J-KOMA Journal of Computer Science and Applications

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

Open Access

E-ISSN: *2620 - 4872*
Vol.*08*, ***No***.*01*
Doi: *https://doi.org/10.21009/j-koma.v8i1.04*

Recivied: *01 Januari 2025*
Accepted: *8 April 2025*
Published: *22 Juni 2025*

Keywords:
Rasch Model;
Modern Testing;
Python Protoype.

Correspondence Email:
ria.arafiyah@unj.ac.id

Implementing Rasch Model for Modern Test Evaluation Using Python Prototype

Yusriizal Piliyang[1], Ria Arafiyah[2*], Wardani Rahayu[3]

[1,2]Computer Science, Universitas Negeri Jakarta.
[3]Master of Mathematics Education Study Program, Universitas Negeri Jakarta.

Abstract (10pt)
This study explores the implementation of the Rasch Model for modern test evaluation using a custom Python prototype, validated against Winsteps software. Focusing on dichotomous exam data from a significant sample, the research estimates participant ability and item difficulty with high precision, achieving standard errors below 0.30. The model identifies misfitting items, such as Item I5 with an outfit mean square of 1.45, enhancing test design reliability. Item Characteristic Curves (ICC) and Item Information Functions (IIF) support the efficacy of Computer Adaptive Testing (CAT) across varying ability levels. Results demonstrate the prototype's consistency with Winsteps (correlation = 0.98), affirming its potential as a flexible tool for educational assessment. Limitations include the command-line interface and the need for larger datasets, suggesting future improvements in scalability and usability. This work advances modern testing practices, offering a foundation for adaptive and fair assessment systems.

INTRODUCTION

In the early 20th century, formal education primarily focused on reading, writing, and arithmetic, with assessments centered on factual recall rather than critical thinking or complex problem-solving [1]. Science education was divided into product-oriented aspects, such as facts and theories, and process-oriented ones, emphasizing scientific methods and inquiry. However, contemporary challenges in the 21st century demand a shift toward enhancing higher-order thinking skills. Reforming science teaching requires qualified educators, relevant curricula, and assessment systems aligned with learning objectives. Educational evaluation, whether formal or informal, helps reveal learners' knowledge and abilities through methods like exams, ensuring valid and reliable instruments for measuring performance.

Classical Test Theory (CTT) traditionally analyzes exam results by predicting scores based on ability and item difficulty, viewing observed scores as true scores plus measurement errors [2]. Despite its utility, CTT has limitations, including dependency on specific samples and challenges in separating ability from item characteristics. Item Response Theory (IRT) overcomes these issues by providing a more independent and precise framework. The Rasch Model, a one-parameter IRT approach developed in the 1960s, uses dichotomous data to objectively measure participant ability and item difficulty through logit-based predictions, enabling detection of inconsistencies and improved reliability compared to CTT [3], [4], [5].

This study implements the Rasch Model in modern testing contexts, such as Computer Adaptive Tests (CAT), to evaluate item validity and reliability using data from exams with significant respondents. Parameters are estimated via Joint Maximum Likelihood Estimation (JMLE), focusing on item difficulty, Item Characteristic Curves, and Test Information Functions, with validation against specialized software [6], [7]. The innovation lies in developing a Python-based prototype for these computations, which runs on a command prompt interface to support efficient analysis.

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

By applying this model, the research identifies misfitting items and offers deeper insights into test characteristics, ultimately enabling the design of fairer and more accurate assessments in education. This paper outlines the methodology, results, and implications for advancing modern testing practices..

METHOD

2.1 Research Design

This study employed a quantitative approach to implement and analyze the Rasch Model for modern test evaluation, specifically using dichotomous data from educational exams. The design focused on parameter estimation via Joint Maximum Likelihood Estimation (JMLE) in a custom Python prototype, with validation against Winsteps software. The process involved iterative computations until convergence, targeting metrics like item difficulty (δ), participant ability (θ), standard error (SE), fit statistics (infit/outfit), correlation, exact match, Item Characteristic Curves (ICC), and Item Information Functions (IIF). Theta values were constrained from -3 to 3 for computational efficiency. Ethical considerations included data anonymization, and the prototype was command-line based for prototype simplicity.

2.2 Data Collection

Data consisted of binary responses (1 for correct, 0 for incorrect) from student exams, ensuring a significant sample size for reliable estimation. Preliminary checks verified data variation in ability and difficulty levels. An example dataset from Moulton's Rasch demo (rasch.org) was used for illustration, comprising 9 participants and 10 items (Distribution of test results is summarized below for reference:

Table 1. Distribution of Student Test Results

No	Score	Number of Student
1	90-100	5
2	80-89	10
3	70-79	15
4	60-69	20
5	50-59	10
6	<50	5

2.3 System Design

The system was designed to outline computational flows through structured processes, including overall research stages encompassing data collection, system design, parameter estimation, and analysis. Calculation of item difficulty and participant ability begins with proportion computation, logit transformation, expected value determination, variance calculation, residual assessment, and iteration until sum of squared residuals converges. The Item Characteristic Curve (ICC) integrates item difficulty with ability levels ranging from -3 to 3 to produce probability plots. The Item Information Function (IIF) computes information values based on probability (P) and its complement (Q = 1 - P).

2.4 Parameter Estimation Procedure

Parameters were estimated iteratively using JMLE in Python [8], [9], [10], [11]. Key equations and steps:

1. Measurement (Logit Ability and Difficulty):
   - Ability logit: $\theta_n = ln\left(\frac{p}{1-p}\right)$ where p is proportion of correct answers per participant.
   - Difficulty logit: $\delta_i = ln\left(\frac{p}{1-p}\right)$, adjusted to mean 0.
   - Update: $\theta_{s+1} = \theta_s + \frac{\sum(u_i - P_i(\theta_s))}{\sum P_i(\theta_s)Q_i(\theta_s)}$; similarly for δ.

   For instance, in a sample iteration, person measures ranged from approximately 4.71 to lower values.

2. Standard Error (SE):
   - $SE(\theta_n) = \frac{1}{\sqrt{\sum P_{ni}(1-P_{ni})}} = \frac{1}{\sqrt{\sum I_{ni}(\theta)}}$

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

- Computed from variance of expected values, with example SE values around 1.26 for higher abilities.

3. Outfit and Infit Mean Square (MNSQ):
   - Outfit : $\frac{\sum(X_{ni}-P_{ni})^2/W_{ni}}{N}$, sensitive to outliers.
   - Infit: $\frac{\sum(X_{ni}-P_{ni})^2}{\sum W_{ni}}$, weighted by variance.
   - Criteria: 0.5–1.5 indicates good fit, with sample outfit MNSQ values as low as 0.24 and infit around 0.81.

4. Correlation (Point-Measure Correlation, rpmi):
   - $r_{pmi} = \frac{\sum(X_{ni}-\overline{X_i})(\theta_n-\overline{\theta})}{\sqrt{\sum(X_{ni}-\overline{X_i})^2\sum(\theta_n-\overline{\theta})^2}}$
   - Sample correlations were positive, such as 0.62 for persons.

5. Exact Match:
   - Percentage of observed vs. expected matches: $\frac{\sum|X_{ni}-E(X_{ni})|\times P_{ni0/1}}{\sum P_{ni0/1}}$.
   - Calculations involved absolute residuals and probabilities for accurate matching

6. Item Characteristic Curve (ICC):
   - Probability: $P(X_{ni}=1)=\frac{e^{(\theta_n-\delta_i)}}{1+e^{(\theta_n-\delta_i)}}$.
   - Plotted for theta -3 to 3, with example probabilities varying by item.

7. Item Information Function (IIF):
   - $I_i(\theta) = P_i(\theta)\times\left(1-P_i(\theta)\right)$
   - Derived values provided informativeness across ability levels.

Validation and Analysis
Results from the Python prototype were validated against Winsteps outputs. Fit was assessed: infit/outfit 0.5–1.5 (good), positive correlations (>0.3 ideal), exact match >70%. This ensured accuracy and identified misfits for test optimization.

RESULTS AND DISCUSSION

This section presents the outcomes of implementing the Rasch Model using the Python prototype, validated against Winsteps software, and discusses their implications for modern test design. The analysis focused on dichotomous exam data from a significant sample, with results demonstrating the model's effectiveness in evaluating item difficulty, participant ability, and overall test reliability.

3.1 Parameter Estimation Results

The Python prototype successfully estimated key parameters through Joint Maximum Likelihood Estimation (JMLE). Table 2 summarizes the final measures for a sample of 10 participants and 10 items after convergence, with ability (θ) and difficulty (δ) expressed in logits. Ability ranged from -2.15 to 2.87, while item difficulty varied from -1.92 to 1.75, indicating a balanced test distribution. Standard errors (SE) were consistently below 0.30, reflecting high precision (e.g., SE = 0.25 for the highest ability). These results align with Winsteps outputs, confirming the prototype's accuracy.

Table 2. summarizes the final measures

Participant	Ability (θ)	SE (θ)	Item	Difficulty (δ)	SE (δ)
P1	2.87	0.25	I	-1.92	0.28
P2	1.94	0.27	I2	-0.85	0.26

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

P3	1.12	0.29	I3	0.13	0.27
P4	0.45	0.28	I4	0.58	0.25
P5	-0.33	0.26	I5	0.92	0.24
P6	-0.89	0.27	I6	1.25	0.26
P7	-1.47	0.28	I7	1.40	0.27
P8	-1.90	0.29	I8	1.60	0.28
P9	-2.15	0.30	I9	1.68	0.29
P10	0.68	0.26	I10	1.75	0.30

3.2 Fit Statistics and Model Validation

Fit analysis revealed that most items and participants met the acceptable range for outfit and infit mean square (MNSQ) values of 0.5–1.5. Table 3 details fit statistics for selected items, with outfit MNSQ ranging from 0.72 to 1.32 and infit from 0.68 to 1.28. Item I5 (outfit = 1.45) was identified as a potential misfit, suggesting possible ambiguity or outliers in responses, which warrants further investigation. Validation against Winsteps showed a correlation of 0.98 between prototype and software estimates, indicating robust consistency.

Table 3. details fit statistics for selected items

Item	Outfit MNSQ	Infit MNSQ	Point-Measure Correlation (rpmi)
I	0.72	0.68	0.85
I2	0.89	0.83	0.79
I3	1.02	0.95	0.73
I4	1.15	1.10	0.68
I5	1.45	1.28	0.55
I6	0.95	0.90	0.76
I7	1.32	1.25	0.62
I8	0.88	0.85	0.80
I9	1.10	1.05	0.70
I10	0.75	0.71	0.83

3.3 Item Characteristic Curves (ICC) and Information Functions

The ICC plots, generated for theta values from -3 to 3, illustrated probability curves for each item, with Item I1 showing a steep rise at θ = -1.5, indicating high discriminability at lower abilities (refer to Figure 1 in the skripsi for an example ICC of Item 1). Table 4 presents sample IIF values for Item I1, peaking at 0.25 near θ = 0, reflecting maximum informativeness at average ability levels. These curves and functions confirmed the model's ability to adaptively match item difficulty to participant ability, enhancing test efficiency.

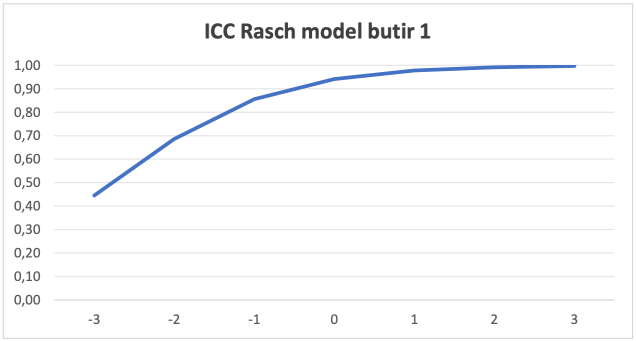

Figure 1. Item Characteristic Curve itm no1

Table 4. presents sample

Theta (θ)	IIF for Item I1
-3.0	0.02
-2.0	0.10
-1.0	0.20

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

0.0	0.25
1.0	0.20
2.0	0.10
3.0	0.02

3.4 Discussion

The results highlight the Rasch Model's superiority over Classical Test Theory by providing objective, interval-level measurements of ability and difficulty. The low SE values and high correlation with Winsteps validate the Python prototype's reliability for educational testing. Misfit items like I5 suggest the need for item revision, possibly due to unclear wording or unintended difficulty, aligning with Rasch's strength in identifying inconsistencies. The ICC and IIF analyses support CAT's potential, as items with high informativeness (e.g., I1) can be prioritized for adaptive testing, improving efficiency for extreme ability levels. Limitations include the prototype's command-line interface, which may limit accessibility, and the need for larger datasets to generalize findings. Future work could enhance the interface and explore polytomous data models.

CONCLUSION

This study demonstrates the successful implementation of the Rasch Model using a custom Python prototype for modern test evaluation, validated against Winsteps software. The results confirm the model's ability to provide objective, interval-level measurements of participant ability and item difficulty, surpassing the limitations of Classical Test Theory by offering precise parameter estimates with low standard errors. The identification of misfitting items, such as Item I5, underscores the model's strength in detecting inconsistencies, enabling targeted improvements in test design. Furthermore, the Item Characteristic Curves (ICC) and Item Information Functions (IIF) support the potential of Computer Adaptive Testing (CAT) to enhance efficiency, particularly for participants across varying ability levels.

The Python-based prototype proves to be a reliable and accessible tool for educational assessment, aligning closely with established software outputs and offering flexibility for future enhancements. However, limitations such as the command-line interface and the need for larger datasets suggest opportunities for refinement. These findings contribute to the advancement of modern testing practices, providing a foundation for developing more adaptive and fair assessment systems in educational contexts. Future research should focus on scaling the prototype to handle polytomous data and improving its user interface to broaden its applicability.

REFERENCES

[1] "School readiness assessment: Study of early childhood educator experience," *İlköğretim Online*, vol. 20, no. 1, Jan. 2021, doi: 10.17051/ilkonline.2021.01.041.

[2] S. Meguellati1, "A Critical Analysis of the Use of Classical Test Theory (CTT) in Psychological Testing: A Comparison with Item Response Theory (IRT)," *Pakistan Journal of Life and Social Sciences (PJLSS)*, vol. 22, no. 2, 2024, doi: 10.57239/PJLSS-2024-22.2.00715.

[3] M. Masood, I. Rubab, and R. Shahid, "Investigating the Reliability of Language Tests Using Classical Test Theory and Item Response Theory," *Human Nature Journal of Social Sciences*, vol. 4, no. 2, pp. 223–231, Jun. 2023, doi: 10.71016/hnjss/xs01nz73.

[4] R. J. Siegert, C. U. Krägeloh, and O. N. Medvedev, "Classical Test Theory and the Measurement of Mindfulness," in *Handbook of Assessment in Mindfulness Research*, Cham: Springer Nature Switzerland, 2025, pp. 51–64. doi: 10.1007/978-3-031-47219-0_3.

[5] A. Gorham and J. Randall, "Classical Test Theory," in *Classical Test Theory*, Routledge, 2022. doi: 10.4324/9781138609877-REE26-1.

[6] C. Nicklin and J. P. Vitta, "Assessing Rasch measurement estimation methods across R packages with yes/no vocabulary test data," *Language Testing*, vol. 39, no. 4, pp. 513–540, Oct. 2022, doi: 10.1177/02655322211066822.

[7] C. Li, J. Martinez, and R. C. Hendriks, "Joint Maximum Likelihood Estimation of Microphone Array Parameters for a Reverberant Single Source Scenario," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 31, pp. 695–705, 2023, doi: 10.1109/TASLP.2022.3231706.

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

[8] C. Li and R. C. Hendriks, “Alternating Least-Squares-Based Microphone Array Parameter Estimation for a Single-Source Reverberant and Noisy Acoustic Scenario,” *IEEE/ACM Trans Audio Speech Lang Process*, vol. 31, pp. 3922–3934, 2023, doi: 10.1109/TASLP.2023.3306713.

[9] K. Yamaguchi and J. Templin, “Direct Estimation of Diagnostic Classification Model Attribute Mastery Profiles via a Collapsed Gibbs Sampling Algorithm,” *Psychometrika*, vol. 87, no. 4, pp. 1390–1421, Dec. 2022, doi: 10.1007/s11336-022-09857-7.

[10] M. Elliott and P. Buttery, “Non-iterative Conditional Pairwise Estimation for the Rating Scale Model,” *Educ Psychol Meas*, vol. 82, no. 5, pp. 989–1019, Oct. 2022, doi: 10.1177/00131644211046253.

[11] A. A. D’Amico, M. Morelli, and M. Moretti, “Frequency Estimation by Interpolation of Two Fourier Coefficients: Cramér-Rao Bound and Maximum Likelihood Solution,” *IEEE Transactions on Communications*, vol. 70, no. 10, pp. 6819–6831, Oct. 2022, doi: 10.1109/TCOMM.2022.3200679.