CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 10 (2) (2025), Pages 970-985
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Deep-Rasch as an Alternative to Rasch Modeling under Assumption Violations and Small Sample Sizes

Agus Santoso[1*], Farit Mochamad Afendi[2], Timbul Pardede[3], Heri Retnawati[4,5], Ibnu Rafi[5], Ezi Apino[4], and Munaya Nikma Rosyada[6]

[1]*Study Program of Mathematics Education, Universitas Terbuka, Indonesia*
[2]*Department of Statistics and Data Science, Institut Pertanian Bogor (IPB University), Indonesia*
[3]*Study Program of Statistics, Universitas Terbuka, Indonesia*
[4]*Department of Educational Research and Evaluation, Universitas Negeri Yogyakarta, Indonesia*
[5]*Department of Mathematics Education, Universitas Negeri Yogyakarta, Indonesia*
[6]*Study Program of PhD in Education, Universitas Islam Internasional Indonesia, Indonesia*

Abstract

In certain situations, it may be challenging to fully exploit the advantages of modern test theory, including Rasch modeling and item response theory (IRT), when applied to real data. Although Rasch modeling tends to be more robust than IRT for small sample sizes, it still requires that the assumptions of unidimensionality and local independence be satisfied. In practice, these assumptions are often violated, which can lead to less accurate analyses and reduced validity of the results. Deep-Rasch, which integrates deep learning with Rasch modeling, has been proposed as an alternative measurement framework to overcome these limitations. This study examines the potential of Deep-Rasch as an alternative to Rasch modeling using student response data from 17 final semester examinations at Universitas Terbuka (UT), with sample sizes ranging from 33 to 11,504 students. Most examinations consisted of 30 multiple-choice items. The analyses showed that several datasets violated one or both assumptions of Rasch modeling. Nevertheless, Deep-Rasch performed comparably to conventional Rasch modeling in estimating item difficulty and student ability parameters, as well as in predicting student responses. Remarkably, for the smallest sample size ($n = 33$), Deep-Rasch exhibited slightly better performance than Rasch modeling.

Keywords: Deep learning; Deep-Rasch; Item response theory; Rasch modeling

1 Introduction

Classical test theory (CTT) has long been recognized and widely used in educational measurement practice to determine the extent of students' knowledge or competence based on the scores they obtain on a test. Under CTT, the score a student obtains, which is associated with the level of knowledge or competence, is considered the sum of the student's true score, reflecting their actual level of knowledge or competence, and a measurement error that is random and normally distributed [1], [2], [3], [4], [5]. Although CTT offers a systematic measurement framework based on relatively weak assumptions, is easy to apply, and produces results that can be directly and

[*] **Corresponding author**. E-mail: aguss@ecampus.ut.ac.id

Submitted: August 27, 2025 Reviewed: October 10, 2025 Accepted: October 15, 2025
DOI: https://doi.org/10.18860/cauchy.v10i2.36276

easily interpreted, it is not without limitations and criticism [3], [5], [6], [7], [8], which has led to the development of new measurement theories associated with modern test theory. In practice, modern test theory may consist of Rasch modeling and item response theory (IRT). Although some literature considers Rasch modeling to be part of IRT [9], primarily due to the mathematical models underlying both approaches, other sources argue that the two differ in their philosophical foundations [4].

Modern test theory differs from CTT in several ways. One key distinction lies in CTT's focus on the test as a whole, in which students' knowledge or ability is reflected solely in the total score obtained on the test, which is the sum of the scores from each individual item. In contrast, modern test theory emphasizes that students' knowledge or ability, as measured by a test, is a function of the relationship between the characteristics of each item that makes up the test and the competence targeted by the measurement. These item characteristics may include difficulty, discrimination, pseudo-guessing, and carelessness. Modern test theory has evolved in both its theoretical foundations and its applications in education, including its use in large-scale assessments [10], [11] and computer-adaptive testing (CAT) [12], [13], [14]. The application of modern test theory in measurement practice has extended to a wide range of purposes. These include calibrating test or questionnaire items to produce high-quality instruments [6], [7], [8], [15], equating test scores across multiple test forms [16], [17], [18], detecting the presence of biased items [19], [20], and identifying students' abilities with greater precision [6], [21].

Measurement practices based on modern test theory, which offer advantages such as independence from sample characteristics and the ability to conduct fair measurement or assessment, are grounded in strict assumptions or requirements for achieving measurement precision. Meeting these assumptions is essential to ensure that the resulting analyses have a high level of accuracy, provide sufficient validity evidence, and can even be generalized to a broader scale within the same context. In modern test theory, represented by Rasch modeling and IRT, the measurement instrument administered to students is assumed to focus on a single dominant competence or ability that cannot be directly observed. In other words, the measurement instrument is assumed to be unidimensional [2], [5], [22], [23], [24], [25], [26].

It has also been noted in the literature that modern test theory assumes that the only factor influencing a student's probability of answering a test item correctly or incorrectly is the ability or competence that the measurement instrument is primarily intended to assess [2], [23], [24], [25], [26]. In other words, this assumption, known as the local independence assumption, implies a statistically independent relationship between the probability of a student answering a given test item correctly or incorrectly and the probability of answering any other item correctly or incorrectly. Furthermore, modern test theory assumes that the relationship between a student's level of competence or ability and the probability of providing a correct response to a given test item is monotonic, following a sigmoid-shaped curve commonly referred to as the item characteristic curve (ICC) [2], [24]. It is also worth noting that, although considered a less critical assumption, the normal distribution of data representing the ability, competence, or latent construct measured by a given instrument remains an important consideration in the application of modern test theory, particularly in Rasch modeling [4].

On the one hand, the assumptions underlying modern test theory offer the advantage of ensuring the accuracy and validity of the analytical results obtained from its application. On the other hand, these assumptions present certain challenges, as under specific conditions some assumptions or requirements for achieving measurement precision may be difficult to satisfy with data collected from the administration of a measurement instrument. For example, in certain situations, an instrument designed to measure mathematical problem-solving ability and administered to students in a final semester examination or a national examination may yield data indicating the presence of more than one dominant construct, such as general mathematical problem-solving ability, spatial ability, and numerical ability [27]. The local independence assumption may also be violated for certain pairs of items, given that in test item construction there is a possibility that one item is closely related to another.

In addition, in the application of modern test theory, calibrating two or more test forms that measure the same construct onto a common scale to allow for comparison requires a stage in which students' abilities are treated as a random sample from a population that is normally distributed [28]. In most educational measurement practices, it is challenging to obtain data that reflect student ability characteristics that are normally distributed. Lastly, the use of modern test theory, despite its advantages, faces challenges related to the sample size required to reveal student abilities with a high degree of accuracy. The study conducted by [29] indicated that a sample size of at least 150 is required to accurately estimate the parameters of items comprising a 10-item test under the one-parameter logistic (1-PL) model. Meanwhile, the study by [30] recommended a sample size of more than 1,000 for a 20-item test, at least 700 for a 40-item test, and more than 1,000 for a 50-item test with the first component variance of less than 10% to accurately estimate item parameters under IRT models. Although Rasch modeling can be applied to small sample sizes, such as 30 respondents, similar to other statistical analyses, it may lead to less precise parameter estimates and less powerful fit analyses [31]. This issue of sample size presents a challenge in many educational settings where class sizes are relatively small, often around 30 students. Small sample sizes also pose challenges for achieving high precision and producing stable item parameter estimates in the process of equating test scores across multiple test forms [32].

Modern test theory offers numerous advantages for application in measurement practice. However, its implementation also faces various challenges arising from its underlying assumptions, the requirements necessary for specific applications, and the adequacy of sample size. One potential alternative to address these challenges is to integrate modern test theory models with deep learning. Among the two measurement modeling approaches under modern test theory, this study focuses on Rasch modeling for dichotomous items. This focus was made in light of previous studies indicating that Deep-IRT can serve as an alternative to IRT, particularly the two-parameter logistic (2-PL) model [33], [34], while little is known about its potential when it comes to the integration of deep learning and Rasch modeling. Moreover, this focus is motivated by the fact that educational research often involves relatively small sample sizes. On one hand, there is a need to obtain precise estimates of students' abilities; on the other hand, the available data conditions often make the use of IRT impractical, leaving Rasch modeling as a viable alternative to the limitations of CTT. Nevertheless, achieving accurate results through Rasch modeling requires that its underlying assumptions are satisfied, although these conditions are not always met in practice. In the present study, we introduce Deep-Rasch as a potential alternative to Rasch modeling by examining its performance in estimating test item parameters and examinee abilities, as well as the accuracy of these estimates based on empirical data.

2 Methods

This section outlines how we demonstrated the potential of the Deep-Rasch approach, beginning with a description of the dataset and continuing through the analytical procedures applied to it. It also provides an overview of the fundamental concepts underlying Rasch modeling, deep learning, and the integration of these two measurement frameworks, along with the operationalization of that integration in the present study. This overview serves to clarify the conceptual and methodological distinctions between the Deep-Rasch and the Rasch modeling.

2.1 Datasets

This study used empirical data in the form of student responses from Final Semester Examinations (*Ujian Akhir Semester*, UAS) in eight courses from four faculties (i.e., FHISIP, FEB, FST, and FKIP) at Universitas Terbuka (UT), with examination periods ranging from the odd semester of 2019 to the even semester of 2023. Combining the number of courses with their respective examination periods resulted in a total of 17 examination datasets. The number of examinees (sample size) in these datasets varied, ranging from 33 to 11,504. Of the 17 examinations, 16 had a test length of 30 multiple-choice items and one had a test length of 40 multiple-choice items. All items used in these examinations were first analyzed using the CTT approach, and the results indicated that each item was able to discriminate between test-takers' abilities, as evidenced by positive point-biserial correlation coefficients. The characteristics of the datasets used in this study are presented in Table 1.

Table 1: Characteristics of the datasets used in the present study

Course code	Subject	Faculty	Exam period	Sample size	Test length
ADBI4201	Business English	FHISIP	2019.1	4,783	30
ESPA4123	Economic Statistics	FEB	2019.1	9,211	30
ESPA4224	Economic and Business Statistics	FEB	2019.1	1,122	30
ISIP4215	Introduction to Social Statistics	FHISIP	2019.1	7,781	30
ADBI4201	Business English	FHISIP	2019.2	9,421	30
ESPA4123	Economic Statistics	FEB	2019.2	6,586	30
ISIP4215	Introduction to Social Statistics	FHISIP	2019.2	8,152	30
SATS4211	Statistical Methods II	FST	2019.2	33	30
PEMA4210	Educational Statistics	FKIP	2022.2	6,709	30
SATS4111	Computer I	FST	2022.2	1,415	30
SATS4111	Computer I	FST	2023.1	2,799	30
ADBI4201	Business English	FHISIP	2023.1	6,293	30
PEMA4210	Educational Statistics	FKIP	2023.1	11,504	30
SATS4211	Statistical Methods II	FST	2023.1	265	30
ESPA4424	Economic and Business Statistics	FEB	2023.2	769	40
SATS4111	Computer I	FST	2023.2	2,840	30
SATS4211	Statistical Methods II	FST	2023.2	197	30

Note. FHISIP = *Fakultas Hukum, Ilmu Sosial, dan Ilmu Politik* (Faculty of Law, Social and Political Sciences), FEB = *Fakultas Ekonomi dan Bisnis* (Faculty of Economics and Business), FST = *Fakultas Sains dan Teknologi* (Faculty of Science and Technology), and FKIP = *Fakultas Keguruan dan Ilmu Pendidikan* (Faculty of Teacher Training and Education).

2.2 Rasch Modeling and Deep-Rasch

This study focuses on two analytical frameworks: the Rasch model and the Deep-Rasch model—the latter being a reformulation of the Rasch model within a deep learning framework. Although the Rasch model can be viewed as either a model-based or data-based measurement approach [35], it is more commonly classified as the former to distinguish it from the 1-PL IRT model. A key implication of adopting a model-based perspective is that if the response patterns of all examinees on a given item do not conform to the model, the response data for that item are excluded from the analysis.

The Rasch model defines the probability of an individual i correctly responding to a dichotomous item j ($X_{ij} = 1$) as a logistic function of the individual's ability level (θ_i) and the estimated difficulty parameter of item j (β_j), both located on a common logit scale (Eq. 1) [22], [36], [37]. The item difficulty parameter represents the level of ability required for an examinee to have a 50% chance of responding correctly. Within the Rasch model, all dichotomous items in a test are assumed to have identical discrimination (fixed at 1). Rasch modeling is built upon several assumptions, two of which are fundamental: unidimensionality and local independence [22], [38]. The unidimensionality assumption requires that the test measures a single dominant latent trait, while local independence posits that an individual's latent ability is the sole determinant of the responses to any pair of dichotomous test items.

$$P_j(\theta_i) = P(X_{ij} = 1 \mid \theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} \tag{1}$$

In this study, we propose a Rasch modeling approach based on deep learning, referred to as *Deep-Rasch*. Deep learning, a branch of machine learning, is grounded in multilayer artificial neural networks. The multiple layers in such architectures are designed to extract features from the input data, with the term “deep” denoting the presence of multiple network layers.

The network framework employed in Deep-Rasch consists of an *examinee layer* and an *item layer*, modeled independently. The outputs of these two subnetworks are then combined to compute the probability that an examinee answers a given item correctly. Fig. 1 presents the network structure of Deep-Rasch, adapted from the Deep-IRT framework proposed by [33]. The implementation in this study follows the analytical procedure systematically described in [33], [34].

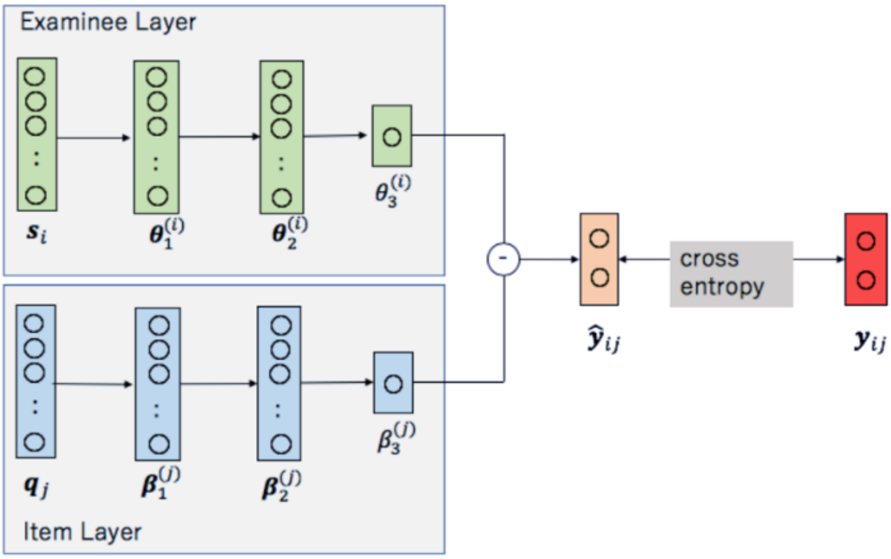

Figure 1: Outline of Deep-Rasch, adapted from Deep-IRT [33, p. 4]

In the examinee layer, $\mathbf{s}_i \in \{0, 1\}^I$ represents examinee i as a one-hot encoded binary vector, with a value of 1 in the i-th position and 0 elsewhere, where I denotes the total number of examinees. The examinee layer consists of three node layers as defined in Eqs. 2–4, where $\mathbf{W}^{(\theta_1)}, \mathbf{W}^{(\theta_2)}, \mathbf{W}^{(\theta_3)}$ are the weight matrices, and $\boldsymbol{\tau}^{(\theta_1)}, \boldsymbol{\tau}^{(\theta_2)}, \tau^{(\theta_3)}$ are the corresponding bias parameters. The final layer output $\theta_3^{(i)}$ represents the estimated ability parameter for examinee i. The first two layers use the hyperbolic tangent activation function defined in Eq. 5.

$$\boldsymbol{\theta}_1^{(i)} = \tanh\left(\mathbf{W}^{(\theta_1)}\mathbf{s}_i + \boldsymbol{\tau}^{(\theta_1)}\right) \tag{2}$$

$$\boldsymbol{\theta}_2^{(i)} = \tanh\left(\mathbf{W}^{(\theta_2)}\boldsymbol{\theta}_1^{(i)} + \boldsymbol{\tau}^{(\theta_2)}\right) \tag{3}$$

$$\theta_3^{(i)} = \mathbf{W}^{(\theta_3)}\boldsymbol{\theta}_2^{(i)} + \tau^{(\theta_3)} \tag{4}$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{5}$$

A similar architecture applies to the item layer, where $\mathbf{q}_j \in \{0, 1\}^J$ represents item j as a one-hot vector, with J being the total number of items. The item layer also comprises three node layers (Eqs. 6–8), where $\mathbf{W}^{(\beta_1)}, \mathbf{W}^{(\beta_2)}, \mathbf{W}^{(\beta_3)}$ denote the weight matrices, and $\boldsymbol{\tau}^{(\beta_1)}, \boldsymbol{\tau}^{(\beta_2)}, \tau^{(\beta_3)}$

represent the bias terms. The final layer output $\beta_3^{(j)}$ represents the estimated difficulty parameter for item j.

$$\boldsymbol{\beta}_1^{(j)} = \tanh\left(\mathbf{W}^{(\beta_1)}\mathbf{q}_j + \boldsymbol{\tau}^{(\beta_1)}\right) \tag{6}$$

$$\boldsymbol{\beta}_2^{(j)} = \tanh\left(\mathbf{W}^{(\beta_2)}\boldsymbol{\beta}_1^{(j)} + \boldsymbol{\tau}^{(\beta_2)}\right) \tag{7}$$

$$\beta_3^{(j)} = \mathbf{W}^{(\beta_3)}\boldsymbol{\beta}_2^{(j)} + \tau^{(\beta_3)} \tag{8}$$

Under the Deep-Rasch framework, Eq. 9 defines the probability that examinee i answers item j correctly through the hidden layer $\mathbf{h}^{(i,j)} = (h_0^{(i,j)}, h_1^{(i,j)})$, where $\mathbf{h}^{(i,j)} = \mathbf{W}^{(y)T}(\theta_3^{(i)} - \beta_3^{(j)}) + \boldsymbol{\tau}^{(y)}$. Here, $\mathbf{W}^{(y)}$ denotes the weight vector and $\boldsymbol{\tau}^{(y)}$ represents the bias term. Unlike the conventional Rasch model, Deep-Rasch does not assume that examinee abilities or item difficulty parameters follow any specific statistical distribution.

$$\widehat{y}_{i,j} = \text{softmax}(\mathbf{h}^{(i,j)}) = \frac{\exp(h_1^{(i,j)})}{\exp(h_0^{(i,j)}) + \exp(h_1^{(i,j)})} \tag{9}$$

For parameter estimation, Deep-Rasch employs the backpropagation algorithm to minimize a cross-entropy-based loss function, which quantifies the classification error. Equation 10 defines this loss, where $u_{i,j}$ and $\widehat{y}_{i,j}$ denote the actual and predicted responses of examinee i to item j, respectively.

$$\mathcal{L}(u_{i,j}, \widehat{y}_{i,j}) = -u_{i,j}\log(\widehat{y}_{i,j}) - (1 - u_{i,j})\log(1 - \widehat{y}_{i,j}) \tag{10}$$

2.3 Data Analysis

The data analysis in this study proceeded from examining the fulfillment of the two fundamental assumptions underlying the Rasch model to evaluating the performance of the Rasch and Deep-Rasch models in predicting examinees' responses to test items. The extent to which the unidimensionality assumption was satisfied was examined using principal component analysis (PCA), yielding the eigenvalues of the first and second factors. The empirical data were considered to support unidimensionality when the ratio of the first to the second eigenvalue exceeded 3 [39], [40].

Meanwhile, the assumption of local independence was assessed using Yen's Q_3 statistic [41], which reflects the correlation between each pair of items in the test. A pair of items was deemed to indicate local dependence when the absolute value of Yen's Q_3 statistic exceeded 0.2 [42]. By examining the extent to which these two assumptions were satisfied, it was anticipated that some datasets might fail to meet at least one of them, thus strengthening the relevance of exploring Deep-Rasch as an alternative to conventional Rasch modeling.

After verifying the two fundamental assumptions of Rasch modeling, the analysis proceeded with the estimation of item difficulty parameters and the evaluation of item fit to the Rasch model. The assessment of fit between observed item responses and model expectations was based on two mean square (MNSQ) residual summary statistics—namely, the information-weighted or inlier-sensitive fit (Infit MNSQ) and the outlier-sensitive fit (Outfit MNSQ) [43], [44]. Acceptable model fit was indicated when MNSQ values ranged between 0.5 and 1.5 [45]. Although both indices are typically reported, greater emphasis is generally placed on Infit MNSQ, as it is less affected by outliers and more sensitive in detecting patterns that indicate inconsistencies between item difficulty and examinee ability estimates within the Rasch model [43], [46], [47]. After item fit was evaluated, ability parameters were subsequently estimated.

The potential of Deep-Rasch as an alternative to the Rasch model was demonstrated through the estimation of item difficulty and ability parameters under two data-splitting scenarios. The first scenario involved splitting the data by examinees. In this scenario, each dataset was randomly divided into two groups containing half (or nearly half) of the total number of examinees, while the number of test items remained identical to that of the original dataset. The correlation between the estimated item difficulty parameters across the two groups was then computed to assess model performance, with higher correlations indicating greater consistency and stability in parameter estimation for both the Rasch and Deep-Rasch models.

The second data-splitting scenario was based on test items. In this scenario, the data were randomly divided into two groups, each consisting of half (or nearly half) of the total test items, while the number of examinees was maintained as in the original dataset. The estimated ability parameters of examinees obtained from the two resulting datasets were then correlated and used as indicators of model performance, with higher correlations reflecting better model stability and generalizability. Each data-splitting scenario was repeated ten times, and the mean of the obtained correlations was computed.

In addition, the area under the curve (AUC) value was calculated for each data-splitting scenario to evaluate the predictive performance of the models in classifying the correctness of examinees' responses. The AUC, derived from the receiver operating characteristic (ROC) curve, is commonly employed to evaluate the performance of binary classification methods. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR). The AUC represents the area under the ROC curve, ranging from 0.5 to 1.0, and is generally expected to exceed 0.8 [48], indicating that the model demonstrates good predictive performance in classifying response correctness.

All analyses were conducted using the R programming language [49] within the RStudio environment [50]. Several R packages were employed to support data processing and model implementation, including *openxlsx* [51], *psych* [52], *EFA.dimensions* [53], *mirt* [54], *Metrics* [55], *tensorflow* [56], and *keras* [57].

The TensorFlow [56] and Keras [57] frameworks were used within the RStudio environment to implement the Deep-Rasch model under two modeling configurations: (1) a standard model using binary cross-entropy loss and (2) a customized model incorporating additional penalty components derived from the response and item distributions. Each configuration consisted of two parallel subnetworks representing examinees and items, respectively, each containing two hidden layers of 50 neurons with tanh activation functions, followed by linear layers that produced latent estimates of examinee ability and item difficulty parameters. These latent representations were then combined through a subtraction layer, and the resulting difference was passed to a final sigmoid output layer representing the probability of a correct response. The models were trained using the Adam optimizer [58], with a batch size of 1, a validation split of 0.2, and 30 to 50 epochs, depending on the training scenario.

3 Results and Discussion

This section presents the results of the analysis conducted on datasets comprising student responses from 17 examinations administered at Universitas Terbuka (UT). The analysis aimed to explore the potential of Deep-Rasch, a measurement framework that integrates deep learning with Rasch modeling. Initial findings revealed that several datasets violated one or both of the key assumptions of Rasch modeling—namely, unidimensionality and local independence—thereby underscoring the relevance of applying Deep-Rasch to the data. The subsequent sections compare the performance of the two measurement frameworks in estimating item parameters (i.e., item difficulty) and examinee ability parameters, followed by an evaluation of their predictive accuracy in modeling student responses. Each set of findings is accompanied by a brief discussion to contextualize the results.

3.1 Assessment of Unidimensionality and Local Independence

A fundamental prerequisite for estimating both item parameters and student abilities within modern test theory, including Rasch modeling, is the fulfillment of two core assumptions. The first is *unidimensionality*, which stipulates that the administered test measures a single dominant latent trait or ability. The second is *local independence*, which asserts that a student's response to a given item is determined solely by their underlying ability and is not influenced by responses to other items. Violation of either assumption undermines the validity and accuracy of the resulting item and ability parameter estimates within the Rasch framework.

Table 2: Assessment of unidimensionality and local independence assumptions

Course code and exam period	λ_1	λ_2	λ_1/λ_2	**Min Q_3**	**Max Q_3**
ADBI4201-2019-1	3.964	0.849	4.669	−0.131	0.106
ADBI4201-2019-2	4.622	0.496	9.313	−0.101	0.079
ADBI4201-2023-1	4.061	0.407	9.980	−0.114	0.075
ESPA4123-2019-1	2.652	0.557	4.766	−0.169	0.186
ESPA4123-2019-2	2.393	0.528	4.530	−0.199	0.107
ESPA4224-2019-1	2.433	0.773	3.146	−0.193	0.145
ESPA4424-2023-2	**2.247**	**1.845**	**1.217**	**−0.392**	**0.389**
ISIP4215-2019-1	1.690	0.388	4.356	−0.182	0.073
ISIP4215-2019-2	1.630	0.421	3.872	−0.115	0.097
PEMA4210-2022-2	**0.848**	**0.720**	**1.177**	−0.136	0.113
SATS4111-2022-2	**3.449**	**1.676**	**2.057**	**−0.370**	**0.253**
SATS4111-2023-1	**3.459**	**1.290**	**2.681**	**−0.234**	**0.387**
SATS4111-2023-2	**3.087**	**1.110**	**2.782**	**−0.236**	**0.322**
SATS4211-2019-2	**2.931**	**1.984**	**1.478**	**−0.447**	**0.557**
PEMA4210-2023-1	**1.533**	**0.731**	**2.097**	**−0.264**	**0.206**
SATS4211-2023-1	**1.154**	**0.741**	**1.557**	**−0.245**	0.197
SATS4211-2023-2	**1.098**	**0.808**	**1.359**	**−0.235**	**0.230**

Table 2 summarizes the results of the assessment of the two assumptions. The findings indicate that data from nine examinations (shown in bold) did not provide sufficient evidence to support the assumption that the corresponding instruments measured a single dominant construct or ability. This conclusion is based on the ratio of the first to the second eigenvalue being less than 3 [39], [40]. Furthermore, several item pairs in eight examinations exhibited local dependence, as indicated by the absolute value of the minimum or maximum Yen's Q_3 statistic exceeding 0.2 [42]. Consequently, the empirical data satisfied both assumptions underlying Rasch modeling in only eight of the seventeen datasets analyzed.

3.2 Comparison of the Consistency of Difficulty Parameter Estimates Between Rasch Modeling and Deep-Rasch

This section presents the results on the performance of Rasch and Deep-Rasch with respect to their consistency in estimating item difficulty parameters across 17 examinations (datasets). Item difficulty parameters were estimated using two measurement frameworks—Rasch and Deep-Rasch—under a data-splitting design based on the number of examinees. In this design, each framework produced difficulty estimates for the first and second halves of the data, which were then correlated. The procedure was repeated ten times, and the mean correlation across replications was computed. The mean correlations of item difficulty estimates derived from Rasch modeling and Deep-Rasch are presented in Table 3.

Table 3: Mean correlation of item difficulty estimates using Rasch and Deep-Rasch under examinee-based splitting

Course code	Subject	Sample size	Rasch	Deep Rasch	Difference
ADBI4201	Business English	4,783	0.998	0.988	0.011
ESPA4123	Economic Statistics	9,211	0.998	0.982	0.016
ESPA4224	Economic and Business Statistics	1,122	0.981	0.949	0.032
ISIP4215	Introduction to Social Statistics	7,781	0.998	0.984	0.014
ADBI4201	Business English	9,421	0.999	0.985	0.015
ESPA4123	Economic Statistics	6,586	0.995	0.981	0.014
ISIP4215	Introduction to Social Statistics	8,152	0.998	0.982	0.015
SATS4211	Statistical Methods II	33	0.133	0.574	-0.441
PEMA4210	Educational Statistics	6,709	0.997	0.979	0.018
SATS4111	Computer I	1,415	0.992	0.970	0.022
SATS4111	Computer I	2,799	0.997	0.977	0.020
ADBI4201	Business English	6,293	0.998	0.983	0.014
PEMA4210	Educational Statistics	11,504	0.996	0.973	0.023
SATS4211	Statistical Methods II	265	0.932	0.883	0.049
ESPA4424	Economic and Business Statistics	769	0.976	0.950	0.026
SATS4111	Computer I	2,840	0.995	0.980	0.015
SATS4211	Statistical Methods II	197	0.930	0.896	0.034

The results in Table 3 indicate that item difficulty estimates obtained from the two halves under the Rasch model yielded higher mean correlations than those estimated under Deep-Rasch for most datasets. For the Statistical Methods II examination with only 33 examinees, the mean correlations between halves were relatively low under both approaches; nevertheless, the correlations from the Rasch model were weak or negligible, whereas those from the Deep-Rasch model were moderate [59]. In addition, the results suggest that Deep-Rasch yields greater consistency than Rasch in estimating item difficulty when applied to a very small sample size. Together with the fact that the Statistical Methods II dataset ($n = 33$) did not satisfy unidimensionality and local independence, these findings highlight the potential of Deep-Rasch for small samples and in contexts where the fundamental Rasch assumptions are violated. Although the Rasch model is often considered usable with samples of around 30, these results align with literature advising caution in small-sample applications [31], [60].

3.3 Comparison of the Consistency of Person Ability Parameter Estimates Between Rasch Modeling and Deep-Rasch

The comparison between Rasch and Deep-Rasch was further extended to their consistency in estimating examinee ability. Datasets were randomly split into two equal halves based on test length (number of items). For each half, examinee ability was estimated using both models. The estimated abilities from the two halves were then correlated. This procedure was repeated ten times, and the mean correlation was computed. Results are reported in Table 4.

As shown in Table 4, mean correlations of examinee abilities across item-based halves are generally lower than those for item difficulty parameters. This pattern is expected given non-homogeneous item difficulty distributions: random item splits often produce halves with different difficulty compositions, weakening between-half ability correlations. When considered relative to sample size, the results suggest that Deep-Rasch provides greater consistency than Rasch in estimating examinee abilities when the number of examinees is relatively small (e.g., fewer than 500). This advantage of Deep-Rasch over Rasch modeling also appears under conditions where the two fundamental Rasch assumptions are not satisfied.

Table 4: Mean correlation of examinee ability estimates using Rasch and Deep-Rasch under item-based splitting

Course code	Subject	Sample size	Rasch	Deep Rasch	Difference
ADBI4201	Business English	4,783	0.629	0.583	0.046
ESPA4123	Economic Statistics	9,211	0.487	0.415	0.071
ESPA4224	Economic and Business Statistics	1,122	0.441	0.410	0.031
ISIP4215	Introduction to Social Statistics	7,781	0.308	0.309	-0.001
ADBI4201	Business English	9,421	0.675	0.625	0.050
ESPA4123	Economic Statistics	6,586	0.457	0.436	0.021
ISIP4215	Introduction to Social Statistics	8,152	0.367	0.341	0.027
SATS4211	Statistical Methods II	33	0.246	0.293	-0.047
PEMA4210	Educational Statistics	6,709	0.180	0.166	0.014
SATS4111	Computer I	1,415	0.625	0.528	0.097
SATS4111	Computer I	2,799	0.617	0.526	0.091
ADBI4201	Business English	6,293	0.606	0.485	0.121
PEMA4210	Educational Statistics	11,504	0.138	0.088	0.051
SATS4211	Statistical Methods II	265	0.152	0.200	-0.048
ESPA4424	Economic and Business Statistics	769	0.440	0.431	0.009
SATS4111	Computer I	2,840	0.554	0.426	0.128
SATS4211	Statistical Methods II	197	0.149	0.181	-0.032

3.4 Evaluation of Examinee Response Predictions

Both the Rasch and Deep-Rasch models allow for predicting examinees' response status on test items, that is, the probability of providing a correct response. This section evaluates the predictive performance of both models. The evaluation was conducted using two data-splitting designs: (1) splitting by test length (number of items) and (2) splitting by examinees. For each split, response status was predicted, and the corresponding area under the curve (AUC) value was computed. Each procedure was repeated ten times, and the mean AUC across replications was calculated. Table 5 and Table 6 respectively present the mean AUC values under the item-based and examinee-based data-splitting scenarios.

Table 5: Mean AUC values under the item-based data-splitting design

Course code	Subject	Sample size	Rasch	Deep Rasch	Difference
ADBI4201	Business English	4,783	0.834	0.787	0.047
ESPA4123	Economic Statistics	9,211	0.738	0.726	0.012
ESPA4224	Economic and Business Statistics	1,122	0.728	0.703	0.025
ISIP4215	Introduction to Social Statistics	7,781	0.732	0.719	0.013
ADBI4201	Business English	9,421	0.825	0.785	0.041
ESPA4123	Economic Statistics	6,586	0.714	0.707	0.007
ISIP4215	Introduction to Social Statistics	8,152	0.729	0.711	0.017
SATS4211	Statistical Methods II	33	0.448	0.729	-0.281
PEMA4210	Educational Statistics	6,709	0.689	0.688	0.001
SATS4111	Computer I	1,415	0.780	0.769	0.011
SATS4111	Computer I	2,799	0.798	0.784	0.014
ADBI4201	Business English	6,293	0.781	0.767	0.014
PEMA4210	Educational Statistics	11,504	0.651	0.650	0.002
SATS4211	Statistical Methods II	265	0.687	0.687	0.000
ESPA4424	Economic and Business Statistics	769	0.712	0.699	0.013
SATS4111	Computer I	2,840	0.769	0.760	0.009
SATS4211	Statistical Methods II	197	0.694	0.691	0.003

Table 6: Mean AUC values under the examinee-based data-splitting design

Course code	Subject	Sample size	Rasch	Deep Rasch	Difference
ADBI4201	Business English	4,783	0.839	0.807	0.031
ESPA4123	Economic Statistics	9,211	0.749	0.747	0.001
ESPA4224	Economic and Business Statistics	1,122	0.740	0.729	0.011
ISIP4215	Introduction to Social Statistics	7,781	0.737	0.739	-0.002
ADBI4201	Business English	9,421	0.833	0.808	0.025
ESPA4123	Economic Statistics	6,586	0.726	0.726	0.000
ISIP4215	Introduction to Social Statistics	8,152	0.737	0.734	0.003
SATS4211	Statistical Methods II	33	0.744	0.725	0.020
PEMA4210	Educational Statistics	6,709	0.696	0.715	-0.020
SATS4111	Computer I	1,415	0.789	0.782	0.007
SATS4111	Computer I	2,799	0.804	0.791	0.013
ADBI4201	Business English	6,293	0.788	0.780	0.008
PEMA4210	Educational Statistics	11,504	0.660	0.684	-0.024
SATS4211	Statistical Methods II	265	0.692	0.713	-0.022
ESPA4424	Economic and Business Statistics	769	0.713	0.708	0.005
SATS4111	Computer I	2,840	0.781	0.776	0.005
SATS4211	Statistical Methods II	197	0.697	0.716	-0.020

Tables 5 and 6 generally show that both Rasch and Deep-Rasch performed well in predicting examinees' response status across the two data-splitting scenarios. When AUC values were derived from the item-based split, Deep-Rasch demonstrated higher accuracy than Rasch for the smallest sample ($n = 33$). In contrast, for the examinee-based split, no consistent pattern was observed across sample size conditions indicating that one model consistently outperformed the other.

This finding contrasts with that of [33], who reported that Deep-IRT outperformed IRT in predicting examinee responses across datasets with varying sample sizes and test lengths, particularly for sample sizes between 26 and 1,139. Although a consistent performance pattern was not observed between Rasch and Deep-Rasch in the present study, the mean AUC differences between the two frameworks were minimal in both data-splitting scenarios. Considering that several datasets did not satisfy the assumptions underlying Rasch modeling, Deep-Rasch again emerges as a viable alternative for estimating item difficulty and examinee ability parameters as well as for predicting response status under non-ideal measurement conditions.

4 Conclusion

This study has demonstrated the potential of Deep-Rasch as an alternative to Rasch modeling for use in measurement practice. To this end, we conducted an analysis of empirical data consisting of student responses to examinations from eight courses, comprising 17 examinations or datasets. The analysis of real data revealed that not all data from the test forms used met the assumptions of unidimensionality and local independence, which are fundamental assumptions of modern test theory, including Rasch modeling. Failure to meet one or both of these assumptions led to less accurate estimates of item parameters and test-taker abilities when the data were still analyzed within the Rasch measurement framework. This issue was further exacerbated in one examination or dataset with the number of examinees (sample size) of fewer than 50, a particularly challenging situation for Rasch analysis, even though among modern test theory models, Rasch is generally more accommodating of small sample sizes compared to IRT. Based on the empirical data, Deep-Rasch was shown to have potential as an alternative for estimating examinee's ability and item difficulty parameters when one or both of the assumptions underlying Rasch modeling are not met and when the sample size is very small. Nonetheless, this study is limited to the use of empirical data, which constrains our ability to gain deeper insights into the performance of Deep-Rasch under a broader range of conditions. Therefore, simulation studies involving varied conditions or factors, such as sample size and test length, are deemed necessary to provide a more comprehensive understanding of Deep-Rasch performance in terms of its consistency or stability in estimating examinee ability and item difficulty and to determine the conditions under which it yields more consistent or stable estimates.

CRediT Authorship Contribution Statement

Agus Santoso: Conceptualization, Investigation, Writing – Review & Editing, Supervision, Funding Acquisition. **Farit Mochamad Afendi:** Conceptualization, Methodology, Software, Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing. **Timbul Pardede:** Writing – Review & Editing, Funding Acquisition. **Heri Retnawati:** Conceptualization, Writing – Review & Editing, Supervision. **Ibnu Rafi:** Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing. **Ezi Apino:** Formal Analysis, Writing – Review & Editing. **Munaya Nikma Rosyada:** Writing – Review & Editing.

Declaration of Generative AI and AI-assisted technologies

ChatGPT was used solely to assist in the initial translation of this manuscript from Bahasa Indonesia into English. The AI-generated text was subsequently reviewed, edited, and revised by the authors to ensure accuracy, clarity, and fidelity to the original meaning.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding and Acknowledgments

This work was financially supported by the Institute for Research and Community Service (*Lembaga Penelitian dan Pengabdian kepada Masyarakat*, LPPM), Universitas Terbuka (UT), through the Applied Research Scheme under the Annual Work Plan and Budget for Fiscal Year 2024 (Grant Number: B/627/UN31.LPPM/PT.01.03/2024). The authors also wish to thank the UT Testing Center (*Pusat Pengujian*) for providing access to student response data from the odd and even semester final examinations and for granting permission to utilize the data in this study.

Data and Code Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request, subject to confidentiality constraints.

References

[1] C. D. Desjardins and O. Bulut, *Handbook of Educational Measurement and Psychometrics Using R*. Boca Raton, FL: CRC Press, 2018.

[2] R. K. Hambleton and R. W. Jones, "Comparison of classical test theory and item response theory and their applications to test development," *Educational Measurement: Issues and Practice*, vol. 12, no. 3, pp. 38–47, 1993. DOI: 10.1111/j.1745-3992.1993.tb00543.x.

[3] I. Rafi, H. Retnawati, E. Apino, D. Hadiana, I. Lydiati, and M. N. Rosyada, "What might be frequently overlooked is actually still beneficial: Learning from post national-standardized school examination," *Pedagogical Research*, vol. 8, no. 1, pp. 1–15, 2023. DOI: 10.29333/pr/12657.

[4] S. E. Stemler and A. Naples, "Rasch measurement v. item response theory: Knowing when to cross the line," *Practical Assessment, Research, & Evaluation*, vol. 26, no. 11, pp. 1–16, 2021. DOI: 10.7275/V2GD-4441.

[5] C. Zanon, C. S. Hutz, H. Yoo, and R. K. Hambleton, "An application of item response theory to psychological test development," *Psicologia: Reflexao e Critica*, vol. 29, no. 1, pp. 1–10, 2016. DOI: 10.1186/s41155-016-0040-x.

[6] T. Pardede et al., "Gaining a deeper understanding of the meaning of the carelessness parameter in the 4pl irt model and strategies for estimating it," *REID (Research and Evaluation in Education)*, vol. 9, no. 1, pp. 86–117, 2023. DOI: 10.21831/reid.v9i1.63230.

[7] A. Santoso, "Karakteristik butir tes pengantar statistika sosial berdasarkan teori respon butir," *Jurnal Pendidikan Matematika dan Sains*, vol. 6, no. 2, pp. 158–168, 2018. DOI: 10.21831/jpms.v6i2.23959.

[8] A. Santoso et al., "From investigating the alignment of a priori item characteristics based on the ctt and four-parameter logistic (4-pl) irt models to further exploring the comparability of the two models," *Practical Assessment, Research & Evaluation*, vol. 29, no. 14, pp. 1–28, 2024. DOI: 10.7275/pare.2043.

[9] L. Tesio, A. Caronni, D. Kumbhare, and S. Scarano, "Interpreting results from rasch analysis 1. the 'most likely' measures coming from the model," *Disability and Rehabilitation*, vol. 46, no. 3, pp. 591–603, 2024. DOI: 10.1080/09638288.2023.2169771.

[10] A. Robitzsch, "On the choice of the item response model for scaling PISA data: Model selection based on information criteria and quantifying model uncertainty," *Entropy*, vol. 24, no. 6, pp. 1–26, 2022. DOI: 10.3390/e24060760.

[11] P. W. van Rijn, S. Sinharay, S. J. Haberman, and M. S. Johnson, "Assessment of fit of item response theory models used in large-scale educational survey assessments," *Large-scale Assessments in Education*, vol. 4, no. 1, pp. 1–23, 2016. DOI: 10.1186/s40536-016-0025-3.

[12] A. Rahim, S. Hadi, D. Susilowati, Marlina, and Muti'ah, "Developing of computerized adaptive test (cat) based on a learning management system in mathematics final exam for junior high school," *International Journal of Educational Reform*, 2023. DOI: 10.1177/10567879231211297.

[13] M. D. Reckase, "The influence of computerized adaptive testing (cat) on psychometric theory and practice," *Journal of Computerized Adaptive Testing*, vol. 11, no. 1, pp. 1–12, 2024. DOI: 10.7333/2403-1101001.

[14] R. Rukli and N. A. Atan, "Simulation of low-high method in adaptive testing," *REiD (Research and Evaluation in Education)*, vol. 10, no. 1, pp. 35–49, 2024. DOI: 10.21831/reid.v10i1.66922.

[15] A. Santoso, K. Kartianom, and G. K. Kassymova, "Kualitas butir bank soal statistika (studi kasus: Instrumen ujian akhir mata kuliah statistika universitas terbuka)," *Jurnal Riset Pendidikan Matematika*, vol. 6, no. 2, pp. 165–176, 2019. DOI: 10.21831/jrpm.v6i2.28900.

[16] B. Kartowagiran, S. Munadi, H. Retnawati, and E. Apino, “The equating of battery test packages of mathematics national examination 2013–2016,” in *SHS Web of Conferences*, A. G. Abdullah, J. Foley, I. G. N. A. Suryaputra, and A. Hellman, Eds., Bali, Indonesia: EDP Sciences, 2018, pp. 1–6. DOI: 10.1051/shsconf/20184200022.

[17] H. Retnawati, “Perbandingan metode penyetaraan skor tes menggunakan butir bersama dan tanpa butir bersama,” *Jurnal Kependidikan: Penelitian Inovasi Pembelajaran*, vol. 46, no. 2, pp. 164–178, 2016. DOI: 10.21831/jk.v46i2.10383.

[18] E. Yusron, H. Retnawati, and I. Rafi, “Bagaimana hasil penyetaraan paket tes usbn pada mata pelajaran matematika dengan teori respons butir?” *Jurnal Riset Pendidikan Matematika*, vol. 7, no. 1, pp. 1–12, 2020. DOI: 10.21831/jrpm.v7i1.31221.

[19] M. Kleinman and J. A. Teresi, “Differential item functioning magnitude and impact measures from item response theory models,” *Psychological Test and Assessment Modeling*, vol. 58, no. 1, pp. 79–98, 2016.

[20] H. Penton, C. Dayson, C. Hulme, and T. Young, “An investigation of age-related differential item functioning in the EQ-5D-5L using item response theory and logistic regression,” *Value in Health*, vol. 25, no. 9, pp. 1566–1574, 2022. DOI: 10.1016/j.jval.2022.03.009.

[21] A. Santoso et al., “The effect of scoring correction and model fit on the estimation of ability parameter and person fit on polytomous item response theory,” *REiD (Research and Evaluation in Education)*, vol. 8, no. 2, pp. 140–151, 2022. DOI: 10.21831/reid.v8i2.54429.

[22] D. Andrich and I. Marais, *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences*. Singapore: Springer Nature Singapore, 2019. DOI: 10.1007/978-981-13-7496-8.

[23] T. G. Bond, Z. Yan, and M. Heene, *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 4th ed. New York, NY: Routledge, 2021.

[24] R. K. Hambleton and H. Swaminathan, *Item Response Theory: Principles and Applications*. New York, NY: Springer Science+Business Media, 1985.

[25] P. Mair, *Modern Psychometrics with R*. Cham, Switzerland: Springer International Publishing, 2018.

[26] S. A. Wind and C. Hua, *Rasch Measurement Theory Analysis in R*. Boca Raton, FL: CRC Press, 2022. DOI: 10.1201/9781003174660.

[27] H. Retnawati, Kumaidi, G. K. Kassymova, M. Socheath, and O. Ndayizeye, “How many dimensions and items were enough for mathematics test in national examination? (application of multidimensional logistic model in item response theory),” in *AIP Conference Proceedings*, Yogyakarta, Indonesia, 2022, p. 050 012. DOI: 10.1063/5.0111066.

[28] M. D. Barrett and W. J. van der Linden, “Estimating linking functions for response model parameters,” *Journal of Educational and Behavioral Statistics*, vol. 44, no. 2, pp. 180–209, 2019. DOI: 10.3102/1076998618808576.

[29] A. Şahin and D. Anıl, “The effects of test length and sample size on item parameters in item response theory,” *Educational Sciences: Theory & Practice*, vol. 17, no. 1, pp. 321–335, 2017. DOI: 10.12738/estp.2017.1.0270.

[30] P. Sopa, P. Tuksino, and P. Makmee, “The appropriate sample size for test length on estimating item parameters in item response theory,” *Journal of Research Methodology*, vol. 36, no. 2, pp. 144–163, 2023.

[31] J. M. Linacre, “Sample size and item calibration [or person measure] stability,” *Rasch Measurement Transactions*, vol. 7, no. 4, p. 328, 1994.

[32] T. R. O'Neill, J. L. Gregg, and M. R. Peabody, "Effect of sample size on common item equating using the dichotomous rasch model," *Applied Measurement in Education*, vol. 33, no. 1, pp. 10–23, 2020. DOI: 10.1080/08957347.2019.1674309.

[33] E. Tsutsumi, R. Kinoshita, and M. Ueno, "Deep item response theory as a novel test theory based on deep learning," *Electronics*, vol. 10, no. 9, pp. 1–19, 2021. DOI: 10.3390/electronics10091020.

[34] E. Tsutsumi, R. Kinoshita, and M. Ueno, "Deep-irt with independent student and item networks," in *Proceedings of the 14th International Conference on Educational Data Mining*, I.-H. Hsiao, S. Sahebi, F. Bouchet, and J.-J. Vie, Eds., Paris, France: International Educational Data Mining Society, 2021, pp. 510–517.

[35] A. Maydeu-Olivares and R. Montaño, "How should we assess the fit of rasch-type models? approximating the power of goodness-of-fit statistics in categorical data analysis," *Psychometrika*, vol. 78, no. 1, pp. 116–133, 2013. DOI: 10.1007/s11336-012-9293-1.

[36] D. Andrich, *Rasch Models for Measurement.* Sage Publications, 1988.

[37] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests.* University of Chicago Press, 1980.

[38] B. Hayat, "Adjustment for guessing in a basic statistics test for indonesian undergraduate psychology students using the rasch model," *Cogent Education*, vol. 9, no. 1, pp. 1–17, 2022. DOI: 10.1080/2331186X.2022.2059044.

[39] J. Hattie, "Methodology review: Assessing unidimensionality of tests and items," *Applied Psychological Measurement*, vol. 9, no. 2, pp. 139–164, 1985. DOI: 10.1177/014662168500900204.

[40] S. L. Slocum-Gori and B. D. Zumbo, "Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis," *Social Indicators Research*, vol. 102, no. 3, pp. 443–461, 2011. DOI: 10.1007/s11205-010-9682-8.

[41] W. M. Yen, "Effects of local item dependence on the fit and equating performance of the three-parameter logistic model," *Applied Psychological Measurement*, vol. 8, no. 2, pp. 125–145, 1984. DOI: 10.1177/014662168400800201.

[42] W.-H. Chen and D. Thissen, "Local dependence indexes for item pairs using item response theory," *Journal of Educational and Behavioral Statistics*, vol. 22, no. 3, pp. 265–289, 1997. DOI: 10.3102/10769986022003265.

[43] J. M. Linacre, "What do infit and outfit, mean-square and standardized mean?" *Rasch Measurement Transactions*, vol. 16, no. 2, p. 878, 2002.

[44] B. D. Wright and J. M. Linacre, "Reasonable mean-square fit values," *Rasch Measurement Transactions*, vol. 8, no. 3, p. 370, 1994.

[45] V. Aryadoust, L. Y. Ng, and H. Sayama, "A comprehensive review of rasch measurement in language assessment: Recommendations and guidelines for research," *Language Testing*, vol. 38, no. 1, pp. 6–40, 2021. DOI: 10.1177/0265532220927487.

[46] F. Quansah, "Item and rater variabilities in students' evaluation of teaching in a university in ghana: Application of many-facet rasch model," *Heliyon*, vol. 8, no. 12, e12548, 2022. DOI: 10.1016/j.heliyon.2022.e12548.

[47] M. Tavakol and R. Dennick, "Psychometric evaluation of a knowledge based examination using rasch analysis: An illustrative guide: Amee guide no. 72," *Medical Teacher*, vol. 35, no. 1, pp. 838–848, 2013. DOI: 10.3109/0142159X.2012.737488.

[48] Ş. K. Çorbacıoğlu and G. Aksel, "Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value," *Turkish Journal of Emergency Medicine*, vol. 23, no. 4, pp. 195–198, 2023. DOI: 10.4103/tjem.tjem_182_23.

[49] R Core Team, *R: A language and environment for statistical computing*, Online; accessed 2024, R Foundation for Statistical Computing, Vienna, Austria, 2024. Available online.

[50] Posit team, *Rstudio: Integrated development environment for r*, Online; accessed 2024, Posit Software, PBC, Boston, MA, 2024. Available online.

[51] P. Schauberger and A. Walker, *Openxlsx: Read, write and edit xlsx files*, R package version 4.2.5.2, 2023. Available online.

[52] W. Revelle, *Psych: Procedures for psychological, psychometric, and personality research*, R package version 2.4.6.26, Evanston, IL, 2024. Available online.

[53] B. P. O'Connor, *Efa.dimensions: Exploratory factor analysis functions for assessing dimensionality*, R package version 0.1.8.4, 2024. Available online.

[54] R. P. Chalmers, "Mirt: A multidimensional item response theory package for the r environment," *Journal of Statistical Software*, vol. 48, no. 6, pp. 1–29, 2012. DOI: 10.18637/jss.v048.i06.

[55] B. Hamner and M. Frasco, *Metrics: Evaluation metrics for machine learning*, R package version 0.1.4, 2018. Available online.

[56] J. J. Allaire and T. Yuan, *Tensorflow: R interface to "tensorflow"*, R package version 2.16.0, 2024. Available online.

[57] T. Kalinowski, J. J. Allaire, and F. Chollet, *Keras: R interface to "keras"*, R package version 2.15.0, 2024. Available online.

[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015, pp. 1–15. DOI: 10.48550/arXiv.1412.6980.

[59] H. Akoglu, "User's guide to correlation coefficients," *Turkish Journal of Emergency Medicine*, vol. 18, no. 3, pp. 91–93, 2018. DOI: 10.1016/j.tjem.2018.08.001.

[60] W.-H. Chen, W. Lenderking, Y. Jin, K. W. Wyrwich, H. Gelhorn, and D. A. Revicki, "Is rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? an example using PROMIS pain behavior item bank data," *Quality of Life Research*, vol. 23, no. 2, pp. 485–493, 2014. DOI: 10.1007/s11136-013-0487-5.