JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)
DOI: http://dx.doi.org/10.15408/jp3i.v15i1.50616
http://journal.uinjkt.ac.id/index.php/jp3i

Applying the Rasch Model as a Diagnostic Tool for Rating Scale Refinement

Asrijanty

Faculty of Psychology, Universitas Islam Negeri Syarif Hidayatullah, Jakarta, Indonesia

asrijanty@uinjkt.ac.id

Abstract

The Rasch model has been widely used in educational, psychological, and health research to evaluate the measurement quality of instruments. In many applications, however, Rasch analysis is primarily reported to support validation or confirm the adequacy of a scale. Although diagnostic analyses may be conducted during instrument development, their role in informing substantive instrument refinement is less explicitly documented and therefore less visible in the literature. This study aims to demonstrate how the Rasch model can be applied as a diagnostic tool to support the refinement of rating scales. Using empirical data from an attitude scale, the study illustrates how detailed Rasch outputs—such as item fit, response category functioning, and threshold ordering—can be interpreted to identify specific sources of measurement problems. These insights provide a basis for targeted revisions, demonstrating how Rasch analysis can contribute not only to validation but also to iterative instrument refinement. This study contributes to the methodological literature by highlighting a more comprehensive use of the Rasch model that integrates validation and diagnostic purposes. It also provides practical guidance for researchers, particularly those less familiar with Rasch analysis, on how to use model outputs to improve measurement instruments, especially rating scales.

Keywords: category functioning, rasch model, rating scale, threshold ordering

Abstrak

Model Rasch telah banyak digunakan dalam penelitian pendidikan, psikologi, dan kesehatan untuk mengevaluasi kualitas pengukuran suatu instrumen. Dalam penerapannya analisis Rasch terutama dilaporkan untuk mendukung keputusan validasi atau mengonfirmasi kelayakan suatu skala. Meskipun analisis diagnostik mungkin dilakukan selama proses pengembangan instrumen, perannya dalam memberikan dasar bagi penyempurnaan instrumen secara substantif kurang secara eksplisit didokumentasikan dan karena itu kurang terlihat dalam literatur. Penelitian ini bertujuan untuk menunjukkan bagaimana model Rasch dapat digunakan sebagai alat diagnostik untuk mendukung penyempurnaan instrumen pengukuran khususnya rating scale. Dengan menggunakan data empiris dari skala sikap, penelitian ini mengilustrasikan bagaimana hasil analisis Rasch yang rinci—seperti item fit, fungsi kategori respon dan urutan threshold,—dapat diinterpretasikan untuk mengidentifikasi sumber spesifik dari permasalahan pengukuran. Hasil analisis memberikan dasar untuk melakukan revisi yang lebih terarah, serta menunjukkan bahwa analisis Rasch dapat berkontribusi tidak hanya pada validasi, tetapi juga pada penyempurnaan instrumen secara iteratif. Penelitian ini berkontribusi pada literatur metodologis dengan menekankan penggunaan model Rasch yang lebih komprehensif, yang mengintegrasikan tujuan validasi dan diagnostik. Selain itu, penelitian ini juga memberikan panduan praktis bagi peneliti, khususnya yang belum familiar dengan analisis Rasch, mengenai cara memanfaatkan hasil analisis Rasch untuk meningkatkan kualitas instrumen pengukuran khususnya rating scale.

***Kata kunci**: fungsi kategori, model Rasch, pengembangan skala, rating scale, urutan threshold*

Introduction

Rating scales are used widely across fields such as healthcare, education, psychology, and the social sciences. These instruments are commonly used to assess constructs such as attitudes, perceptions, behavioural tendencies, and performance. In many applied contexts, rating scales also serve as the basis for critical decision-making, including determining whether an individual meets specific criteria, qualifies for a program, or demonstrates acceptable performance.

Rating scales are structured response formats that allow respondents or raters to assign ordered categories to represent the degree or intensity of a particular attribute. The data may be obtained either through self-report or through external raters, such as in performance assessments. In both cases, the results are expected to accurately reflect an individual's standing on the underlying construct. Likert (1932) introduced the Likert scale as a method for measuring attitudes by asking respondents to indicate their level of agreement with a series of statements along ordered categories. In the context of rater-mediated assessment, Engelhard and Wind (2018) emphasise that rating scales serve as tools through which raters translate observations into quantitative scores using predefined categories. These categories are assumed to represent increasing levels of the latent construct being measured.

Despite their widespread use across educational, psychological, health, and social research contexts, rating scales remain vulnerable to a range of challenges that can compromise the validity and interpretability of the resulting data. These challenges largely stem from variability in how individuals interpret and apply the elements of the scale, whether as respondents providing self-reports or as raters evaluating others (Engelhard & Wind, 2028; Schwarz, 1999; Tourangeau, et al., 2000)

A fundamental issue concerns item functioning. Ideally, each item should clearly represent a specific aspect or level of the construct being measured. In practice, however, both respondents and raters may interpret item content differently due to variations in language comprehension, prior experience, or contextual understanding. Cognitive theories of survey response suggest that individuals engage in complex interpretation processes when responding to items, which may lead to inconsistencies even when items are carefully designed (Schwarz, 1999; Tourangeau et al., 2000). As a result, responses may reflect differing interpretations of the same item rather than consistent judgments of the intended construct, introducing variability not attributable to the construct itself.

Closely related to this is the issue of category functioning. Even when item content is relatively clear, the response options themselves may be interpreted inconsistently. Labels such as "agree" and "strongly agree," or performance levels such as "adequate" and "good," do not always carry uniform meaning across individuals (Engelhard Jr. & Wind, 2028; Linacre, 2002; Schwarz, 1999). Consequently, categories may be used in overlapping or inconsistent ways, with some options rarely selected or failing to represent progressively higher levels of the construct (eq. Alford, et al., 2025; Al-Qerem et al., 2025; Argo et al., 2021; Colledani et al., 2025; Lu et al., 2025; Provan, et al., 2026; Yamashita, 2022). This indicates that the intended ordering and distinction between categories may not be fully realised in practice.

Importantly, these issues are not limited to self-report contexts. In performance-based assessments, similar challenges arise when external raters evaluate individuals. Variability in ratings is often not solely a matter of severity or leniency but reflects differences in how raters interpret item criteria, task expectations, and category descriptors (Humphry & Heldsinger, 2014; Humphry & Mantuoro, 2023; Myford & Wolfe, 2004). When such elements are not sufficiently explicit, raters may rely on their own internal standards, leading to inconsistencies that parallel those found in self-report data. Thus, both self-reported responses and rater-based evaluations are subject to a common underlying issue: variation in how items and categories are interpreted.

Taken together, these considerations highlight that measurement problems in rating scales are fundamentally linked to how scale components are perceived and used in practice. Variability in interpretation—across respondents as well as raters—can affect both item-level responses and the functioning of response categories. Recognising this shared source of inconsistency provides an important foundation for developing more robust approaches to evaluating and improving rating scale instruments.

Although a substantial body of research has shown that rating scales do not always function as intended, it remains common practice to use them as ready-to-use instruments without prior examination of how well their components operate in practice. In many applications, limited attention is paid to whether respondents interpret and utilise response categories consistently and meaningfully, despite the potential impact of such inconsistencies on measurement quality (DeVellis, 2017; Wind, 2023).

At the same time, the evaluation of rating scales tends to focus predominantly on the instrument level rather than the item level. Analytical approaches such as confirmatory factor analysis (CFA) are widely employed to establish evidence of construct validity by assessing the fit between observed data and a hypothesised factor structure. While these approaches are valuable for examining dimensionality, they offer only limited insight into how individual items function in practice—particularly regarding how response categories are interpreted and used (DeVellis, 2017; Wind, 2023). As a result, potential problems at the level of item responses and category usage may remain undetected, even when overall model fit appears acceptable.

In this context, the Rasch model provides a rigorous measurement framework grounded in the principle of objective measurement, where comparisons between persons are independent of the specific items used, and comparisons between items are independent of the sample of respondents (Andrich, 1988; Rasch, 1960/1980; Wright & Masters, 1982). This invariance property distinguishes the Rasch model from many traditional approaches and enables the construction of interval-level measures from ordinal data. By placing persons and items on a common latent continuum, the Rasch model supports more precise and meaningful interpretation of measurement outcomes.

The Rasch model has developed substantially over the past five decades (e.g. Andrich, 1978, 1982, 1988; Andrich & Marais 2019; Wright & Stone, 1979; Wright & Masters, 1982) and has been widely applied across a range of fields, including education, psychology, health sciences, and the social sciences. (Engelhard & Wind, 2018; Tennant& Küçükdeveci, 2023; Wind, 2023).

However, in the published literature, Rasch analysis is most often reported in relation to validation outcomes, particularly for confirming the adequacy of a scale (e.g., Aldahi et al., 2026; Alnahdi et al., 2021; Deviana, et al., 2020; deSouza Oliveira-Kumakura et al., 2018; Duhn et al., 2021; Dwiliesanti & Yudiarso, 2022; Lee et al., 2023; Syahputra, et al. 2024). The use of the Rasch model as a diagnostic tool—providing substantive information for instrument refinement—is less explicitly documented and therefore less visible in the literature.

This relative lack of emphasis on substantive or diagnostic reporting suggests an opportunity to more clearly articulate how Rasch outputs can be used not only to evaluate instruments but also to deepen understanding of item functioning and guide instrument refinement. For example, when misfitting items are removed without further elaboration, important information about the sources of misfit—such as unclear wording, construct underrepresentation, or inappropriate response formats—may remain underexplored. Consequently, instruments may be simplified rather than substantively refined.

This diagnostic capability is particularly important for rating scales, where measurement quality depends not only on item fit but also on the functioning of response categories. Rasch analysis enables detailed examination of whether response categories operate in an ordered and meaningful manner, whether thresholds are properly structured, and whether respondents can reliably distinguish between adjacent categories (Andrich, 2011; Andrich & Marais, 2019; Linacre, 2002; Wind, 2023).

A more explicit articulation of the diagnostic use of the Rasch model can contribute to both methodological and substantive advancements. It encourages a more comprehensive use of model

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

outputs beyond overall fit statistics and supports a deeper understanding of how constructs are operationalised through items and response categories, particularly in fields that rely on rating scales, such as education, psychology, and health.

This study aims to demonstrate how the Rasch model can be used not only for validation but also as a diagnostic tool to identify and address weaknesses in measurement instruments. In doing so, it seeks to make the diagnostic use of Rasch analysis more visible in the literature and to highlight its potential for supporting systematic instrument refinement in applied research contexts.

Methods

Participants

The dataset consisted of N = 1414 senior high school students. The sample included students from both science (IPA) and social science (IPS) programs of study, as well as a representation of male and female students. The attitude scale data used for the illustrative analyses in this article were drawn from pilot testing conducted as part of the instrument development process at the Centre for Educational Assessment (Pusat Penilaian Pendidikan), and no additional sampling procedures were implemented for the purpose of this study.

These characteristics are reported to provide a general description of the dataset; however, they are not intended to support population-level inferences. The primary purpose of this study is methodological illustration rather than generalisation.

Intruments

The instrument analysed in this study is part of a broader Likert-type attitude scale toward mathematics developed by a team organised by Pusat Penilaian Pendidikan. The full instrument consists of 150 items covering five dimensions: Self-Confidence, Anxiety, Enjoyment, Motivation, and Effort.

For this study, only the motivation dimension was selected for analysis. This dimension comprises 25 items, including 13 favourable and 12 unfavourable items. The selection of a single dimension was intended to provide a focused and manageable illustration of Rasch analysis, particularly in examining item functioning and response category performance.

Responses were collected using a four-point Likert scale: *Strongly Agree, Agree, Disagree,* and *Strongly Disagree*.

Consistent with the methodological purpose of this study, the instrument is used solely as an empirical example to demonstrate the application of the Rasch model.

Data Collection

The data were originally collected in 2011 by the Pusat Penilaian Pendidikan through standard classroom-based administration procedures. As this study relies on secondary data, no additional data collection was conducted by the author.

The dataset is used solely to demonstrate the application of Rasch measurement analysis.

Data Analysis

The data were analysed using the Rasch measurement model to illustrate how rating scale instruments can be evaluated to improve measurement quality.The analysis focused on several key aspects, including: (a) targeting by examining the alignment between item locations and person locations on the latent continuum; (b) reliability, as an indicator of the instrument's ability to distinguish between different levels of the latent construct; (c) item functioning, including fit statistics; and (d) the functioning of response

categories, including threshold ordering and category utilization. The analysis was conducted using RUMM2030 software as a tool to operationalize the Rasch model.

This approach enables the identification of potential measurement issues across multiple aspects of the instrument, particularly those involving interactions among respondents, items, and response categories, which are often not detected by conventional analyses (e.g., Andrich, 2011; Linacre, 2002). In line with the purpose of this study, the analysis is intended to demonstrate the diagnostic capabilities of the Rasch model.

Given the polytomous nature of the data, the analysis was conducted using a Rasch polytomous model (Andrich & Marais, 2019), which allows for the simultaneous examination of targeting, reliability, item functioning, and category functioning within a unified framework. Since the analysis was conducted using RUMM2030, the evaluation criteria adopted in this study are consistent with those operationalised in the software and remain grounded in Rasch measurement principles.

Targeting and reliability

Targeting was evaluated by examining the alignment between item locations and respondents' levels on the latent construct using the person–item distribution. Adequate targeting is indicated when the range and mean of item locations correspond reasonably well to the distribution of respondents, with minimal gaps or clustering along the latent continuum.

Reliability was evaluated using the Person Separation Index (PSI), which reflects the instrument's ability to distinguish between different levels of the latent construct. PSI values of 0.70 or higher were considered acceptable for group-level interpretations, with higher values indicating better discrimination.

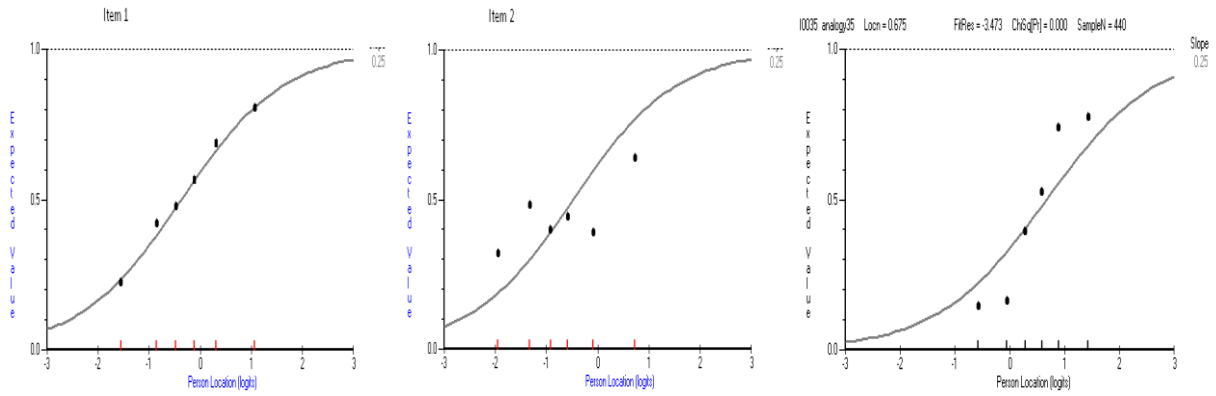
Item functioning

Item functioning was assessed using fit statistics, including residuals and chi-square statistics. Items were considered to exhibit acceptable fit when fit residuals fell within the range of approximately -2.5 to +2.5 and when chi-square values were non-significant after adjustment for multiple testing (Bonferroni adjustment to reduce the type 1 error, divided by the number of items), indicating consistency with model expectations. Items with fit residuals greater than +2.5 indicate underdiscrimination. This suggests that the item may not be functioning consistently across the latent trait continuum, potentially due to issues such as multidimensionality, poorly defined constructs, or ambiguity in item wording. Such items may threaten the validity of measurement and require further investigation. In contrast, items with fit residuals less than -2.5 indicate overdiscriminating. This may indicate redundancy or local dependency with other items. This differs from Classical Test Theory (CTT), where over-discriminating items are generally considered desirable. From a CTT perspective, the presence of more highly discriminating items tends to increase the reliability of the instrument, as reflected in stronger item–total correlations.

Graphically, item functioning is examined using the Item Characteristic Curve (ICC). This involves comparing the observed mean for each group (class interval) with the theoretical mean. Items that fit the model show observed means across all groups that closely align with the theoretical mean.

Items indicating low discrimination are characterised by observed means in the lower groups (low class intervals) that are higher than the theoretical mean, while observed means in the higher groups (high class intervals) are lower than the theoretical mean. In contrast, items indicating overdiscriminating are characterised by observed means in the lower groups that are lower than the theoretical mean, and observed means in the higher groups that are higher than the theoretical mean, but do not distinguish between persons in lower and higher groups.

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



Source: Personal

Figure 1. ICCs of three items indicating fit (left), under discrimination (middle), and over discrimination (right)

Category functioning

Category functioning was evaluated by the extent to which response categories align with the underlying latent construct. This was examined using the thresholds and Category Characteristic Curves (CCCs).

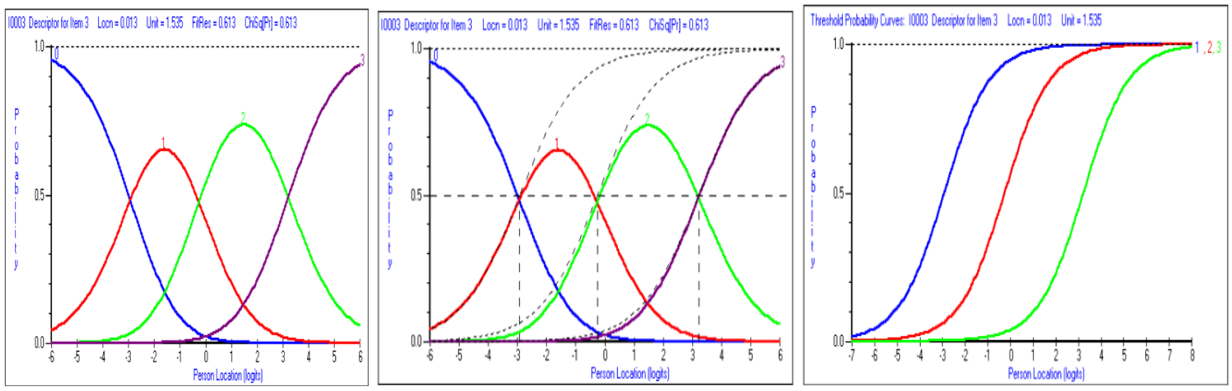
Thresholds represent the points at which the probability of endorsing two adjacent categories is equal (50%). In dichotomous data, the threshold corresponds to item difficulty, indicating the point at which the probabilities of responding 0 and 1 are equal. For items with four response categories, there are three thresholds: threshold 1 is the intersection between categories 1 and 2; threshold 2 is between categories 2 and 3; and threshold 3 is between categories 3 and 4 probability

CCCs display the probability of endorsing each response category. The first category (score 0) shows a monotonic decreasing pattern, whereas the fourth category (score 3) shows a monotonic increasing pattern. The curves for the middle categories (scores 1 and 2) are not monotonic but exhibit a single peak. As the latent trait increases, the probability of selecting a score of 1 or 2 initially increases; however, beyond a certain point, it begins to decrease.

Figure 2 presents the CCCs for four categories (left panel), the three CCC thresholds (middle panel), and the Threshold Probability Curves (TCCs), which show the probability of each threshold (right panel).

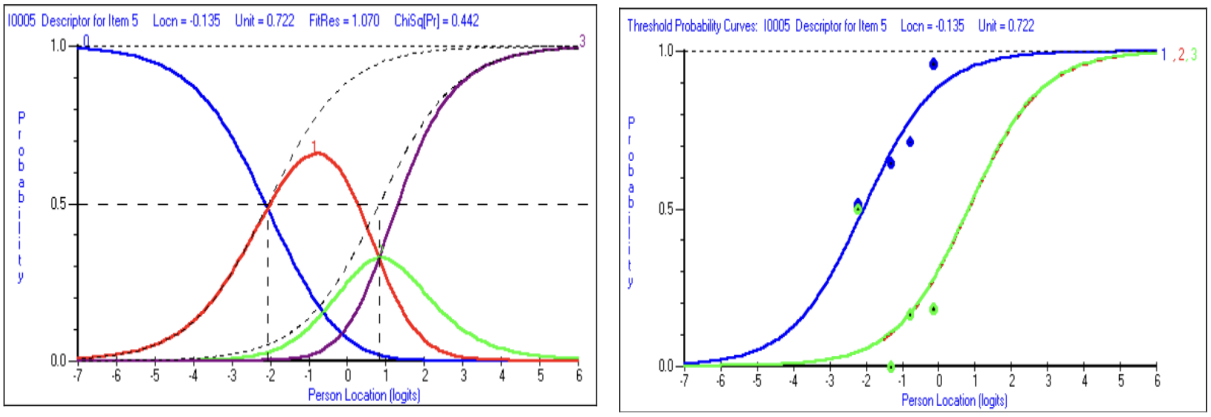
Categories are considered to function properly when the thresholds are ordered. This means that the location of threshold 1 is lower than that of thresholds 2 and 3, and threshold 2 lies between thresholds 1 and 3. Proper functioning can be evaluated based on the ordering of threshold locations. Figure 2 illustrates an example of well-functioning categories, as indicated by the expected ordering of thresholds, with threshold 1 located below thresholds 2 and 3, and threshold 2 positioned between thresholds 1 and 3. The distances between thresholds are also not excessively small.

Categories do not function as expected when thresholds are disordered. Figure 3 presents CCCs and TCCs for an item with disordered thresholds. It can be observed that thresholds 2 and 3 are very close or overlapping. In addition, in the CCC for category 3 (score 2), the probability never reaches a maximum. This indicates that respondents at any level of the latent trait are unlikely to select category 3 or obtain a score of 2. Instead, the most probable responses are categories 2 or 4. This suggests that category 3 is not functioning properly.



Source: Interpreting RUMM2030 Part II., 2019

Figure 2. CCCs and TCCs for polytomous responses with four-category responses and with thresholds ordered



Source: Interpreting RUMM2030 Part II., 2019

Figure 3. CCCs and TCCs for an item with disordered thresholds

Results and Discussion

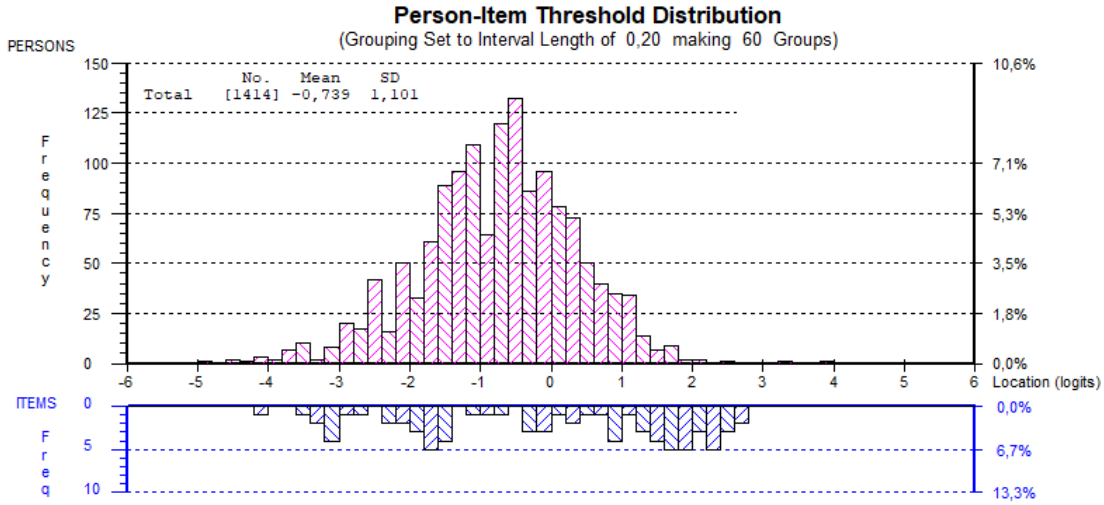
Targeting and Reliability

Targeting was examined by comparing the distribution of person locations and item locations along the latent continuum. The person–item distribution indicated that the items generally covered the range of respondents' positions on the construct; however, some mismatch was observed (see Figure 4). The mean person location was -0.739 logits, indicating that, on average, respondents were located below the centre of the item distribution. This suggests that the items tended to represent relatively higher levels of the construct compared to the respondents' overall positions.

Reliability was assessed using the Person Separation Index (PSI). The PSI value of 0.919 indicates a high level of reliability, suggesting that the instrument can effectively distinguish among respondents with different levels of the underlying construct. This level of reliability supports its use for group-level interpretations and indicates strong internal consistency within the scale.

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

Overall, although the scale is reasonably well targeted with an adequate spread of items, targeting could be further improved by including items that more closely match the distribution of respondents along the construct, particularly those located around the mean person location.



Source: Personal, Result Analysis using RUMM2030 (2026)

Figure 4. Person-Item Threshold Distribution

Item Functioning

Fit statistics

Item functioning was evaluated using fit residuals and chi-square statistics. Most items showed acceptable fit to the model expectations, with fit residuals falling within the acceptable range. However, several items—specifically Items 4, 28, 127, and 147—displayed high positive fit residuals, indicating potential deviation from the underlying construct.

In contrast, a number of items (Items 125, 80, 23, 56, 50, 91, and 88) exhibited relatively large negative fit residuals, suggesting possible redundancy among items capturing very similar aspects of the construct. Despite this indication, examination of residual correlations did not reveal substantial local dependence among these items. The complete item fit statistics are presented in Table 1.

Table 1. Item Fit Statistics

Item	Location	SE	FitResid	DF	ChiSq	DF	Prob
I0004	-0,658	0,036	11,658	1354,23	249,890	9	0,000
I0010	1,263	0,047	2,093	1350,39	25,531	9	0,002
I0016	-1,106	0,044	0,418	1354,23	5,755	9	0,764
I0023	0,494	0,045	-5,684	1352,31	69,583	9	0,000
I0028	0,458	0,043	5,469	1353,27	34,570	9	0,000
I0041	-0,824	0,045	1,408	1354,23	15,118	9	0,088
I0046	0,035	0,042	1,086	1352,31	5,530	9	0,786
I0050	0,422	0,043	-3,842	1344,65	30,000	9	0,000
I0056	0,501	0,045	-5,081	1354,23	74,583	9	0,000
I0061	-0,037	0,041	0,808	1352,31	11,168	9	0,264
I0080	-0,459	0,046	-5,728	1348,48	84,506	9	0,000
I0081	-0,654	0,041	-1,605	1351,35	12,428	9	0,190
I0087	1,035	0,048	0,000	1353,27	18,655	9	0,028
I0088	0,972	0,048	-3,086	1351,35	35,464	9	0,000
I0091	-0,423	0,042	-3,465	1354,23	36,109	9	0,000
I0103	-1,390	0,047	0,367	1352,31	21,380	9	0,011
I0104	0,839	0,048	-0,347	1352,31	15,306	9	0,083
I0111	-0,227	0,045	-3,332	1350,39	28,143	9	0,001
I0115	0,317	0,042	2,895	1332,20	12,905	9	0,167
I0125	-0,299	0,046	-6,026	1325,49	81,046	9	0,000
I0127	-0,030	0,041	9,250	1348,48	122,367	9	0,000
I0133	0,689	0,049	-2,993	1341,77	19,803	9	0,019
I0138	-0,576	0,044	-1,506	1342,73	22,893	9	0,006
I0141	0,167	0,044	-1,262	1343,69	19,988	9	0,018
I0147	-0,511	0,042	6,693	1340,82	125,174	9	0,000

Figure 5 presents the ICCs for four items with very large positive fit residuals. Among these items (Items 4, 127, 147, and 28), Item 4, which exhibits the largest positive fit residual, consistently shows the greatest deviation from the expected ICC.

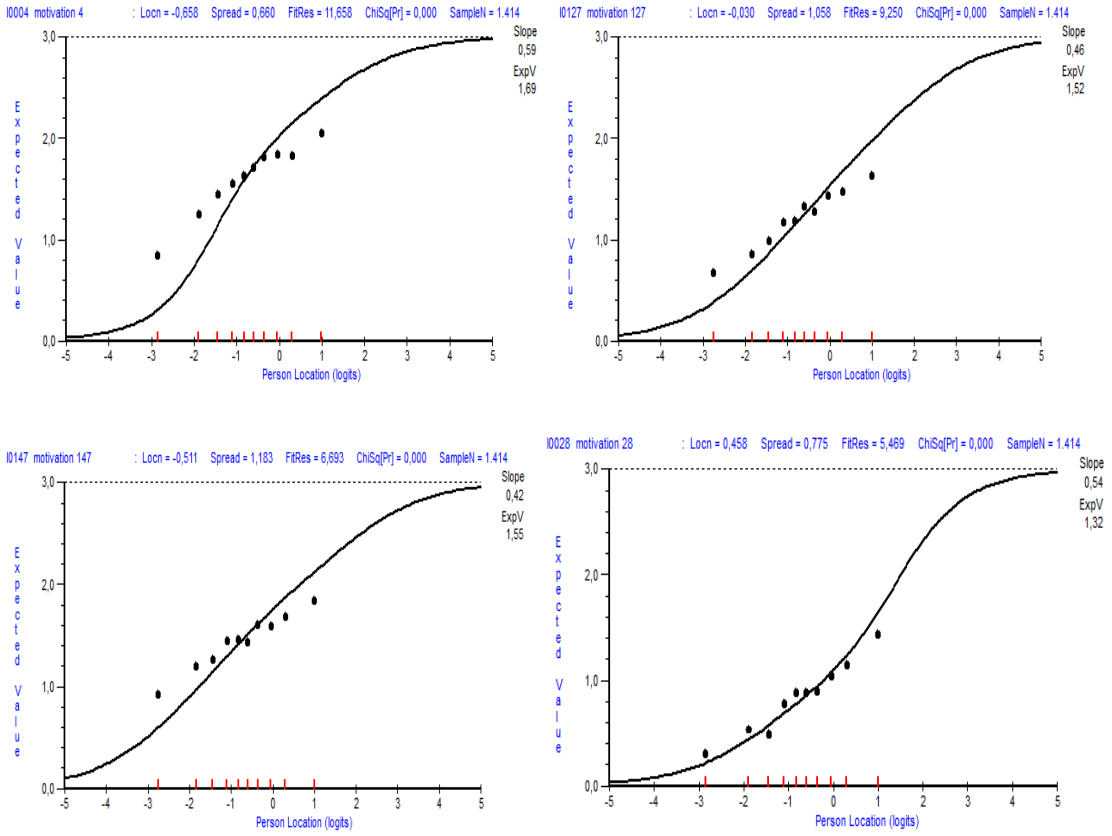


Figure 5. ICCs for the least discriminating Items

To provide a comparison, Figure 6 presents two items that fit the model well (Items 46 and 16), with fit residuals of 1.06 and 0.418, respectively. For these items, the observed means closely align with the theoretical mean, indicating good model-data fit.

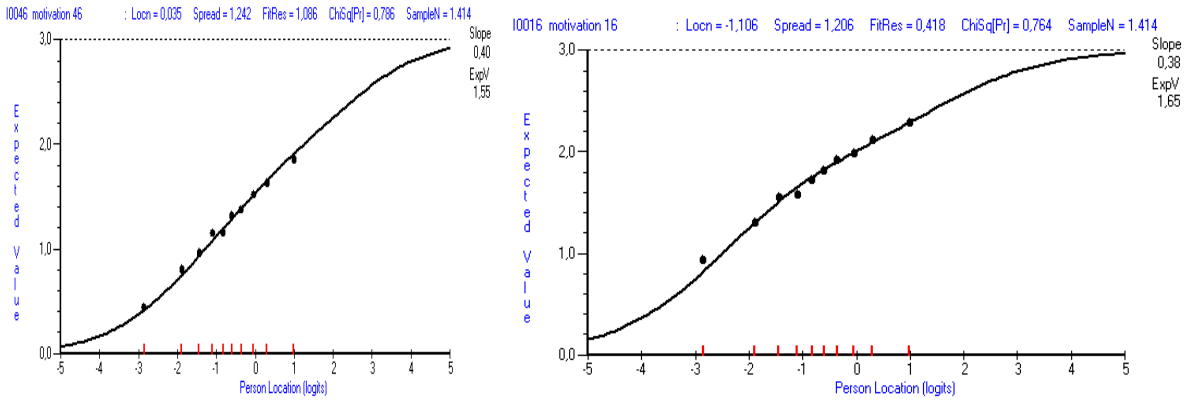


Figure 6. ICCs for Well-Fitting Items

Figure 7 illustrates examples of overdiscriminating items, specifically Item 125 (fit residual = -6.026) and Item 80 (fit residual = -5.728). The observed means for the lower groups tend to fall below the model expectations, whereas those for the higher groups exceed them.

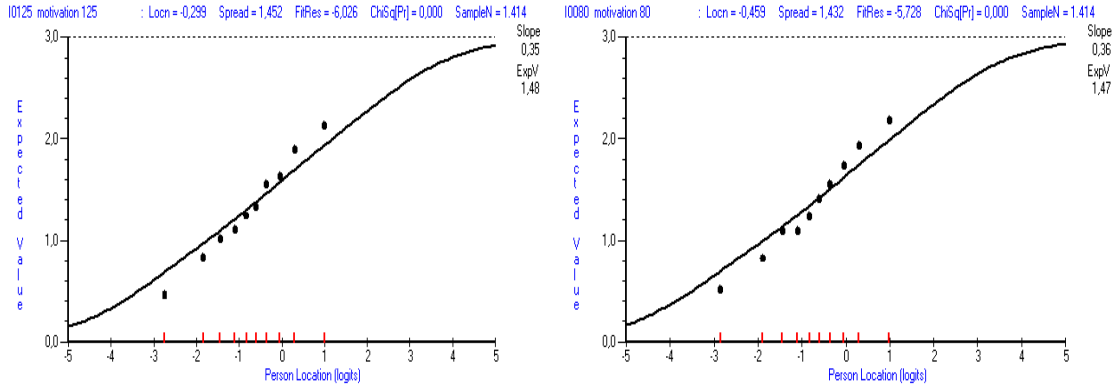


Figure 7. ICCs for Overdiscriminating Items

Substantive Diagnosis Based on Item Wording

To gain further insight into items that do not conform to the model, the wording of the items was reviewed. The wording of the four least discriminating items is presented in Table 2, whereas Table 3 presents the wording of the seven overdiscriminating items.

It was observed that the 4 least discriminating items tend to reflect behaviours or external conditions rather than the intended construct of motivation toward mathematics. For example, Item 4 appears to capture a specific behavioural tendency (i.e., sitting in the front row during mathematics class) rather than an evaluative attitude toward mathematics. Such behaviour is likely influenced by situational factors (e.g., classroom conditions, seating availability), making it inconsistent with the underlying latent trait. This misalignment likely explains the high residual.

Item 127 appeared to confound students' motivation to learn mathematics with their access to external resources, such as private tutoring. As a result, responses to this item may reflect opportunity or socioeconomic factors rather than the intended attitudinal construct.

Item 147 represented a behavioural or resource-based indicator (i.e., ownership of additional mathematics books) rather than an attitude. Similar to Item 4, this item is influenced by external conditions, such as resource availability, and therefore does not align well with the latent construct being measured.

Item 28, which was negatively worded, may have introduced additional cognitive complexity due to its double-negative structure. This can lead to inconsistent interpretation among respondents and is a well-documented source of misfit in Rasch analysis, often referred to as a method effect.

Meanwhile, for overdiscriminating items, it appears to express highly similar semantic content, often using the same key terms (e.g., "enthusiastic") across slightly different contexts (see Table 3). This redundancy suggests that the items capture nearly identical aspects of the construct, leading to overly predictable response patterns and limited additional information from the measurement.

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

Table 2. Wording of Least Discriminating Items

No item	Item Statement
4	I always sit in the front row during mathematics class *(Saya selalu duduk di barisan paling depan pada saat jam pelajaran matematika*)
127	I want to deepen my understanding of mathematics by attending extra lessons *(Saya ingin mendalami pelajaran matematika dengan mengikuti les)*
147	I have additional books for studying mathematics besides the textbook used at school. *(Saya memiliki buku-buku lain selain buku pelajaran matematika yang digunakan di sekolah)*
28	Poor mathematics grades do not discourage me from studying mathematics. *(Nilai matematika yang jelek tidak akan membuat saya malas belajar matematika)*

Table 3. Wording of Overdiscriminating Items

No item	Item Statement
125	I feel enthusiastic when studying mathematics. (*Saya merasa bersemangat saat belajar matematika*)
80	I am enthusiastic about doing mathematics homework. (*Saya bersemangat mengerjakan PR matematika*)
23	I am not enthusiastic when I must do mathematics homework. (*Saya tidak bersemangat ketika harus mengerjakan PR pelajaran matematika*)
56	I am not enthusiastic about attending mathematics classes. (*Saya tidak bersemangat mengikuti pelajaran matematika*)
50	Whenever I am about to attend a mathematics class at school, my motivation immediately decreases.

	Setiap akan menghadapi pelajaran matematika di sekolah, semangat saya langsung turun
91	I would rather attend other classes than mathematics at school. (*Lebih baik mengikuti pelajaran lain daripada pelajaran matematika di sekolah*)
88	Bagi saya mengerjakan soal matematika hanya membuang waktu. *(For me, working on mathematics problems is just a waste of time)*

Category Functioning

Threshold Ordering

The results indicate that several items exhibited disordered thresholds, specifically Items 4, 10, 87, and 88 (see Table 4). Disordered thresholds suggest that respondents did not use the response categories in a consistent and ordered manner. Among these items, item 4 is also the least discriminating, as indicated by the item fit.

In a properly functioning rating scale, thresholds between adjacent categories are expected to increase monotonically, reflecting increasing levels of the latent trait. However, for these items, the thresholds do not follow the expected order.

Table 4. Category Thresholds Estimate for Each Item

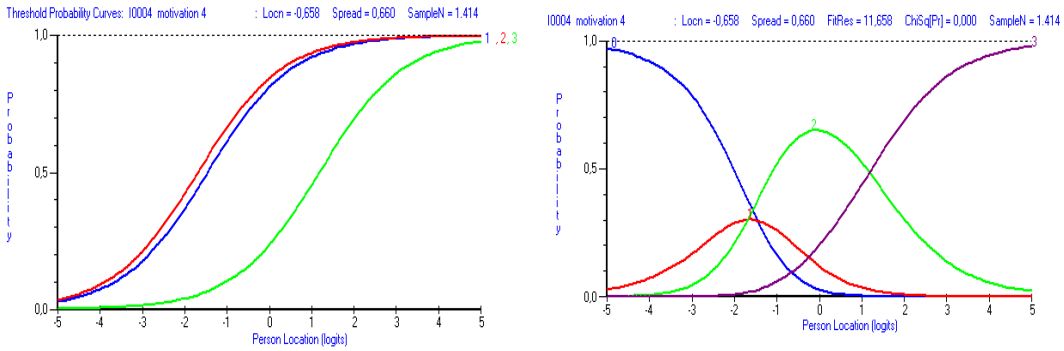
Item	Location	Mean	UnCThr 1	UnCThr 2	UnCThr 3
I0004	-0,66	-0,66	**-1,46**	**-1,69**	**1,18**
I0010	1,26	1,26	**0,23**	**1,84**	**1,72**
I0016	-1,11	-1,11	-3,19	-1,77	1,64
I0023	0,49	0,49	-1,99	0,76	2,71
I0028	0,46	0,46	-1,55	1,38	1,54
I0041	-0,82	-0,82	-3,05	-1,55	2,13
I0046	0,04	0,04	-2,34	-0,19	2,63
I0050	0,42	0,42	-1,76	0,97	2,05
I0056	0,50	0,50	-1,95	0,98	2,47
I0061	-0,04	-0,04	-2,24	0,20	1,93
I0080	-0,46	-0,46	-3,39	-0,33	2,34
I0081	-0,65	-0,65	-3,08	-0,22	1,33

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

I0087	1,04	1,04	**-0,89**	**2,37**	**1,62**
I0088	0,97	0,97	**-1,15**	**2,22**	**1,85**
I0091	-0,42	-0,42	-2,97	0,45	1,25
I0103	-1,39	-1,39	-4,13	-1,62	1,58
I0104	0,84	0,84	-1,59	1,89	2,21
I0111	-0,23	-0,23	-3,08	-0,16	2,55
I0115	0,32	0,32	-1,67	1,00	1,62
I0125	-0,30	-0,30	-3,24	-0,23	2,57
I0127	-0,03	-0,03	-2,10	-0,12	2,13
I0133	0,69	0,69	-1,92	1,59	2,40
I0138	-0,58	-0,58	-3,42	0,15	1,55
I0141	0,17	0,17	-2,20	0,93	1,77
I0147	-0,51	-0,51	-2,77	-0,73	1,97

Graphical Analysis of Category Functioning

To further examine category functioning, Threshold Characteristic Curves (TCCs) and Category Characteristic Curves (CCCs) were presented. Figure 8 displays TCCs and CCCs for items with disordered thresholds.



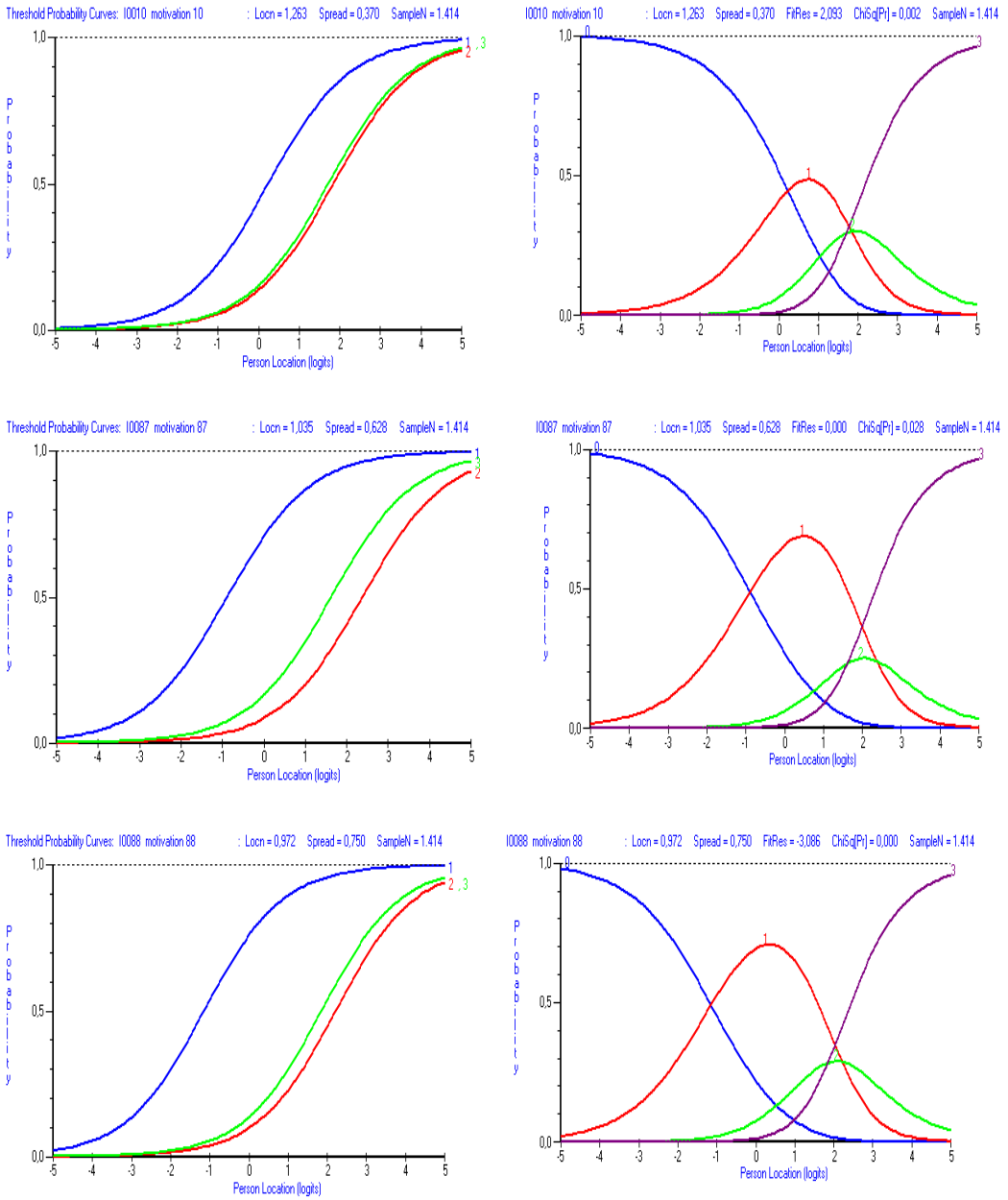


Figure 8. TCCs and CCCs for Items with Disordered Threshold

Figure 8 shows that for Item 4, Threshold 1 and Threshold 2 are nearly overlapping, and the CCC indicates a low probability that respondents will endorse Category 2. This suggests that the category is not functioning as intended. In contrast, for Items 10, 87, and 88, Thresholds 3 and 4 are closely spaced, with Category 3 exhibiting suboptimal functioning.

Substantive Diagnosis Based on Item Wording

To better understand the items with disordered thresholds, the wording of the affected items was examined. Item 4, as discussed previously, reflects habitual behaviour ("I always sit in the front row during mathematics class"), which may not vary systematically across levels of motivation. Item 10 expresses avoidance behaviour ("I would like to skip mathematics class whenever possible"), which may elicit extreme or polarised responses rather than gradual variation. Item 87 represents a general belief ("I believe that learning mathematics is beneficial") that may be endorsed by most respondents, thereby limiting category differentiation. Item 88, a negatively worded statement ("Working on mathematics

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

problems is a waste of time"), may introduce additional cognitive processing, leading to inconsistent use of response categories.

These findings suggest that respondents may experience difficulty mapping their perceptions onto the available response categories when items are behaviourally specific, extreme in tone, or cognitively complex. As a result, the intended ordinal structure of the rating scale may be weakened.

Table 5: Wording of Items with Disordered Thresholds

No item	Item Statement
4	I always sit in the front row during mathematics class *(Saya selalu duduk di barisan paling depan pada saat jam pelajaran matematika*)
10	I would like to skip mathematics class whenever possible. *(Saya berharap bisa membolos saat ada kelas matematika)*
87	I believe that learning mathematics is beneficial. *(Saya yakin belajar matematika adalah hal yang bermanfaat)*
88	Working on mathematics problems is a waste of time. *(Mengerjakan soal matematika membuang waktu)*

Item Refinement: Removal and Rescoring

Following the analysis, instrument refinement can be undertaken through item revision or item removal. Based on the diagnostic results, several refinements were identified. Items 4, 28, 127, and 147 were flagged for substantial misfit, suggesting they may not align well with the primary construct being measured. These items were therefore identified as candidates for revision or removal.

In addition, items with high negative fit residuals were examined for potential redundancy. Although these items did not exhibit problematic residual correlations, content similarities suggested that some items might be redundant and could be reduced to improve instrument efficiency. For example, Items 125, 80, 56, and 23 all use similar wording (e.g., "enthusiastic") within closely related contexts. Among these items, it may be sufficient to retain only one while revising or removing the others.

For items exhibiting disordered thresholds (Items 10, 87, and 88), revision should be considered to ensure that each response category functions as intended. When items are confusing or overly extreme, respondents may have difficulty distinguishing between categories or may respond inconsistently across them.

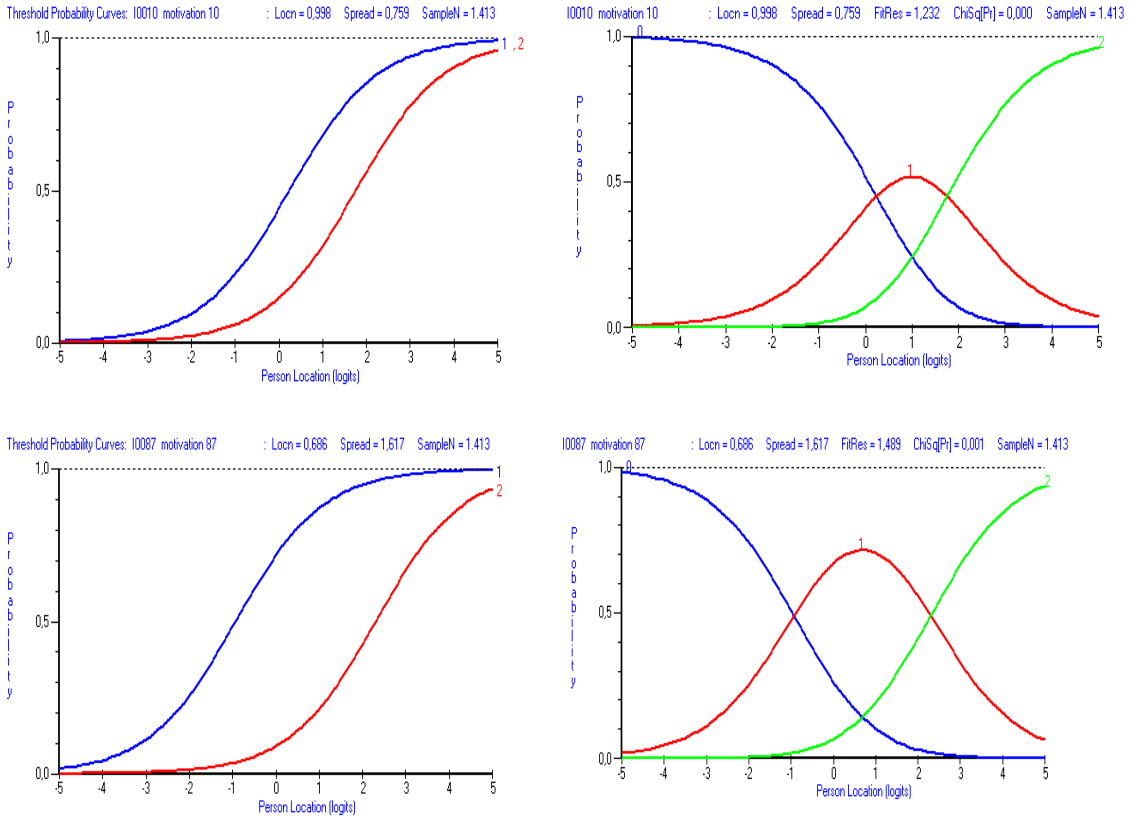
In situations where item revision is not feasible, analysis may proceed by removing items that do not align with the construct and rescoring items with disordered thresholds.

In this study, the analysis examined the impact of removing misfitting items and rescoring items with disordered thresholds, particularly on the Person Separation Index (PSI) and threshold ordering.

Re-analysis After Item Removal and Rescoring

A re-analysis was conducted by removing the four least discriminating items (Items 4, 127, 147, and 28), excluding three overdiscriminating items (Items 125, 80, and 23), and rescoring items with disordered thresholds (Items 10, 87, and 88). The results showed that the Person Separation Index (PSI) remained high, although it decreased slightly from 0.907 to 0.886. This decrease is understandable, given that three overdiscriminating items were excluded from the analysis. For comparison, when the three overdiscriminating items were retained—while still removing the four misfitting items and rescoring the three items with disordered thresholds—the PSI increased slightly from 0.907 to 0.910.

Importantly, rescoring response categories yielded ordered thresholds for previously problematic items. The category probability curves showed clearer separation between adjacent categories, indicating improved functioning of the response scale (Figure 9).



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

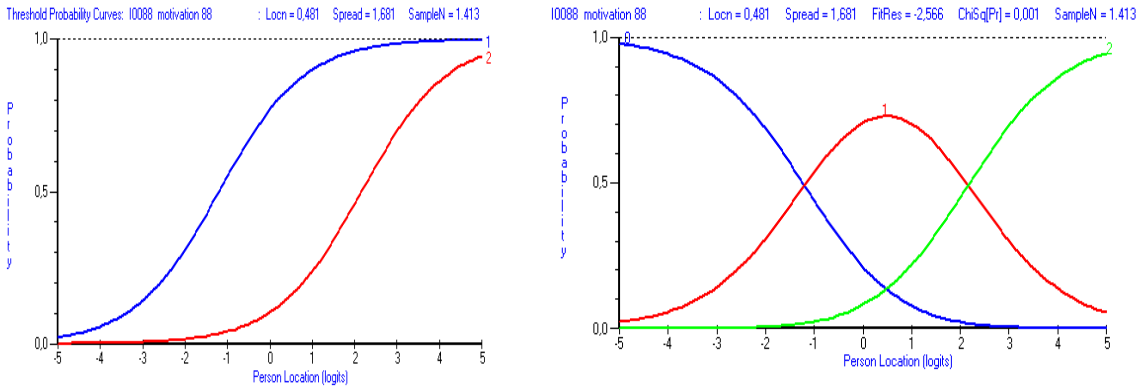


Figure 9. TCCs and CCCs for Three Items with Categories Collapsed

Discussion

This study focuses on two aspects of measurement quality within the Rasch framework—item functioning and category functioning—while treating targeting and reliability as indicators of the scale's overall performance. The results show that, although the instrument demonstrates good reliability and reasonably adequate targeting, more meaningful insights emerge from a detailed examination of how individual items and response categories function.

At the item level, the findings highlight the diagnostic value of Rasch fit statistics in distinguishing between different types of misfit. Several items (Items 4, 28, 127, and 147) exhibited large positive fit residuals, indicating under-discrimination and weak alignment with the latent construct. Substantive examination of item wording suggests that these items do not primarily capture the intended construct of motivation toward mathematics, but instead reflect behavioural tendencies or external conditions. For example, items related to classroom seating, access to tutoring, or ownership of learning resources are likely influenced by situational or socioeconomic factors. This supports the interpretation that item misfit in this case reflects construct-irrelevant variance rather than random error.

In contrast, items with large negative fit residuals indicate over-discrimination, a phenomenon that is particularly informative within the Rasch framework. While such items might appear desirable from a Classical Test Theory perspective, Rasch analysis reveals that they violate the model. The observed pattern—where several over-discriminating items share highly similar wording (e.g., repeated use of "enthusiastic")—suggests redundancy in content. These items provide only limited additional information and may reduce the scale's efficiency. Importantly, this finding illustrates that good statistical fit in conventional terms does not necessarily imply optimal measurement quality, reinforcing the added value of Rasch analysis as a diagnostic approach.

Graphical analyses using Item Characteristic Curves (ICCs) further support these interpretations. Misfitting items show clear deviations from model expectations, while well-fitting items demonstrate close alignment between observed and expected responses. Over-discriminating items exhibit steeper-than-expected curves, reflecting overly predictable response patterns. Together, these results demonstrate how statistical and graphical evidence can be combined to provide a more nuanced understanding of item functioning.

At the category level, the presence of disordered thresholds in several items indicates that respondents experienced difficulty using the response scale in an ordered and consistent manner. This issue is further

illustrated by the Category Characteristic Curves (CCCs), which show overlapping response categories and limited differentiation between adjacent options. Notably, one of the items with disordered thresholds (Item 4) was also identified as misfitting at the item level, suggesting that problems in item functioning and category functioning may be interrelated.

Substantive analysis of item wording provides additional insight into category dysfunction. Items that are behaviourally specific, extreme in tone, or cognitively complex—such as negatively worded statements—appear to hinder respondents' ability to map their perceptions onto the available response categories. For example, avoidance-oriented or strongly worded items may elicit polarised responses, reducing the effective use of intermediate categories. Similarly, general or widely endorsed statements may limit response variability, leading to poorly differentiated categories. These findings suggest that category dysfunction is not solely a property of the rating scale format, but arises from the interaction between item content and response structure.

The improvement observed after rescoring categories and removing or revising problematic items further supports this interpretation. The restoration of ordered thresholds and clearer separation between categories indicates that relatively simple modifications—such as collapsing categories or refining item wording—can substantially improve measurement quality. At the same time, the slight changes in reliability following item removal highlight the trade-off between scale length and measurement precision, particularly when over-discriminating (but redundant) items are excluded.

Although targeting and reliability indicate that the scale functions well overall, these indices alone would not reveal the specific sources of measurement problems identified in this study. This underscores the importance of going beyond global fit and reliability indices to examine item- and category-level functioning in detail.

Overall, the findings demonstrate that Rasch analysis provides a powerful framework for diagnosing measurement issues in rating scale instruments. By integrating statistical indicators, graphical analysis, and substantive evaluation of item content, researchers can identify not only whether an instrument works, but also how and why specific components may fail to function as intended. This supports a more iterative and evidence-based approach to instrument refinement, in which both item quality and response category design are systematically evaluated and improved.

Conclusion

This study demonstrates that the Rasch model provides a powerful diagnostic framework for evaluating rating-scale instruments, particularly at the levels of item and category functioning. While overall indicators such as targeting and reliability suggest that the scale performs adequately, detailed Rasch analysis reveals important measurement issues that would not be evident from global indices alone.

At the item level, the findings show that both under- and over-discriminating items require attention. Misfitting items were found to reflect construct-irrelevant content, such as behavioural tendencies or external conditions, whereas over-discriminating items indicated redundancy in item wording. At the category level, disordered thresholds and overlapping response categories highlight difficulties respondents face in using the rating scale as intended. These issues were shown to arise not only from the response format but also from the interaction between category structure and item characteristics.

Importantly, the study illustrates that such measurement problems can be addressed through targeted refinement, including item revision, removal of problematic or redundant items, and rescoring of response categories. These modifications led to improved category functioning while maintaining acceptable reliability, thereby supporting a more efficient and conceptually coherent instrument.

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

Overall, the findings reinforce the value of using Rasch analysis not only for validation but also as a diagnostic tool to guide instrument refinement. By integrating statistical evidence with substantive evaluation of item content, researchers can enhance both the quality and interpretability of rating scale measurements.

Acknowledgment

The attitude scale data used for the illustrative analyses in this article were drawn from pilot testing conducted as part of the instrument development process at the Pusat Penilaian Pendidikan. The author gratefully acknowledges the opportunity to use these data.

Conflict of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

References

Aldhahi, M. I., Tesio, L., Scarano, S., Bakhsh, H. R., Alhasani, R., bin Sheeha, B. H., & Caronni, A. (2026). Comparative psychometric evaluation of the Arabic version of four patient-reported outcome measures for sleep assessment: a construct validity study using Rasch analysis. *Sleep and Breathing*, 30(1). https://doi.org/10.1007/s11325-025-03554-2

Alnahdi, G. H., Goldan, J., & Schwab, S. (2021). Psychometric properties and Rasch validation of the teachers' version of the perception of resources questionnaire. *Frontiers in psychology*, *12*, 633801. https://doi.org/10.3389/fpsyg.2021.633801

Andrich D. (1988). *Rasch models for measurement*. Sage.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigm. *Medical Care, 42*(1), 7-16.

Andrich D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research*, 11(5), 571–585. https://doi.org/10.1586/erp.11.59

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences.* Springer.

Alford, A. J., Casteleijn, D., & Robertson, L. J. (2025). Brief psychiatric rating scale – expanded version: construct validity using Rasch model analysis. *South African Journal of Psychiatry*, 31.a2343. https://doi.org/10.4102/sajpsychiatry.v31i0.2343

Al-Qerem, W., Basem, D., Khdair, S., Jarab, A., & Eberhardt, J. (2025). Validation of the Arabic version of the Multiple Sclerosis Impact Scale (MSIS-29): A Rasch analysis study. *Archives of Clinical Neuropsychology*, 40(3), 520-528. https://doi.org/10.1093/arclin/acae121

Argo, A. R. B., Yulianto, H., & Nuryanto, D. (2021). Evaluating psychometric properties of the stress measurement instrument (the Operational and Organizational Police Stress Questionnaires) with the application of Rasch Model in the Indonesian Nasional Police (INP). JP3I (*Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia)*, 10 (1), 39-59. https://doi.org/10.15408/jp3i.v10i1.17557

Colledani, D., González Pizzio, A. P., Devita, M., & Anselmi, P. (2025). Investigating the functioning of rating scales with Rasch models. *Assessment*, 32(3). https://doi.org/10.1177/10731911241245792

DeSouza Oliveira-Kumakura, A.R., Caldeira , S., Prado Simão, T., Camargo-Figuera ,F.A., de Almeida Lopes Monteiro da Cruz, D., & Campos de Carvalho, E. (2018). The contribution of the Rasch model to the clinical validation of nursing diagnoses: Integrative literature review. *International Journal of Nursing Knowledge*, 29(2), 89-96. doi:10.1111/2047-3095.12162

DeVellis, R.F. (2017). *Scale development: Theory and applications*. Sage.

Deviana, T., Hayat, B., & Tresniasari, N. (2020). Female Hedonistic Behavior Questionnaire (FHBQ): Psychometric properties based on the Rasch model. JP3I (*Jurnal Pengukuran Psikologi dan Pendidikan Indonesia*), 9(2), 44-56.

Duhn, P. H., Amris, K., Bliddal, H., & Wæhrens, E. E. (2023). The validity of the Danish version of the Fibromyalgia Impact Questionnaire–revised applied in a clinical setting: a Rasch analysis. *Scandinavian Journal of Rheumatology*, 52(4). https://doi.org/10.1080/03009742.2022.2098631

Dwiliesanti, W.G. & Yudiarso, A. (2022). Rasch analysis of the Indonesian version of Individual Work Performance Questionnaire (IWPQ). JP3I (*Jurnal Pengukuran Psikologi dan Pendidikan Indonesia).* 11 (2), 153-167.

Engelhard,Jr, G. & Wind, S.A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge.

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher,* 43(5), 253–263. https://doi.org/10.3102/0013189X1454215

Humphry, S., & Montuoro, P. (2023). Judging similarity versus judging difference. *Frontiers in Education*, 8.1117410. https://doi.org/10.3389/feduc.2023.1117410

Lee, S., Yun, H.-J., Jeon, M., & Kang, M. (2023). Validating athletes' subjective performance scale: A Rasch model analysis. IJASS (*International Journal of Applied Sports Sciences*), 35 (2), 238-250. https://doi.org/10.24985/ijass.2023.35.2.238

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22 140,* 55.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3(1), 85–106.

Lu, Y. M., Wu, Y. Y., & Lue, Y. J. (2025). Rasch Analysis of the QuickDASH in patients with neck pain. *Journal of Clinical Medicine*, 14.1870. https://doi.org/10.3390/jcm14061870

Myford, C. & Wolfe, E.W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement : Part 1. *Journal of Applied Measurement, 4* (4), 386–422.

Piussi, R., Thomeé, R., Samuelsson, K., & Hamrin Senorski, E. (2025). Measurement properties of the Swedish version of the Knee Self-Efficacy Scale: A Rasch analysis. *Journal of Experimental Orthopaedics,* 12 (2) e70306. https://doi.org/10.1002/jeo2.70306

Provan, S. A., Berner-Hammer, H., & Kleppang, A. L. (2026). Psychometric evaluation of the Norwegian version of the revised Fibromyalgia Impact Questionnaire. *Scandinavian Journal of Rheumatology*, 55 (2), 143-150. https://doi.org/10.1080/03009742.2025.2573532

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. The University of Chicago Press.

Syahputra, W., Widhiastuti, I., Baydhowi, Falah, S., Yundianto, D., Ali, M.M. (2024). Examining the psychometric properties of the prosocial behavior scale using Indonesian pesantren (Islamic boarding education system) sample. JP3I (*Jurnal Pengukuran Psikologi dan Pendidikan Indonesia),* 13 (2), 181-199.

Schwarz, N. (1999). Self-Reports: How the questions shape the answer. *American Psychologist*. 54 (2), 93-105.

Tennant, A., & Küçükdeveci, A. A. (2023). Application of the Rasch measurement model in rehabilitation research and practice: early developments, current practice, and future challenges. *Frontiers in rehabilitation sciences*, *4*, 1208670. https://doi.org/10.3389/fresc.2023.1208670

Tourangeau, R., Rips, L.J, Rasinski, K. (2000). *The Psychology of survey response*. Cambridge University Press.

Wind, S. A. (2023). *Exploring Rating Scale Functioning for Survey Research.* Sage.

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

Wright, B. D., & Stone, M. K. (1979). *Best Test Design: Rasch Measurement*. MESA Press.

Yamashita, T. (2022)**.** Analyzing Likert scale surveys with Rasch models. *Research Methods in Applied Linguistics.* 1 (3), 100022. https://doi.org/10.1016/j.rmal.2022.100022