JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)
DOI: http://dx.doi.org/10.15408/jp3i.v15i2.51647
http://journal.uinjkt.ac.id/index.php/jp3i

# Ensuring Fairness in Educational Book Evaluation: A Many-Facet Rasch Analysis of Rater Severity and Item Functioning

**Suprananto[1], Bambang Suryadi[2], Jenny Lewis[3], Valendra Granitha Shandika Puri[2]**

Faculty of Teacher Training and Education, Universitas Singaperbangsa Karawang, Indonesia[1]
Faculty of Psychology, UIN Syarif Hidayatullah Jakarta, Indonesia[2]
School of Education and Creative Arts, University of Southern Queensland, Australia[3]

supranantо@fkip.unsika.ac.id

## Abstract

Maintaining the quality of educational textbooks requires valid, reliable, and transparent evaluation. However, multi-rater assessments may be affected by differences in judgment and rater severity. This study examined a textbook evaluation system using the Many-Facet Rasch Model (MFRM). Six evaluators assessed three books using a 12-item instrument. The findings revealed substantial variation in rater severity, including one unusually severe rater, indicating a need to improve consistency in evaluators' judgments. Most items functioned adequately, but two showed misfit and required revision or removal. The analysis also distinguished differences in the estimated quality of the three books while accounting for rater effects. These findings demonstrate the practical value of MFRM for strengthening the fairness and credibility of textbook evaluation. Institutions responsible for textbook assessment should regularly review evaluation instruments, clarify rating criteria, and provide training and calibration to promote consistent judgments. They should also investigate unusually severe or inconsistent rating patterns and revise or remove items that do not function as intended. Incorporating MFRM into evaluation procedures can help institutions identify weaknesses in instruments and rating practices, support evidence-based textbook selection, and improve quality assurance. These steps can help policymakers establish more transparent and rigorous processes for evaluating educational materials.

**Keywords**: Assessment validity, book evaluation, Many-Facet Rasch Model, rater severity.

## *Abstrak*

*Menjaga kualitas buku teks pendidikan memerlukan evaluasi yang valid, reliabel, dan transparan. Namun, perbedaan pertimbangan dan tingkat keparahan penilai dapat memengaruhi hasil evaluasi. Penelitian ini mengkaji sistem evaluasi buku teks menggunakan Many-Facet Rasch Model (MFRM). Enam penilai mengevaluasi tiga buku menggunakan instrumen dengan 12 butir. Hasil menunjukkan variasi tingkat keparahan penilai yang cukup besar, termasuk satu penilai dengan tingkat keparahan yang sangat tinggi. Temuan ini mengindikasikan perlunya meningkatkan konsistensi dalam penilaian. Sebagian besar butir berfungsi dengan baik, tetapi dua butir menunjukkan ketidaksesuaian (misfit) sehingga perlu ditinjau untuk kemungkinan revisi atau penghapusan. Analisis juga mengidentifikasi perbedaan estimasi kualitas ketiga buku dengan memperhitungkan pengaruh penilai. Temuan ini menunjukkan bahwa MFRM dapat memperkuat keadilan dan kredibilitas dalam evaluasi buku teks. Lembaga terkait perlu meninjau instrumen secara berkala, memperjelas kriteria penilaian, serta melaksanakan pelatihan dan kalibrasi bagi para penilai. Pola penilaian yang tidak biasa perlu ditelaah, sementara butir yang tidak berfungsi sebagaimana mestinya perlu dievaluasi lebih lanjut. Integrasi MFRM ke dalam prosedur evaluasi dapat membantu mengidentifikasi kelemahan instrumen dan praktik penilaian, mendukung pemilihan buku berbasis bukti, serta meningkatkan penjaminan mutu. Penerapan langkah-langkah tersebut diharapkan dapat mendukung proses evaluasi bahan pendidikan yang lebih transparan, konsisten, dan akuntabel.*

***Kata kunci:*** *Evaluasi buku pendidikan, ketegasan penilai, Many-Facet Rasch Model, validitas asesmen.*


This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

## Introduction

Evaluating educational materials is essential to maintaining the quality and effectiveness of learning resources in educational settings. Textbooks and instructional books shape students' learning experiences and influence the teaching process; therefore, systematic and reliable evaluation procedures are necessary to ensure that educational materials meet standards of content accuracy, pedagogical relevance, clarity, and curriculum alignment. Previous studies have emphasised that rigorous evaluation systems are important for improving educational quality and supporting evidence-based decision-making in the selection of learning resources (Brookhart, 2013; Kane, 2013). Recent studies further highlight that educational evaluation systems must ensure fairness, transparency, and psychometric quality, particularly in assessments involving expert judgment and rubric-based scoring (Adie et al., 2021; Wind & Jones, 2022). However, evaluation processes involving human raters often introduce measurement error from differences in judgment, scoring standards, and inconsistencies in rating practices (de Klerk et al., 2023).

In Indonesia, Law Number 3 of 2017 on the Book System and Government Regulation Number 75 of 2019 on its implementation regulate the evaluation of educational books. These regulations classify educational books as textbooks or non-textbooks and require all educational books to undergo formal evaluation before approval for instructional use. The evaluation process is conducted according to the principles of professionalism, objectivity, and accountability, and it covers four major aspects: content, presentation, design, and graphics. Through this regulatory framework, the government seeks to ensure that educational books meet academic and pedagogical standards while remaining aligned with national values and educational objectives (Republic of Indonesia, 2017, 2019).

Historically, educational book assessment in the Indonesian context, particularly within the government evaluation framework, has relied primarily on qualitative discussions and expert judgments. These approaches have provided valuable insights into the appropriateness and quality of educational books; however, they have generally focused on descriptive evaluation rather than further psychometric analysis of the measurement process. Consequently, research has paid limited attention to aspects such as rater consistency, measurement reliability, item functioning, and interactions among evaluators, criteria, and assessed books. The growing demand for more objective, evidence-based evaluation systems highlights the need for analytical approaches that complement qualitative judgments and strengthen the validity of educational book assessment.

Many educational evaluation systems employ multiple raters to improve fairness and credibility by incorporating different perspectives and expertise into the assessment process. Nevertheless, multi-rater evaluation introduces methodological challenges because raters may differ in how they interpret criteria, scoring standards, and judgment tendencies (Engelhard, 2013; Linacre, 1989). Variations in rater severity and leniency may produce score inconsistencies that do not necessarily reflect the true quality of the evaluated objects (Myford & Wolfe, 2003). Research in educational measurement has shown that multi-rater assessments frequently encounter problems related to rater severity, halo effects, and scoring inconsistency, which may threaten the validity of evaluation outcomes if not analysed using appropriate psychometric models (Jin & Chiu, 2021; Wind, 2023). In addition, rater cognition, professional background, and interpretation of scoring rubrics may substantially influence evaluation consistency, particularly in subjective assessment contexts (Tai et al., 2022; Trace et al., 2020). These issues indicate that evaluation quality depends not only on the assessment instrument but also on raters' behaviour and consistency.

To address these challenges, advanced psychometric approaches have been developed to analyse rating data more systematically. One widely used approach is the Many-Facet Rasch Model (MFRM), which extends the Rasch measurement framework to simultaneously analyse multiple facets involved in assessment processes, such as raters, items, and evaluated objects (Linacre, 1989). The Rasch model provides a probabilistic measurement framework based on invariant measurement principles (Rasch, 1960; Bond & Fox, 2015), while the Many-Facet extension enables the calibration of different components of an evaluation system on a common measurement scale. Through this approach, researchers can identify misfitting items, detect differences in rater severity, and estimate the relative quality of evaluated objects (Myford & Wolfe, 2003; Wright & Masters, 1982). Recent studies have



This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-

demonstrated the effectiveness of MFRM in improving scoring reliability and identifying latent sources of measurement error in writing assessment, peer evaluation, teacher assessment, and competency-based evaluation systems (Liu & Jin, 2022; Wind & Guo, 2021).

Although previous studies have extensively examined rater bias, item functioning, and measurement reliability in educational assessment, most have focused on standardised testing, writing assessment, or performance evaluation contexts (Engelhard, 2013; Wolfe & McVay, 2012). Comparatively limited attention has been given to educational book evaluation systems that rely on expert raters. In many educational institutions, book evaluations still depend primarily on descriptive statistics or average scores, which do not adequately account for potential measurement error arising from item characteristics, rater severity, and interactions among facets. Furthermore, despite the growing application of MFRM in educational measurement contexts (Bond & Fox, 2015; Wind, 2023), its use in policy-driven textbook evaluation systems remains relatively limited, particularly within developing educational systems (de Klerk et al., 2023). Consequently, empirical evidence regarding the interaction among raters, items, and book quality within an integrated psychometric framework remains limited.

Addressing this gap is important because evaluation decisions based solely on raw scores may mask systematic bias and reduce the reliability of assessment outcomes. Without a measurement model that explicitly accounts for rater behaviour and item functioning, it becomes difficult to determine whether score differences truly reflect book quality or merely artifacts of the rating process.

Therefore, this study examines the quality of a book evaluation system using the Many-Facet Rasch Model. Specifically, the study analyses interactions among raters, evaluation items, and books within a unified measurement framework to identify potential sources of measurement error, including misfitting items and variations in rater severity, and to estimate the relative quality of the evaluated books. The novelty of this study lies in its application of MFRM to educational book evaluation by simultaneously examining raters, items, and books within an integrated psychometric framework. Unlike conventional descriptive scoring approaches, this study provides a more comprehensive measurement perspective for identifying bias, inconsistency, and reliability issues in educational book evaluation. The findings are expected to contribute theoretically and practically to developing more objective, reliable, and evidence-based educational evaluation systems. Accordingly, this study addresses the following research questions:

1. How well do the evaluation items function within the book evaluation system, and which items require further review?
2. To what extent do evaluators differ in their severity when assessing educational books?
3. How does the MFRM estimate the relative quality of the evaluated books after accounting for differences in rater severity and item functioning?

## Methods

### Research Design

This study employed a quantitative measurement approach using the Many-Facet Rasch Model (MFRM) within an exploratory case study design to examine the quality of an educational book evaluation system. The evaluation process was conducted by the Ministry of Education and Culture in 2024 as part of the national quality assurance system for educational materials. In Indonesia, textbooks must undergo formal evaluation before the government approves them for use in schools to ensure compliance with academic, pedagogical, and content standards.

The study focused on evaluation data from three books assessed by six evaluators using a 12-item evaluation instrument. This case was selected to explore the feasibility and analytical value of applying MFRM to examine item functioning, rater severity, and differences in estimated book quality within an authentic institutional evaluation context. The sample was intended to support an in-depth examination of the evaluation process rather than to provide a statistically representative sample of the national textbook population. Accordingly, the findings should be interpreted as exploratory evidence of the model's application in this specific context, not as generalizable estimates of the quality of educational textbooks nationally.



This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-

The authors of this study were members of the official book evaluation committee involved in the assessment process. Their involvement enabled access to authentic evaluation data while maintaining adherence to the official evaluation procedures and ethical standards established by the Ministry. All data analysed in this study were anonymised, and no personal identifying information of raters was disclosed. The study used secondary evaluation data collected as part of an institutional quality assurance process; therefore, no direct intervention or identifiable human subject data were involved.

The Many-Facet Rasch Model extends the Rasch measurement framework by simultaneously incorporating multiple facets involved in an assessment process, including raters, items, and evaluated objects (Linacre, 1989). The model is particularly appropriate for multi-rater evaluation contexts in which differences in rater severity, item functioning, and object quality may influence scoring outcomes. In this study, we used the MFRM to estimate item functioning, rater severity, and the relative quality of books within a unified measurement framework.

The MFRM probability model applied in this study can be expressed as follows:

$$log\ (\frac{P_{nijk}}{P_{nijk-1}}) = \theta_n - \delta_i - \rho_j - \tau_k$$

where:

$\theta_n$ represents the quality measure of book (n),
$\delta_i$ represents the difficulty of item (i),
$\rho_j$ represents the severity of rater (j),
$\tau_k$ represents the rating scale threshold between category (k-1) and (k).

This formulation calibrates all facets on a common logit scale, enabling direct comparison among books, raters, and items.

**Data Source**

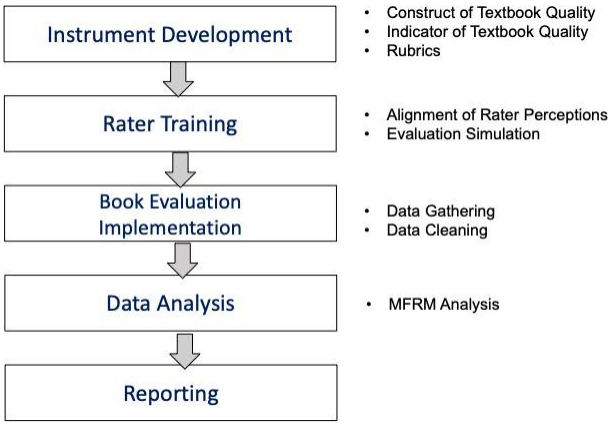
The data were obtained from a book evaluation process in which six raters assessed three educational books using a structured evaluation instrument with twelve items. The raters were selected purposively based on their expertise, professional experience, and familiarity with the subject matter and textbook evaluation procedures. All raters had prior experience in educational material assessment and were officially appointed by the Ministry as members of the evaluation panel.

The selection of three books and six raters followed the institutional evaluation design used in the national textbook assessment process. A fully crossed rating design was employed, in which each rater evaluated all books using the same set of items. This design produced a complete dataset suitable for Many-Facet Rasch analysis because it enabled simultaneous estimation of rater severity, item difficulty, and book quality.

Within the MFRM framework, the study involved three primary facets: raters represented the severity facet, Books as the objects of measurement, and Assessment items as evaluation criteria representing the difficulty facet. The overall research procedure, including the rating process and stages of MFRM analysis, is summarised in Figure 1.



This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-




Sources: Personal data

**Figure 1.** Research Procedure of the Many-Facet Rasch Measurement Analysis

## Assessment Instrument

This study used a structured rubric to evaluate the quality of educational books across several academic and pedagogical dimensions. The instrument was developed based on national educational textbook evaluation regulations and guidelines concerning textbook quality standards. The development process involved several stages. First, the researchers established a construct framework for textbook quality based on regulatory standards for educational textbooks. Second, indicators were formulated to represent four major aspects of textbook quality: (a) content/material appropriateness, (b) presentation appropriateness, (c) language appropriateness, and (d) graphical appropriateness. Each aspect included several indicators designed to capture key dimensions of educational book quality. Table 1 shows the textbook evaluation criteria.

**Table 1.** The Criteria of Textbook Evaluation

| CODE | Textbook Criteria |
|---|---|
| ITEM-05 | Learning outcomes are aligned with learning objectives |
| ITEM-06 | Assessments are able to measure the achievement of learning objectives |
| ITEM-07 | Material is explained validly and in accordance with the field of study |
| ITEM-08 | Information presented is relevant to real life |
| ITEM-09 | Material and examples are appropriate to students' development |
| ITEM-10 | Material, examples, illustrations, and activities support the norms, values, and morals |
| ITEM-11 | Presentation of material is consistent, logical, sequential, and coherent |
| ITEM-12 | Cover design, book contents, illustrations, and infographics are relevant and aesthetically pleasing |
| ITEM-13 | Learning materials and assessments motivate learning and encourage critical thinking |
| ITEM-14 | The content of each chapter includes all essential elements |
| ITEM-15 | Language used is appropriate for the students' age |
| ITEM-16 | Uses standard language that is easy to understand |

Sources: Personal data

Those indicators were selected from the first three aspects of the framework, while the graphical appropriateness aspect was excluded from the analysis.

Each item was scored using a four-category rating scale:
1 = Poor,
2 = Fair,
3 = Good,
4 = Excellent.


http://journal.uinjkt.ac.id/index.php/jp3i
This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-

The use of a multi-category rating scale enabled raters to provide graded judgments regarding the quality of each aspect of the evaluated books and generated ordinal data appropriate for Rasch-based analysis.

### Data Analysis

The rating data (1-4 from six raters) were analysed using the Many-Facet Rasch Model to estimate the measurement parameters associated with books, raters, and items. The analysis was conducted using MINIFAC 4.5.0 (https://www.winsteps.com/minifac.htm), which is a widely used software package for Rasch and MFRM analyses.

All measurement estimates were expressed on a logit (log-odds unit) scale, allowing books, raters, and items to be placed on a common linear continuum. This aligns with the probabilistic measurement principles proposed in Rasch Measurement Theory (Rasch, 1960). The analysis generated several diagnostic statistics, including item fit statistics, rater severity estimates, category functioning, reliability indices, separation indices, Wright maps, and facet calibration measures.

Item and rater fit were evaluated using Infit and Outfit Mean Square (MNSQ) statistics. Following common Rasch measurement guidelines, acceptable fit values ranged from 0.5 to 1.5 (Bond & Fox, 2015; Linacre, 1989). Values above 1.5 indicate excessive randomness or noise, whereas values below 0.5 suggest overfit or redundancy.

The analysis also examined category functioning to evaluate whether the rating scale categories operated appropriately. We assessed category effectiveness based on category frequencies, average measures, outfit statistics, and threshold ordering. Ordered thresholds and acceptable outfit values indicated that the rating scale categories functioned appropriately.

Reliability and separation indices were further used to evaluate the consistency and discriminatory power of the measurement model. Higher reliability and separation values indicate greater stability and stronger differentiation among facets within the evaluation system.

## Results and Discussion

### Results

#### *Summary Statistics of the MFRM Analysis*

To examine the influence of an extremely severe rater on model functioning, two MFRM analyses were conducted. The first model included all six raters, whereas the second model excluded one rater identified as substantially more severe than the others. This comparison was not intended to eliminate a problematic rater, since the fit statistics indicated that P-1 still demonstrated acceptable model fit, but rather to explore the extent to which an extremely severe rater might influence overall model stability, parameter estimation, and the psychometric characteristics of the rating system. The comparison between the six-rater and five-rater models therefore served as a sensitivity analysis to evaluate the robustness of the MFRM results under different rater compositions. Several psychometric indicators were examined, including parameter estimates, reliability, fit statistics, separation indices, chi-square statistics, standard errors, and strata indices. Table 2 summarises the comparison between the six-rater and five-rater models to identify which model demonstrated stronger measurement properties and greater stability in evaluating educational book quality.

**Table 2.** Summary Statistics of the MFRM Models

| Parameter | SIX-Rater Model | FIVE-Rater Model |
|---|---|---|
| **Rater logits range** | -1.44 to 6.89 | -1.78 to 0.86 |
| Most severe rater | P-1 (6.89) | P-5 (0.86) |
| Most lenient rater | P-2 (-1.44) | P-2 (-1.78) |
| **Book logits range** | -0.47 to 0.38 | -0.49 to 0.35 |



This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-

| Parameter | SIX-Rater Model | FIVE-Rater Model |
|---|---|---|
| Highest quality book | Book B | Book B |
| Lowest quality book | Book C | Book C |
| **Rater reliability (pop/sample)** | 0.98 / 0.98 | 0.81 / 0.85 |
| **Book reliability (pop/sample)** | 0.47 / 0.65 | 0.40 / 0.60 |
| **Rater Infit MnSq** | 0.60–1.33 | 0.71–1.21 |
| **Rater Outfit MnSq** | 0.54–1.19 | 0.59–1.18 |
| **Rater separation (pop/sample)** | 7.02 / 7.70 | 2.06 / 2.35 |
| **Book separation (pop/sample)** | 0.95 / 1.36 | 0.82 / 1.23 |
| **Rater strata (pop/sample)** | 9.69 / 10.60 | 3.07 / 3.47 |
| **Book strata (pop/sample)** | 1.60 / 2.15 | 1.43 / 1.98 |
| **Rater $\chi^2$** | 224.0* | 25.8* |
| **Book $\chi^2$** | 5.4 | 4.7 |
| **Rater S.E. range** | 0.29–0.50 | 0.29–0.51 |
| **Book S.E. range** | 0.24–0.28 | 0.26–0.30 |

$p < .001$

Sources: Personal data

Table 2 summarises the overall MFRM statistics for both the six-rater and five-rater models. Overall, the six-rater model showed superior psychometric performance, particularly in rater reliability, separation, and strata indices. The substantially higher separation values indicate clearer differentiation in rater severity levels, although the presence of an extremely severe rater (P-1) contributed to wider variability within the model. Despite these differences, both models showed acceptable fit statistics, with all Infit and Outfit Mean Square values remaining within the recommended Rasch range of 0.5–1.5. In addition, the relatively stable book logit measures across both models suggest that book-quality estimates remained consistent after excluding the outlier rater. These findings support the usefulness of the MFRM approach in detecting rater effects and improving the validity and objectivity of educational book evaluation systems.

### *Category Functioning Statistics*

Category functioning analysis was conducted to evaluate whether the rating scale categories operated appropriately within the MFRM. In Rasch-based measurement, effective category functioning requires that categories be used consistently by raters, demonstrate ordered thresholds, and reflect increasing levels of the latent construct along the measurement continuum (Linacre, 2002; Bond & Fox, 2015). Table 3 presents the category functioning statistics, while Figure 1 illustrates the category probability curves across the latent trait continuum.

**Table 3**. Category Functioning Statistics

| Category | Frequency | Percentage | Average Measure | Expected Measure | Outfit MnSq | Threshold Calibration | Interpretation |
|---|---|---|---|---|---|---|---|
| 1 | 6 | 3% | -2.24 | -1.76 | 0.60 | — | Rarely used, but functioning appropriately |



http://journal.uinjkt.ac.id/index.php/jp3i

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-

| Category | Frequency | Percentage | Average Measure | Expected Measure | Outfit MnSq | Threshold Calibration | Interpretation |
|---|---|---|---|---|---|---|---|
| 2 | 51 | 24% | -0.77 | -0.95 | 1.20 | -3.52 | Acceptable category functioning |
| 3 | 126 | 58% | 0.32 | 0.41 | 0.70 | -1.29 | Most frequently used and stable category |
| 4 | 33 | 15% | 6.78 | 6.61 | 0.50 | 4.81 | Highest category, relatively difficult to attain |

Sources: Personal data

The results presented in Table 3 indicate that the rating scale categories functioned appropriately within the MFRM framework. The average measures increased monotonically across categories, suggesting that higher rating categories consistently reflected higher levels of perceived book quality. This finding confirms that the rating scale operated logically along the latent measurement continuum.

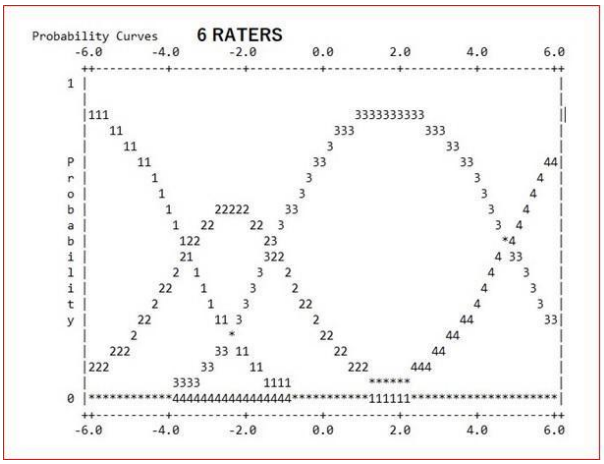
In addition, all Outfit Mean Square (MnSq) statistics remained within the acceptable Rasch range, indicating that each category contributed meaningfully to the measurement process without introducing excessive noise or randomness. The Outfit MnSq values ranged from 0.50 to 1.20, demonstrating stable and predictable category usage among raters.

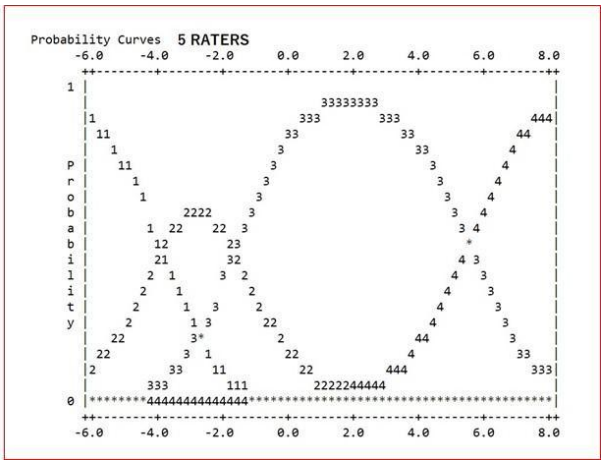
The threshold calibrations were also properly ordered, indicating no category disordering. This result suggests that raters could clearly distinguish between adjacent score categories when evaluating book quality. Ordered thresholds are considered an important indicator of rating scale validity because they reflect progressive transitions between response categories along the latent construct continuum.

Category 3 emerged as the most frequently used category, accounting for 58% of all ratings. This finding suggests that raters tended to assign moderate evaluations across the assessed books. In contrast, Category 1 was rarely used, representing only 3% of responses, indicating that extremely low evaluations were uncommon in the assessment process.

Furthermore, Category 4 demonstrated the highest threshold calibration (4.81 logits), indicating that the highest rating category was relatively difficult to attain and likely reserved for books perceived as demonstrating exceptionally strong quality characteristics. Overall, these findings support the validity and effectiveness of the rating scale structure used in the educational book evaluation instrument.

To further examine the effectiveness of the rating scale categories reported in Table 3, Figure 3 displays the category probability curves across the latent continuum. The figure visually evaluates category ordering, category dominance, and threshold distribution within the Many-Facet Rasch Model framework.




Sources: Personal data



This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-

**Figure 3**. Category Probability Curves for the Rating Scale Categories

Figure 3 further illustrates how the rating scale categories function across the latent continuum. Each category displayed a distinct probability region, indicating that the categories operated in an ordered and meaningful manner within the measurement model.

Category 3 occupied the broadest region along the continuum, confirming that it was the most frequently used category among raters. In contrast, Category 1 appeared only within the lower end of the continuum, reflecting the relatively infrequent use of extremely low ratings.

The probability curves also demonstrate that Category 4 was located at the upper end of the logit scale, indicating that the highest category was relatively difficult to achieve and reserved for books perceived as having superior quality characteristics.

***Item Functioning Analysis***

Item functioning analysis was conducted to evaluate the psychometric performance of each item in the assessment instrument. The analysis examined item validity using fit statistics, particularly the Infit and Outfit Mean Square (MNSQ) indices. In Rasch-based measurement, acceptable fit statistics generally fall within 0.5–1.5, indicating that an item contributes meaningfully to the measurement construct and functions consistently within the model's probabilistic expectations (Bond & Fox, 2015; Linacre, 2002). Fit statistics outside this range may indicate excessive randomness, redundancy, or unexpected response patterns that threaten construct validity and measurement precision (Engelhard, 2013; Boone et al., 2014).

To provide a clearer overview of item performance, Table 4 presents the item fit statistics generated from the MFRM analysis for both the six-rater and five-rater models.

**Table 4.** Item Fit Statistics for the Six-Rater and Five-Rater Models

| Item | Six-Rater Infit | Six-Rater Outfit | Status | Five-Rater Infit | Five-Rater Outfit | Status |
|---|---|---|---|---|---|---|
| Item 05 | 1.00 | 0.78 | Fit | 1.07 | 0.88 | Fit |
| Item 06 | 0.62 | 0.52 | Fit | 0.64 | 0.57 | Fit |
| Item 07 | 1.80 | 1.40 | Misfit | 1.97 | 1.64 | Misfit |
| Item 08 | 0.57 | 0.47 | Fit | 0.64 | 0.57 | Fit |
| Item 09 | 0.95 | 1.04 | Fit | 0.99 | 1.22 | Fit |
| Item 10 | 1.19 | 0.74 | Fit | 1.24 | 0.83 | Fit |
| Item 11 | 0.76 | 0.78 | Fit | 0.66 | 0.60 | Fit |
| Item 12 | 1.50 | 1.67 | Fit/Marginal | 1.08 | 1.13 | Fit |
| Item 13 | 0.72 | 0.74 | Fit | 0.76 | 0.88 | Fit |
| Item 14 | 0.98 | 0.98 | Fit | 0.92 | 0.84 | Fit |
| Item 15 | 0.08 | 0.06 | Overfit | 0.06 | 0.04 | Overfit |
| Item 16 | 0.96 | 1.07 | Fit | 1.13 | 1.55 | Marginal Misfit |

*Note. Acceptable fit statistics generally fall within the range of 0.5–1.5. The analysed indicators retained their original numbering (Items 5–16) because Items 1–4, related to graphical appropriateness, were excluded from the present analysis.*
Sources: Personal data

The results show that most items demonstrated acceptable Infit and Outfit MNSQ values, indicating that they functioned appropriately within the measurement framework and contributed consistently to the assessment of book quality. As shown in Table 4, most items in both models fell within the recommended fit range, suggesting that the evaluation criteria captured the intended construct and supported the instrument's unidimensionality. Similar findings have been reported in previous Rasch-


http://journal.uinjkt.ac.id/index.php/jp3i
This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-

based educational assessment studies, which emphasised that well-fitting items enhance the stability and interpretability of measurement results (Wind & Jones, 2019; Eckes, 2015).

However, several items exhibited fit statistics outside the acceptable range. In both the six-rater and five-rater models, Item 07 demonstrated clear misfit behaviour, with Infit MNSQ values of 1.80 and 1.97, respectively. In the five-rater model, Item 07's Outfit MNSQ also exceeded the acceptable threshold (1.64), indicating response patterns inconsistent with model expectations. In addition, Item 12 in the six-rater model showed marginal misfit with an Outfit MNSQ value of 1.67, while Item 16 in the five-rater model slightly exceeded the recommended Outfit criterion (1.55). These findings suggest that raters may have interpreted some evaluation criteria inconsistently or that the items did not align optimally with the latent construct being measured.

Misfitting items may introduce measurement noise or construct-irrelevant variance (Wright & Linacre, 1994). Such conditions may occur when items are ambiguously worded, interpreted differently by raters, or fail to align conceptually with the underlying construct (Myford & Wolfe, 2003). In educational material evaluation contexts, misfit may also reflect inconsistencies in how evaluators interpret indicators related to content quality, presentation, language clarity, or graphical design aspects of educational books.

***Rater Severity and Leniency***

The analysis also examined raters' scoring behaviour by estimating their severity levels on the logit scale. The results revealed noticeable differences among raters in terms of severity and leniency. Some raters displayed higher logit estimates, indicating stricter scoring tendencies, while others showed lower logit estimates, reflecting more lenient evaluation behaviour. In the MFRM, differences in rater severity are considered an important source of measurement variability because raters may apply scoring criteria inconsistently across evaluation contexts (Linacre, 1989; Eckes, 2015).

The variability in rater severity suggests that evaluators interpreted and applied the assessment criteria with different levels of strictness. Such differences are common in performance-based and judgment-based assessment systems where human evaluation plays a central role (Engelhard, 2013). Recent studies also indicate that rater severity may be influenced by professional background, scoring experience, cognitive interpretation of rubrics, and individual judgment tendencies (de Klerk et al., 2023; Wind, 2023). Consequently, variability in rater behaviour may introduce construct-irrelevant variance that affects the fairness and consistency of assessment outcomes.

Table 5 presents the rater calibration results, including logit severity measures and fit statistics for each evaluator. The results indicate substantial variation in rater severity across the evaluation process.

**Table 5.** Rater Severity and Fit Statistics in the Many-Facet Rasch Model

| Rater | Logit Measure | Severity Category | Infit MnSq | Outfit MnSq | Interpretation |
|---|---|---|---|---|---|
| P-1 | 6.89 | Extremely Severe | 1.33 | 0.80 | Outlier |
| P-5 | 1.05 | Severe | 0.60 | 0.54 | Fit |
| P-3 | 0.07 | Moderate | 0.64 | 0.56 | Fit |
| P-6 | -0.06 | Moderate | 0.78 | 0.84 | Fit |
| P-4 | -0.63 | Lenient | 1.07 | 1.19 | Fit |
| P-2 | -1.44 | Most Lenient | 1.24 | 1.19 | Fit |

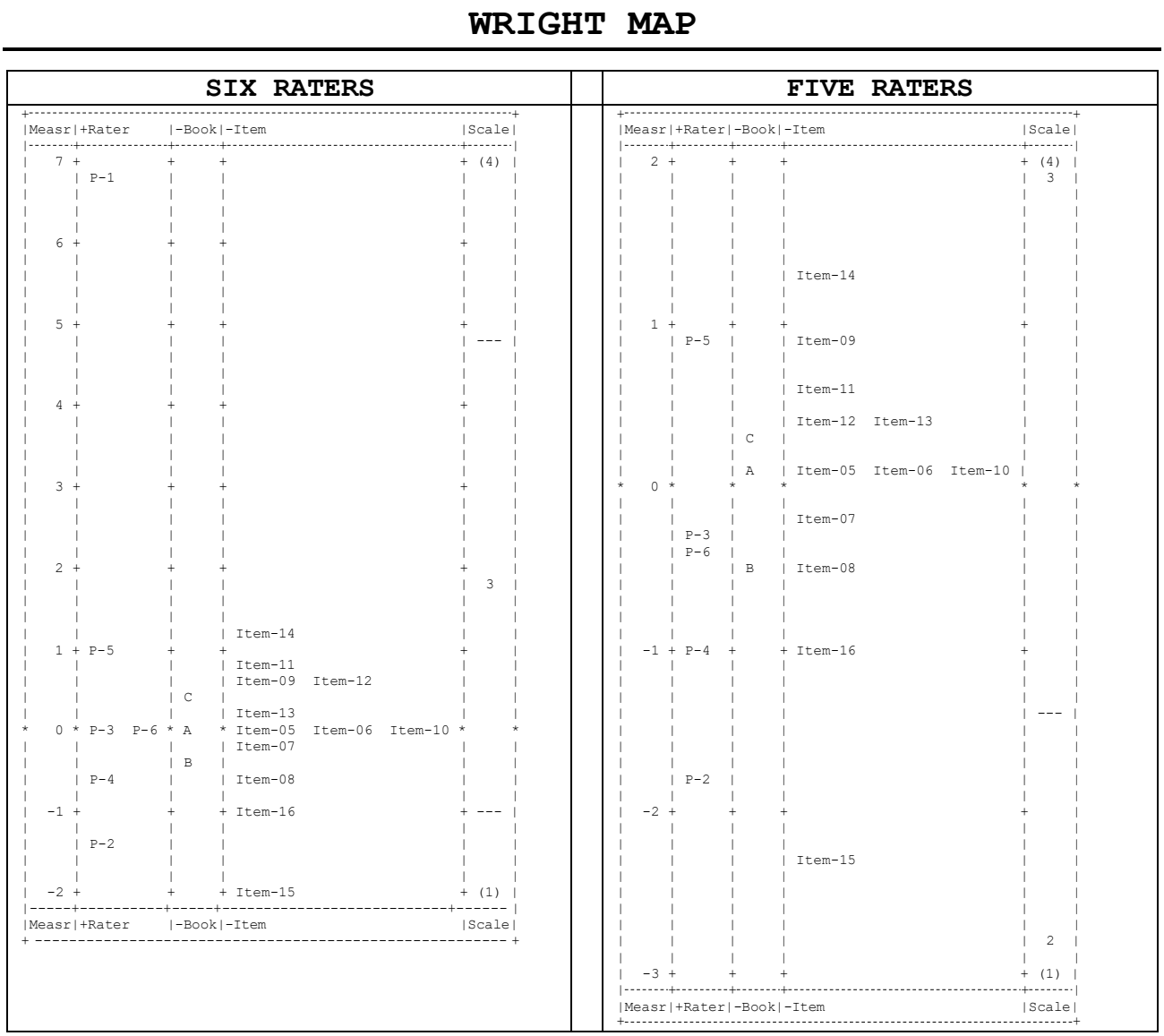
Sources: Personal data

As shown in Table 5, rater P-1 demonstrated an extremely high severity estimate (6.89 logits), substantially distant from the other raters. This finding indicates that P-1 applied the evaluation criteria far more strictly than the remaining evaluators and was therefore identified as a potential outlier. In contrast, rater P-2 showed the lowest logit measure (-1.44 logits), reflecting the most lenient scoring tendency. Despite these differences, most raters demonstrated acceptable Infit and Outfit MnSq statistics within the recommended range of 0.5–1.5, indicating generally consistent scoring behaviour.



This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-

In addition, the analysis identified a rater whose severity estimate was substantially distant from the cluster of other raters on the logit scale. This rater exhibited extreme scoring behaviour and was therefore categorised as an outlier. Extreme raters may distort measurement results because their scoring patterns systematically differ from the model's probabilistic expectations (Myford & Wolfe, 2003). Previous MFRM studies have shown that highly severe or highly lenient raters may reduce inter-rater reliability and threaten the validity of evaluation outcomes if not properly identified and controlled (Wind & Guo, 2021; Jin & Chiu, 2021). Figure 4 further visualises the relative positions of raters on the common logit scale together with items and books within the Many-Facet Rasch Model framework.




Sources: Personal data

**Figure 4**. Wright Map of Raters, Items, and Books in the Many-Facet Rasch Model

Figure 4 presents the Wright Map generated from the Many-Facet Rasch Model analysis. The map visually represents the relative positions of raters, items, and books on a common logit scale. By placing all facets within a unified measurement continuum, the Wright Map allows simultaneous examination of rater severity, item difficulty, and book quality within the same probabilistic framework (Bond & Fox, 2015; Boone et al., 2014).


http://journal.uinjkt.ac.id/index.php/jp3i
This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-

In the FACETS output, positive (+) and negative (−) logit measures indicate relative positioning on the latent continuum; however, the interpretation differs across facets depending on the conceptual meaning of the construct being measured. For the rater facet, higher positive logit measures indicate greater rater severity. For the book facet, lower or more negative logit measures indicate higher book quality. Similarly, for the item facet, lower or more negative logit measures indicate lower item difficulty, meaning that the criteria were easier for books to satisfy.

The map shows clear variation in rater severity across evaluators. Rater P-1 was positioned substantially higher on the logit scale than the other raters, indicating extremely severe scoring behaviour. In contrast, raters positioned lower on the scale demonstrated more lenient scoring tendencies. This finding supports the statistical results reported in the rater calibration analysis and confirms the presence of systematic differences in scoring behaviour among evaluators.

The Wright Map also illustrates the relative distribution of item difficulty and book quality. Items located higher on the scale represent more demanding evaluation criteria, whereas lower-positioned items indicate relatively easier criteria to satisfy. Similarly, books positioned higher on the continuum demonstrated stronger overall performance across evaluation criteria compared with books located lower on the scale. Overall, the Wright Map enhances the interpretability of the evaluation system by visually demonstrating the interaction among raters, items, and books within the Many-Facet Rasch Model framework.

***Book Quality Measurement***

The quality of the evaluated books was estimated using logit measures generated by the MFRM analysis. These logit values represent the relative position of each book on a latent continuum of book quality and enable the estimation of book performance within a probabilistic measurement framework (Bond & Fox, 2015; Linacre, 1989). In Rasch-based measurement, logit measures provide interval-level estimates that allow more precise interpretation of the relative quality of assessed objects compared with raw ordinal scores (Wright & Masters, 1982).

Books with higher logit measures represent higher estimated quality levels, indicating that they received consistently stronger evaluations relative to the assessment criteria and rater severity. Conversely, books with lower logit values indicate relatively lower levels of quality within the evaluated sample. These findings show that the MFRM framework can differentiate book quality while accounting for variation in item difficulty and rater behaviour (Eckes, 2015; Engelhard, 2013). Table 6 presents the calibration results for book quality in both the six-rater and five-rater models. The table summarises logit measures and fit statistics for each evaluated book.

**Table 6.** Book Quality Calibration and Fit Statistics in the Many-Facet Rasch Model

| Book | Model | Logit Measure | Category | Infit MnSq | Outfit MnSq | Interpretation |
|---|---|---|---|---|---|---|
| Book A | Six-Rater | 0.09 | Moderate | 0.78 | 0.65 | Fit |
| Book B | Six-Rater | -0.47 | Moderate | 1.16 | 0.99 | Highest Quality |
| Book C | Six-Rater | 0.38 | Moderate | 1.00 | 0.91 | Lowest Quality |
| Book A | Five-Rater | 0.14 | Moderate | 0.80 | 0.71 | Fit |
| Book B | Five-Rater | -0.49 | Moderate | 1.23 | 1.17 | Highest Quality |
| Book C | Five-Rater | 0.35 | Moderate | 0.92 | 0.82 | Lowest Quality |

Sources: Personal data

As shown in Table 6, Book B obtained the lowest logit measure in both the six-rater (-0.47 logits) and five-rater (-0.49 logits) models, indicating the highest estimated quality among the evaluated books. In contrast, Book C demonstrated the highest logit values (0.38 logits and 0.35 logits), suggesting relatively



This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-

lower quality within the evaluated sample. Book A was positioned near the centre of the continuum, reflecting moderate quality performance across both models.

All books demonstrated acceptable Infit and Outfit Mean Square statistics within the recommended range of 0.5–1.5, indicating that the observed ratings were consistent with model expectations. These findings suggest that the evaluated books functioned appropriately within the measurement framework and that the estimated quality measures were psychometrically stable.

Compared with raw scores, the logit scale provides a more meaningful representation of book quality because it adjusts for differences in rater severity and item difficulty. This approach allows for more accurate comparisons among books and ensures that the estimated quality levels reflect the underlying measurement construct rather than artifacts of scoring variation (Myford & Wolfe, 2003). Recent psychometric studies have also emphasised that Rasch-based measurement models improve the fairness, interpretability, and objectivity of educational evaluation systems by separating object quality from systematic measurement error associated with raters and assessment instruments (Wind, 2023; Liu & Jin, 2022).

Furthermore, the findings indicate that MFRM provides a more robust framework for educational book evaluation than conventional descriptive scoring approaches. By calibrating books, raters, and items on a common measurement scale, the model enhances the credibility of evaluation outcomes and supports more evidence-based decision-making in educational material assessment (Boone et al., 2014; Wolfe & McVay, 2012).

***Additional Rasch Diagnostics***

To strengthen the psychometric evaluation of the MFRM analyses, additional Rasch diagnostic indices were examined, including dimensionality evidence, local independence, residual analysis, point-measure correlations, and standardised residual statistics.

**Table 7**. Additional Rasch Diagnostic

| Rasch Diagnostic | SIX-RATERS Model | FIVE-RATERS Model | Interpretation |
|---|---|---|---|
| Dimensionality Evidence | Explained variance: 56.98%<br>Residual variance: 43.02% | Explained variance: 28.70%<br>Residual variance: 71.30% | The SIX-RATERS model showed stronger evidence of unidimensionality because the explained variance exceeded 50% and was higher than the residual variance. |
| Local Independence | Pearson Chi-square = 184.17, p = 0.8491 | Pearson Chi-square = 161.44, p = 0.6686 | Both models demonstrated acceptable model-data fit and provided support for local independence. However, the lower residual variance in SIX-RATERS indicated stronger evidence of independence. |
| Residual Analysis | Mean residual = 0.00<br>SD = 0.92 | Mean residual = −0.01<br>SD = 0.95 | Both models showed residuals centered near zero. The SIX-RATERS model |


http://journal.uinjkt.ac.id/index.php/jp3i
This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-

| | | | |
|---|---|---|---|
| | Residual variance = 43.02% | Residual variance = 71.30% | showed lower unexplained variance and greater measurement stability. |
| Point-Measure Correlation (Book Facet) | Correlation range: 0.70–0.78;<br><br>P-4 = 0.06 | weak correlations, particularly P-4 (0.11) and P-5 (0.17), suggesting moderate inconsistency in scoring behavior | The SIX-RATERS model demonstrated strong consistency between observed ratings and Rasch measures for the book facet. |
| Standardized Residuals (ZStd) | Most values within −2 to +2 | Most values within −2 to +2 | Both models demonstrated acceptable residual behavior, the SIX-RATERS model exhibited slightly better residual stability and overall fit consistency than the FIVE-RATERS model |
| Overall Psychometric Performance | Higher explained variance, lower residual variance, stronger correlations, and more stable residual patterns | Lower explained variance, higher residual variance, and weaker measurement stability | The SIX-RATERS model demonstrated stronger psychometric performance and provided better measurement quality than the FIVE-RATERS model. |

Sources: Personal data

Overall, the additional Rasch diagnostics indicated that the SIX-RATERS model provided stronger psychometric evidence than the FIVE-RATERS model. The model demonstrated better dimensionality, lower unexplained variance, stronger measurement consistency, and more stable residual patterns, supporting its suitability for the evaluation framework.

## Discussion

The present study demonstrates the value of the Many-Facet Rasch Model (MFRM) for examining the psychometric quality of an educational book evaluation system. By jointly analysing books, raters, and evaluation items, MFRM provides information that conventional scoring methods based on raw or average scores may overlook. Given the limited number of books included, the findings should be interpreted as an exploratory methodological application rather than as evidence representative of the entire national textbook population.

### *Item Functioning and Evaluation Criteria*

Most evaluation items functioned adequately, suggesting that the instrument generally supported the intended evaluation process. However, the misfit identified in Item 07 indicates that this criterion did not operate as consistently as expected within the measurement model. From the perspective of Rasch Measurement Theory, item misfit may signal ambiguity in wording, overlap with other criteria, or differences in how raters interpret the intended standard (Bond & Fox, 2015; Engelhard, 2013). These possibilities are particularly relevant when evaluation criteria require professional judgment rather than straightforward factual verification.



This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-

The available MFRM results identify the inconsistency associated with Item 07 but do not establish its underlying cause. Determining whether the misfit reflects unclear wording, conceptual overlap, or inconsistent rubric interpretation requires a substantive review of the item, its scoring descriptors, and any available rater comments. Therefore, Item 07 should be reviewed before a decision is made to revise or remove it. This finding highlights the importance of combining psychometric diagnostics with content-based review to strengthen the validity and defensibility of educational book evaluation.

### *Rater Severity and Evaluation Consistency*

The substantial variation in rater severity, particularly the unusually high severity of rater P-1, highlights how individual scoring practices can influence multi-rater evaluation. Previous research suggests that ratings may be shaped by professional experience, internal standards, and interpretations of scoring rubrics (Engelhard, 2013; Wind & Guo, 2019). In this case, P-1's extreme severity could reflect stricter evaluation standards, a different interpretation of the rubric, or other factors affecting judgment. However, the MFRM results alone cannot determine which explanation is correct.

Accordingly, P-1's severity should be treated as a diagnostic signal rather than conclusive evidence of rater incompetence or bias. Further investigation could examine the rater's scoring pattern across individual items and books, compare the ratings with the rubric descriptors, and review any written comments or available documentation. Such evidence would help distinguish legitimate differences in professional judgment from inconsistencies arising from unclear criteria or inconsistent rubric application.

The reported improvement in model reliability and parameter stability after excluding the outlier further demonstrates the importance of examining the influence of extreme raters. Nevertheless, exclusion should not be an automatic response to statistical extremity. Institutions should first investigate the source of the unusual pattern and document the rationale for any decision to retain, retrain, or exclude a rater. This approach supports fairness and accountability while reducing the risk of removing legitimate but stringent professional judgments.

## Implications

### *Implications for Educational Book Evaluation*

Taken together, the findings indicate that an evaluation system’s quality depends not only on the instrument itself but also on how its criteria are interpreted and applied. MFRM offers a systematic means of identifying problematic items and unusual rater patterns that may remain hidden in conventional average scores (Linacre, 1989; Myford & Wolfe, 2003). Its value therefore lies not simply in producing adjusted estimates, but also in providing diagnostic evidence to guide improvements in evaluation procedures.

Institutions responsible for textbook evaluation should periodically review item wording and scoring descriptors, provide structured training and calibration for raters, and establish procedures for investigating unusual rating patterns. Where feasible, psychometric findings should be considered alongside content review and rater feedback before consequential decisions are made. These measures can strengthen the transparency, consistency, and defensibility of textbook evaluation. Although the present study is limited to a small set of books and raters, it illustrates how MFRM can support evidence-informed quality assurance and provides a basis for further investigation in broader evaluation settings.

### *Implications for Book Evaluation Policy*

Beyond methodological considerations, our findings suggest important implications for designing and implementing educational book evaluation policies. National textbook evaluation programs typically rely on expert reviewers who assess educational materials based on multiple criteria related to content quality, pedagogical effectiveness, and presentation. The results indicate that the reliability of such evaluation



This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-

systems depends not only on the quality of the books being assessed but also on the quality of the evaluation instruments and the consistency of rater judgments.

Identifying misfitting items suggests that evaluation frameworks used in textbook assessment should be reviewed periodically and empirically validated. Without such validation, evaluation criteria may not function effectively as measurement indicators. Previous studies in educational measurement emphasise that evaluation standards must be supported by empirically tested instruments to ensure the credibility and validity of assessment outcomes (Bond & Fox, 2015).

Findings on rater variability also highlight the importance of structured rater management in textbook evaluation systems. Research on rater-mediated assessments indicates that rater training, calibration sessions, and continuous monitoring of scoring patterns are essential strategies for improving scoring consistency and reducing bias (Engelhard, 2013; Eckes, 2015). Implementing these strategies can strengthen the reliability and fairness of book evaluation processes conducted at institutional or national levels.

Furthermore, applying the Many-Facet Rasch Model provides a scientifically grounded framework for improving quality assurance in educational publishing. By enabling evaluators to examine item functioning, rater severity, and book quality within a unified measurement system, the model offers valuable diagnostic insights into the evaluation process. As Linacre (1989) notes, such multifaceted measurement approaches are particularly useful in complex evaluation contexts where multiple sources of variation must be accounted for.

Taken together, these implications suggest that integrating modern measurement approaches into textbook evaluation systems can enhance the transparency, reliability, and credibility of policy decisions related to the selection and approval of educational materials. In this way, advances in educational measurement can directly strengthen evidence-based policymaking in educational publishing.

## Conclusion

This study demonstrates the value of the Many-Facet Rasch Model (MFRM) in strengthening educational book evaluation by identifying potential weaknesses in evaluation items and differences in rater severity that may affect the fairness and credibility of assessment outcomes. The findings support using psychometric analysis to complement expert judgment, rather than relying solely on aggregated evaluation scores.

For practical implementation, the government agency responsible for national book assessment should integrate MFRM analysis into its annual evaluation procedures. The proposed SOP should require evaluators' ratings to be compiled in a standardised digital dataset and analysed using MFRM software after each evaluation cycle. Before final approval decisions are made, a designated technical team should review items flagged for misfit and investigate unusually severe or inconsistent rater patterns. The review findings should inform decisions on item revision, rubric clarification, and targeted rater retraining, and the resulting actions should be documented for the next evaluation cycle. This procedure would make MFRM a practical quality-control mechanism within the existing assessment process.

This study is limited by its exploratory case study design, involving three books and six evaluators; therefore, its findings should not be generalised to the entire population of national textbooks. Future studies should examine the proposed procedure using larger datasets across multiple evaluation cycles and institutions to assess its feasibility and contribution to evaluation quality. Such evidence would help policymakers refine implementation guidelines and establish a sustainable, evidence-based approach to national book assessment.

## Acknowledgment

The authors sincerely thank the Director of the Textbook Centre, Agency for Standards, Curriculum, and Educational Assessment, Ministry of Primary and Secondary Education, Indonesia, for granting permission to use the instrument and data. Appreciation is also extended to the experts and raters whose



This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-

contributions to the development of the evaluation instrument and the implementation of the study were invaluable

## Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Authors' Contributions

Author 1 conceived and designed the study, developed the research framework, conducted the data analysis, interpreted the findings, and prepared the manuscript. Author 2 contributed to manuscript writing and supported content refinement. Author 3 critically reviewed the manuscript and provided substantive feedback on the study and its interpretation. Author 4 revised the manuscript based on feedback and discussions among the authors. All authors read and approved the final version of the manuscript.

## References

Adie, L., Klenowski, V., & Wyatt-Smith, C. (2021). Moderation and assessment in higher education: A review of the literature. *Assessment & Evaluation in Higher Education, 46*(2), 185–197. https://doi.org/10.1080/02602938.2020.1729424

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.

de Klerk, S., Bosker, R. J., & van de Grift, W. (2023). Rater effects and reliability issues in performance assessment: Recent developments in many-facet Rasch measurement. *Educational Measurement: Issues and Practice, 42*(2), 45–57. https://doi.org/10.1111/emip.12532

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Jin, K. Y., & Chiu, M. M. (2021). Exploring rater effects in performance assessments using many-facet Rasch measurement. *Educational Psychology, 41*(7), 879–897. https://doi.org/10.1080/01443410.2021.1882050

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press.

Linacre, J. M. (2022). *A user guide to FACETS* (Version 3.84.0) [Computer software]. Winsteps.com. https://www.winsteps.com/a/Facets-Manual.pdf

Liu, Y., & Jin, Y. (2022). Applications of many-facet Rasch measurement in educational assessment: A systematic review. *Frontiers in Psychology, 13*, 845327. https://doi.org/10.3389/fpsyg.2022.845327

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4* (4), 386–422.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

Republik Indonesia. (2017). *Undang-Undang Republik Indonesia Nomor 3 Tahun 2017 tentang Sistem Perbukuan*. Pemerintah Republik Indonesia.

Republik Indonesia. (2019). *Peraturan Pemerintah Republik Indonesia Nomor 75 Tahun 2019 tentang Peraturan Pelaksanaan Undang-Undang Nomor 3 Tahun 2017 tentang Sistem Perbukuan*. Pemerintah Republik Indonesia.

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2022). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education, 83*(2), 433–448. https://doi.org/10.1007/s10734-021-00648-4

Trace, J., Meier, S., & Janssen, G. (2020). Rater cognition and scoring validity in educational assessment. *Assessing Writing, 45*, 100469. https://doi.org/10.1016/j.asw.2020.100469

Wind, S. A. (2023). Advances in many-facet Rasch measurement for educational assessment. *Educational Measurement: Issues and Practice, 42*(1), 5–17. https://doi.org/10.1111/emip.12530



This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-

Wind, S. A. (2023). Advances in rater-mediated assessment and many-facet Rasch measurement. *Frontiers in Education, 8*, 1187452. https://doi.org/10.3389/feduc.2023.1187452

Wind, S. A., & Guo, W. (2019). Exploring the combined effects of rater misfit and differential rater fun ctioning in performance assessments. Educational and Psychological Measurement, 79(5), 962–987. https://doi.org/10.1177/0013164419834613 https://doi.org/10.1016/j.edurev.2019.100279

Wind, S. A., & Guo, F. (2021). Examining rater effects using many-facet Rasch measurement in educational assessment contexts. *Journal of Educational Measurement, 58*(3), 344–366. https://doi.org/10.1111/jedm.12298

Wind, S. A., & Jones, E. (2019). The many-facet Rasch model. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of personality and individual differences* (pp. 1–4). Springer. https://doi.org/10.1007/978-3-319-28099-8_1348-1

Wind, S. A., & Jones, E. (2022). Psychometric issues in rater-mediated assessment systems. *Frontiers in Education, 7*, 873214. https://doi.org/10.3389/feduc.2022.873214

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying rater effects. *Educational Measurement: Issues and Practice, 31*(2), 30–39. https://doi.org/10.1111/j.1745-3992.2012.00232.x

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370. https://www.rasch.org/rmt/rmt83b.htm

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.



This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-