AI for Multiple-Choice Tests Optimization: Towards Scores Regeneration Based on Effective Item Selection

Najoua HRICH[1,2], Charafeddin ElHADDOUCHI[2], Mohamed AZEKRI[3] and Mohamed KHALDI[1]

[1]Research team in Computer Science and University Pedagogical Engineering, [3]Higher Normal School, Abdelmalek Essaadi University, Tetouan, Morocco

[2] Regional Center of Education & Training Professions, Institutions for Higher Executive Training, Tangier, Morocco

[3] Regional Academy of Education & Training, Ministry of National Education Preschool and Sports, Tetouan, Morocco

nhrich@uae.ac.ma

Abstract. In the assessment field, the synergy between artificial intelligence (AI) and item response theory (IRT) opens new revolutionary perspectives. AI brings unprecedented data analysis and processing capabilities, while IRT provides a robust theoretical framework for modeling learners' responses to test items. This combination enables addressing the complex challenges of assessment more effectively and precisely than ever before.

This research explores the application of IRT in conjunction with AI in the assessment area. Focusing on score regeneration following the removal of improper items, our study employs AI to effectively identify and remove such items, thereby influencing candidate selection. Using key parameters of IRT, such as difficulty, discrimination, and chance, this research aims to develop a deep learning (DL) model for classifying and selecting items. The study specifically applies the Artificial Neural Network (ANN) method via binary classification models for this task. Experimental results show that the proposed model performs very well, with a high accuracy rate for item selection.

Keywords: Artificial Intelligence, Item Response Theory, Deep learning, Artificial Neural Network, Assessment, Multiple Choice Tests, Item Analysis.

1 Introduction

In educational assessment, the synergy between AI and IRT marks a decisive moment, offering unprecedented potential.

IRT provides a solid theoretical basis for overcoming the challenges associated with the credibility of assessments based on Multiple-Choice Tests (MCTs). At the core of IRT is the idea that a respondent's probability of providing the correct answer to an item depends both on their own skill level and on the characteristics of the item. By modeling the relationship between the examinees' latent traits and item parameters, IRT offers a robust means of analyzing test items and drawing meaningful inferences about individual abilities.[1]

M. Khaldi et al. (eds.), Proceedings of the E-Learning and Smart Engineering Systems (ELSES 2024), Atlantis Highlights in Social Sciences, Education and Humanities 36,
https://doi.org/10.2991/978-2-38476-408-2_8

On the other hand, AI has emerged as a transformative force in various fields, including education. With its capability to handle massive amounts of data, recognize intricate patterns, and make intelligent decisions, AI offers major potential for improving the assessment process. By exploiting AI techniques, educators and researchers can open new avenues for the creation of MCTs.[2]

The synergy between AI and IRT represents a paradigm shift in assessment practices, offering unprecedented opportunities to improve the validity, reliability, and fairness of assessments. By integrating AI-based approaches with IRT principles, researchers can address long-standing challenges such as item selection, test equivalence, and score interpretation with greater accuracy and efficiency.[3]

This research seeks to investigate the complementary relationship between AI and IRT within the realm of educational assessment, with particular emphasis on score regeneration of scores following the removal of inappropriate items. By using AI algorithms to identify and eliminate such items, this research explores how the fusion of AI and IRT principles can optimize test construction and improve the accuracy of assessment results. Through a systematic investigation of AI-based item selection techniques and their alignment with key IRT parameters, this investigation seeks to advance our understanding of how emerging technologies can augment traditional psychometric methodologies in assessment.

2 Background

2.1 Multiple-choice tests Based Assessments.

MCTs-based assessments are widely used in various educational and professional domains to assess individuals' knowledge and skills. However, despite their popularity, this approach has some notable limitations. These include the MCTs' limited ability to assess the depth of understanding, their tendency to promote rote memorization over conceptual comprehension, and their susceptibility to response strategy rather than true competence. In addition, item quality, in terms of difficulty and discriminatory ability, may vary, thus impacting the results accuracy and consistency [4]. Additionally, the random guessing effect can influence candidates' responses, introducing an additional source of error into the assessment [5] [6]. To address these challenges, the integration of IRT offers a robust methodological approach. By assessing the quality of each item, IRT enables precise analysis of difficulty and discrimination, helping to improve the overall quality of the assessment [7]. Consequently, the use of IRT complements MCTs by guaranteeing more precise and objective measures of individuals' skills. In response to these concerns, the application of IRT offers potential solutions. Item analysis can be used to assess the quality and relevance of each test question, including its difficulty and discriminative ability [8]. By incorporating IRT, which models the relationship between individuals' skills and their responses to items, it is possible to obtain more precise and objective measures of learners' skills, considering both item difficulty and respondents' abilities [9]. Thus, item analysis and the use of IRT offer a more robust

approach to assessing individuals' skills, enhancing the accuracy and consistency of assessment results

2.2 Item Response Theory

IRT is a psychometric model introduced relatively recently, in the later decades of the 20th century. to address issues that traditional psychometrics may not always satisfactorily addressed. Unlike classical psychometrics, IRT offers an estimate of item properties that is not specific to a particular sample of individuals [10]. For example, assessing an item's technical properties, such as difficulty or discrimination, provides results that are no longer merely relative to a specific sample, but rather generalizable for a wider population. IRT rests on the principle that a respondent's answer to an item is determined by two factors: the subject's attributes, known as "latent traits", and the properties of the item itself, such as its difficulty and discrimination power, as well as the influence of guessing in some situations [11]. Technically, IRT uses single- or multi-parameter models to establish the relationship between an individual's latent trait and the probability of succeeding on an item, graphically represented by an item characteristic curve. The aim of the method is twofold: to estimate the metric properties of items, such as difficulty and discrimination, as well as the latent skill level of individuals. Moreover, these estimates are assumed to be independent of the samples and items used in the study, thus guaranteeing n invariance property [1]. In summary, IRT offers a more robust and generalizable approach to assessing individuals' skills and abilities, with wide applications in various fields, including education and psychology.

Item parameters. IRT is founded on the principle that the likelihood of an individual answering an item correctly is influenced by both the individual's ability and the characteristics of the item. This probability is determined based on the item's parameters, including its difficulty, discrimination, and chance of random success. [12]. The following Table 1 summarizes of IRT item parameters.

Table 1. IRT Item parameters

Item parameter	Description
Difficulty	The difficulty parameter refers to the level of ability needed for an individual to have a 50% chance of answering an item correctly. Easy items are likely to be answered correctly even by individuals with lower ability, whereas difficult items require higher ability for a correct response.
Discrimination	Assesses the item’s capacity to differentiate between individuals with varying skill levels. Items with high discrimination effectively distinguish between those with high and low skills
The probability of success on an item	The likelihood that an individual will correctly answer an item without having the required ability. This probability is typically lower for multiple-choice items compared to open-response items.

Using sophisticated mathematical models such as the two-parameter item response model (2-PLM) or the three-parameter item response model (3-PLM), the IRT estimates individuals' latent ability and item parameters, enabling in-depth analysis of test-taker performance and item characteristics.

IRT Models. In IRT, three main models are widely used to assess individuals' responses to items [1-5].
Rasch model (1PL): This model assumes that there is only one parameter for each item, representing its difficulty. It does not consider item discrimination.
Two-parameter model (2PL): This model adds a second parameter for each item, representing its discrimination ability.
Three-parameter model (3PL): This model includes a third parameter for each item, representing the probability of success by chance.

2.3 Overview of Artificial Intelligence in Assessment

Artificial Intelligence. AI includes various techniques and algorithms aimed at creating systems capable of performing tasks that usually require human intelligence. In education, AI is increasingly utilized to enhance teaching and learning processes, with a particular focus on improving assessment methods.

Applications of AI in assessment include the automatic generation of exam questions, the automatic analysis of student responses, the personalization of tests to suit individual learners' needs, and the detection of fraud and plagiarism. By applying techniques such as machine learning, deep learning, and natural language processing, AI can automate many assessment-related tasks, enabling teachers and assessors to save time and provide feedback more quickly and efficiently [13] [14].

Synergy between AI and IRT in Addressing Assessment Challenges. The convergence of AI and IRT presents unique opportunities to address educational assessment challenges. By combining IRT's sophisticated modeling capabilities with AI's machine learning and deep learning algorithms, it is possible to develop more accurate, efficient, and equitable assessment systems.

By using AI to identify problematic items in tests and to select items according to test-takers characteristics, it is possible to improve the validity and reliability of assessments. In addition, by using AI to personalize tests according to learners' individual needs, it is possible to provide fairer and more inclusive assessments.

By exploiting the synergy between AI and IRT, it becomes feasible to push back the boundaries of traditional assessment methods and develop new approaches that better meet the diverse needs of learners in the modern world.[15]

3 Methodology Design

3.1 Description of the Research Design

This study adopts an experimental approach to explore the application of AI in conjunction with IRT in the field of educational assessment. The main objective is to develop and evaluate a deep learning model for the automatic score regeneration, based on IRT parameters for selecting items, using artificial neural networks (ANNs).

Data collection and preparation. The data used in this study consist of a set of candidates' test results from a recruitment competition. The test comprises 100 items with 4 distractors each. Each item is scored as 1 point for a correct answer, with no penalty for incorrect answers.

The dataset consists of responses from 500 individuals who participated in recruitment competitions for prospective computer science teachers, hosted by the Ministry of National Education, Pre-school, and Sports during the 2022 and 2023 academic years.

Results are stored in an Excel file with two sheets, each for one academic year. Columns represent individual test items, and rows correspond to candidates. Each cell shows a candidate's score on an item (1 for a correct answer and 0 for an incorrect one). A sample of the dataset is illustrated in Figure 1.

Items	Id_C1	Id_C2	Id_C3	Id_C4	Id_C5	Id_C6	Id_C7	Id_C8	Id_C9
Item1	0	0	1	1	1	0	1	0	1
Item2	0	0	1	1	0	0	1	0	1
Item3	1	0	1	1	1	0	0	1	0
Item4	0	0	0	0	0	0	0	1	0
Item5	0	0	0	0	1	0	0	0	1
Item6	1	1	0	0	0	1	0	1	0
Item7	1	0	0	1	0	0	0	1	0

Fig. 1. Extract of dataset

Deep learning model. The proposed model is based on Artificial Neural Networks (ANNs), a powerful technique that can capture complex patterns in data. Specifically, we have used the binary classification model for test item selection.

As shown in figure 2, the model will take as input the difficulty parameter generated by the PyIRT library in Python. We trained the model using a training dataset comprising examples of selected and non-selected items and evaluated its performance on a separate dataset.

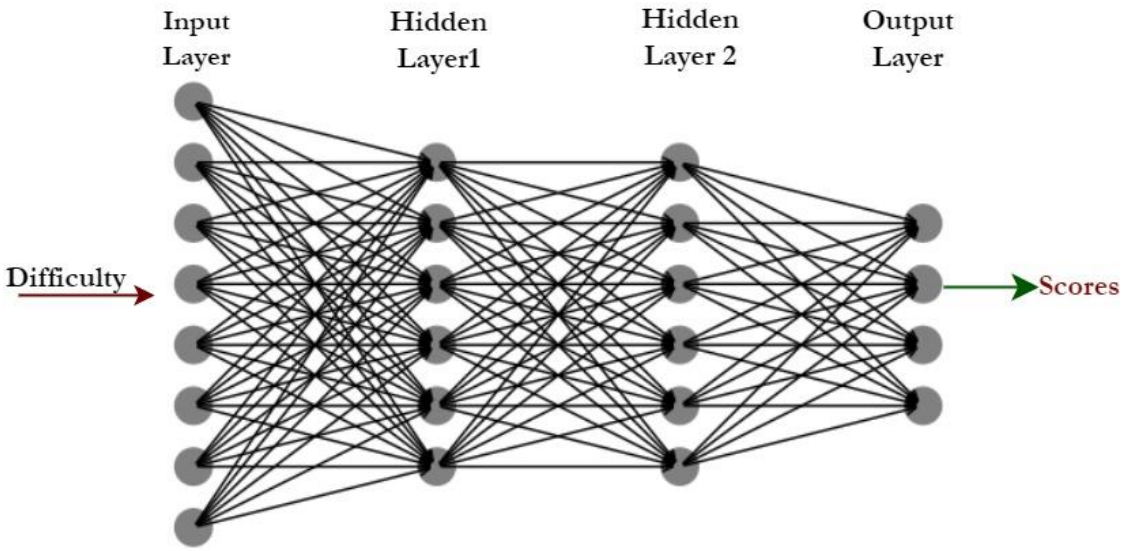

Fig. 2. Deep Neural Network Architecture

Training Procedure. The model is trained using Adam's optimization algorithms, by minimizing the loss function defined for the binary classification task. We also employed regularization techniques such as dropout to prevent model overfitting

Testing Procedure.

Evaluating the model on a real data set is critical to assessing its effectiveness. During this phase, the model will be validated to determine its ability to accurately identify relevant items and regenerate examinee scores. The results will provide critical insight into the model's performance and practical relevance in real-world applications.

For model testing, data from the first Excel spreadsheet is used for training, which provides essential information for the model. Data from the second Excel sheet is reserved for testing, which allows us to gain an independent assessment of the model's performance on new data.

We have chosen the sigmoid function as our activation function and the binary cross-entropy as our loss function. We have selected the Adam optimizer with a learning rate of 0.001. This sequential approach allows us to implement the model in a structured and efficient manner.

Evaluation procedure. The model is evaluated by measuring its performance on an independent test dataset, using performance metrics to assess its ability to correctly select test items. The most frequently used metrics are presented in Table 2.

Table 2. Performance Metrics

Performance Metric	Definition	Formula

Accuracy	Accuracy measures the ratio of correct predictions to the total number of predictions, offering an overall view of the model's ability to classify both positive and negative instances. Although it indicates how often the model's predictions are correct, accuracy may be less meaningful in situations with class imbalance, where classes are not evenly distributed.	$\frac{\text{True Positive} + \text{True negative}}{Total\ number\ of\ predictions}$
Precision	Precision evaluates the accuracy of positive predictions by dividing the number of true positives by the total number of positive predictions (true positives plus false positives).	$\frac{\text{True positif}}{\text{True positif} + \text{False Positif}}$
Recall (Sensitivity)	It assesses the model's capacity to detect all positive class instances	$\frac{\text{True positif}}{\text{True positif} + \text{False Negative}}$
F1-Score	The F1-Score is the harmonic mean of precision and recall, offering a balanced measure of a model's performance, especially with imbalanced classes. It integrates both precision and recall into a single metric, reflecting the trade-off between false positives and false negatives. A higher F1-Score indicates a better balance between precision and recall, making it valuable for evaluating models on imbalanced datasets.	$\frac{2 \times \text{precision} \times \text{recal}}{\text{precision} + \text{recall}}$
Area under the Receiver Operating Characteristic Curve (AUC-ROC)	The AUC-ROC measures the model's overall ability to differentiate between positive and negative classes. It represents the area under the ROC curve, with a value of 1.0 indicating perfect classification and 0.5 suggesting random guessing. A higher AUC-ROC value reflects better performance in distinguishing between the classes	
Area under the Precision-Recall curve (AUC-PRC)	The AUC-PRC evaluates the model's overall performance by measuring the area under the precision-recall curve. A higher AUC-PRC value signifies better performance, with a value of 1.0 indicating perfect precision and recall at all thresholds.	

We have used the Scikit-learn libraries to calculate these metrics with a view to evaluating the performance of our neural network model. The results are presented in the following section.

Parameter optimization. We have also explored different model configurations, adjusting hyperparameters such as the layers' number, neurons per layer, and learning rates, enabling us to adopt a 4-layer network.

4 Results and Discussion

The model's evaluation uses a range of important metrics to offer a comprehensive overview of its performance. The confusion matrix gives a detailed breakdown of accurate and inaccurate predictions, demonstrating the model's capability to differentiate between relevant and irrelevant items.

A Python script utilizing the "sklearn. metrics" library was used to obtain these metrics, and the results are depicted in Figures 3 and 4.

A thorough analysis of these results validates the outstanding performance of the item selection model based on the difficulty parameter. The confusion matrix reveals that the model accurately identified 499 out of 540 items in class 0 (irrelevant items) and correctly predicted 612 items in class 1 (relevant items).

```
Confusion Matrix:
[[499  24]
 [ 41 612]]

Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.95      0.94       523
           1       0.96      0.94      0.95       653

    accuracy                           0.94      1176
   macro avg       0.94      0.95      0.94      1176
weighted avg       0.95      0.94      0.94      1176


Accuracy: 0.9447278911564626

Area under the ROC curve (AUC-ROC): 0.9456618811837701

Area under the Precision-Recall curve (AUC-PR): 0.9671704801138
```

Fig. 3. Detailed Performance Metrics

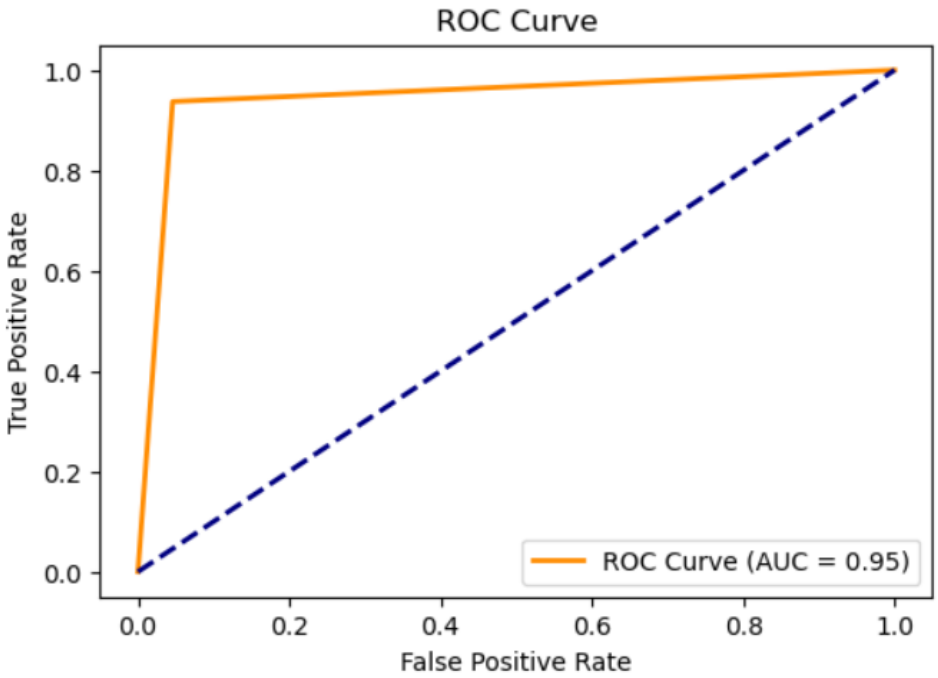

Fig. 4. ROC curve: The Model Performance.

The classification report provides detailed insights into the metrics of precision, recall, and F1-score for each class. It can be seen that the model is effective in distinguishing between items that are more challenging and those that are more distinctive. The overall accuracy of 94.44% suggests that the model is reliable across the entire dataset.

In addition, the AUC-ROC and AUC-PRC are useful for assessing the model's ability to generalise. Achieving scores around 95%, these metrics suggest that the model performs consistently well across different datasets, highlighting its robustness and adaptability

5 Conclusion

In conclusion, our research highlights the significant impact of integrating DL techniques with IRT in competitive test administration. This synergistic approach could greatly enhance the quality of test outcomes

By combining DL techniques with IRT principles, we have introduced an innovative methodology to solve the problem of score regeneration after the removal of problematic items. This innovative approach not only improves the accuracy of score adjustments, but also helps to identify and correct potential problems within test items. As a result, educators and test designers can draw on data to refine and optimize their assessments.

This combined approach offers several key advantages. Firstly, it provides a deeper understanding of candidate performance, enabling more accurate and actionable

feedback. Secondly, it supports the development of more reliable and valid tests by ensuring that the remaining items accurately reflect the intended measurement criteria. Finally, it improves decision-making processes linked to pedagogical strategies, enabling educators to adapt their teaching methods on the basis of comprehensive performance data.

Overall, the integration of DL and IRT represents a substantial improvement in the area of educational assessment. It provides stakeholders with robust tools to improve test quality, make informed instructional decisions and, ultimately, foster better learning outcomes. This research lays the foundation for more effective and equitable assessment practices, demonstrating the value of advanced analytical techniques in the educational field.

References

1. Hambleton, R.K., Swaminathan, H., Rogers, H.J.: Fundamentals of item response theory. Sage Publications, Newbury Park, Calif (1991).
2. Baker, R., Siemens, G.: Educational Data Mining and Learning Analytics. In: Sawyer, R.K. (ed.) The Cambridge Handbook of the Learning Sciences. pp. 253–272. Cambridge University Press, Cambridge (2014). https://doi.org/10.1017/CBO9781139519526.016.
3. Liu, R., Huggins-Manley, A.C., Bulut, O.: Retrofitting Diagnostic Classification Models to Responses From IRT-Based Assessment Forms. Educational and Psychological Measurement. 78, 357–383 (2018). https://doi.org/10.1177/0013164416685599.
4. Haladyna, T.M.: Developing and Validating Multiple-choice Test Items. Routledge (2004).
5. Zanon, C., Hutz, C.S., Yoo, H. (Henry), Hambleton, R.K.: An application of item response theory to psychological test development. Psicol. Reflex. Crit. 29, (2016). https://doi.org/10.1186/s41155-016-0040-x.
6. Reise, S.P., Waller, N.G.: Item response theory for dichotomous assessment data. In: Measuring and analyzing behavior in organizations: Advances in measurement and data analysis. pp. 88–122. Jossey-Bass/Wiley, Hoboken, NJ, US (2002).
7. de Ayala, R.J.: The theory and practice of item response theory. Guilford Press, New York, NY, US (2009).
8. Baker, F.B., Kim, S.-H.: Item Response Theory: Parameter Estimation Techniques, Second Edition. CRC Press (2004).
9. Shute, V.J., Becker, B.J. eds: Innovative Assessment for the 21st Century: Supporting Educational Needs. Springer US, Boston, MA (2010). https://doi.org/10.1007/978-1-4419-6530-1.
10. TRI_DICHO.pdf, https://www.irdp.ch/data/secure/1952/document/TRI_DICHO.pdf.
11. Lord, F.M., Novick, M.R.: Statistical Theories of Mental Test Scores. IAP (2008).
12. Embretson, S.E., Reise, S.P.: Item Response Theory. Psychology Press (2013).

13. Hrich, N., Azekri, M., Khaldi, M.: ARTIFICIAL INTELLIGENCE FOR EDUCATIONAL ASSESSMENT. Presented at the 16th annual International Conference of Education, Research and Innovation , Seville, Spain November (2023). https://doi.org/10.21125/iceri.2023.0598.
14. Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J.M., Milligan, S., Selwyn, N., Gašević, D.: Assessment in the age of artificial intelligence. Computers and Education: Artificial Intelligence. 3, 100075 (2022). https://doi.org/10.1016/j.caeai.2022.100075.
15. Yeung, C.-K.: Deep-IRT: Make Deep Learning Based Knowledge Tracing Explainable Using Item Response Theory, http://arxiv.org/abs/1904.11738, (2019). https://doi.org/10.48550/arXiv.1904.11738.