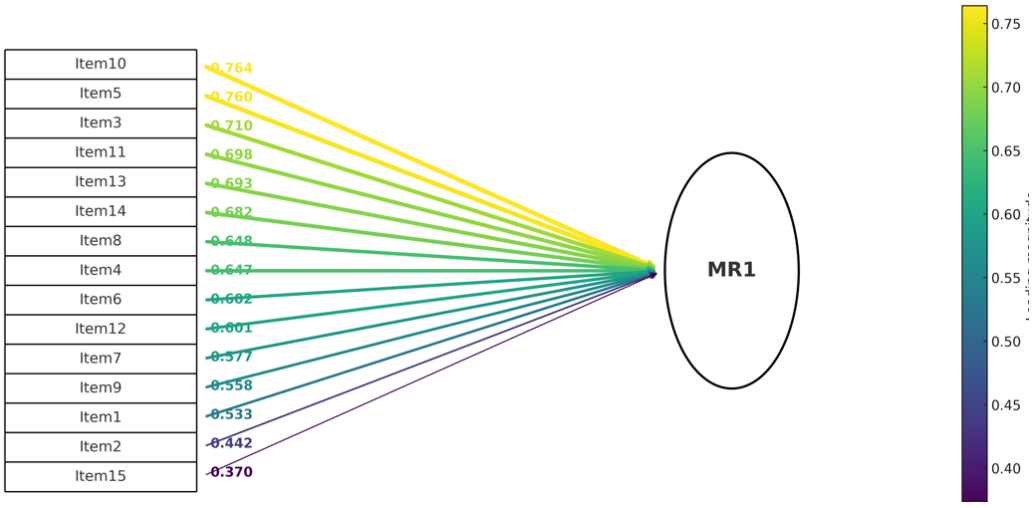
Faktor 1 menjelaskan 43 persen dari total varians, yang menunjukkan dominasi satu faktor utama dalam memengaruhi keseluruhan struktur data. Sementara itu, 14 faktor lainnya hanya menyumbang proporsi varians yang sangat kecil secara individual, dan secara kolektif diperlukan untuk menjelaskan sisa 57 persen varians. Kondisi ini memperkuat kesimpulan bahwa struktur data yang dianalisis cenderung unidimensional, dengan satu faktor dominan yang secara substansial menguasai pemodelan variabel laten yang diukur.

Hasil tersebut menunjukan bahwa instrumen kuesioner yang dianalisis mengukur satu faktor laten utama, memenuhi persyaratan unidimensi sehingga dapat menggunakan analisis IRT seperti seperti GRM. Unidimensional terpenuhi jika rasio nilai *eigen* faktor pertama terhadap yang kedua melebihi 4,0 (Reckase, 1979). Selain itu, dominasi satu faktor utama dalam analisis nilai *eigen* mendukung penggunaan model IRT unidimensional (Reise et al., 2007). Unidimensionalitas merupakan syarat penting dalam validasi skala psikometrik, terutama ketika kita menggunakan pendekatan IRT (Meijer & Tendeiro, 2018). Dengan demikian, kita bisa menyimpulkan bahwa hasil ini memberikan dasar teoritis dan empiris yang kuat, menunjukkan bahwa instrumen yang digunakan dalam penelitian ini memenuhi kriteria unidimensionalitas. Oleh karena itu, instrumen tersebut layak untuk dianalisis lebih lanjut menggunakan pendekatan GRM dalam kerangka Teori Respons Butir.

Untuk memastikan bahwa konstruk yang diukur bersifat unidimensional, dilakukan analisis faktor eksploratori terhadap seluruh item dalam instrumen. Visualisasi hasil analisis menunjukkan hubungan antar item dengan satu faktor utama (MR1; Gambar 2).

Hasil analisis faktor menunjukkan bahwa seluruh item dalam instrumen memuat ke satu faktor utama yang diberi label MR1, mengindikasikan bahwa konstruk yang diukur bersifat unidimensional. Hal ini menunjukkan bahwa seluruh item mengukur satu dimensi laten yang sama, yang merupakan prasyarat penting dalam pengembangan dan validasi instrumen psikometrik.

Nilai faktor *loading* dari masing-masing item berkisar antara 0,37 hingga 0,764. Kriteria item dengan *loading* di atas 0,30 dianggap berkontribusi terhadap faktor yang sama (Hair Jr et al., 2009). Secara umum, *loading* antara 0,30–0,49 dianggap cukup, 0,50–0,69 dianggap kuat, dan nilai ≥ 0,70 dianggap sangat kuat. Hasil analisis menunjukkan bahwa seluruh item



Gambar 2. Visualisasi Struktur Unidimensional Hasil Analisis Faktor Eksploratori

berkontribusi pada satu faktor utama kualitas beras premium dengan tingkat kekuatan berbeda. Tiga indikator dengan kontribusi sangat kuat (≥0,70) adalah nasi tetap pulen meskipun dingin (Item 10, *loading* 0,764), butiran beras utuh dengan sedikit patah (Item 5, 0,760), dan warna putih alami tanpa pewarna (Item 3, 0,710). Indikator dengan kontribusi kuat (0,50–0,69) meliputi tidak mudah rusak/ berkutu (Item 11, 0,698), tidak perlu pencucian berkali-kali (Item 13, 0,693), tidak berbau apek/ kimia (Item 14, 0,682), daya tahan simpan baik (Item 8, 0,648), bersih dari benda asing (Item 4, 0,647), ukuran butiran seragam (Item 6, 0,602), tekstur nasi konsisten tiap masak (Item 12, 0,577), varietas terkenal dan berkualitas (Item 7, 0,558), serta nasi tidak mudah basi di suhu ruang (Item 9, 0,533). Sementara itu, kontribusi cukup (0,30–0,49) ditunjukkan oleh nasi lembut dan tidak keras (Item 1, 0,442), aroma nasi harum dan segar (Item 2, 0,370), dan rasa enak meski tanpa lau*k* (Item 15, 0,442). Temuan ini menegaskan bahwa aspek tekstur pascamasak, bentuk fisik, dan penampilan visual merupakan penentu paling dominan dalam persepsi kualitas beras premium.

Berdasarkan temuan tersebut, tidak terdapat item yang perlu dieliminasi secara langsung, karena semua item memiliki nilai *loading* ≥ 0,30. Oleh karena itu, instrumen ini dapat dinyatakan layak untuk dianalisis lebih lanjut, terutama dalam konteks penggunaan model IRT unidimensional seperti *Graded Response Model* (GRM). Namun demikian, item-item dengan kontribusi rendah tetap perlu menjadi perhatian untuk evaluasi lebih lanjut atau perbaikan tata bahasanya.

Secara teoretis, hasil ini sejalan dengan pandangan Meijer & Tendeiro (2018), yang menekankan pentingnya unidimensionalitas dalam validasi skala psikologis. Selain itu, menurut Embretson & Reise (2000), identifikasi struktur faktor laten merupakan dasar utama untuk memastikan bahwa model IRT dapat digunakan secara tepat dan valid dalam pengukuran konstruk laten.

4.1.2 *Local independent*

Mengacu pada nilai absolut Q3 ≤ 0,30 menandakan bahwa asumsi ini terpenuhi (Yen, 1993). Hasil menunjukkan bahwa seluruh pasangan item memiliki nilai Q3 dalam batas toleransi, didukung pula oleh hasil uji χ^2 dan G^2 yang tidak signifikan secara statistik. Ini menunjukkan bahwa tidak ada korelasi residual yang berarti antar item.

Dengan demikian, dapat disimpulkan bahwa instrumen memenuhi asumsi independensi lokal, memungkinkan setiap item mengukur konstruk secara mandiri. Oleh karena itu, instrumen ini layak digunakan dalam analisis IRT unidimensional. Hasil ini juga memperkuat validitas struktural instrumen dalam konteks pemodelan laten.

Temuan ini sejalan dengan literatur, di mana Chen & Thissen (1997) memperkenalkan Q3 sebagai alat utama deteksi lokal dependen, dan Embretson & Reise (2000) menegaskan

Tabel 2. Hasil Uji Monotonisitas dan Koefisien Loevinger (H) untuk Setiap Item

Item	ItemH	#ac	#vi	#vi/#ac	*maxvi*	*sum*	*sum/#ac*	*zmax*	*#zsig*	*crit*
Item1	0,36	24	0	0	0	0	0	0	0	0
Item2	0,31	24	0	0	0	0	0	0	0	0
Item3	0,5	16	0	0	0	0	0	0	0	0
Item4	0,47	24	0	0	0	0	0	0	0	0
Item5	0,54	16	0	0	0	0	0	0	0	0
Item6	0,41	24	0	0	0	0	0	0	0	0
Item7	0,39	24	0	0	0	0	0	0	0	0
Item8	0,45	24	0	0	0	0	0	0	0	0
Item9	0,38	24	1	0,04	0,03	0,03	0,0013	0,24	0	8
Item10	0,5	24	0	0	0	0	0	0	0	0
Item11	0,48	24	0	0	0	0	0	0	0	0
Item12	0,41	24	0	0	0	0	0	0	0	0
Item13	0,48	21	0	0	0	0	0	0	0	0
Item14	0,53	14	0	0	0	0	0	0	0	0
Item15	0,26	24	0	0	0	0	0	0	0	0

bahwa pelanggaran independensi lokal dapat menyebabkan bias estimasi dan menurunkan kualitas pengukuran. Maka, hasil uji ini menjadi indikator penting bagi kualitas dan keabsahan psikometrik suatu instrumen.

4.1.3 Monotonisasi

Berdasarkan hasil analisis monotonisitas pada Tabel 2, seluruh item instrumen memenuhi asumsi monotonik yang ditunjukkan oleh tidak adanya pelanggaran dan proporsi pelanggaran sebesar 0 persen. Nilai koefisien Loevinger (*ItemH*) juga mendukung kesimpulan ini, di mana sebagian besar item memiliki skalabilitas memadai hingga baik. Secara khusus, Item 5 (Beras Premium memiliki butiran beras yang utuh dengan sedikit butir patah) menunjukkan nilai skalabilitas tertinggi, menegaskan pentingnya bentuk fisik butiran sebagai indikator mutu. Demikian pula, Item 14 (Beras Premium tidak mengandung bau apek atau bau kimia) juga menempati posisi teratas, yang memperlihatkan bahwa aroma alami merupakan penanda kualitas penting dalam persepsi konsumen. Secara teknis, nilai *sum/#ac* ≠ 0 dan *zmax* > 0 pada Item9 (Nasi dari Beras Premium tidak mudah basi saat disimpan pada suhu ruang) berarti bahwa ini item sedikit fluktuatif terhadap pola monotonik, artinya respons tiap item tidaklah sempurna dalam sumbu monoton. Namun, mengingat bahwa #vi = 0 dan *#zsig* = 0, maka tidak ada penambahan yang signifikan atau aktual berdasarkan pemilih terhadap asumsi monotonisitas.

Dalam beberapa kasus, fluktuasi semacam ini merupakan akibat dari variasi kecil dalam data atau kelompok responden yang memberikan jawaban konsisten namun tidak cukup kuat untuk dianggap sebagai pelanggaran nyata. Dengan demikian, Item 9 (Nasi dari Beras Premium tidak mudah basi saat disimpan pada suhu ruang) tetap memenuhi asumsi monotonisitas, dengan nilai ItemH = 0,45 yang menunjukkan bahwa butir ini masih informatif dan bermanfaat dalam pemutusan psikometrik instrumen. Oleh karenanya, perbedaan nilai dalam fungsi monotonik *sum/#ac* = 0,0013 ini tidak memengaruhi validitas item secara keseluruhan. Temuan ini menunjukkan bahwa responden dengan tingkat kemampuan laten yang lebih tinggi cenderung memberikan respons pada kategori yang lebih tinggi. Hal ini mencerminkan bahwa fungsi respons setiap item meningkat secara monoton, sesuai dengan asumsi dasar model IRT. Selain fakta bahwa monotonisitas harus dipenuhi dalam skala nonparametrik untuk memastikan ketertiban respons (Mokken, 1997; Van Der Ark, 2012), penekanan bahwa nilai Loevinger H ≥ 0,30 menunjukkan daya diskriminasi yang cukup (Sijtsma & Molenaar, 2002). Lebih lanjut, monotonisitas merupakan asumsi dasar yang esensial dalam penerapan model IRT unidimensional (Junker, 1996).

Tabel 3. Hasil Uji Kecocokan Model IRT (GRM, GPCM, PCM, dan NRM) Berdasarkan Analisis ANOVA pada Faktor Persepsi Kualitas

Jenis IRT	AIC	SABIC	HQ	BIC	*logLik*	X2	df	p
GRM	10.677,17	10.717,10	10.788,34	10.955	-5.263,60			
GPCM	10.748,42	10.788,35	10.859,59	11.026	-5.299,20	-71,25	0	NaN
PCM	10.986,43	11.018,90	11.076,85	11.212	-5.432,20	-266,00	-14	NaN
NRM	10.751,89	10.815,77	10.929,76	11.196	-5.255,90	352,54	59	0

4.1.4 Pemilihan model IRT

Dalam kerangka *Item Response Theory* (IRT), terdapat beberapa model yang digunakan untuk menganalisis data skala kategori. GRM dikembangkan oleh Samejima (1969) dan umumnya digunakan untuk item dengan skala bertingkat (ordinal), di mana setiap kategori jawaban merepresentasikan tingkat intensitas tertentu. GPCM merupakan pengembangan dari PCM yang lebih fleksibel karena mengestimasi parameter diskriminasi tiap item, sehingga mampu membedakan sejauh mana suatu item sensitif terhadap variasi kemampuan responden. PCM adalah bentuk khusus dari Rasch Model untuk item kategori bertingkat, dengan asumsi bahwa semua item memiliki parameter diskriminasi yang sama. Sementara itu, NRM digunakan untuk item dengan kategori jawaban nominal (tidak berurutan), sehingga model ini lebih sesuai jika kategori tidak mencerminkan tingkatan tertentu melainkan hanya pilihan alternatif.

Pemilihan model analisis ini didasarkan pada hasil uji kecocokan model (*model fit*) yang membandingkan beberapa model dalam kerangka Teori Respons Butir (*Item Response Theory*), yaitu GRM, GPCM, PCM, dan NRM.

Dalam penelitian ini, perbandingan kecocokan model dilakukan menggunakan tiga kriteria informasi, yaitu *Akaike Information Criterion* (AIC), *Bayesian Information Criterion* (BIC), dan *Sample-Size Adjusted Bayesian Information Criterion* (SABIC). Pemilihan model terbaik didukung oleh kriteria information criteria seperti AIC, BIC, dan SABIC, yang secara teoritis menunjukkan bahwa model dengan nilai terkecil adalah yang paling *parsimonious* tanpa mengorbankan kesesuaian data (Dziak et al., 2020; Sen, et al 2017). Berdasarkan Tabel 3, GRM menunjukkan nilai AIC (10.677,17), SABIC, dan *log-likelihood* tertinggi secara relatif, yang mengindikasikan model ini memiliki keseimbangan terbaik antara akurasi dan kompleksitas model. Dibandingkan dengan GPCM dan PCM, GRM menghasilkan *logLik* yang lebih tinggi (–5.263,60) dan nilai penalti kompleksitas (BIC dan HQ) yang lebih rendah. Model NRM meskipun memiliki X^2 signifikan (p = 0), menunjukkan peningkatan kompleksitas (df = 59) yang tidak sebanding dengan perbaikan *fit,* sehingga berisiko *overfitting.*

Tabel 4. Parameter Diskriminasi

Item	a1	Keterangan
Item1	1,234	Sangat Baik
Item2	0,972	Baik
Item3	1,862	Sangat Baik
Item4	1,946	Sangat Baik
Item5	2,610	Sangat Baik
Item6	1,579	Sangat Baik
Item7	1,443	Sangat Baik
Item8	1,837	Sangat Baik
Item9	1,371	Sangat Baik
Item10	2,392	Sangat Baik
Item11	2,086	Sangat Baik
Item12	1,596	Sangat Baik
Item13	2,209	Sangat Baik
Item14	2,187	Sangat Baik
Item15	0,672	Baik

Selanjutnya, dari nilai X^2 GRM lebih rendah dari lain, GRM memperlihatkan penjelasan terkuat sebagai model yang tidak kompleks, dengan model yang paling realistis. Juga, mengingat efisiensi parameter, struktur ordinal data yang dibangun dalam instrumen yang menggunakan skala Likert, dan asumsi dasar model yang dipenuhi, seperti monotonisitas, unidimensionalitas dan LD, GRM merupakan model terbaik yang dapat digunakan untuk menganalisis data penelitian. Instrumen tersebut sudah layak untuk dianalisis lebih lanjut dengan model IRT parametrik GRM.

4.1.5 Parameter diskriminasi

Untuk menilai seberapa baik setiap item dapat membedakan responden berdasarkan tingkat kemampuan laten, dilakukan analisis parameter diskriminasi (a1) dengan menggunakan pendekatan *Graded Response Model* (GRM). Parameter a1 menunjukkan kemampuan item dalam membedakan individu berdasarkan tingkat kemampuan laten. Teori menyatakan bahwa makin tinggi nilai a1, makin tajam item membedakan individu pada rentang kemampuan yang berbeda. Item dengan a1 > 1,0 dianggap sangat informatif sedangkan a1 < 0,5 tidak efektif dalam diskriminasi (Baker & Kim, 2004). Hasil estimasi parameter diskriminasi dari semua item dapat dilihat dalam Tabel 4.

Dari Tabel 4, sebagian besar item memperoleh nilai diskriminasi di atas 1,0, menandakan kemampuan sangat baik dalam membedakan responden dengan tingkat persepsi kualitas beras premium yang berbeda. Misalnya Item 5 (a1 = 2,610), butiran beras utuh dengan sedikit butir patah, menunjukkan daya diskriminasi sangat tinggi. Artinya konsumen sangat sensitif terhadap kondisi fisik butiran beras. Begitu pula Item 10 yaitu nasi tetap pulen meskipun dingin dan Item 13 yaitu tidak perlu pencucian berkali-kali, yang keduanya juga memperlihatkan daya diskriminasi tinggi.

Sebaliknya, Item 15 (a1 = 0,672), rasa enak meski dimasak tanpa lauk, hanya menunjukkan diskriminasi mendekati batas bawah efektivitas, sehingga substansi atau redaksinya mungkin perlu dievaluasi lebih lanjut. Mengacu pada klasifikasi parameter diskriminasi (a1) menurut Hambleton et al. (2011) dan interpretasi Himelfarb (2019), sebanyak 13 butir item dapat dikategorikan sangat baik, sedangkan Item 2 yaitu aroma nasi harum dan segar dan Item 15 yaitu rasa enak meski dimasak tanpa lauk, termasuk kategori baik.

4.1.6 Ambang batas kumulatif *(threshold)*

Hasil analisis menggunakan GRM menunjukkan estimasi parameter diskriminasi dan ambang batas kategori untuk masing-masing item dalam instrumen (Tabel 5).

Selain itu, parameter ambang batas menggambarkan titik-titik pada kontinum kemampuan di mana responden beralih dari satu kategori jawaban ke kategori yang lebih tinggi. Harusnya, ambang tersebar merata dan tidak tumpang tindih. Item5, misalnya, memiliki ambang batas yang cukup lebar dan negatif, menunjukkan bahwa peralihan kategori terjadi terutama pada responden dengan kemampuan rendah sampai sedang. Namun, jika ambang terlalu sempit atau saling tumpang tindih, masing-masing item, hal tersebut dapat menunjukkan bahwa responden sulit untuk membedakan perbedaan antar kategori jawaban, yang mengarah pada ketidakjelasan fungsi item. Analisis parameter ambang batas kumulatif dalam Tabel 5 memberikan hasil yang baik.

Tabel 5. Ambang Batas Kumulatif (*Threshold*)

Item	a1	Keterangan
Item1	1,234	Sangat Baik
Item2	0,972	Baik
Item3	1,862	Sangat Baik
Item4	1,946	Sangat Baik
Item5	2,610	Sangat Baik
Item6	1,579	Sangat Baik
Item7	1,443	Sangat Baik
Item8	1,837	Sangat Baik
Item9	1,371	Sangat Baik
Item10	2,392	Sangat Baik
Item11	2,086	Sangat Baik
Item12	1,596	Sangat Baik
Item13	2,209	Sangat Baik
Item14	2,187	Sangat Baik
Item15	0,672	Baik

Dalam keseluruhan, analisis dapat mendukung bahwa mayoritas item dalam instrumen ini layak digunakan karena memiliki kualitas psikometrik yang baik. Namun demikian, item-item seperti Item15, memiliki parameter yang relatif lemah sehingga tetap memerlukan tinjauan lanjutan untuk menjamin keandalan konstruk instrumen secara keseluruhan.

Temuan ini sejalan dengan kerangka teori yang dikembangkan oleh Samejima (1969) yang memperkenalkan GRM sebagai model IRT yang tepat untuk skala politomus terurut, seperti skala Likert. Selain itu, Embretson dan Reise (2000) menekankan bahwa distribusi ambang batas yang merata dan diskriminasi tinggi merupakan indikator penting dari kualitas instrumen dalam

Tabel 6. Ambang Batas Aktual Antar Kategori

Item	b1	b2	b3	b4	Rata-rata	Keterangan
Item1	-0.999	-0.966	0.075	0.571	-0.330	Sedang
Item2	-0.839	0.312	0.397	0.689	0.140	Sedang
Item3	-2.025	-1.046	0.517	1.766	-0.197	Sedang
Item4	-1.376	-0.864	-0.77	1.27	(0.435	**Tinggi**
Item5	-1.817	0.027	0.768	2.084	0.266	Sedang
Item6	-1.37	-0.819	-0.652	0.274	-0.642	**Rendah**
Item7	-0.992	-0.554	-0.236	0.426	-0.339	Sedang
Item8	-1.528	-1.32	0.519	0.657	-0.418	**Rendah**
Item9	-1.512	-0.089	0.124	2.209	0.183	Sedang
Item10	-1.295	-0.778	0.078	0.671	-0.331	Sedang
Item11	-0.453	-0.044	0.885	1.567	0.489	**Tinggi**
Item12	-0.847	-0.426	-0.059	0.332	-0.250	Sedang
Item13	-2.148	-0.395	-0.352	0.023	-0.718	**Rendah**
Item14	-2.006	-1.967	-1.258	0.395	-1.209	**Rendah**
Item15	-1.31	-1.083	0.147	1.909	-0.084	Sedang

mengukur variasi kemampuan responden secara sensitif dan akurat.

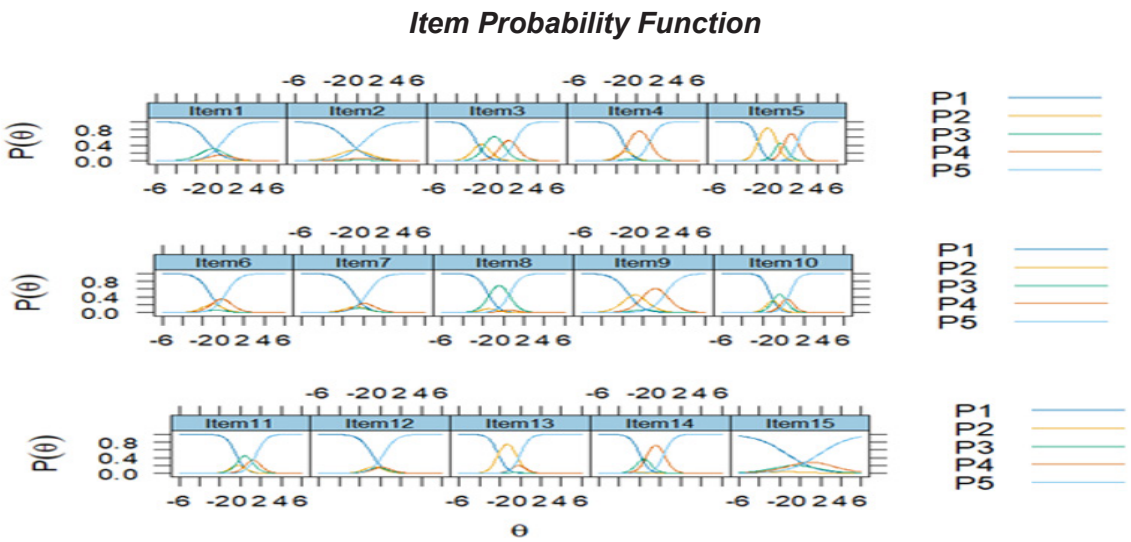
Endorsement/ambang batas aktual antar kategori data yang digunakan dalam analisis ini ada hasil skala kumulatif yang dikonversi ke skala probabilistik. Pada model *Graded Response Model* (GRM), ambang batas muncul sebagai b1–b4 menggambarkan titik kemampuan laten (θ) di mana subjek beralih dari satu kategori respons ke kategori respons berikutnya (Tabel 6).

Rata-rata nilai ambang atau location di dalamnya menandakan ukuran tingkat *endorsement* atau kemudahan butir dengan dipilih atau disetujui sembilan angka *location* tersebut menggambarkan seberapa sukar atau mudahnya item untuk setiap subjek. Ketika nilai *location* besar, kemampuan yang dibutuhkan juga besar sampai bisa memilih item yang rata-ratanya tinggi (*endorsement* tinggi), sedangkan jika nilai *location* rendah atau negatif, maka item mudah dipilih (*endorsement* rendah). Berdasarkan klasifikasi ini, item dengan rata-rata *location* ≤ –0,40 dikategori sebagai rendah, antara –0,39 hingga 0,39 adalah sedang, dan jika ≥ 0,40 adalah tinggi

Hasil simulasi menunjukkan bahwa sebagian besar item 8 dari 15 berada dalam kategori sedang yang menandakan *location endorsement* setara dengan kemampuan responden rata-rata. Sebanyak 5 item berada dalam kategori rendah, contohnya adalah Item13 dan Item14, menandakan item yang sangat mudah dijawab positif. Hanya 2 item berada dalam kategori tinggi, yaitu Item4 dan Item11, yang berarti memerlukan kemampuan tinggi untuk mencapai kategori respons atas. Klasifikasi ini berguna untuk menilai dan meningkatkan ketajaman butir serta dapat merekontruksi instrumen.

Sebaran kategori lokasi ini penting dalam konteks GRM karena menunjukkan cakupan kemampuan yang diukur oleh setiap item. Item dengan lokasi tinggi mengukur responden berkemampuan tinggi, sedangkan item dengan lokasi rendah mengukur responden dengan kemampuan lebih rendah. Keberagaman ini penting untuk memastikan bahwa instrumen dapat mengevaluasi spektrum penuh preferensi atau kemampuan yang dimaksud.

Dari sisi psikometri, hasil tersebut mencerminkan bahwa instrumen memiliki distribusi tingkat kesulitan yang baik dan beragam. Meski demikian, Item6, Item8, Item13, dan terutama, Item14 perlu direvisi karena memiliki nilai *location* yang sangat rendah. Hal ini dapat diartikan bahwa item-item tersebut terlalu mudah disetujui oleh responden, bahkan responden yang punya tingkat kemampuan laten yang rendah. Inilah yang dapat menjadi risiko besar, yaitu bahwa item-item tersebut kurang mampu membedakan antara responden



Gambar 3. Kurva Karakteristik Tiap Item (ICC) Berdasarkan Model GRM

yang memiliki tingkat kemampuan laten tertentu dari yang lain. Oleh karena itu, perlu dilakukan revisi yang pertama adalah dengan menjadikan pernyataan lebih spesifik, sulit dan menghindari formulasi yang bersifat normatif atau umum. Revisi kedua adalah bias dalam membuat kalimat yang ambigu serta terlalu umum yang mudah disetujui oleh semua responden tanpa pertimbangan. Dengan begitu, alasan revisi yang dilakukan adalah untuk masing-masing item harus memiliki daya beda *artifacts*, sehingga instrumen dapat membedakan persepsi lebih tajam dan mencakup rentang kemampuan laten yang lebih luas dalam kasus populasi.

Sesuai dengan Hambleton & Swaminathan, (1985); Rubio et al., (2007); dan Matteucci & Stracqualursi, (2006), ambang batas yang terlalu tinggi atau rendah dapat menurunkan validitas diskriminatif item. Oleh karena itu, revisi terhadap item-item dengan *endorsement* ekstrem sangat direkomendasikan demi memperkuat kualitas instrumen dalam mengukur persepsi atau preferensi secara lebih akurat dan seimbang.

4.1.7 Grafik ICC per Item

Gambar 3 menampilkan kurva karakteristik tiap item (ICC) berdasarkan model GRM. Instrumen pengukuran persepsi kualitas beras premium disusun menggunakan skala Likert bertingkat 5 kategori respons, yaitu 1 = sangat tidak setuju, 2 = tidak setuju, 3 = netral, 4 = setuju, dan 5 = sangat setuju. Skala ini dipilih karena mampu merepresentasikan variasi intensitas sikap responden terhadap setiap pernyataan secara lebih terukur. Selain itu, dalam kerangka *Item Response Theory* (IRT), penggunaan skala ordinal seperti Likert memungkinkan penerapan model *Graded Response Model* (GRM) yang dirancang untuk menguji kinerja butir (item) secara lebih detail, mencakup parameter diskriminasi dan ambang batas kategori (*threshold*). Dengan demikian, selain memperoleh informasi deskriptif, analisis ini juga dapat mengidentifikasi apakah kategori respons yang digunakan benar-benar berfungsi optimal dalam membedakan persepsi konsumen.

Berdasarkan interpretasi kurva Karakteristik Item (*Item Characteristic Curve*/ICC) hasil pemodelan *Graded Response Model* (GRM), diperoleh gambaran performa masing-masing item dalam skala Likert 5 poin. Item 3, 4, 5, 13, dan 14 menunjukkan kurva yang terdistribusi secara simetris dan berlapis, yang menandakan bahwa setiap kategori respons berfungsi secara optimal. Dengan kata lain, masing-masing kategori memiliki rentang kemampuan laten (θ) tertentu yang jelas, dan digunakan secara proporsional oleh responden dari berbagai tingkat kemampuan. Pola ini mencerminkan kualitas psikometrik yang baik dan sangat ideal untuk instrumen berbasis skala Likert.

Sebaliknya, item 1, 2, 6, 7, 10, 11, dan 15 menunjukkan dominasi kurva P1 (kategori terendah) di sebagian besar rentang kemampuan θ. Hal ini mengindikasikan bahwa

mayoritas responden cenderung memilih kategori terbawah, bahkan pada kemampuan sedang. Artinya, item-item tersebut relatif terlalu mudah dan tidak cukup sensitif untuk membedakan responden dengan preferensi atau kemampuan menengah ke atas. Akibatnya, item-item tersebut perlu dikaji ulang karena kurang mampu memberikan informasi yang kaya dalam rentang konstruksi yang lebih luas.

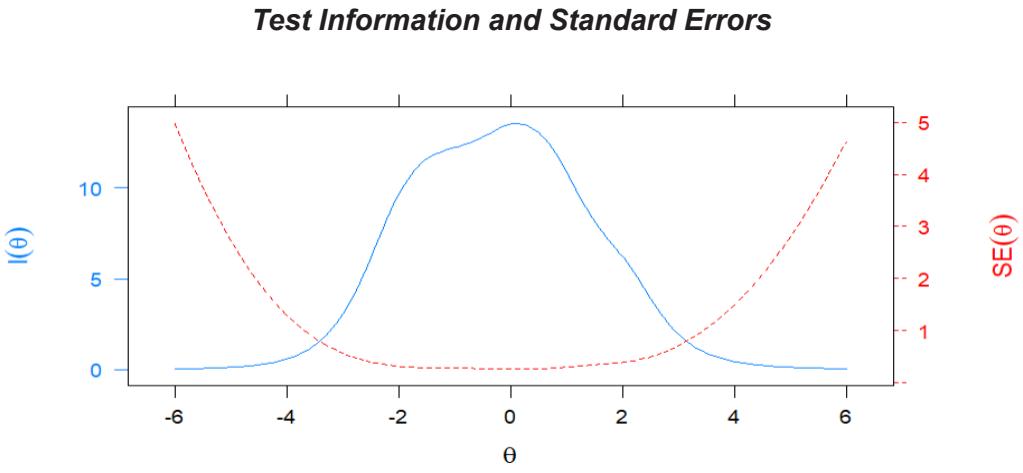
Adapun item 8 dan 9 memperlihatkan bentuk kurva yang asimetris dengan tumpang tindih antar kategori. Kondisi ini menunjukkan bahwa batas antar kategori tidak cukup jelas atau terlalu sempit, sehingga membingungkan responden dalam membedakan pilihan respons. Fenomena ini bisa mengarah pada kesalahan klasifikasi atau hilangnya akurasi pengukuran dalam tingkat kemampuan tertentu.

Secara keseluruhan, dapat disimpulkan bahwa kualitas psikometrik terbaik ditunjukkan oleh item 3, 4, 5, 13, dan 14 karena memiliki kurva yang terdistribusi seimbang dan responsif terhadap variasi kemampuan responden. Sebaliknya, item 1, 2, 6, 7, 10, 11, dan 15 memerlukan revisi atau perampingan karena dominasi kategori rendah yang berlebihan, sedangkan item 8 dan 9 perlu diperbaiki dari segi struktur kategori agar tidak terjadi redundansi atau tumpang tindih antar pilihan respons.

(GRM). Kurva informasi menunjukkan bahwa puncak informasi berada di sekitar θ = 0, yang menandakan bahwa tes paling informatif dan sensitif bagi responden dengan tingkat kemampuan laten menengah. Di luar rentang –2 hingga +2, kurva informasi menurun tajam, menunjukkan bahwa tes ini menjadi kurang sensitif dalam mendeteksi perbedaan kemampuan pada individu dengan kemampuan sangat rendah atau sangat tinggi.

Sebaliknya, kurva galat standar (SE) menunjukkan pola yang berbanding terbalik dengan kurva informasi. Di area di mana informasi tinggi (sekitar θ = 0), nilai SE rendah (sekitar 0,3–0,5), yang berarti estimasi kemampuan laten paling akurat berada di sekitar nilai tengah. Namun, pada rentang ekstrem (θ < -3 atau θ > +3), SE meningkat secara drastis, menandakan bahwa estimasi kemampuan untuk responden dengan kemampuan ekstrem menjadi kurang presisi.

Berdasarkan pola kedua kurva tersebut, dapat disimpulkan bahwa instrumen ini memiliki performa pengukuran optimal untuk populasi dengan kemampuan rata-rata, namun kurang efektif dalam mengukur individu dengan kemampuan yang sangat rendah atau sangat tinggi. Oleh karena itu, instrumen ini cocok digunakan untuk penilaian umum, bukan untuk diagnostik ekstrem, kecuali ditambahkan butir-butir yang secara khusus menyasar kelompok ekstrem tersebut (misalnya, butir yang sangat mudah atau sangat sulit).



Gambar 4. *Total Information Function*

4.1.8 *Total Information Function*

Gambar 4 memperlihatkan kurva Informasi Tes (I(θ)) dan Galat Standar (SE(θ)) dari hasil estimasi model *Graded Response Model*

Informasi tes dan galat standar memiliki hubungan terbalik makin banyak informasi yang diberikan oleh item atau tes, makin akurat estimasi kemampuan yang dihasilkan (Lord, 2008). Selain itu, menurut Embretson dan Reise (2000), distribusi informasi yang merata sesuai dengan kebutuhan pengukuran adalah karakteristik tes yang baik. Ditegaskan pada kurva informasi tes yang sempit dan terpusat di sekitar $\theta = 0$ mengindikasikan bahwa tes tersebut lebih cocok untuk penilaian sumatif standar, bukan untuk mengidentifikasi kemampuan ekstrem dalam suatu populasi (Baker, 2001).

4.1.9 Alternatif pilihan/opsi kuesioner

Kuesioner dengan 5 opsi, setelah dilakukan uji asumsi IRT yaitu dimensionalitas, *local independent* serta monotonisasi, diperoleh model yang cocok adalah GRM. Selanjutnya dengan model GRM dilakukan alternatif opsi yaitu Opsi 4 dan opsi 3. Penyesuaian dari opsi 5 (12345) ke opsi 4 (12234) dilakukan, sedangkan 5 pilihan kategori 12345 dilakukan rekategori 3 pilhan (11233) (Chong et al., 2022). Dari ke 3 kategori pilihan tersebut masing-masing memenuhi persyaratan unidimensi, *local independent* dan monotonisasi dengan model IRT yang terpilih tetap sama yaitu GRM.

Untuk membandingkan efektivitas instrumen berdasarkan jumlah opsi jawaban, dilakukan analisis terpisah terhadap tiga versi instrumen, yaitu versi dengan 3 opsi, 4 opsi, dan 5 opsi jawaban. Masing-masing versi diuji menggunakan pendekatan eksperimental sebanyak 12 replikasi, dengan setiap replikasi melibatkan pengambilan sampel acak berjumlah 150 responden dengan pengembalian. Variabel terikat yang dianalisis adalah *Root Mean Square Error* (RMSE), yang digunakan sebagai indikator ketepatan estimasi kemampuan dalam pemodelan IRT. Hasil perhitungan nilai RMSE untuk ketiga versi instrumen disajikan pada Tabel 7.

Hasil replikasi menunjukkan bahwa rata-rata nilai RMSE untuk versi 5 pilihan berkisar antara 0,2000 hingga 0,2350, versi 4 pilihan antara 0,1867 hingga 0,2138, dan versi 3 pilihan antara 0,2004 hingga 0,2540. Hal ini mengindikasikan bahwa instrumen dengan 4 opsi jawaban cenderung menghasilkan nilai RMSE yang lebih rendah dan stabil, dibandingkan versi lainnya (Linacre, 2002)

Tabel 7. Nilai *Root Mean Square Error* (RMSE) pada Instrumen dengan 3, 4, dan 5 Opsi

Replikasi	5 Pilihan	4 Pilihan	3 Pilihan
1	0.2255	0.2224	0.2238
2	0.2231	0.2082	0.2243
3	0.2350	0.2121	0.2267
4	0.2225	0.2089	0.2045
5	0.2238	0.2105	0.2540
6	0.2064	0.1990	0.2085
7	0.2132	0.2025	0.2004
8	0.2000	0.1867	0.2311
9	0.2113	0.1869	0.2329
10	0.2065	0.1909	0.2234
11	0.2340	0.2138	0.2353
12	0.2309	0.2086	0.2086

Selanjutnya, perbedaan signifikan antar kondisi jumlah opsi jawaban diuji dengan analisis varians satu arah. Berdasarkan hasil uji ANOVA menunjukkan terdapat perbedaan yang signifikan ($p<0,001$) antara rata-rata nilai *Root Mean Square Error* (RMSE) dari model dengan jumlah pilihan jawaban yang berbeda, yaitu 3 pilihan, 4 pilihan, dan 5 pilihan. Akurasi ini diukur melalui nilai RMSE yang mencerminkan kesesuaian antara estimasi kemampuan dan data aktual responden.

Hasil uji lanjutan Tukey HSD (Tabel 8) menunjukkan bahwa terdapat perbedaan yang signifikan antara beberapa pasangan kelompok jumlah pilihan jawaban dalam memengaruhi nilai RMSE.

Secara spesifik, perbandingan antara model dengan 5 pilihan dan 4 pilihan menghasilkan selisih nilai RMSE yang signifikan sebesar 0,0151 (*p-value* 0,0182). Demikian pula, perbandingan antara model dengan 3 pilihan dan 4 pilihan menunjukkan selisih RMSE sebesar 0,0186 yang signifikan (*p-value* 0,0033). Sebaliknya, perbandingan antara model dengan 3 pilihan dan 5 pilihan tidak menunjukkan perbedaan signifikan, dengan selisih sebesar 0,0034.

Untuk mendukung temuan kuantitatif terkait perbandingan efektivitas estimasi kemampuan, visualisasi dalam bentuk *boxplot* disajikan

Tabel 8. Hasil Uji Tukey HSD terhadap Perbedaan Nilai RMSE antar Jumlah Opsi Jawaban

	Perbedaan	Batas Bawah	Batas Atas	*Adjusted P-value*
Lima_Pilihan-Empat_Pilihan	0,01514	0,00228	0,02801	0,01815
Tiga_Pilihan-Empat_Pilihan	0,01858	0,00572	0,03145	0,00335
Tiga_Pilihan-Lima_Pilihan	0,00344	-0,00942	0,01631	0,79002

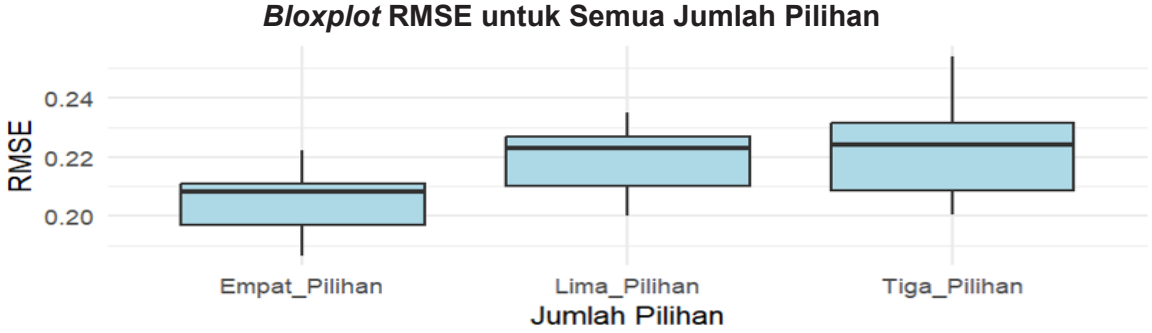
pada Gambar 5. *Boxplot* ini menggambarkan distribusi nilai RMSE dari masing-masing versi instrumen berdasarkan jumlah opsi jawaban, yaitu 3, 4, dan 5 pilihan.

Gambar 5 memperlihatkan visualisasi *boxplot* nilai *Root Mean Square Error* (RMSE) dari ketiga versi instrumen berdasarkan jumlah opsi jawaban, yaitu 3, 4, dan 5 pilihan. Terlihat bahwa distribusi RMSE pada versi Empat_Pilihan berada pada posisi terendah, dengan median yang paling kecil dan rentang antar-kuartil yang lebih sempit dibandingkan versi lainnya. Ini mengindikasikan bahwa model dengan 4 opsi jawaban tidak hanya memiliki estimasi kemampuan yang lebih akurat, tetapi juga lebih konsisten antar replikasi. Sementara itu, versi Tiga_Pilihan menunjukkan variasi yang lebih besar, sedangkan versi Lima_Pilihan berada di tengah dengan sebaran yang lebih stabil dibandingkan versi 3 opsi, namun tidak sebaik versi 4 opsi. Pola ini konsisten dengan hasil analisis statistik sebelumnya yang menunjukkan keunggulan model dengan 4 pilihan dalam hal efisiensi dan stabilitas estimasi.

(GRM). Pertama, seperti yang ditunjukkan pada Tabel 3, menunjukkan bahwa seluruh instrumen memiliki reliabilitas yang diperkirakan termasuk dalam kategori baik (α > 0,7) kecuali Item15 (Widyaningsih et al., 2021; Himelfarb, 2019).

Dalam konteks CTT, seluruh item menunjukkan indeks diskriminasi yang sangat baik dengan nilai korelasi > 0,70, serta indeks *endorsement* berada pada kategori sedang hingga tinggi. Hasil ini mengindikasikan bahwa instrumen mampu membedakan responden berdasarkan tingkat pemahaman mereka terhadap atribut kualitas beras premium. Namun, keterbatasan CTT terletak pada ketergantungan parameter item terhadap sampel yang digunakan (Yuan et al., 2021), serta ketidakmampuannya memberikan estimasi individual dari kesalahan pengukuran.

Untuk mengatasi keterbatasan tersebut, analisis dilanjutkan dengan GRM yang merupakan salah satu model dalam Teori Respons Butir (IRT) yang relevan untuk data ordinal (Samejima, 1969; Embretson & Reise,



Gambar 5. Perbandingan Distribusi RMSE Berdasarkan Jumlah Opsi Jawaban (3, 4, dan 5 Pilihan)

4.2 Pembahasan

Penelitian ini mengevaluasi properti psikometrik dari instrumen persepsi kualitas beras premium dengan pendekatan teori tes klasik (CTT) dan *Graded Response Model* 2000). GRM memberikan estimasi parameter diskriminasi (a) dan ambang batas kategori (b) yang menunjukkan kualitas masing-masing item secara lebih mendalam. Sebagian besar item menunjukkan kemampuan diskriminatif yang

baik, dengan a ≥ 1,0, meski di beberapa item kemperilakuan ini tidak tercapai, seperti pada Item15.

Validitas struktural, instrumen juga memenuhi tiga asumsi dasar IRT yaitu unidimensionalitas, independensi lokal, dan monotonisitas. Rasio *eigenvalue* antara faktor pertama dan kedua sebesar 7,2 mengindikasikan struktur unidimensional yang kuat (Reise et al., 2007). Nilai Q3 residual seluruh item berada dibawah 0,30 (Sijtsma & Molenaar, 2002) yang menandakan tidak ada pelanggaran independensi lokal (Chen & Thissen, 1997). Selain itu, analisis Mokken menunjukkan semua item memenuhi syarat monotonisitas Loevinger H ≥ 0,30. Dari sisi item *evaluation parameter* terhadap diskriminasi, ambang batas serta kurva Karakteristik Item dilakukan evaluasi menyeluruh terhadap 15 item instrumen preferensi beras menggunakan GRM. GRM dipilih karena kemampuannya menangani data ordinal dengan penggunaan skala likert, serta mendestimasi parameter item secara independen dari karakteristik sampel. Letak ambang batas antar kategori serta tingkat akurasi item dalam membedakan respons diperoleh melalui pendekatan GRM.

Berdasarkan hasil analisis parameter diskriminasi, ambang batas (b1–b4), serta kurva Karakteristik Item (*Item Characteristic Curve*/ICC), dilakukan evaluasi menyeluruh terhadap 15 item dalam instrumen preferensi beras menggunakan pendekatan *Graded Response Model* (GRM). Model ini dipilih karena kemampuannya menangani data ordinal seperti skala Likert dan mengestimasi parameter item secara independen dari karakteristik sampel. GRM juga memungkinkan identifikasi ambang batas antar kategori dan tingkat akurasi item dalam membedakan tingkat kemampuan responden (Samejima, 1969; Hambleton & Swaminathan, 1985).

Secara umum, 13 dari 15 item memiliki nilai parameter diskriminasi di atas 1,0 yang menunjukkan bahwa sebagian besar item memiliki daya beda yang baik dalam membedakan preferensi atau kemampuan responden (Rubio et al., 2007). Item seperti Item3 (a = 1,862), Item4 (a = 1,946), Item10 (a = 2,392), dan Item13 (a = 2,209) memperlihatkan performa psikometrik yang optimal, dengan ambang kategori yang berurutan dan kurva ICC yang simetris, menandakan bahwa responden dengan tingkat kemampuan berbeda memberikan respons yang konsisten dan sesuai ekspektasi.

Namun ditemukan dua item yang memerlukan perhatian khusus, yaitu Item5 (a = 2.610), meskipun memiliki diskriminasi tinggi menunjukkan pola ambang yang tidak beraturan dan kurva ICC yang tumpang tindih. Hal tersebut mengindikasikan potensi masalah pada label kategori atau tata bahasa item5 yang membingungkan, bukan daya beda. Item15 (a = 0,672) memiliki nilai diskriminasi rendah dan kurva ICC mendatar, dengan dominasi kategori P1 (terendah), yang menunjukkan bahwa item ini kurang sensitif dalam menangkap variasi preferensi responden (Hambelton et al., 2011). Selain itu, beberapa item lain seperti Item2 dan Item11 memiliki ambang batas yang sebagian besar berurutan, namun masih perlu perbaikan terutama pada kategori tengah untuk meningkatkan presisi informasi. Sementara itu, Item6 dan Item9 memperlihatkan dominasi pada satu kategori, yang dapat menurunkan efektivitas dalam mendeteksi perbedaan antar responden (Reise et al., 2007).

Lebih lanjut, Item14 menunjukkan nilai lokasi sebesar –1,209, yang merupakan nilai terendah di antara semua butir dalam instrumen. Nilai ini menunjukkan bahwa responden dengan kemampuan laten (θ) yang sangat rendah pun cenderung memberikan respons pada kategori tinggi untuk item ini. Fenomena ini dikenal sebagai *over-endorsement*, yaitu kondisi di mana sebuah item terlalu mudah disetujui oleh sebagian besar responden, tanpa memperhatikan variasi tingkat kemampuan mereka. Dalam konteks *Graded Response Model* (GRM), hal ini menjadi indikasi bahwa item tersebut kurang efektif dalam membedakan responden berdasarkan kemampuan laten, karena respons yang diberikan tidak mencerminkan variasi kemampuan secara akurat (Samejima, 1969; Embretson & Reise, 2000). *Over-endorsement* sering kali disebabkan oleh redaksi item yang terlalu umum, bersifat normatif, atau mengandung bias sosial yang mendorong kesepakatan universal.

Beberapa item juga menunjukkan tumpang tindih pada ambang kategori, yang berpotensi menurunkan akurasi estimasi kemampuan laten responden (Hair et al., 2017). Meskipun demikian, kurva informasi tes menunjukkan bahwa puncak informasi berada pada $\theta \approx 0$ dengan rentang efektif antara -2 hingga +2, yang berarti bahwa instrumen ini paling akurat dalam mengukur responden dengan kemampuan sedang.

Secara keseluruhan, temuan ini menegaskan bahwa GRM merupakan pendekatan yang efektif dalam mengevaluasi kualitas psikometrik skala preferensi, karena mampu memberikan informasi rinci tentang fungsi tiap item dan kategori respons. Kombinasi GRM dan teori tes klasik (CTT) memungkinkan analisis yang lebih komprehensif dan mendalam terhadap kualitas item (Bellamkonda & Pattusamy, 2022; Yuan et al., 2021; Rubio et al., 2007). Untuk meningkatkan akurasi dan efisiensi instrumen, direkomendasikan agar dilakukan revisi redaksi atau rekategori terhadap Item5, Item15 dan item 14, serta pengujian ulang untuk mengevaluasi perbaikan tersebut secara empiris.

V. KESIMPULAN

Penelitian ini berhasil mengembangkan dan mengevaluasi instrumen persepsi kualitas beras premium dengan pendekatan *Graded Response Model* (GRM) sebagai model yang tepat untuk mengukur preferensi konsumen terhadap kualitas beras premium, karena mampu memetakan variasi respons secara akurat di sepanjang spektrum kemampuan laten. Instrumen menunjukkan performa psikometrik yang baik, ditandai dengan terpenuhinya asumsi unidimensionalitas, independensi lokal, dan monotonisitas. Mayoritas item memiliki daya diskriminasi tinggi (a > 1,0), ambang respons yang logis, serta fungsi kategori yang berfungsi optimal. Beberapa item yang menunjukkan kelemahan dan perlu mendapat perhatian untuk perbaikan. Item 14 (Beras Premium tidak mengandung bau apek atau bau kimia saat belum dimasak) memiliki daya diskriminasi yang relatif rendah, sehingga kurang efektif dalam membedakan responden dengan tingkat persepsi laten yang berbeda terkait kualitas beras. Sementara itu, Item 15 (Nasi dari Beras Premium memberikan rasa enak meskipun dimasak tanpa lauk) memperlihatkan kelemahan pada aspek ambang batas kategori (*threshold*), di mana beberapa kategori respons tumpang tindih sehingga tidak berfungsi secara optimal. Adapun Item 5 (Beras Premium memiliki butiran beras yang utuh dengan sedikit butir patah) meskipun substantifnya relevan dengan dimensi kualitas, namun memiliki nilai diskriminasi yang lebih rendah dibandingkan mayoritas item lainnya, sehingga kontribusinya terhadap pengukuran dimensi laten kualitas beras premium menjadi kurang tajam.

Analisis efektivitas skala respons menunjukkan bahwa versi dengan 4 opsi menghasilkan nilai RMSE paling rendah dan stabil daripada 3 dan 5 opsi, sehingga direkomendasikan sebagai format terbaik dalam pengukuran ini.

DAFTAR PUSTAKA

Baker, F. B. (2001). *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). CRC Press. https://doi.org/10.1201/9781482276725

Bellamkonda, N., & Pattusamy, M. (2022). Validation of Fear of COVID-19 Scale in India: Classical test theory and item response theory approach. *International Journal of Mental Health and Addiction, 20*(4), 2400–2407. https://doi.org/10.1007/s11469-021-00521-2

Brentari, E., & Golia, S. (2007). Unidimensionality in The Rasch Model: How to Detect and Interpret. *Statistica*, *LXVII*(3), 253–261.

Cai, L., & Monroe, S. (2014). *A new statistic for evaluating item response theory models for ordinal data* (CRESST Report 839). National Center for Research on Evaluation, Standards, and Student Testing (CRESST). https://eric.ed.gov/?id=ED555726

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chamhuri, N., & Batt, P. J. (2013). Exploring the Factors Influencing Consumers' Choice of Retail Store When Purchasing Fresh Meat in Malaysia. *International Food and Agribusiness Management Review*, *16*.

Chen, W.-H., & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behavioral*

Statistics Fall, 22(3), http://jebs.aera.net

Chong, J., Mokshein, S. E., & Mustapha, R. (2022). Applying the Rasch Rating Scale Model (RSM) to Investigate The Rating Scales Function in Survey Research Instrument. *Cakrawala Pendidikan*, *41*(1), 97–111. https://doi.org/10.21831/cp.v41i1.39130

De Ayala, R. J., & Hertzog, M. A. (1991). The Assessment of Dimensionality for Use in Item Response Theory. *Multivariate Behavioral Research*, *26*(4), 765–792. https://doi.org/10.1207/s15327906mbr2604_9

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2020). Sensitivity and specificity of information criteria. *Briefings in bioinformatics*, *21*(2), 553–565. https://doi.org/10.1093/bib/bbz016

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Fokkema, M., & Greiff, S. (2017). How performing PCA and CFA on the same data equals trouble: Overfitting in the assessment of internal structure validity. *European Journal of Psychological Assessment, 33*(6), 399–402. https://doi.org/10.1027/1015-5759/a000460.

Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate Data Analysis* (7th ed.). Prentice Hall.

Hambelton, R. K., van der Linden, W. J., & Wells, C. S. (2011). IRT Models for The Analysis of Polytomously Scored Data: Brief and Selected History of Model Building Advances. In M. L. Nering & R. Ostini (Eds.), *Handbook of Polytomous Item Response Theory Models* (pp. 21–42). Routledge.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Springer Science+Business Media LLC.

Meijer, R. R., & Tendeiro, J. N. (2018). Unidimensional item response theory. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 413–433). Wiley. https://doi.org/10.1002/9781118489772.ch15

Himelfarb, I. (2019). A Primer on Standardized Testing: History, Measurement, Classical Test Theory, Item Response Theory, and Equating. *Journal of Chiropractic Education*, *33*(2), 151–163. https://doi.org/10.7899/JCE-18-22

Junker, B. W. (1996). *Monotonicity and Conditional Independence in Models for Student Assessment and Attitude Measurement*. Johnson.

Linacre, J. M. (2002). Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement*, 3(1), 85–106.

Lord, F. M. (2008). *Applications of Item Response Theory to Practical Testing problems*. Routledge.

Matteucci, M., & Stracqualursi, L. (2006). Student Assesment Via Graded Response Model. *Statistica*, *LXVI*(4).

Mokken, R. J. (1997). Nonparametric Models for Dichotomous Responses. In W. J. Van der Linden & R. K. Hambelton (Eds.), *Handbook of Modern Item Response Theory*. Springer New York. https://doi.org/10.1007/978-1-4757-2691-6

Natasya, D. V., Setiawan, B., & Eka Hardana, A. (2024). Prefensi Konsumen terhadap Atribut Beras Premium. *VIGOR: Jurnal Ilmu Pertanian Tropika Dan Subtropika*, *9*(1), 33–43.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207–230. https://doi.org/10.3102/10769986004003207

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The Role of The Bifactor Model in Resolving Dimensionality Issues in Health Outcomes Measures. *Quality of Life Research*, 16(SUPPL 1), https://doi.org/10.1007/s11136-007-9183-7

Revelle, W. (2023). *Psych: Procedures for psychological, psychometric, and personality research*. R package version 2.3.9. https://CRAN.R-project.org/package=psych

Rubio, V. J., Aguado, D., Hontangas, P. M., & Hernández, J. M. (2007). Psychometric Properties of An emotional adjustment measure: An application of the Graded Response Model. *European Journal of Psychological Assessment*, *23*(1), 39–46. https://doi.org/10.1027/1015-5759.23.1.39

Samejima, F. (1969). *Estimation of Latent Ability Using A Response Pattern of Graded Scores*. The William Byrd Press.

Sen, S. & Bradshaw, L. (2017). Comparison of Relative Fit Indices for Diagnostic Model Selection. *Applied Psychological Measurement*. 41,014662161769552. https://doi.org/10.1177/0146621617695521.

Sijtsma, K., & Molenaar, I. M. (2002). Introduction to Nonparametric Item Response Theory. *Quality of Life Research*, 4.

Van Der Ark, L. A. (2012). New Developments in Mokken Scale Analysis in R. *Journal of Statistical Software*, *48*(5), 1–27. http://www.jstatsoft.org/

Van der Ark, L. A. (2007). *Mokken scale analysis in R. Journal of Statistical Software*, 20(11), 1–19. https://doi.org/10.18637/jss.v020.i11

Widyaningsih, S. W., Yusuf, I., Prasetyo, Z. K., & Istiyono, E. (2021). The Development of The Hots Test of Physics Based on Modern Test Theory: Question Modeling Through E-learning of Moodle LMS. *International Journal of Instruction*, *14*(4), 51–68. https://doi.org/10.29333/iji.2021.1444a

Yen, W. M. (1993). Scaling Performance Assessments: Strategies for Managing Local Item Dependence. *Journal of Educational Measurement Fall*, *30*(3), 187–213.

Yuan, T., Honglei, Z., Xiao, X., Ge, W., & Xianting, C. (2021). Measuring Perceived Risk in Sharing Economy: A Classical Test Theory and Item Response Theory approach. *International Journal of Hospitality Management*, *96*. https://doi.org/10.1016/j.ijhm.2021.102980

Zhou, H., Xia, D., & He, Y. (2020). Rice Grain Quality—Traditional Traits for High Quality Rice and Health-Plus Substances. *Molecular Breeding*, *40*(1). https://doi.org/10.1007/s11032-019-1080-6.

BIODATA PENULIS:

Eny Cahyaningsih, dilahirkan di Klaten, 25 November 1977. Penulis menyelesaikan S1 Program Studi Statistik di Universitas Gadjahmada tahun 2001, S2 Manajemen Bisnis Institut Pertanian Bogor tahun 2013 .

Nurul Qomariyah Ahmad dilahirkan di Jakarta, 10 Juli 1982. Penulis menyelesaikan pendidikan S1 jurusan Teknologi Pangan Institut Pertanian Bogor tahun 2005 dan S2 jurusan Penelitian dan Evaluasi Pendidikan Universitas Negeri Jakarta tahun 2013.

Wardani Rahayu, dilahirkan pada 6 Maret 1964. Penulis menyelesaikan S1 Program Studi Pendidikan Matematika di Institut Keguruan dan Ilmu Pendidikan (IKIP) Jakarta tahun 1988, S2 Matematika di Universitas Negeri Jakarta tahun 1994, dan S3 Program Studi Penelitian dan Evaluasi Pendidikan (PEP) di Universitas Negeri Jakarta tahun 2008. Dan dikukuhkan sebagai Guru Besar dalam bidang evaluasi pendidikan matematika di FMIPA Universitas Negeri Jakarta pada tahun 2021.

Achmad Ridwan, dilahirkan di Jakarta pada 17 Agustus 1963. Penulis menyelesaikan S1 Pendidikan Kimia di Institut Keguruan dan Ilmu Pendidikan (IKIP) Jakarta tahun 1987, S2 Kimia Fisika di Universitas Gadjah Mada tahun 1993, serta meraih dua gelar doktor, yaitu Dr. bidang Pengelolaan Sumber Daya Alam dan Lingkungan di Institut Pertanian Bogor (2007) dan Dr. bidang Pendidikan (*Education Research and Assessment*) di Universitas Negeri Jakarta (2016).