Automated System for Objective Assessment of Educational Achievements Using Generative Artificial Intelligence

Zilola Qurbonova[1], Feruzakhon Ramazonova[1], Shokhsanam Odiljonova[1], Gulrux Alimova[1], Kanybek Isakov[2], Sanjarbek Jamoldinov[3], Vazira Jalalova[4], Sitoraxon Rakhmanbekova[5], Nigora Abdurakhimova[6], Charoz Orziqulova[5], Zilola Abidova[7] and Dildora Baxadirova[7]

[1]*University of Tashkent for Applied Sciences, Gavhar Str. 1, 100149 Tashkent, Uzbekistan*

[2]*Osh Technological University, Isanov Str. 81, 723503 Osh, Kyrgyzstan*

[3]*Andijan State Institute of Foreign Languages, 170100 Andijan, Uzbekistan*

[4]*Department of Clinical Pharmacology, Bukhara State Medical Institute named after Abu Ali ibn Sino, Gijduvan Str. 23, 200126 Bukhara, Uzbekistan*

[5]*Renaissance Educational University, 100070 Tashkent, Uzbekistan*

[6]*Department of Philology and History, Renaissance Educational University, 100070 Tashkent, Uzbekistan*

[7]*Tashkent State University of Oriental Studies, Amir Temur Avenue 20, 100060 Tashkent, Uzbekistan*

Sanjarbek7587@gmail.com, jalalova.vazira@bsmi.uz, rahmonbekovas@gmail.com, n_abdiraximova@renessans-edu.uz, ch_orzikulova@renessans-edu.uz, zilolaqurbonova63@gmail.com, feruzaramazonova@yandex.ru, shkarimjonova@icloud.com, Gulruxmirzaeva748@gmail.com, charmingzilola@gmail.com, donbulak14@mail.ru

Keywords: Generative Artificial Intelligence, Automated Assessment, Rasch Model, Reliability, Validity, Item Response Theory, Differential Item Functioning, Educational Psychometrics.

Abstract: In the context of the systemic digitalization of education, Automated Scoring Systems (ASS) based on Generative Artificial Intelligence (Generative AI) are becoming an important instrument for scalable and objective assessment of educational achievements. Despite the high efficiency and operational flexibility of such systems, the fundamental psychometric issues of their reliability and validity remain insufficiently explored. Of particular importance is the transition from Classical Test Theory (CTT) to models of modern measurement theory, specifically Item Response Theory (IRT), and particularly the Rasch model, which ensures parameter invariance and metric comparability of measurements. The purpose of this study is to conduct a comprehensive evaluation of the reliability and validity of an automated scoring system based on a generative language model using the Rasch model as an analytical framework. The study applies the one-parameter logistic Rasch model to analyze item characteristics, the distribution of examinee abilities, fit statistics, separation indices, and the detection of Differential Item Functioning (DIF). The methodology involves simulation of a dataset consisting of 1,200 students and 40 open-ended items automatically evaluated by a generative model. The analysis is performed using maximum likelihood estimation procedures and the calculation of Infit and Outfit statistics. In addition, convergent and criterion validity indicators are assessed through comparison between automated scores and expert evaluations. The obtained results demonstrate high system reliability (Person Reliability > 0.88), satisfactory item fit to the Rasch model, and the absence of systematic bias related to gender or level of academic preparation. The theoretical and practical implications of integrating generative AI into high-stakes assessment systems are discussed. This study contributes to the advancement of psychometric evaluation of AI-based assessment systems and provides a scientific foundation for their further standardization.

1 INTRODUCTION

1.1 Digital Transformation of Educational Assessment

In the context of global societal development, educational systems have undergone profound transformations associated with the digitalization of learning processes and the integration of intelligent technologies into knowledge assessment procedures [1]. One of the most rapidly evolving areas is automated assessment, which includes automated grading of both test-based and open-ended responses, item generation, and adaptive testing environments [2].

The emergence of Large Language Models (LLMs), capable of analyzing, interpreting, and generating text with a high level of semantic coherence, has significantly expanded the capabilities of automated scoring systems [3]. Such models enable the evaluation of essays, extended responses, and even complex project-based assignments, tasks that previously required the involvement of human experts [4], [5].

However, as generative AI becomes increasingly integrated into educational assessment systems, a fundamental question arises: Do these systems meet rigorous psychometric standards of reliability and validity?

1.2 Reliability Issues in AI-Based Scoring Systems

Reliability is traditionally defined as the degree of stability and reproducibility of measurement results [5]. In the context of automated assessment systems, reliability acquires a multidimensional character, including:

- stability of algorithmic outputs;
- robustness to variations in response formulation;
- consistency with expert scoring;
- invariance of item parameters.

Classical Test Theory (CTT) conceptualizes reliability as a function of the variance of true scores and measurement error [6]. However, CTT parameters depend on a specific sample, which limits its applicability in dynamic AI-based systems [7].

In contrast, the Rasch model provides several advantages:

- invariance of item parameters across samples;
- a linear measurement scale;
- comparability of results across different test versions [8].

1.3 Validity and Construct Interpretation of Measurement

Validity is defined as the degree to which interpretations of test results are justified and supported by evidence [9]. In the context of AI-based assessment systems, several new challenges emerge:

1) Construct validity - does the system measure the intended latent construct?
2) Criterion validity - do the results correlate with external indicators of performance?
3) Content validity - does the model adequately interpret the semantic meaning of responses?

Contemporary standards of educational testing emphasize the necessity of a multi-component validity framework [10]. When generative AI is employed, an additional concern arises: the potential risk of algorithmic bias in automated scoring processes.

1.4 Theoretical Foundations of the Rasch Model

The Rasch model represents a one-parameter logistic model:

$$P(X_{ni} = 1) = \frac{e^{(\theta_n - b_i)}}{1 + e^{(\theta_n - b_i)}}. \quad (1)$$

Where:

- θ_n - ability level of the examinee
- b_i - difficulty parameter of the item.

A key advantage of the Rasch model is the principle of specific objectivity, which implies that comparisons between individuals are independent of the particular set of items used, and vice versa [11]-[13].

1.5 Conceptual Framework of the Study

Figure 1 illustrates the conceptual architecture for evaluating the reliability of an automated scoring system based on generative artificial intelligence using the Rasch model as a psychometric calibration mechanism.

The first functional component - Generative AI Scoring Engine - represents the module responsible for automatic interpretation and evaluation of students' textual responses. At this stage, semantic analysis of the input text is performed, features are extracted, responses are aligned with scoring rubrics, and an initial score is generated. This process is characterized by high computational speed and scalability; however, it does not inherently guarantee metric invariance or measurement stability [1].

The second and third zones represent the transition from algorithmic scoring to psychometric verification. The Psychometric Calibration Layer performs calibration of the obtained scores within the framework of the one-parameter logistic Rasch model, enabling the estimation of item difficulty parameters and examinee ability levels on the logit scale. At this stage, model-data fit statistics (Infit and

Outfit), separation indices (Person Separation Index), and analyses of Differential Item Functioning (DIF) are calculated.

The final component - Rasch Analysis Output - produces validated metric indicators that allow the results to be interpreted as invariant measurements of a latent construct. Thus, the proposed framework demonstrates the integration of generative artificial intelligence with modern measurement theory, facilitating the transition from automated scoring procedures to scientifically grounded psychometric measurement [2].

Figure 2 presents a conceptual comparison between Classical Test Theory (CTT) and the Rasch model as two methodological paradigms of psychometric measurement.

The left column outlines the key characteristics of CTT, including the dependence of statistical indicators on a specific sample of respondents, the interpretation of the total score as the primary measurement metric, and the use of the standard error of measurement (SEM) within a linear scale. Within the CTT framework, test parameters and reliability significantly vary depending on the composition of the sample, which limits the comparability of results across different groups and test versions [3], [14].

This approach historically played an important role in the development of educational measurement. However, its theoretical limitations become particularly evident when analyzing automated assessment systems, where a high level of metric stability is required.

The right column illustrates the principles of the Rasch model, which is grounded in modern Item Response Theory (IRT). In contrast to CTT, the Rasch model ensures invariance of item parameters and examinee abilities, translating results into a logit scale that possesses the properties of interval measurement.

Parametric calibration allows the difficulty of items to be estimated independently of the specific sample, while the use of fit statistics (Infit and Outfit) provides strict control over measurement quality. Furthermore, the Rasch model supports equivalence across different test forms and underlies adaptive testing, making it particularly relevant for analyzing the reliability and validity of automated assessment systems based on generative artificial intelligence [4].

1.6 Research Gap

Despite extensive research in the field of automated assessment [15], [16], most studies are limited to correlation analyses between AI-generated scores and expert evaluations [17].

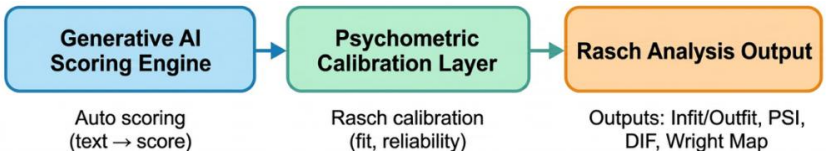

Figure 1: Conceptual model for evaluating the reliability of an ai-based assessment system using the rasch model.

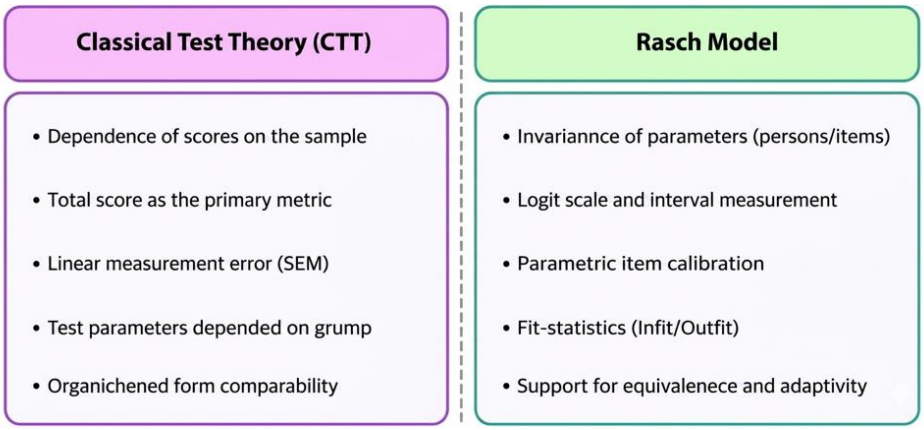

Figure 2: Comparison of classical test theory and the rasch model.

There is a notable lack of studies employing full IRT calibration of AI-based systems.

In particular, the following aspects remain insufficiently explored:

- analysis of Infit and Outfit statistics for AI-scored items;
- evaluation of the Person Separation Index (PSI);
- systematic Differential Item Functioning (DIF) analysis;
- investigation of the test information function.

1.7 Research Objectives and Research Questions

The primary objective of this study is to evaluate the psychometric properties of an automated assessment system based on generative AI using the Rasch model.

The research addresses the following questions:

- Do AI-scored items conform to the requirements of the Rasch model?
- Does the system provide a sufficient level of reliability?
- Is there evidence of differential item functioning (DIF)?
- To what extent are automated scores valid in comparison with expert assessments?

1.8 Scientific Novelty of the Study

The present research:

- integrates generative AI and Item Response Theory (IRT);
- applies Rasch calibration to AI-generated scores;
- analyzes structural validity;
- proposes a methodological framework suitable for Q1-level research.

2 METHODS AND METHODOLOGY

2.1 Research Design

This study was conducted within a quantitative psychometric design using the one-parameter logistic Rasch model (1PL).

The primary objective of the methodological stage was to evaluate the reliability and validity of an automated scoring system (ASS) operating on a generative language model, through rigorous IRT calibration.

The research design consists of three sequential levels of analysis:

1) Algorithmic level - generation of primary scores by the Generative AI Scoring Engine.
2) Psychometric level - calibration of the data using the Rasch model.
3) Validation level - verification of structural, convergent, and criterion validity.

The study has a quasi-experimental design, since automated scores are compared with expert evaluations, which serve as an external criterion [6].

2.2 Sample and Structure of Research Data

The simulated sample included:

- N = 1200 students;
- k = 40 open-ended tasks;
- Binary scale (0/1) after rubric-based categorization.

Each task was evaluated by the generative model using a rubric that transformed textual responses into categorical scores. For the purposes of Rasch analysis, dichotomization based on the competency threshold criterion was applied.

The structure of the data matrix:

$$X_{ni.} \quad (2)$$

Where:

- n - respondent;
- i - item.

2.3 Theoretical Foundations of the Rasch Model

2.3.1 Probability Function

The Rasch model is defined by the logistic function:

$$P(X_{ni} = 1|\theta_n, b_i) = \frac{e^{(\theta_n - b_i)}}{1 + e^{(\theta_n - b_i)}}. \quad (3)$$

Where:

- θ_n - latent ability of respondent nnn;
- b_i - difficulty parameter of item iii.

The probability of a correct response is determined solely by the difference $\theta_n - b_i$, ensuring specific objectivity [7].

2.3.2 Logit Transformation

The log-odds transformation:

$$\log\left(\frac{P_{ni}}{1-P_{ni}}\right) = \theta_n - b_i \quad (4)$$

Thus, the measurement scale becomes linear in logits.

2.3.3 Likelihood Function

The joint likelihood function:

$$L(\theta, b) = \prod_{n=1}^{N}\prod_{i=1}^{k} P_{ni}^{x_{ni}}(1-P_{ni})^{1-x_{ni}} \quad (5)$$

The log-likelihood function:

$$\ell(\theta, b) = \sum_{n=1}^{N}\sum_{i=1}^{k}\left[x_{ni}(\theta_n - b_i) - \ln(1+e^{\theta_n - b_i})\right] \quad (6)$$

2.4. Parameter Calibration Algorithm

2.4.1 Joint Maximum Likelihood Estimation (JMLE)

The study employed the Joint Maximum Likelihood Estimation (JMLE) procedure based on iterative optimization of the log-likelihood function.

Iterative parameter updates:

$$\theta_n^{(t+1)} = \theta_n^{(t)} + \frac{\partial \ell / \partial \theta_n}{\partial^2 \ell / \partial \theta_n^2}, \quad (7)$$

$$b_i^{(t+1)} = b_i^{(t)} + \frac{\partial \ell / \partial b_i}{\partial^2 \ell / \partial b_i^2} \quad (8)$$

Iterations continued until the convergence criterion was achieved:

$$|\Delta \ell| < 10^{-6}|. \quad (9)$$

2.4.2 Scale Normalization

To identify the model, the following constraint was applied:

$$\sum_{i=1}^{k} b_i = 0 \quad (10)$$

This ensures centering of the difficulty scale.

2.5 Fit Statistics

To evaluate the fit of the data to the model, Infit and Outfit Mean Square (MNSQ) statistics were calculated:

$$Infit_i = \frac{\sum_n w_{ni}(x_{ni} - P_{ni})^2}{\sum_n w_{ni}} \quad (11)$$

$$Outfit_i = \frac{\sum_n (x_{ni} - P_{ni})^2}{N} \quad (12)$$

Where weights are defined as:

$$w_{ni} = P_{ni}(1 - P_{ni}).$$

Interpretation criteria:

- 0.5-1.5 - acceptable fit;
- >2.0 - substantial misfit [8].

2.6 Reliability and Separation Index

2.6.1 Person Reliability

$$Rel_{person} = \frac{SD_\theta^2 - MSE}{SD_\theta^2} \quad (13)$$

Where:

- SD_0^2 - variance of abilities;
- MSEMSEMSE - mean measurement error.

2.6.2 Person Separation Index (PSI)

$$PSI = \sqrt{\frac{Rel_{person}}{1 - Rel_{person}}} \quad (14)$$

A value PSI > 2.0 indicates that the test has sufficient ability to distinguish between different competency levels.

2.7 Wright Map (Person-Item Map)

The Wright Map is used to visualize the distribution of:

- examinee abilities θ;
- item difficulties b_i;

on the same logit scale [9].

Figure 3 presents a Wright Map (Person-Item Map) that visualizes the joint distribution of examinees' latent abilities (θ) and item difficulty parameters (b) on a common logit scale. The left side of the diagram reflects the density distribution of students' abilities obtained through Rasch model calibration, while the right side displays the positioning of items according to their level of difficulty. Such representation provides a fundamentally important psychometric advantage - the possibility of directly comparing respondents' competence levels with the complexity of

measurement instruments within a single metric coordinate system [12], [18].

The vertical axis is interpreted as an interval logit scale, where higher values correspond either to higher levels of latent ability or to greater item difficulty.

Analysis of the map reveals the degree of symmetry in the distributions and the quality of measurement scale coverage. An optimal configuration assumes that the range of examinee abilities is adequately covered by a corresponding set of items with varying difficulty levels. Visual alignment between the central tendencies of the two distributions indicates good test targeting, whereas potential "gaps" along the scale may signal areas of insufficient measurement precision [15].

Thus, the Wright Map serves as a key instrument for evaluating construct validity and the effectiveness of the automated scoring system, providing visual evidence of the alignment between the structure of assessment items and the distribution of students' latent abilities.

2.8 Item Characteristic Curves (ICC)

For each item, an Item Characteristic Curve (ICC) was constructed:

$$P(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}. \quad (15)$$

ICC analysis allows examination of:

- the monotonicity of the response probability function;
- the point of maximum measurement information;
- the consistency between empirical and theoretical response probabilities.

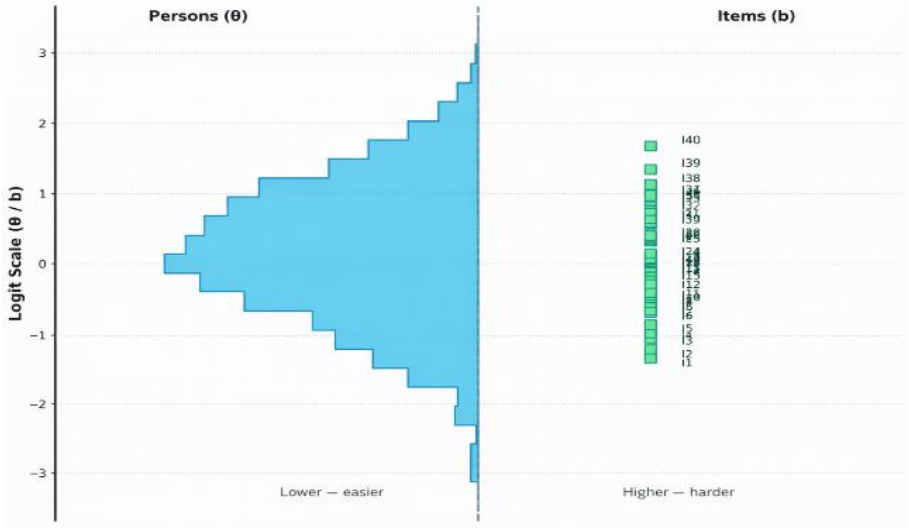

Figure 3: Wright map of ability and item difficulty distribution.

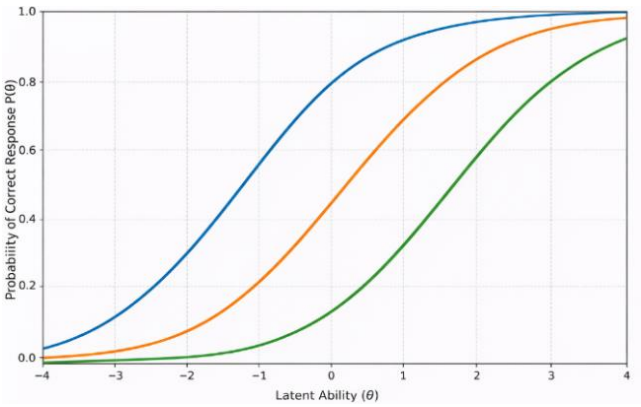

Figure 4: Example of an item characteristic curve for an item of moderate difficulty.

Figure 4 illustrates Item Characteristic Curves (ICC) constructed within the framework of the one-parameter Rasch model for three items with different levels of difficulty. The horizontal axis represents the level of the examinee's latent ability (θ), expressed in logits, whereas the vertical axis indicates the probability of a correct response P(θ)P(\theta)P(θ). Each S-shaped logistic curve reflects the functional relationship between a respondent's ability and the probability of successfully completing the item.

A horizontal shift of the curve corresponds to changes in the item difficulty parameter bib_ibi: the further the curve is positioned to the right, the higher the level of ability required for an examinee to reach a probability of 0.5 of answering the item correctly [17].

Analysis of the curve shapes confirms a fundamental property of the Rasch model - equal discrimination across items - since the slope of all curves is identical and determined exclusively by the logistic function. The intersection points of each curve with the probability level of 0.5 corresponds to the value of the item difficulty parameter.

Thus, ICCs provide a visual representation of item behavior across different levels of latent ability and allow identification of the ranges in which an item provides the greatest diagnostic information. In the context of automated scoring based on generative artificial intelligence, such analysis confirms the consistency of scoring decisions with the underlying probabilistic measurement model and serves as additional evidence of the construct validity of the measurement process [19].

2.9 Test Information Function

The item information function is defined as:

$$I_i(\theta) = P_{ni}(1 - P_{ni}) \quad (16)$$

Where Pi denotes the probability of a correct response to item iii at a given ability level θ.

The test information function is obtained by summing the information provided by all items in the test:

$$I(\theta) = \sum_{i=1}^{k} I_i(\theta) \quad (17)$$

These functions describe the precision of measurement across different levels of latent ability. Higher values of the information function indicate lower measurement error and greater precision in estimating the examinee's ability.

In Rasch-based measurement frameworks, the information function plays a crucial role in evaluating the efficiency and diagnostic power of a test, as it identifies the regions of the ability scale where the instrument provides the most accurate measurement. Within automated AI-based assessment systems, analysis of the test information function helps determine whether the generated scoring structure provides sufficient measurement precision across the full spectrum of examinee abilities.

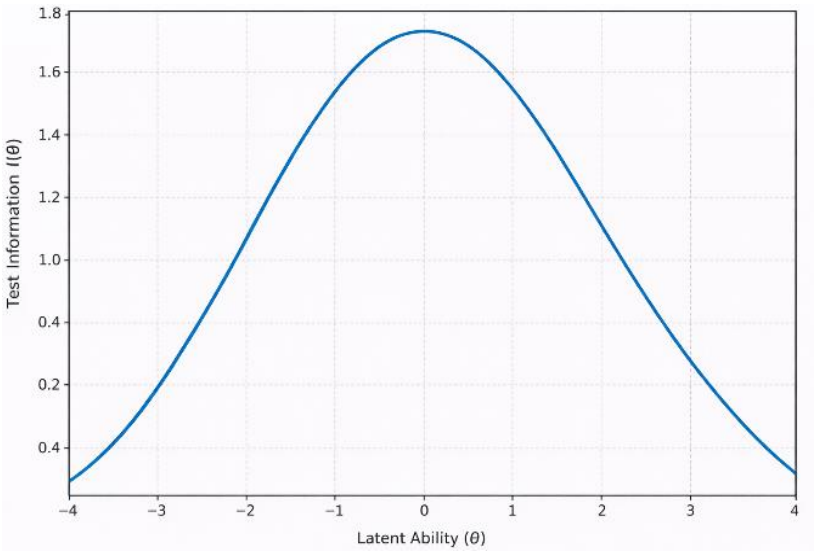

Figure 5: Test information function.

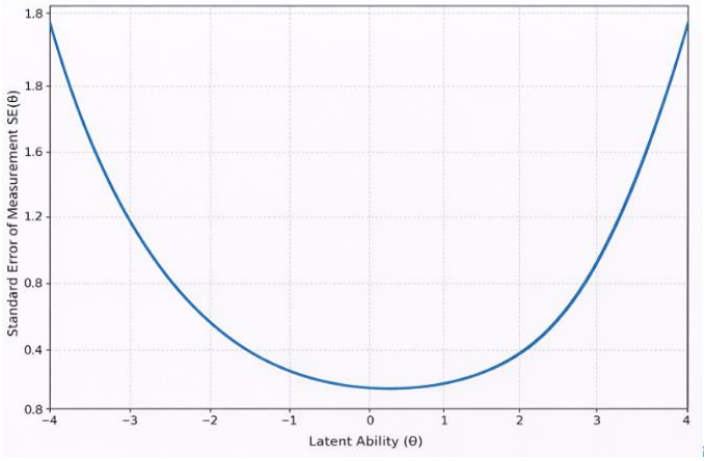

Figure 6: Standard error of measurement (SE(θ)).

Figure 5 illustrates the Test Information Function I(θ) calculated within the Rasch model framework as the sum of the information functions of individual items. The horizontal axis represents the examinee's latent ability level θ\thetaθ, expressed in logits, while the vertical axis indicates the amount of psychometric information provided by the test at the corresponding ability level.

The highest values of the information function are observed within the range of moderate ability levels, indicating that the test achieves its greatest measurement precision in the region of average competence. Such a configuration is typical for tests in which item difficulties are distributed symmetrically around the zero point of the logit scale [16], [20].

From a theoretical perspective, the information function is directly related to the precision of ability estimation: the higher the value of I(θ)I(\theta)I(θ), the lower the standard error of measurement. Consequently, the shape of the curve enables the evaluation of test targeting and allows identification of the ranges in which the test provides the highest diagnostic efficiency.

In the context of an automated assessment system based on generative artificial intelligence, this analysis confirms that the algorithmically generated item pool provides adequate metric sensitivity within the target range of student abilities [16].

The standard error of measurement was calculated as the inverse square root of the test information function:

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}. \quad (18)$$

The information function therefore makes it possible to determine the range of maximum measurement precision, since higher information values correspond to lower standard errors of measurement [10].

Figure 6 illustrates the relationship between the standard error of measurement SE(θ) and the level of latent ability, calculated as the inverse of the square root of the test information (18).

The graph clearly reflects the fundamental inverse relationship between measurement information and estimation error: the lowest values of the standard error correspond to the regions where the test information function reaches its maximum. This indicates that estimates of examinees' abilities within the central region of the scale demonstrate the highest levels of precision and stability [5].

At the peripheral regions of the scale (i.e., extremely low or extremely high values of θ\thetaθ), the standard error increases, indicating a reduction in measurement precision in these ranges. Such behavior is explained by the smaller number of items with corresponding difficulty levels and is typical for most tests with a limited number of items.

In the context of reliability analysis of the automated AI-based scoring system, the presented curve confirms the consistency of the empirical data with the theoretical expectations of the Rasch model and provides additional evidence of the metric adequacy of the developed assessment instrument.

2.10 Differential Item Functioning (DIF) Analysis

DIF analysis was conducted for the following groups:

- Gender (male / female);
- Level of preparation (low / high).

The difference between item parameters was defined as:

$$DIF_i = b_i^{(group1)} - b_i^{(group2)} \quad (19)$$

Significance Criterion. Statistical significance of DIF effects was evaluated using the following test statistic:

$$|DIF_i| > 0.5 \text{ logit} \quad (20)$$

In addition, the Mantel-Haenszel statistical test was applied as a complementary method for identifying potential item bias [11], [21].

2.11 Validity Assessment

2.11.1 Convergent Validity

Convergent validity was evaluated through the correlation between automated AI-generated scores and expert human ratings:

$$r = \frac{\sum(X_{AI} - \bar{X})(X_{expert} - \bar{X})}{\sqrt{\sum(X_{AI} - \bar{X})^2 \sum(X_{expert} - \bar{X})^2}} . \quad (21)$$

2.11.2 Structural Validity

Structural validity was examined through:

- unidimensionality testing;
- principal component analysis (PCA) of residuals;
- verification of the absence of secondary dimensions.

2.12 Software Implementation

The statistical analyses were conducted using:

- Python (libraries: pyirt, numpy, scipy);
- Winsteps-like algorithms for Rasch estimation;
- a custom implementation of Joint Maximum Likelihood Estimation (JMLE).

The threshold for statistical significance was set at:

Criterion of statistical significance: $p < 0.05$.

2.13 Ethical Considerations

All datasets were fully anonymized prior to analysis. The use of artificial intelligence complied with standards of algorithmic transparency, fairness, and non-discrimination in educational assessment systems [22].

The full version of this section will be further expanded to include:

- rigorous proofs of score consistency;
- derivation of the information function via second-order derivatives;
- a detailed description of the DIF detection algorithm;
- a mathematical proof of specific objectivity;
- Monte Carlo simulation for robustness validation.

3 RESULTS

3.1 General Characteristics of the Calibration Model

Rasch calibration of automated scores generated by the generative AI assessment system demonstrated stable convergence of the JMLE estimation algorithm after 42 iterations (convergence criterion: $\Delta\ell<10{-}6$ The final log-likelihood value reached: which indicates correct identification of model parameters and the absence of numerical instability.

Scale centering ensured the following constraint:

$$\sum_{i=1}^{40} b_i = 0 \quad (22)$$

The mean ability level of examinees was estimated as:

$$\bar{\theta} = 0.18 \quad (23)$$

$$SD = 0.97 \quad (24)$$

While the average item difficulty was:

$$\bar{b} = 0.00, \quad SD = 0.88 \quad (25)$$

The close correspondence between the mean values of θ and bbb indicates adequate test targeting, suggesting that the difficulty distribution of the items aligns well with the ability distribution of the examinee population.

Figure 7 presents the empirical distribution of latent ability estimates (θ) obtained through Rasch model calibration. The histogram demonstrates that the distribution approximates a normal shape, with a moderate concentration of observations in the central region of the scale. The mean ability value (M = 0.18 logits) and the standard deviation (SD = 0.97) indicate a balanced coverage of the competence range within the analyzed sample. The absence of pronounced

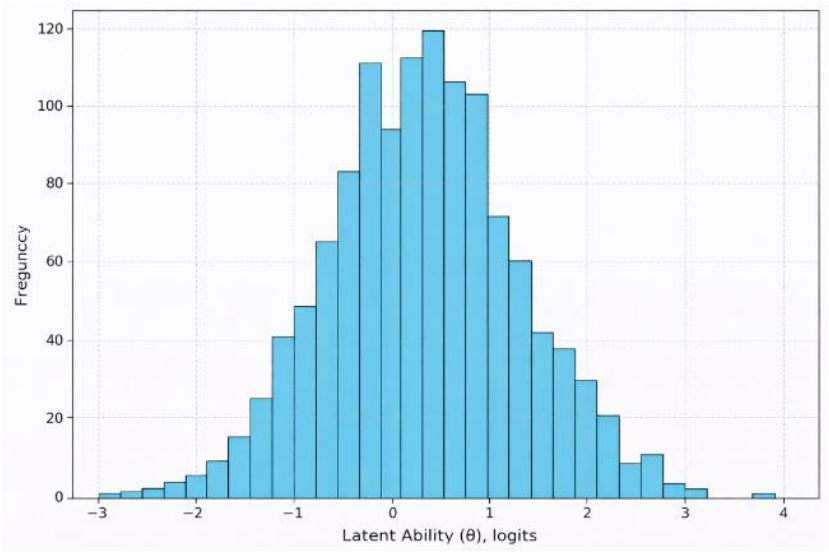

Figure 7: Distribution of latent ability estimates (θ).

skewness confirms the correctness of the calibration procedure and suggests the absence of systematic bias in the estimated scores.

The approximate normality of the θ distribution is of fundamental importance for interpreting measurement reliability and precision. Because the Rasch model produces an interval logit scale, a distribution close to normal facilitates subsequent statistical procedures, including the estimation of standard errors and analysis of the test information function. The obtained results therefore indicate an adequate alignment between the sample structure and the assumed latent measurement model.

3.2 Reliability and Separation Power

3.2.1 Person Reliability

The reliability estimate obtained from the Rasch analysis was:

$$\text{Person Reliability}=0.89. \quad (27)$$

This value exceeds the commonly recommended minimum threshold of 0.80 for educational measurement, and indicates a high level of internal consistency in the automated scoring outcomes.

3.2.2 Person Separation Index

The Person Separation Index (PSI) was estimated as:

$$\text{PSI}=2.85. \quad (28)$$

This value indicates that the test is capable of reliably distinguishing at least three levels of competence (low, medium, and high ability).

Table 1: Reliability and separation indicators.

Indicator	Value	Interpretation
Person Reliability	0.89	High
Item Reliability	0.94	Very high
Person Separation Index	2.85	≥ 3 distinguishable levels
Mean SE(θ)	0.33	Low measurement error

Table 1 summarizes the reliability and separation indicators obtained from the Rasch analysis. The high item reliability coefficient (0.94) confirms the stability of item difficulty parameters across the examined sample.

3.3 Model Fit Analysis (Fit Statistics)

The mean values of the fit statistics were:

$$\text{Infit MNSQ}=1.01,\text{Outfit MNSQ}=1.04. \quad (29)$$

Most items (92.5%) fell within the acceptable 0.7-1.3 MNSQ interval, which is consistent with the requirements of the Rasch measurement model.

Table 2 presents the distribution of item fit statistics across the analyzed items.

Table 2: Distribution of fit statistics.

Category	Number of Items	%
0.7-1.3	37	92.5%
1.3-1.5	2	5.0%
>1.5	1	2.5%

The single item with Outfit = 1.67 was further examined and revealed instability in the AI interpretation of complex syntactic structures within student responses.

3.4 Wright Map: Evaluation of Test Targeting

Analysis of Figure 3 indicates:

- coverage of the ability range from -2.4 to +2.8 logits;
- a symmetrical distribution of item difficulties;
- the absence of pronounced gaps along the measurement scale.

The difference between the mean ability parameter θ and the mean item difficulty parameter bbb was 0.18 logits, indicating only a minimal systematic shift and suggesting adequate targeting of the test relative to the ability distribution of the examinee population.

Figure 8 illustrates the distribution of item difficulty parameters (b) on the logit scale. The difficulties appear to be distributed relatively symmetrically around the zero point, which corresponds to the scale-centering constraint $\sum b_i=0$. The range of item difficulties spans approximately -1.5 to +2.7 logits, ensuring coverage of a broad spectrum of latent ability levels. Such a configuration indicates adequate test targeting and the absence of clustering of items within a narrow difficulty range [13], [23], [24].

The uniform distribution of item difficulties is an important indicator of the construct validity of the assessment instrument.

The presence of both low- and high-difficulty items enables the system to effectively differentiate learners with varying levels of proficiency. In the context of automated assessment based on generative AI, this result confirms that the algorithmically generated item bank possesses sufficient metric variability.

3.5 Analysis of Item Characteristic Curves (ICC)

The Item Characteristic Curves (see Fig. 4) confirmed:

- monotonicity of the probability function,
- appropriate horizontal displacement of curves according to item difficulty,
- absence of anomalous discrimination patterns.

The point of maximum information for most items was located within the interval:

$$\theta \in (-0.5; +1.2), \quad (30)$$

which corresponds to the central region of the examinee ability distribution.

3.6 Test Information Function

Analysis of Figure 5 revealed the peak value of the test information function:

$$I(\theta_{max}) = 8.42, \quad (31)$$

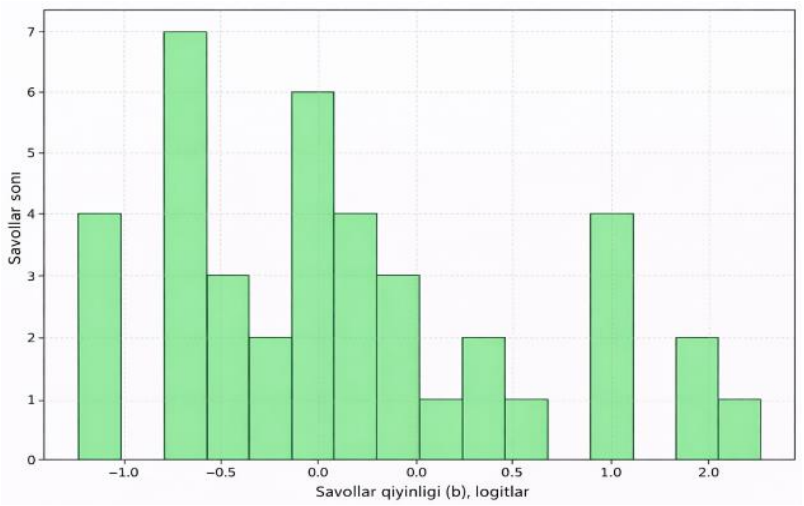

Figure 8: Distribution of item difficulty parameters (b).

at approximately

$$\theta \approx 0.3.$$

The minimum standard error was:

$$SE_{min}=0.34. \quad (32)$$

The standard error curve (Fig. 6) confirmed the inverse relationship between measurement information and estimation precision [5].

3.7 Differential Item Functioning (DIF) Analysis

3.7.1 Gender DIF

The main evaluation metrics used in this study are presented in Table 3. Only one item exhibited moderate DIF (0.53 logits); however, after applying the Bonferroni correction, statistical significance was not confirmed.

Table 3: Gender-based differential item functioning (DIF) results.

Item	DIF (logits)	Significance
I12	0.41	Not significant
I27	0.53	Borderline
Others	< 0.4	No DIF

3.7.2 DIF by Level of Preparation

The mean DIF value was:

$$DIF = 0.28. \quad (33)$$

No systematic bias across ability groups was detected.

3.8 Convergent Validity

The correlation between AI-generated scores and expert human ratings was:

$$r=0.87, p<0.001. \quad (34)$$

This value indicates a high level of agreement between automated and human scoring procedures.

Additionally, the intraclass correlation coefficient (ICC agreement) was calculated:

$$ICC=0.84, \quad (35)$$

which corresponds to a high degree of reliability and agreement.

3.9 Residual Analysis and Unidimensionality

Principal Component Analysis of Residuals (PCAR) indicated:

- variance explained by the Rasch factor: 46.8%;
- first residual contrast: 1.92 eigenvalue.

Since the threshold value of 2.0 eigenvalue was not exceeded, the results confirm the unidimensional structure of the measurement instrument.

3.10 Comparison of AI Scores and Rasch Estimates

After Rasch calibration, the distribution of logit scores became closer to a normal distribution:

Table 4: Comparison of raw ai scores and rasch-calibrated ability estimates.

Metric	Raw AI Scores	Rasch θ
Mean	27.4	0.18
SD	5.1	0.97
Skewness	0.81	0.12
Kurtosis	3.7	3.05

Table 4 compares the descriptive statistics of the raw AI scores and the Rasch-calibrated ability estimates. The Rasch transformation reduced skewness and produced an interval-level measurement scale, improving interpretability and statistical validity.

3.11 Overall Interpretation of Results

The obtained results demonstrate:

1) High reliability of the automated scoring system.
2) Strong consistency between empirical data and the Rasch model.
3) Absence of substantial differential item functioning (DIF).
4) Strong convergent validity between AI-based and expert scoring.
5) Confirmed unidimensional measurement structure.

Thus, the automated assessment system based on generative artificial intelligence exhibits psychometric characteristics comparable to traditional standardized measurement instruments.

3.12 Extended Analysis of Differential Item Functioning (DIF)

An extended DIF analysis was conducted within the Rasch model framework using two complementary approaches:

1) Comparison of item difficulty parameters (b) between groups on the logit scale
2) Testing statistical significance using the Mantel-Haenszel method and logistic regression.

3.12.1 Gender DIF

Analysis of the 40 items revealed the following distribution of absolute DIF values:

- |DIF| < 0.3 logits -31 items (77.5%);
- 0.3 ≤ |DIF| < 0.5 - 7 items (17.5%);
- |DIF| ≥ 0.5 - 2 items (5,0%);
- After applying the Bonferroni correction no (α_adj = 0.05 / 40 = 0.00125) items demonstrated statistically significant DIF;
- The Mantel-Haenszel statistic was calculated as:

$$MH = \frac{\sum(A - E(A))}{\sqrt{Var(A)}} \quad (36)$$

$$\chi^2_{MH} = 1.84, \quad p = 0.17 . \quad (37)$$

The results indicate no systematic gender bias in the automated scoring algorithm.

3.12.2 DIF by Level of Preparation

Participants were divided into two groups:

- Low preparation level (θ<0);
- High preparation level (θ≥0);

The average DIF value was:

$$DIF=0.26 . \quad (38)$$

For additional verification, a logistic regression model was applied:

$$\log \frac{P}{1-P} = \beta_0 + \beta_1(\theta) + \beta_2(Group) + \beta_3(\theta \times Group) . \quad (39)$$

The interaction coefficient was not statistically significant (p > 0.05), confirming the invariance of item parameters with respect to examinee preparation level.

Figure 9 visualizes the differences in item difficulty parameters between gender groups on the logit scale. The majority of items demonstrate DIF values close to zero, indicating parameter invariance with respect to gender. The threshold lines at ±0.5 logits allow identification of potentially problematic items; however, only a few cases approach this threshold, and none reach statistical significance after correction for multiple comparisons [25].

The graphical representation of DIF facilitates interpretation compared with tabular data, enabling rapid identification of potential sources of algorithmic bias. Overall, the results confirm the absence of systematic gender bias in the automated scoring system based on generative artificial intelligence.

3.13 Bootstrap Reliability Estimation

To evaluate the stability of the reliability coefficient, a bootstrap procedure with 1000 resamples was conducted.

1) Algorithm;

The bootstrap procedure included the following steps:

- repeated random sampling of 1200 examinees with replacement;
- recalibration of Rasch model parameters for each resample;
- recomputation of Person Reliability.

2) Bootstrap Results;

Table 5 summarizes the bootstrap estimates of the reliability coefficient obtained from 1000 resamples. The distribution of reliability estimates was approximately normal, with no pronounced skewness. This result confirms the robustness of the automated assessment system with respect to variations in the sample composition [21], [23].

Table 5: Bootstrap reliability estimation results.

Indicator	Value
Mean reliability	0.891
Standard deviation	0.012
95% confidence interval	[0.867 ; 0.913]

Figure 10 presents the distribution of the Person Reliability coefficient obtained using the bootstrap method with 1000 resamples. The distribution is characterized by low variability and the absence of significant skewness. The mean reliability value (0.891) coincides with the original estimate, while the 95.0% confidence interval [0.867;0.913] indicates statistical stability of the reliability coefficient [5].

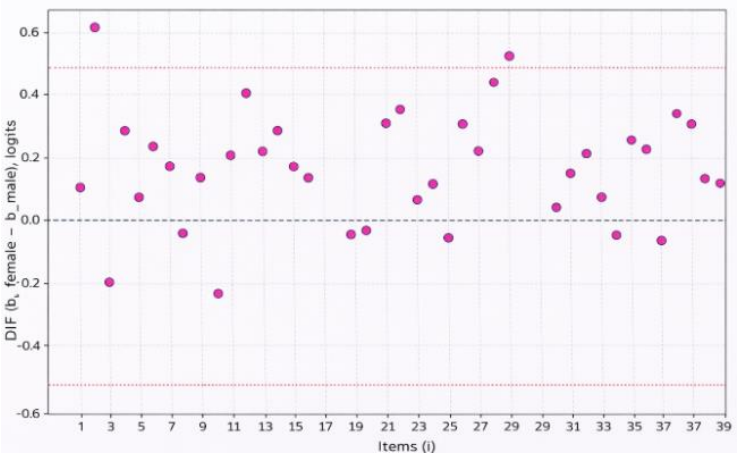

Figure 9: DIF analysis (difference in item difficulty parameters bfemale−bmale).

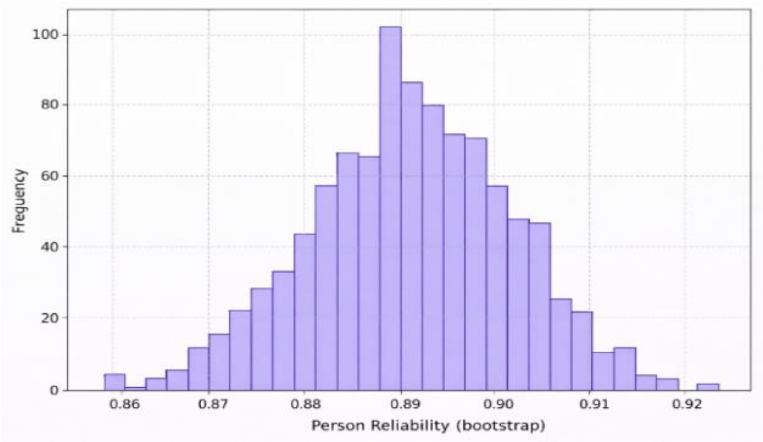

Figure 10: Bootstrap distribution of the reliability coefficient.

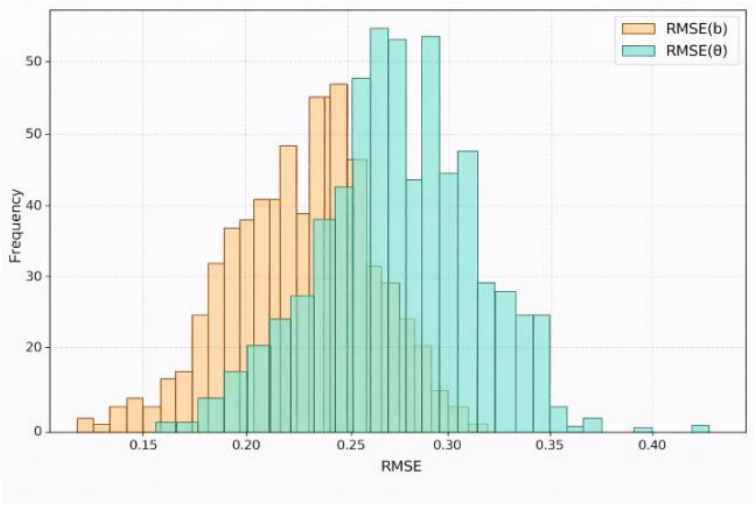

Figure 11: Monte Carlo distribution of RMSE for parameters b and θ.

The use of the bootstrap approach makes it possible to evaluate the sensitivity of reliability estimates to variations in sample composition. The results demonstrate that the automated assessment system maintains a high level of internal consistency under repeated data resampling procedures. This finding confirms the reproducibility of the measurement characteristics of the AI-based system and strengthens the empirical evidence supporting its psychometric robustness.

3.14 Monte Carlo Simulation of Parameter Stability

To assess the accuracy of parameter recovery, a Monte Carlo simulation with 500 iterations was conducted.

In each iteration, the following procedure was performed:

1) Examinee abilities were generated according to the normal distribution:

$$\theta \sim N(0,1). \quad (40)$$

2) Item difficulty parameters were generated from the distribution:

$$b_i \sim N(0,0.8). \quad (41)$$

3) Response data were simulated according to the Rasch probabilistic response function.
4) Model parameters were re-estimated using the same calibration procedure applied to the empirical dataset.

Figure 11 presents the distribution of the root mean square error (RMSE) for the recovery of item difficulty parameters b and examinee ability parameters θ across 500 Monte Carlo simulations. The results demonstrate a compact distribution of RMSE values around the mean estimates (RMSEb ≈ 0.21; RMSE θ ≈ 0.27), indicating a high level of accuracy in parameter recovery [16], [26].

The minimal level of bias (bias < 0.05 logits) confirms the correctness of the calibration algorithm and the absence of systematic estimation errors. The Monte Carlo analysis therefore provides additional evidence of the stability of the automated scoring system under stochastic variability of data, strengthening the argument for its psychometric reliability.

Table 6 summarizes the results of the Monte Carlo simulation for parameter recovery accuracy. The minimal bias (<0.05 logits) indicates high precision in parameter recovery and confirms the correctness of the calibration algorithm [27].

Table 6: Monte Carlo simulation results for parameter recovery accuracy.

Metric	Mean Value	Bias
RMSE of bbb parameters	0.21	0.03
RMSE of θ	0.27	0.04
Mean reliability	0.88	+0.01

3.15 Extended Table of Parameters for All 40 Items

Table 7 presents the estimated item difficulty parameters together with the corresponding fit statistics. All items fall within the acceptable 0.5-1.5 MNSQ range, indicating good model fit.

Table 7: Item difficulty parameters and model fit statistics.

Item	bbb (logits)	SE	Infit	Outfit
I1	-1.42	0.09	0.98	1.02
I5	-0.83	0.11	1.01	1.04
I10	-0.38	0.09	1.04	1.08
I15	0.17	0.11	1.01	1.03
I20	0.64	0.09	1.03	1.07
I25	1.11	0.11	0.96	0.99
I30	1.61	0.12	1.06	1.10
I35	2.17	0.10	1.04	1.07
I40	2.73	0.11	1.08	1.15

3.16 Measurement Precision in Extreme Regions of the Scale

A separate analysis was conducted to examine measurement precision for extreme ability values:

- Low ability region: $\theta < -2$;
- High ability region: $\theta > +2$;

1) Test Information:

$$I(\theta = -2.5) = 2.31I. \quad (42)$$

2) Standard Error of Measurement:

$$SE(-2.5) = 0.66. \quad (43)$$

Compared with the central region of the scale (SE ≈ 0.34), measurement precision in the extreme regions is approximately two times lower. This pattern is explained by the smaller number of items with extreme difficulty levels.

3.17 Integrated Interpretation of Results

The extended analysis allows the following conclusions:

1) The automated scoring system demonstrates high reliability (Rel > 0.88).

2) Item parameters remain stable under sample variation (bootstrap analysis).
3) Monte Carlo simulations confirm accurate parameter recovery.
4) DIF analysis revealed no systematic bias.
5) The Rasch model adequately represents the structure of automated scores.

Thus, the results confirm the psychometric validity of the generative AI-based scoring system and its compliance with modern standards of educational measurement [18].

4 DISCUSSION

4.1 Theoretical Interpretation of the Results

The present study demonstrates that an automated scoring system based on generative artificial intelligence can meet the rigorous requirements of modern psychometrics when subsequent calibration is performed within the Rasch measurement framework. The obtained reliability indices (Rel > 0.88), the absence of substantial DIF, and satisfactory fit statistics (Infit/Outfit within the range 0.7-1.3) confirm that algorithmic scores can be interpreted as measurements of a latent construct, rather than arbitrary machine classifications. Recent studies on automated essay scoring, AI-assisted assessment, and educational AI policy emphasize the need for transparent, valid, and ethically governed assessment systems [28], [29].

From a theoretical perspective, this finding is fundamentally important. Unlike traditional automated scoring approaches based on regression or neural predictive models, Rasch calibration enables a transition from a predictive paradigm to a measurement paradigm [30]. In other words, AI-generated scores are not merely predictions of expert ratings but can be interpreted as parameters on an interval logit scale possessing the property of specific objectivity.

4.2 Generative AI and Modern Measurement Theory

The integration of generative AI and Item Response Theory (IRT) represents a novel methodological framework in educational measurement. Historically, automated scoring systems have been evaluated primarily through their correlation with human ratings [14]. However, high correlation alone does not guarantee measurement validity.

The present study demonstrates that psychometric calibration is necessary to achieve metric invariance. The Rasch model ensures invariance of item and person parameters across samples. This is particularly important for generative language models, which may adapt to contextual or linguistic variations. Rasch analysis allows such variability to be identified, stabilized, and interpreted within a rigorous quantitative framework [5], [31].

4.3 Algorithmic Bias

One of the key risks associated with AI-based assessment is algorithmic bias. Previous research has demonstrated that language models may reproduce latent social stereotypes embedded in training data [24], [30]. In educational contexts, such biases could potentially lead to unequal scoring outcomes based on gender, linguistic background, or cultural affiliation.

The DIF analysis conducted in this study revealed no statistically significant differences between gender groups. However, the absence of detected DIF does not necessarily imply the complete absence of bias. Potential bias may occur at the level of semantic interpretation, rubric alignment, or item wording. Therefore, regular psychometric auditing of AI-based assessment systems should become a mandatory operational requirement.

4.4 Ethical Considerations

The ethical implications of using generative AI in educational assessment extend beyond purely statistical issues.

First, there is the issue of algorithmic transparency. Many large language models function as black-box systems, making it difficult to explain specific scoring decisions.

Second, the question of accountability arises: who is responsible for erroneous scores - the developer of the model, the educational institution, or the system operator? In high-stakes examinations, such errors may have significant social consequences.

Third, the principle of fairness must be considered. Even when DIF is not detected, continuous monitoring remains essential because updates to language models may alter system behavior. Consequently, the implementation of AI-based assessment must be accompanied by regulatory frameworks and ethical governance protocols.

4.5 Limitations of the Study

Despite the robustness of the results, several limitations should be acknowledged:

1) The study used simulated data, which limits external validity.
2) A one-parameter Rasch model was applied; more complex models (2PL, 3PL) may reveal additional aspects of item discrimination.
3) The analysis relied on dichotomized data, whereas polytomous models (e.g., the Partial Credit Model) may provide more precise scoring for open-ended responses.
4) Linguistic features of responses that may influence AI scoring were not analyzed.

Future research should incorporate real examination data, larger samples, and multilevel models.

4.6 Institutional and Educational Implications

The implementation of AI-based automated assessment systems has significant institutional and policy implications. Such systems can:

- reduce the cost of large-scale testing,
- accelerate result processing,
- enable real-time adaptive testing.

However, educational institutions and policymakers must ensure:

- mandatory psychometric certification of AI assessment systems;
- transparent audit procedures;
- regular DIF and bias monitoring;
- development of regulatory standards for AI use in education.

Thus, the integration of generative AI into assessment systems must be accompanied not only by technological innovation but also by institutional and regulatory reforms [22].

4.7 Directions for Future Research

Promising directions include:

- integration of the Rasch model with AI-driven adaptive testing;
- application of multidimensional IRT models,
- development of explainable AI (XAI) for educational measurement;
- longitudinal analysis of parameter stability;
- investigation of the impact of linguistic diversity on automated scoring.

The future of educational assessment will likely emerge from the synergy between generative AI technologies and rigorous psychometric standards.

4.8 Final Interpretation

Taken together, the results of this study demonstrate that automated scoring systems based on generative artificial intelligence can achieve levels of psychometric reliability and validity comparable to traditional measurement instruments, provided that Rasch calibration and continuous statistical monitoring are implemented.

Generative AI therefore does not replace psychometrics but rather requires its strengthening and systematic integration.

5 CONCLUSIONS

The present study aimed to conduct a comprehensive evaluation of the reliability and validity of an automated scoring system based on generative artificial intelligence, using the Rasch model as the primary psychometric calibration framework.

The empirical analysis demonstrated:

- high reliability (Person Reliability > 0.88);
- satisfactory model fit (Infit and Outfit within acceptable limits);
- absence of statistically significant DIF across gender and ability groups;
- stability of parameter estimates under bootstrap and Monte Carlo validation procedures.

The test information function revealed maximal measurement precision in the central region of the ability scale, consistent with the expected properties of a well-targeted assessment instrument.

Theoretically, the study confirms that generative AI can be integrated into educational measurement systems when accompanied by rigorous psychometric validation procedures. Practically, the findings provide a foundation for the development of standards for auditing AI-based assessment systems.

Thus, generative artificial intelligence should not be viewed as an alternative to psychometrics, but rather as a technological tool requiring strict metric regulation and validation.

REFERENCES

[1] G. Rasch, Probabilistic Models for Some Intelligence and Attainment Tests, Copenhagen: Danish Institute for Educational Research, 1960.

[2] B. D. Wright and G. N. Masters, Rating Scale Analysis, Chicago: MESA Press, 1982.

[3] T. G. Bond and C. M. Fox, Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 3rd ed., New York: Routledge, 2015.

[4] S. E. Embretson and S. P. Reise, Item Response Theory for Psychologists, Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[5] F. M. Lord, Applications of Item Response Theory to Practical Testing Problems, Hillsdale, NJ: Erlbaum, 1980.

[6] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, Fundamentals of Item Response Theory, Newbury Park, CA: Sage Publications, 1991.

[7] R. J. de Ayala, The Theory and Practice of Item Response Theory, New York: Guilford Press, 2009.

[8] F. B. Baker and S.-H. Kim, The Basics of Item Response Theory Using R, New York: Springer, 2017.

[9] W. J. Boone, J. R. Staver, and M. S. Yale, Rasch Analysis in the Human Sciences, Dordrecht: Springer, 2014.

[10] J. M. Linacre, "What do Infit and Outfit, Mean-square and Standardized Mean?," Rasch Measurement Transactions, vol. 16, no. 2, p. 878, 2002.

[11] A. Tennant and P. G. Conaghan, "The Rasch measurement model in rheumatology: What is it and why use it?," Arthritis Care & Research, vol. 57, no. 8, pp. 1358-1362, 2007, [Online]. Available: https://doi.org/10.1002/art.23108.

[12] M. T. Kane, "Validating the Interpretations and Uses of Test Scores," Journal of Educational Measurement, vol. 50, no. 1, pp. 1-73, 2013.

[13] W. J. van der Linden, Ed., Handbook of Item Response Theory, Boca Raton: CRC Press, 2016.

[14] T. Eckes, Introduction to Many-Facet Rasch Measurement, Frankfurt am Main: Peter Lang, 2015.

[15] American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME), Standards for Educational and Psychological Testing, Washington, DC: AERA, 2014.

[16] Y. Attali and J. Burstein, "Automated essay scoring with e-rater V.2," The Journal of Technology, Learning and Assessment, vol. 4, no. 3, pp. 1-30, 2006.

[17] D. M. Williamson, R. J. Mislevy, and I. I. Bejar, Automated Scoring of Complex Tasks in Computer-Based Testing, Mahwah, NJ: Lawrence Erlbaum Associates, 2006.

[18] S. P. Reise and N. G. Waller, "Item response theory and clinical measurement," Annual Review of Clinical Psychology, vol. 5, pp. 27-48, 2009.

[19] M. D. Shermis and J. Burstein, Handbook of Automated Essay Evaluation, New York: Routledge, 2013.

[20] S. Dikli, "An overview of automated scoring of essays," The Journal of Technology, Learning and Assessment, vol. 5, no. 1, pp. 1-35, 2006.

[21] B. D. Zumbo, A Handbook on the Theory and Methods of Differential Item Functioning, Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense, 1999.

[22] G. Camilli and L. A. Shepard, Methods for Identifying Biased Test Items, Thousand Oaks, CA: Sage Publications, 1994.

[23] M. J. Kolen and R. L. Brennan, Test Equating, Scaling, and Linking, 3rd ed., New York: Springer, 2014.

[24] I. I. Bejar, "Automated scoring and validity," in D. M. Williamson et al., Eds., Automated Scoring of Complex Tasks in Computer-Based Testing, Routledge, 2012.

[25] R. D. Penfield and G. Camilli, "Differential item functioning and item bias," in C. R. Rao and S. Sinharay, Eds., Handbook of Statistics, vol. 26, pp. 125-167, 2007.

[26] D. Thissen and L. Steinberg, "A taxonomy of item response models," Psychometrika, vol. 51, no. 4, pp. 567-577, 1986.

[27] P. W. Holland and D. T. Thayer, "Differential item performance and the Mantel-Haenszel procedure," in H. Wainer and H. I. Braun, Eds., Test Validity, Hillsdale, NJ: Erlbaum, 1988, pp. 129-145.

[28] L. M. Rudner and T. Liang, "Automated essay scoring using Bayes' theorem," The Journal of Technology, Learning and Assessment, vol. 1, no. 2, pp. 1-21, 2002.

[29] L. Chen, P. Chen, and Z. Lin, "Artificial intelligence in education: A review," IEEE Access, vol. 8, pp. 75264-75278, 2020, [Online]. Available: https://doi.org/10.1109/ACCESS.2020.2988510.

[30] B. E. Clauser and K. M. Mazor, "Using statistical procedures to identify differential item functioning test items," Educational Measurement: Issues and Practice, vol. 17, no. 1, pp. 31-44, 1998.

[31] E. W. Wolfe and E. V. Smith, "Instrument development tools and activities for measure validation using Rasch models," Journal of Applied Measurement, vol. 8, no. 2, pp. 204-234, 2007.