Comparative Evaluation of Three Item Response Theory Models within an Educational Measurement System: Application to the Moroccan National Student Assessment Program in Science

SARA BOUHAZZAMA, CHIADMI MOHAMED SALAH
Study and Research Laboratory in Applied Mathematics (LERMA),
Mohammadia School of Engineering, Mohammed V University,
Ibn Sina Avenue B.P 765, Agdal Rabat 10090,
MOROCCO

Abstract: - Morocco faces the dual challenge of enhancing the quality of education while ensuring equity of opportunities. Established in 2008, the National Program for Assessing Students' Achievements (PNEA) was conceived as a policy instrument to systematically address this concern. This article focuses on the 2019 edition of the program and uses IRT (Item Response Theory) to sequentially apply the one-, two-, and three-parameter models. The aim is to assess their appropriateness in capturing the characteristics of the dataset and enhancing the precision of student evaluations. The study contributes to the refinement of the framework for assessing academic achievement, in accordance with the overarching goal of enhancing educational quality.

Key-Words: - Educational Measurement, Educational Quality, Educational assessment in Morocco, Psychometric analysis, Item Response Theory (IRT), comparison of IRT models, Model selection and Goodness-of-Fit.

Received: November 4, 2025. Revised: February 7, 2026. Accepted: June 5, 2026. Published: July 21, 2026.

1 Introduction

Education constitutes the fundamental bedrock of a nation's development and future, standing resolutely as an indisputable pillar that imparts decisive momentum across all its essential sectors. Morocco grapples with the imperative task of enhancing the quality of education and ensuring equal opportunities for all students. To meet this objective, the assessment of students' learning outcomes in Morocco has become an essential lever for the development of effective educational policies; Therefore, the deployment of a robust psychometric tool for analyzing student performance becomes crucial.

In this regard, the National Evaluation Authority (INE), under the Higher Council for Education, Training, and Scientific Research of Morocco, has initiated the National Program for Assessing Students' Achievements (PNEA), launched in 2008 and having since undergone several developments. In its initial version, the program focused on evaluating the achievements of students in the fourth and sixth grades of primary school, as well as the second and third grades of middle school. In its second iteration in 2016, the program shifted its attention to assessing the level of students in the common core trunk. Subsequently, since 2019, the program has settled on assessing the achievements of students in the sixth grade of primary school and the third grade of middle school, as these represent critical stages in the educational journey, marking the end of primary education and the end of compulsory education. This article will study the case of the last principal test of the program administered in 2019.

This article aims to apply IRT to the PNEA dataset through a sequential implementation of the one-, two-, and three-parameter models. The study seeks to evaluate and compare the suitability of three IRT models in capturing the distinct characteristics of the PNEA data. The results will contribute to a more precise and robust framework for assessing students' academic achievements.

2 Problem Formulation

In Morocco, the National Evaluation Authority (INE), under the CSEFRS, has implemented the National Program for Assessing Students' Achievements (PNEA), which since 2019 has focused on key educational transition points, namely the end of 6th grade (primary school) and the end of 9th grade (lower secondary school). Fully realizing the potential of the data derived from these

assessments requires the implementation of rigorous and robust analytical frameworks. This study addresses the specific problem of identifying which IRT model (one-, two-, or three-parameter) best represents the psychometric properties of science items in the 2019 PNEA dataset for the third year of middle school in science.

IRT provides a systematic framework for analyzing assessment data, offering item-level parameters such as difficulty, discrimination, and guessing, while accounting for students' latent abilities. Unlike classical test theory, IRT produces results that are invariant across populations and test forms, making it particularly well-suited for large-scale national and international assessments.

3 Problem Solution

3.1 Theoretical Overview

Psychometric analysis is a crucial step in the validation process of any measurement instrument. In the context of assessing academic achievement, this involves validating the construct, which means the cognitive test. It relies on mathematical models derived primarily from two measurement theories: Classical Test Theory (CTT) and Item Response Theory (IRT).

3.1.1 Classical Item Psychometric Analysis

Charles Spearman is widely regarded as the founder of Classical Test Theory (CTT). Through his pioneering work on correlation and measurement, he introduced the concepts that gave rise to the true score model and laid the foundations of modern psychometric measurement, [1], [2], [3]. Over the following decades, CTT underwent continuous refinement through the contributions of numerous researchers. Among the most influential developments were advances in psychological measurement and test construction introduced in 1936 [4], followed by one of the first comprehensive and systematic presentations of the theory in 1950, which provided a book-length exposition that organized and consolidated the principles of Classical Test Theory for subsequent generations of researchers and practitioners, [5]. Its theoretical framework was further strengthened through an axiomatic and mathematical formalization in 1966, [6]. Later, in 1986, the theoretical foundations and practical applications of both classical and modern test theory were comprehensively synthesized, further consolidating CTT as a rigorous and widely adopted framework for educational and psychological measurement, [7].

Classical psychometric analysis defines an individual's true score on a test as the mathematical expectation of the observed score. It follows that the observed score X is the result of two unobserved components, namely the true score V and a measurement error E (Equation (1)).

$$X = V + E \tag{1}$$

The classical model relies on the following postulates:

- Measurement error in a normal distribution has a mean of zero.
- The error is uncorrelated with the true score.
- The correlation between errors in two different tests is zero.
- Two tests are parallel if their true scores are equal and the variances of their measurement errors are equal.
- The correlation between the errors of one test and the true score of another test is zero.
- Two tests are said to be k-equivalent if their true scores differ by a constant.

Analyzing items through the classical approach enables a comprehensive evaluation of the measurement instrument's quality, both at the global level: by assessing overall test reliability, and at the local level: by examining individual item psychometric properties specifically difficulty and discrimination.

- **Difficulty of an Item:**

The difficulty of an item, or its success rate, which refers to the proportion of respondents who answer this item correctly, has a significant impact on the test score. Therefore, item selection should consider this parameter: an item with either a very low or very high success rate is not very informative and should be excluded from the calculation of the overall test score. Furthermore, knowing the difficulty index (or p-index) of the items in a test can also help to prioritize them in the order they are administered. The easiest items should then be placed at the beginning of the test for a smooth progression. Thus, apart from a minimum number of very difficult or very easy items, the majority of items should have a p-index that fluctuates slightly around an average of 0.5.

- **Discrimination Power of an Item:**

When assessing individuals based on a given criterion (e.g., academic performance), items with high discrimination power are preferred. Various methods are used to assess this characteristic, with the point-biserial correlation coefficient (r_pbis)

being among the most commonly used. This coefficient, also known as the item-total correlation (rit), measures the correlation between passing or failing an item and the total score.

Due to the item's inclusion in the total score calculation, there is a risk of a biased upward r_pbis. Therefore, it is recommended to use a corrected r_pbis called the item-rest correlation (rir), achieved by removing the studied item from the total score. By recoding items as 0 (in case of failure) and 1 (in case of success), the calculation formula for item j is as follows (Equation (2)):

$$r_{it} = \mathrm{r_{pbis}} = \frac{\mathrm{M_r - M_e}}{\sigma_t}\sqrt{\mathrm{p_j q_j}} \qquad (2)$$

where:
M_r: Mean of total scores obtained across all items of the test by students who answered item j correctly.
M_e: Mean of total scores obtained across all items of the test by students who did not answer item j correctly.
σ_t: Standard deviation of the distribution of total scores.
p_j: Proportion of students who answered item j correctly.
q_j: Proportion of subjects who did not answer item j correctly.

Like any correlation coefficient, rpbis ranges from -1 to +1: the higher the rpbis, the more discriminatory the item, whereas a negative rpbis reflects an inconsistency, indicating the paradox where “the best students fail the item while the weakest ones pass”.

Furthermore, when the item is multiple-choice, it is also important to examine responses to distractors by assessing, for each of them, the percentage of individuals who chose it, along with its rpbis. Those selected by few individuals or not selected at all are considered irrelevant, while those with a positive rpbis represent inconsistencies (distractors chosen by the strongest individuals).

- **Test Reliability**

Reliability refers to the precision with which an instrument measures the aptitude or ability being estimated. In classical theory, it is defined as the correlation between the observed score and the true score. This definition highlights the close relationship between measurement error and reliability. A stronger correlation leads to reduced measurement errors and higher reliability. In this study, reliability is approximated using internal consistency, which evaluates how much each test item measures the same dimension. The most commonly used indicator for measuring test reliability in terms of internal consistency is Cronbach's Alpha (α) (Equation (3)):

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k}\sigma_{Y_j}^2}{\sigma_X^2}\right) \qquad (3)$$

where:
k : Number of items
σ_X^2 : Variance of total score
$\sigma_{Y_j}^2$: Variance of item j

On a scale from 0 to 1, the higher this index, the more reliable the measuring instrument. Generally, the homogeneity of the test is considered satisfactory when the Alpha is greater than or equal to 0.70.

3.1.2 Limitations of Classical Test Theory

Classical Test Theory (CTT) has been in use for about a century and is still widely employed due to its suitability for specific scenarios and its simplicity, making it accessible to individuals without formal training in psychometrics. However, its simplicity comes with limitations in addressing several critical measurement challenges, [8], [9]. Here are some of these limitations:

Sample Dependency: CTT heavily relies on the characteristics of the sample used in the test. This means that test results and their psychometric parameters (such as reliability and validity) may be limited to the population on which the test was applied.

Test Dependency: A person's "true score" and measurement error are defined relative to a specific test, making it difficult to compare scores from different tests measuring the same construct.

Guessing: CTT has no mechanism to handle guessing on multiple-choice items. A low-ability student who guesses correctly on a hard item gets the same score as a high-ability student who knows the answer.

Scoring: In CTT, scoring is based on raw sums or proportions. That means every item contributes equally to the total score, regardless of how well it discriminates or how much information it provides.

Scaling: CTT lacks the capability to perform vertical scaling because of the absence of the item parameter invariance property, which is important for comparing assessments over time.

3.1.3 Item Response Theory (IRT)

Item Response Theory (IRT) emerged in the 1950s and 1960s as a revolutionary framework in psychometrics and educational measurement. It was pioneered by Frederic M. Lord and Georg Rasch, with Rasch's one-parameter logistic model (1PL)

being an early foundational concept, [10], [11]. The 1PL model, later known as the Rasch model, introduced the idea of using a single difficulty parameter for items and a single ability parameter for respondents. This simplicity was further enhanced in the 1960s with the introduction of the two-parameter logistic model (2PL), which added item discrimination as a second parameter.

IRT quickly found its place in educational testing, leading to the development of standardized tests. Over time, the 3PL model, introduced in the 1970s, included a third parameter to account for guessing behavior, improving the precision of item calibration and ability estimation. Advances in estimation methods, including maximum likelihood estimation and Bayesian estimation, made IRT models more accurate in fitting to real-world data, [12], [13].

Item parameters in IRT, such as difficulty, discrimination, or pseudo-chance, remain invariant relative to the distribution of the group of individuals used for their estimation. This property of invariance justifies the use of IRT:

- Estimating an individual's ability is independent of the items they respond to.
- Estimations of item characteristics are independent of the characteristics of individuals responding to the items.
- Only one trait is measured by the test items (Unidimensionality).
- A candidate's response to one item is not influenced by their response to another item (Local independence).
- The probability of answering correctly increases with the individual's ability level (Monotonicity).

In fact, the first condition to ensure the property of invariance is, of course, that the model fits the observed data. If this is not the case, either the model needs to be changed or the item should be modified, and the parameter calibration process for that item should be repeated.

Thus, IRT allows us, on the one hand, to examine the psychometric properties of the items comprising the tests to select those that meet the required criteria for difficulty, discrimination, and fit. On the other hand, it enables us to place students' proficiency levels and item difficulty on the same scale.

The fundamental assumption of Item Response Models (IRMs) is that a subject's performance on an item can be explained by a factor known as the latent trait. This latent trait could be a personality trait, a cognitive ability, or a general skill. Regardless of its nature, it is always denoted as θ (theta). As in the probabilistic model, the relationship between item performance and the latent trait can be described using a function called the Item Characteristic Curve (ICC). Since the function is usually asymptotic, the curve typically takes the following form:

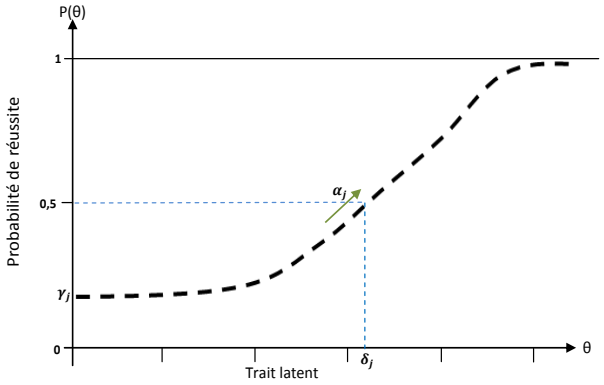

Fig. 1: Example of an Item Characteristic Curve (ICC)
Source: created by the authors

This curve is increasing and monotonic, thus showing that as the individual's ability, represented by the latent trait θ, increases, their probability of responding to the item also increases.

These are probabilistic models in which the probability that an individual responds correctly to an item is a function of their characteristics and those of the item. This function is called the characteristic function and is translated by the following Equation (4):

$$P(j = k \setminus \theta, \xi) = F(\theta, \xi, k) \qquad (4)$$

where:
$\boldsymbol{P}$: Probability that an individual with latent attribute θ answers item j with response k.
$\boldsymbol{\theta}$: Attribute or characteristic of the individual.
$\boldsymbol{\xi}$: Represents the characteristics of the item.

In educational assessment, θ (theta) denotes a student's proficiency level in a given subject. Since proficiency cannot be observed directly, it is estimated from students' actual responses to test items using IRT models, which translate observable response patterns into calibrated measures of underlying ability. At the same time, the item parameters ξ, such as difficulty, discrimination, and guessing, are estimated jointly with θ, ensuring that both student ability and item characteristics are placed on the same measurement scale.

- **One-Parameter Models**

This model is represented by the following logistic function (Equation (5)):

$$P_j(\theta) = \frac{1}{1 + e^{-D(\theta - \delta_j)}} \quad (5)$$

where:

$\boldsymbol{P_j(\theta)}$: Probability that an individual with characteristic θ responds correctly to item j;

$\boldsymbol{D}$: A constant, which in practice is typically assigned the value of 1.7 to make the characteristic curve closely resemble the normal ogive.

When the constant D is set to 1, the model is referred to as the Rasch model.

$\boldsymbol{\delta_j}$: Item j's difficulty parameter, as illustrated in Figure 1, is the θ value corresponding to a probability of success equal to 0.5.

- **Two-Parameter Models**

The probability function (Equation (6)) of this model involves, in addition to the difficulty, the discrimination parameter $\boldsymbol{\alpha_j}$.

$$P_j(\theta) = \frac{1}{1 + e^{-D\alpha_j(\theta - \delta_j)}} \quad (6)$$

$\boldsymbol{\alpha_j}$**:** A parameter that measures, as in the case of classical theory, the discriminative power of the item or its ability to distinguish individuals according to their ability. In the context of IRT, discrimination is expressed by the maximum slope of the characteristic curve.

- **Three-Parameter Models**

In the case of multiple-choice items, individuals have a nonzero probability of choosing the correct answer purely by chance. In IRT, this phenomenon is called pseudo-chance (guessing), and the models that take it into consideration are known as three-parameter models (Equation (7)):

$$P_j(\theta) = \frac{1 - \gamma_j}{1 + e^{-D\alpha_j(\theta - \delta_j)}} \quad (7)$$

where $\boldsymbol{\gamma_j}$ is the pseudo-chance (Guessing) parameter representing the probability of success for low levels of competence.

IRT models provide the following information:

- Measures of difficulty, discrimination, and pseudo-chance, depending on the adopted model.
- The measurement of students' abilities.
- Indicators of data fit to the model's quality (outfit, infit, Chi-Square Fit Statistics for the IRT Models...)
- Item characteristic curves: A curve that relates an individual's ability (latent trait) to the probability of passing the item.

3.1.5 Model Selection and Goodness-of-Fit

The selection of the optimal IRT framework necessitates a rigorous evaluation of the trade-off between statistical parsimony and measurement precision. This process is structured around three analytical levels:

- **Global Model Selection:** The 1PL, 2PL, and 3PL models are compared primarily on the basis of the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These indices serve as penalized likelihood measures where lower values denote superior model-data fit while accounting for over-parameterization. To complement these heuristics, a Likelihood Ratio Test (LRT) provides a statistical basis for determining whether the increased complexity of multi-parameter models yields a significant improvement in representing the observed response patterns.
- **Item-Level Diagnostics:** Beyond global fit, the structural integrity of the scale is verified using Infit and Outfit mean-square (MNSQ) residuals. These statistics detect items exhibiting anomalous behavior: Infit focuses on unexpected responses near the item's difficulty level, while Outfit is sensitive to outlying observations at the extremes of the ability scale. In line with psychometric standards, values between 0.7 and 1.3 are considered acceptable. Additionally, Chi-Square (X^2) Fit Statistics are employed to formally test the null hypothesis that observed item frequencies align with model expectations.
- **Measurement Precision and Reliability:** The Test Information Function (TIF) offers a localized view of precision across the latent trait continuum (θ), identifying where the assessment is most informative. By integrating the TIF, we derive the IRT empirical reliability. This model-based metric serves as a robust counterpart to the traditional Cronbach's Alpha, allowing for a direct comparison of measurement consistency and ensuring that the final model maximizes the accuracy of student ability estimates.

3.2 Methodology and Results

3.2.1 Data

The database for this study comprises a set of 17071 observations and 30 variables. It consolidates the scores of third-year college students, assessed on 27

items related to the sciences. The data is coded using a binary scale, where the value 1 represents a correct response.

3.2.2 Software

The study was conducted using R, with psychometric packages (TAM and ltm) and general-purpose packages for data handling and visualization (readxl, writexl, dplyr, tidyverse, ggplot2, etc.). R Markdown was used to generate reproducible outputs integrating code, results, and tables.

3.2.3 Determination of the Appropriate Item Response Theory (IRT) Model

The present study applies Item Response Theory (IRT) to the PNEA dataset with the primary objective of identifying the model that best represents the psychometric structure of the data. To this end, three IRT models were systematically evaluated: the Rasch model (1PL), the Two-Parameter Logistic model (2PL), and the Three-Parameter Logistic model (3PL). Each of these models offers a different approach to modeling item characteristics and estimates the parameters underlying individual responses, [14].

Analyzing the obtained results, we conducted a comparison of model performances using evaluation criteria such as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and other measures of fit quality. This process allowed us to identify the model that exhibits the best fit and is most suitable for our specific data. Below, we present the results of the comparison of the three IRT models for the database.

Table 1. Results of the comparison of the performance of 1-parameter and 2-parameter (IRT) models

Model	AIC	BIC	log.Lik	LRT	df	p.value
IRT1pl	580808	581017	-290377			
IRT2pl	573338	573756	-286615	7524	27	<0.001

Source: created by the authors

Table 2. Results of the comparison of the performance of 2-parameter and 3-parameter (IRT) models

Model	AIC	BIC	log.Lik	LRT	df	p.value
IRT2pl	573338	573756	-286615			
IRT3pl	571547	572174	-285692	1845	27	<0.001

Source: created by the authors

The model comparison results reported in Table 1 and Table 2 provide a clear basis for selecting the most appropriate IRT framework for the PNEA dataset. Two complementary criteria were used to guide this selection.

- The Akaike Information Criterion (AIC) evaluates the goodness of fit of a statistical model while penalizing for the number of estimated parameters, with lower values indicating a more favorable fit. The obtained values follow a consistent decreasing pattern across models: 580808 for the Rasch model, 573338 for the 2PL model, and 571547 for the 3PL model.
- The Bayesian Information Criterion (BIC) operates on the same principle but applies a more stringent penalty for model complexity, making it a particularly conservative criterion when comparing models of differing parameterization. The BIC values mirror the trend observed for the AIC: 581017 for the Rasch model, 573756 for the 2PL model, and 572174 for the 3PL model.

Based on these results, we can therefore conclude that the Tree-Parameter IRT model is the most suitable for parameter estimation in the IRT model.

3.2.4 Comparative Analysis of Model Fit: Test Information Function and Reliability

To further evaluate the relative performance of the three models, a comparative analysis of their Test Information Functions and associated reliability estimates was conducted, the results of which are presented in Figure 2.

Figure 2 displays the Test Information Functions (TIF) for the three IRT models under comparison — the 1PL (Rasch), 2PL, and 3PL — together with their respective IRT-based reliability estimates and a common reference value for Cronbach's Alpha ($\alpha = 0.70$), computed on the full dataset.

In terms of reliability, all three models perform at a broadly comparable level: 0.726 for the 1PL model, 0.725 for the 2PL model, and 0.729 for the 3PL model. Although the differences are numerically modest, the 3PL model attains the highest reliability estimate, a finding that aligns with and reinforces the advantage already observed through the AIC and BIC criteria.

Turning to the shape of the information curves, the 3PL model concentrates its peak measurement precision in the upper range of the ability continuum ($\theta \approx 1$ to 2). This pattern suggests that the model is particularly well-suited to distinguishing among students whose proficiency exceeds the average, capturing fine-grained differences in ability that the simpler models are less equipped to detect. In

contrast, the 1PL model provides more information in the central ability range ($\theta \approx 0$), indicating a more uniform measurement across ability levels. The 2PL model occupies an intermediate position. These differences reflect the influence of the discrimination and guessing parameters, which are progressively incorporated in the 2PL and 3PL models.

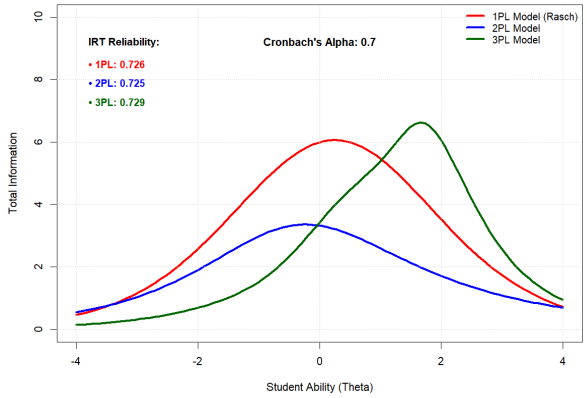

Fig. 2: Test Information Function and Reliability of the 1PL, 2PL, and 3PL IRT Models (science test, 2019)
Source: created by the authors

Overall, Figure 2 confirms that the 3PL model not only fits the data best according to information criteria but also achieves the best measurement precision for the population assessed in the PNEA 2019.

3.2.5 Estimation of Item Response Theory (IRT) Model Parameters (3-Parameter Model)

The estimation of parameters for the 3PL model is carried out using advanced statistical techniques, such as the maximum likelihood method. This process enables us to obtain precise parameter estimates, enhancing our understanding of the nature of items and the underlying skills measured by these items.

Below are the results of the 3-parameter model estimation. This output presented in Table 3 includes the difficulty, discrimination, and guessing parameters for the remaining 27 items. Items Q5, Q8, and Q29 were removed following the initial estimation due to poor psychometric properties.

In addition to estimating the difficulty, discrimination, and guessing parameters, the Three-Parameter Response Theory (IRT) model also allows for the estimation of a crucial parameter: the individual proficiency level (θ). Represented by the variable θ (theta), this parameter reflects the latent level of competence of an individual assessed by the model.

Table 3. Results of the estimation of item parameters using the 3-parameter IRT model

Item	Guessing	Difficulty	Discrimination
Q1	0.39	0.10	1.81
Q2	0.02	-1.45	0.79
Q3	0.00	0.40	0.62
Q4	0.45	0.52	1.30
Q6	0.13	0.04	0.91
Q7	0.31	0.37	1.71
Q9	0.18	-0.02	0.63
Q10	0.26	0.06	1.87
Q11	0.22	2.12	0.85
Q12	0.01	-0.49	1.14
Q13	0.00	-0.19	1.38
Q14	0.41	1.33	1.48
Q15	0.32	1.08	2.00
Q16	0.00	0.78	0.29
Q17	0.10	0.89	1.11
Q18	0.23	0.80	2.08
Q19	0.32	3.09	0.83
Q20	0.29	1.71	2.14
Q21	0.31	2.22	1.94
Q22	0.00	0.99	0.88
Q23	0.24	2.38	1.10
Q24	0.22	1.62	2.93
Q25	0.18	1.92	1.77
Q26	0.24	1.80	1.95
Q27	0.24	0.86	1.54
Q28	0.10	2.70	1.07
Q30	0.01	1.34	0.62

Source: created by the authors

The parameter θ is often interpreted as a measure of mastery in a specific competency domain. It indicates the level of knowledge, skill, or performance of an individual in the domain assessed by the items. Higher values of θ correspond to a higher level of proficiency, while lower values indicate a lower level of proficiency. Table 4 presents an excerpt of the students' estimated proficiency levels (Ability) computed in R using a 3-parameter IRT model:

Table 4. Excerpt from the results of the estimation of student ability levels using 3-parameter IRT model

ID	Ability
Student_1	1.10
Student_2	-0.73
Student_3	-1.23
Student_4	-0.66
Student_5	-0.14
Student_6	0.86
Student_7	0.18
Student_8	1.43
Student_9	-0.73
Student_10	-0.70
Student_11	-0.62

Source: created by the authors

3.2.6 Item Fit Statistics

To assess the degree to which individual items conform to the expectations of each model, item-level fit statistics were examined for both the Rasch and the 3PL models, the results of which are summarized in Table 5.

Table 5 presents two complementary sets of fit statistics: Infit and Outfit mean-square statistics derived from the Rasch model on the left, and Chi-Square fit statistics for the 3PL model on the right. For the Rasch model, values between 0.7 and 1.3 are considered acceptable (Wilson, 2005).

Table 5. Item Fit Statistics for the Rasch Model (Infit and Outfit) and Chi-Square Fit Statistics for the 3PL Model

Item	Infit	Outfit	X^2 (3PL)	p-value
Q1	0.956	0.925	258.88	< 0.001
Q2	0.988	0.973	391.07	< 0.001
Q3	1.012	1.015	199.72	< 0.001
Q4	0.997	0.987	106.97	< 0.001
Q6	0.986	0.981	293.32	< 0.001
Q7	0.959	0.944	213.31	< 0.001
Q9	1.021	1.022	156.95	< 0.001
Q10	0.937	0.912	401.88	< 0.001
Q11	1.041	1.060	70.23	< 0.001
Q12	0.946	0.926	593.18	< 0.001
Q13	0.923	0.908	690.40	< 0.001
Q14	1.025	1.027	55.03	< 0.001
Q15	0.987	0.986	130.89	< 0.001
Q16	1.075	1.089	136.81	< 0.001
Q17	0.968	0.965	190.82	< 0.001
Q18	0.941	0.934	204.90	< 0.001
Q19	1.084	1.106	28.06	< 0.001
Q20	1.034	1.059	199.76	< 0.001
Q21	1.067	1.097	61.68	< 0.001
Q22	0.964	0.960	197.38	< 0.001
Q23	1.054	1.080	53.43	< 0.001
Q24	1.008	1.040	213.60	< 0.001
Q25	1.015	1.043	100.10	< 0.001
Q26	1.022	1.047	100.79	< 0.001
Q27	0.974	0.972	143.54	< 0.001
Q28	1.014	1.058	40.22	< 0.001
Q30	1.009	1.011	167.93	< 0.001

Source: created by the authors

All 27 items fall within this range, with Infit values spanning from 0.923 (Q13) to 1.084 (Q19) and Outfit values from 0.908 (Q13) to 1.106 (Q19), indicating that no item exhibits severely misfit behavior. It is worth noting, however, that items Q13, Q10, and Q12 display slightly lower values (below 0.93), which may point to a degree of redundancy or over-discrimination relative to the Rasch model's expectations.

For the 3PL model, all Chi-Square statistics reach statistical significance ($p < 0.001$), a result that is largely attributable to the large sample size ($N = 17071$), where even negligible deviations from model expectations become detectable. This finding should therefore be interpreted with caution, as statistical significance alone is not indicative of practical misfit. Among the items examined, Q11, Q14, Q19, Q21, and Q23 yield the smallest Chi-Square values, reflecting a closer empirical alignment with the 3PL model predictions.

Taken together, these fit statistics corroborate the selection of the 3PL model as the best-fitting and most informative framework for this dataset.

3.2.7 Representation of Item Characteristic Curves

Now that we have precise parameter estimates and student ability θ, we are ready to move on to a crucial step: representing the Item Characteristic Curves (ICCs). Item Characteristic Curves offer a direct visual representation of the psychometric properties of each item in the test. By plotting the probability of a correct response as a function of the estimated proficiency level of examinees, these curves make it possible to appreciate, in an intuitive and interpretable manner, how difficult an item is, how sharply it discriminates between students of differing ability, and how sensitive it is to variations along the proficiency continuum. When representing the ICCs, we will use the parameter estimates from the 3PL model to plot the characteristic curves for each item. The ICCs will provide insights into the expected performance of individuals based on their proficiency level, as well as differences in item difficulty and discrimination.

This visualization of Item Characteristic Curves plays a crucial role in our study. It allows us to identify the most discriminant and informative items, those that are the most challenging or easiest, and understand how each item contributes to measuring the skills of assessed individuals.

As depicted in Figure 3 and Figure 4, as the proficiency level (Ability) increases, the student is more likely to answer the item correctly, and vice versa, which is entirely logical. This curve also demonstrates that estimating the variable “Ability” allows us to determine whether the student will respond correctly or incorrectly to another item with similar characteristics.

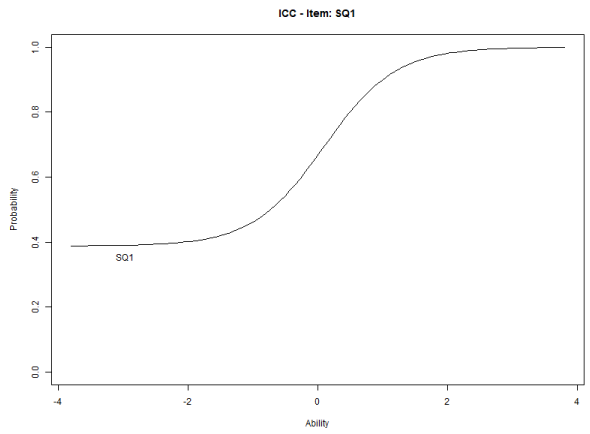

Fig. 3: Item Characteristic Curves for an Item of science for the third year of secondary school 2019
Source: created by the authors

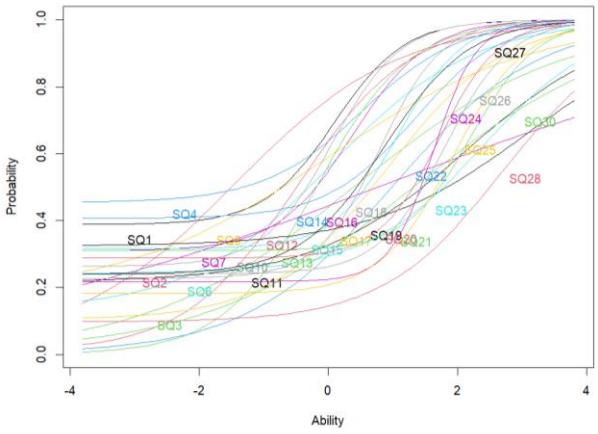

Fig. 4: Item Characteristic Curves for all items of science for the third year of secondary school 2019
Source: created by the authors

In other words, comparing the discrimination and difficulty parameters of the two items helps determine if they are similar. In the case of similarity, it is sufficient to plot the ICC curve of the new item and determine the probability of a correct response to that item by projecting the student's 'Ability,' even if the student has not seen it. This helps determine whether the student will respond correctly or not based on the obtained probability of a correct response.

4 Conclusion

The comparative analysis of the one-, two-, and three-parameter IRT models applied to the PNEA data highlights the importance of appropriately selecting psychometric models to accurately assess student achievements. The results of this study carry several implications for the assessment of student achievement in Morocco. Across all three IRT models tested on the PNEA science data, the 3PL model consistently provided the best fit, confirming that accounting for item discrimination and guessing yields a more accurate representation of student performance than the simpler Rasch framework alone. This has practical consequences: more precise ability estimates translate into fairer assessments and, by extension, into more targeted policies.

It's worth noting that a parallel analysis carried out on the mathematics component of the same PNEA (2019) led to identical conclusions. The convergence of findings across two different subjects strengthens confidence in the robustness of the 3PL model for this type of large-scale, nationally representative data. That said, these results should not be automatically extended to other subject areas or educational levels without further investigation. Subjects relying on open-ended items, or assessments designed for younger or older cohorts, may present different psychometric properties that could affect which model fits best. Future work should therefore examine whether these findings hold across the full range of PNEA assessments, with a view to building a comprehensive and coherent psychometric reference framework for Moroccan national evaluations.

References:

[1] C. Spearman, The Proof and Measurement of Association between Two Things, *The American Journal of Psychology*, Vol.15, No.1, 1904, pp.72–101. https://doi.org/10.2307/1412159.

[2] C. Spearman, Correlations of Sums and Differences, British Journal of Psychology, Vol. 6, No. 2, 1913, pp. 109–133. https://doi.org/10.1111/j.2044-8295.1913.tb00116.x.

[3] R. E. Traub, Classical Test Theory in Historical Perspective, *Educational Measurement: Issues and Practice*, Vol.16, No.4, 1997, pp.8–14. https://doi.org/10.1111/j.1745-3992.1997.tb00603.x.

[4] J. P. Guilford and R. B. Guilford, Personality Factors S, E, and M, and Their Measurement, *The Journal of Psychology: Interdisciplinary and Applied*, Vol.4, No.1, 1936, pp.67–102.

[5] H. O. Gulliksen, Theory of Mental Tests, John Wiley & Sons, New York, 1950, 486 p.

[6] M. R. Novick, The Axioms and Principal Results of Classical Test Theory, *Journal of Mathematical Psychology*, Vol.3, 1966, pp.1–

18. https://doi.org/10.1016/0022-2496(66)90002-6.
[7] L. Crocker and J. Algina, Introduction to Classical and Modern Test Theory, *Holt, Rinehart and Winston*, 1986.
[8] 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, 1993, pp.38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x.
[9] O. A. Awopeju and E. R. I. Afolabi, Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics Examination, *European Scientific Journal*, Vol.12, No.28, 2016, pp.264–282.
[10] F. M. Lord and M. R. Novick, Statistical Theories of Mental Test Scores, Addison-Wesley, Reading, MA, 1968.
[11] Andrich, D., Rasch Models for Measurement, *SAGE Publications*, 1988.
[12] M. D. Toland, Practical Guide to Conducting an Item Response Theory Analysis, *The Journal of Early Adolescence,* Vol.34, No.1, 2014, pp.120–151. https://doi.org/10.1177/0272431613511332.
[13] L. Tay, A. W. Meade and M. Cao, An Overview and Practical Guide to IRT Measurement Equivalence Analysis, *Organizational Research Methods*, Vol.18, No.1, 2015, pp.3–46. https://doi.org/10.1177/1094428114555994.
[14] X. Wang and S. W. Sympson, A Comparison of 1-, 2-, and 3-Parameter Item Response Models for Assessing Student Ability in Mathematics, *Educational and Psychological Measurement*, Vol.81, No.4, 2021, pp.785–806. https://doi.org/10.1177/0013164421991363.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.