Enhancing Educational Assessments: Score Regeneration through Post-Item Analysis with Artificial Intelligence

Najoua Hrich[1,2,*], Mohamed Azekri[3], Charafeddin Elhaddouchi[1], and Mohamed Khaldi[2]

[1]Regional Center for Education and Training Professions, Institutions for Higher Executive Training, Tangier, Morocco
[2]Computer Science and University Pedagogical Engineering Research Team, Normal Higher School, Abdelmalek Essaadi University, Morocco
[3]Regional Academy of Education and Training, Ministry of National Education Preschool and Sports, Morocco
Email: hrnajouaofficiel@gmail.com (N.H.); medazekri@gmail.com (M.A.); charaff4@yahoo.fr (C.E.); medkhaldi@yahoo.fr (M.K.)
*Corresponding author

Manuscript received March 3, 2024; revised May 30, 2024; accepted August 2, 2024; published October 15, 2024

Abstract—**Over the years, the significant increase in candidates taking part in examinations and competitive exams has prompted the relevant bodies and institutions to reconsider their assessment approaches. This evolution largely stems from the growing prevalence of Artificial Intelligence (AI) and continuous technological advances. The primary aim of this research is to evaluate the effectiveness of Multiple-Choice Tests (MCTs) enhanced by AI through comprehensive item analysis followed by score regeneration focusing on the most discriminant items, aiming to strengthen assessment accuracy. The predominant adoption of MCTs has emerged, offering a practical and efficient solution for rigorously assessing a wide range of candidates. The success of this method hinges on the quality of its items. Therefore, ensuring the validity of such exams relies heavily on statistical analysis to select relevant items that are balanced in terms of difficulty and discriminatory power. Given the challenges of this analysis during test development, a new approach using score regeneration through an AI tool is proposed. This approach is based on a posterior statistical analysis of candidate performance with adjusted scores by eliminating the least discriminating items. The research sample was intentionally selected, consisting of computer science trainee teachers spread over the last three academic years. To validate this approach, a comparative study was conducted using the t-student's test and the Spearman correlation coefficient on the grades obtained in algorithmic and programming training modules each year. The results demonstrate that incorporating this score regeneration phase considerably improves the credibility of MCTs-based assessments, providing a solid foundation for educational decision-making. The findings affirm the research objective by showing that AI-enhanced MCTs offer a reliable and valid method for large-scale candidate assessment.**

Keywords—**assessments, items analysis, Artificial Intelligence (AI), competitive exams, Multiple-Choice Tests (MCTs)**

I. Introduction

The constant increase in the number of candidates for various exams and competitions, and the technological evolution of assessment methods, have prompted many educational establishments and institutions to favor Multiple Choice Tests (MCTs) massively. This strategy is emerging as a practical and effective response to the logistical and organizational challenges of this massive influx of participants. MCTs provide a standardized assessment method that enables automated grading, thus reducing the administrative burden. Furthermore, this approach allows a quick and unbiased assessment of candidates' skills, therefore aiding in managing deadlines in time-restricted situations [1]. While the predominance of MCTs may be perceived as a practical solution, it also raises questions about their quality [2]. Thus, while representing a pragmatic response to the challenges inquired by the influx of candidates, it is essential to consider the nuances related to the fairness and validity of these assessments [3]; it should be noted that the impact of randomness can have a notable effect on outcomes. Candidates may inadvertently select the correct answer to the question for which they lack prior knowledge [4].

Constructing a test involves several crucial steps to ensure its validity, reliability, and relevance. In this process, the pre-test, also known as a pilot test, holds a pivotal position [5]. It entails selecting participants representative of the target population to take a preliminary test version under conditions like those intended for the final test. Participants' responses are then collected and analyzed to assess the quality of the items, including their difficulty, discriminative power, and internal consistency [6]. Based on the pre-test results, items may be revised or modified to enhance their validity and reliability. This step identifies potential issues and necessary adjustments before the final test implementation, ensuring its quality and relevance for the target population [7, 8].

Integrating a pre-test into the MCT development process, enabling in-depth statistical analysis of items, is therefore of crucial importance in guaranteeing the validity and reliability of assessments. This systematic approach evaluates each item from various perspectives [9]. It allows for assessing the difficulty of each item by analyzing the overall success rate. Equally essential, this analysis examines the discriminatory capacity of each item, identifying those that significantly distinguish competent from less competent candidates [6]. In addition, this approach ensures that each item contributes to the measurement of the targeted skills. In short, integrating a pre-test into the design of an MCT is a rigorous approach that aims to perfect the quality of the test, ensuring a balanced and accurate assessment of participants' knowledge and skills. This pre-test involves administering a provisional version of the test to a group of individuals and then revising the set of items according to the initial results obtained [10]. In contexts where test confidentiality is paramount, such as professional certification exams or sensitive organizational assessments, traditional methods of pre-test item analysis with experimental groups pose significant challenges and may not be feasible. This necessitates exploring alternative methodologies to maintain test integrity while enhancing

doi: 10.18178/ijiet.2024.14.10.2172

assessment effectiveness.

An effective approach involves utilizing AI-driven item analysis for regenerating scores. By applying AI, assessments can retrospectively analyze candidate performance data to determine each item's difficulty and discriminative power through post-statistical analysis [11]. This approach bypasses the need for pre-test experimental groups and overcomes previous logistical and time constraints. With AI, comprehensive item analysis and score adjustment can be accomplished swiftly, within seconds, revolutionizing assessment practices in confidential testing environments.

This innovative method ensures the validity and reliability of assessments while establishing a new benchmark for adaptive and efficient techniques in various educational and professional contexts.

To validate the approach, a study was conducted involving seventy-two trainee students of the computer science discipline who enrolled in a training program at the Regional Center for Education and Training Professions. Upon completing their training modules, the trainees underwent rigorous evaluation through two distinct assessment methodologies. Firstly, they faced the Final Module Examinations (FME), following specific assessment procedures adopted by the Ministry, enabling them to be selected and ranked based on their results. Additionally, they were subjected to MCTs.

The MCT was subjected to a rigorous analysis, and an AI model used the test results to estimate the difficulty and discrimination indexes, assess the quality of the items, regenerate the scores after eliminating inappropriate items, and reproduce the final list of results [12]. A comparative study was conducted between the scores of the MCTs before and after regeneration, as well as with the results of the FME. Student's t-test [13, 14] and Spearman's Coefficient Analysis [15, 16] were used, revealing a significant difference between the different sets of marks and comparing the performance of the examinees.

The hypothesis suggests that the regeneration of scores, conducted after eliminating non-discriminatory and difficult items for MCTs improves its validity by attenuating the potential bias effects introduced by these items. The peaceful evaluation of items likely to present distortions in assessment aims to adjust participants' scores to ensure a more accurate and equitable measure of their skills. The selective elimination of problematic items, followed by the regeneration of results, is envisaged as a proactive means of improving test quality and minimizing undesirable influences, thus contributing to the enhanced validity of MCTs.

This research aims to verify the impact of regenerating the scores of candidates who have taken an MCT after eliminating non-discriminatory items on their overall performance.

II. Theoretical Framework

A. Test Development Process

The Test development process involves several key stages, each crucial to ensuring the creation of a valid and effective assessment [17]. Fig. 1 provides an overview of these stages:

- Exam Specifications: This initial stage involves defining the purpose of the examination, identifying the target audience, and outlining the content areas to be covered. It includes a comprehensive consideration of learning outcomes and learning objectives to ensure alignment with educational goals. Moreover, this stage determines the types of items to be included.
- Item Edition: Skilled item writers develop questions or tasks based on the specifications outlined in the previous stage. Items undergo rigorous review and editing to ensure clarity, accuracy, and relevance.
- Pilot Tests/Experimentation: A preliminary version of the examination is administered to a small group of test-takers to evaluate item performance and gather data for further analysis and improvement.
- Item Revision: Based on feedback from pilot tests and expert reviews, items may be revised or edited to address any identified issues or concerns.
- Final Test: This is constructed by selecting items that meet predetermined criteria based on item analysis results. The examination is thoroughly reviewed for accuracy and consistency before administration.

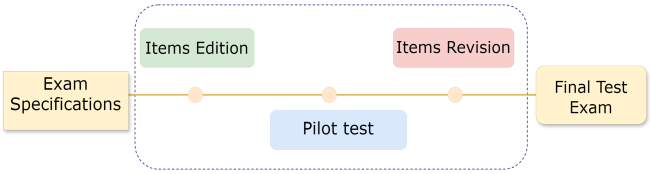

Fig. 1. Test development process.

Each stage of the test development process plays a critical role in ensuring that the assessment accurately measures the test-takers intended knowledge, skills, or abilities [18].

B. MCTs Based Assessment

MCTs have been widely adopted in educational assessment due to their numerous advantages also present challenges. On the benefits side, MCTs offer a practical and objective assessment of learners' knowledge and skills, enabling a wide range of topics to be tested relatively quickly and cost-effectively. Their standardized format facilitates comparing learners' performance and simplifies the grading process. Moreover, MCTs enable reliable and fair assessment, reducing the potential biases associated with other assessment formats. However, despite their advantages, MCTs also present challenges. They can sometimes encourage memorization of information rather than deep comprehension, and some critics argue that they are not as effective at assessing complex skills or critical thinking abilities. Furthermore, the development of high-quality MCTs requires significant expertise in item writing and psychometrics [19–21] to ensure their validity and reliability. Consequently, although MCTs are widely used in educational assessment, it is important to recognize their advantages and limitations and take action to alleviate the potential drawbacks of this assessment format [22].

C. AI in the Assessment Process

The development of assessment tests is a crucial process in educational assessment. With the advent of AI, new opportunities arise to improve and optimize this process [23]. Table 1 examines the integration of AI into each phase of assessment test development, highlighting the most promising approaches and applications.

Table 1. AI for development test process

Phase	AI Approaches	AI Algorithms	Description
Item Edition	Natural Language Processing (NLP)	Word Embeddings (Word2Vec, GloVe)	Word2Vec is a neural network model used to produce word embeddings by learning distributed representations of words based on their context in a continuous vector space [24].
			GloVe constructs word embeddings based on global word-word co-occurrence statistics and factorizes a co-occurrence matrix to generate embeddings capturing both local and global semantic relationships [25].
		Bidirectional Encoder Representations from Transformers (BERT)	BERT is used to automatically generate items from source texts by capturing the semantic and contextual relationships between words [26].
		Generative Pre-trained Transformer (GPT)	GPT is used to generate items using a pre-trained language model that generates coherent and relevant text [27].
Pilot test	Recommendation Systems	Neural Networks	Neural network-based recommendation systems analyze participants' performance and recommend the following items based on their skills and preferences [28].
	Adaptive Algorithms	Item Response Theory	IRT is used to model the probability of a participant answering an item correctly, based on his or her skills and the characteristics of the item. Adaptive algorithms dynamically adjust the test according to participants' responses [29].
Item Revision	Neural Networks	Convolutional Neural Networks	Convolutional neural networks detect potentially biased items by analyzing item characteristics and comparing them to predefined norms [30].
		Autoencoders	Autoencoders are used to detect similarities between items and identify groups of items with similar characteristics, facilitating item review and categorization [31].

III. Materials and Methods

An experimental approach was adopted to validate the hypotheses. The main objective of this approach is to assess the validity of the scoring method, illustrated in Fig. 2, for MCTs.

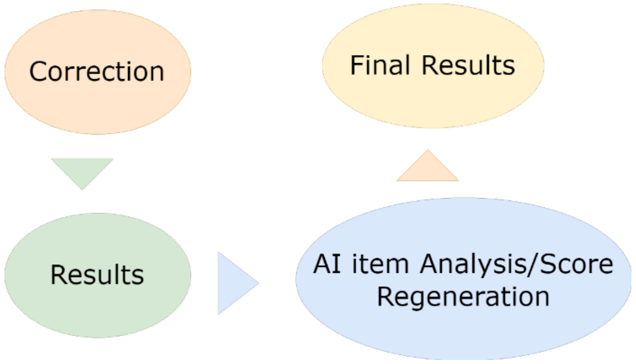

Fig. 2. Approach proposed for scoring MCTs.

With this method, the final score assigned to an examinee is regenerated after the selection of the appropriate items by an AI model based on statistical analysis of the candidates' responses. The indexes considered for this selection are [12]:

- Item difficulty (*d*-index): The proportion of students who answered the item correctly. Items with very high or very low difficulty may be problematic.
- Item discrimination (*p*-index): The degree to which an item differentiates between high-performing and low-performing students. High discrimination is desirable, indicating that the item measures the intended construct.

These indexes provide an in-depth understanding of an item's ability to discriminate among candidate performances and its inherent complexity.

A specific methodology is adopted to achieve the objective. Initially, two types of examination are administered: a FME and an MCT. Both exam forms are described in the next sections. The t-test Student is then used to compare the candidates' score series before and after score regeneration, aiming to detect any significant disparity between the two series. These data series are represented as box plots, facilitating visual observation of their distribution and dispersion. Finally, the Spearman correlation coefficient is calculated to assess the correlation between the series after treatment and the series obtained from the results of the FME. This measure evaluates the correlation between the ranks of candidates in both series, providing additional validation of the proposal. Fig. 3 describes the research methodology:

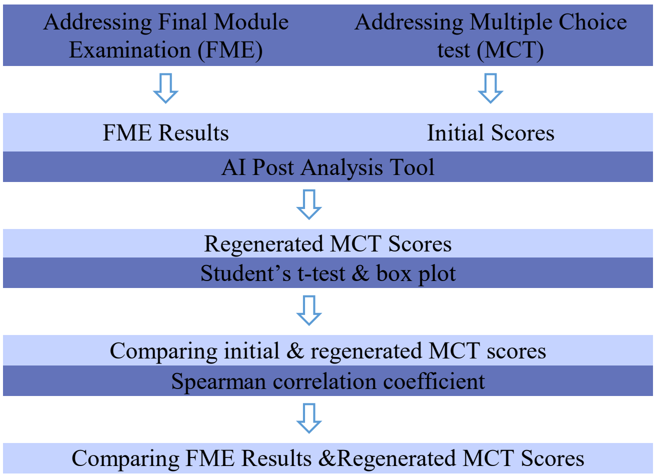
Addressing Final Module Examination (FME)
Addressing Multiple Choice test (MCT)
FME Results
Initial Scores
AI Post Analysis Tool
Regenerated MCT Scores
Student's t-test & box plot
Comparing initial & regenerated MCT scores
Spearman correlation coefficient
Comparing FME Results &Regenerated MCT Scores

Fig. 3. Research methodology.

A. Description of FME

These exams consist of two distinct parts: a written test and a practical assignment. The written section comprises a series of questions designed to assess mastery of fundamental algorithmic concepts, data structures, and programming principles. The practical section consists of a task in which students design, implement, and test a computer program to solve a specific problem.

B. Description of MCTs

The MCTs adopted consist of 100 items each. Each item offered 4 possible choices, and trainees had to select a single answer from these four options. Correct answers were awarded one point, while incorrect answers were not penalized.

These tests underwent a rigorous validation process. From initial conception to completion, every step was planned and executed. Two qualified specialists in the field reviewed each question, assessing its relevance, clarity, and the quality of distractors. Their feedback was carefully considered, and adjustments were made to ensure the validity and reliability of the tests.

C. AI Post-Item Analysis Tool for Educational Assessments

The model used in this study integrates statistical techniques with Deep Learning (DL) to enhance the selection of effective test items based on their difficulty and discrimination indexes. Previously developed and rigorously tested, this model combines traditional statistical methods for item analysis with the computational power of Deep Learning (DL) algorithms.

At its core, the model conducts comprehensive item analysis to assess each item's difficulty level and discriminatory power based on candidates' performance data. It employs advanced statistical techniques to compute difficulty indexes, which indicate the challenge level of items, and discrimination indexes, which gauge their effectiveness in distinguishing between high and low-performing candidates.

Moreover, the model utilizes DL algorithms to regenerate candidates' scores by systematically eliminating scores attributed to inappropriate items identified through the analysis. This post-analysis refinement ensures that assessment results accurately reflect candidates' true abilities, enhancing the validity and reliability of the assessment process.

In summary, the model represents an innovative approach that leverages both statistical methodologies and deep learning advancements to optimize item selection and score regeneration in educational and professional assessments [12].

Fig. 4 illustrates the AI Post item analysis tool for educational assessments which combines statistical methods to assess each item's difficulty and discriminative power. It utilizes artificial neural networks to select the most suitable items based on these assessments. This model, previously validated and effective in item selection for evaluations, has been integrated into our research based on its proven success, this model has demonstrated its effectiveness in item selection for evaluations. These past successes have motivated its integration into this research.

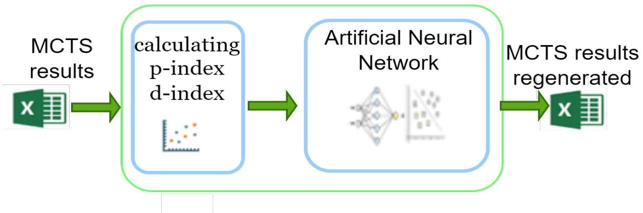

Fig. 4. AI item analysis & regenerating scores process.

D. Student T-Test and Spearman Correlation Coefficient

The student's t-test is a statistical method used to determine whether the mean of one data group is significantly different from that of another. It calculates the p-value. In this study, the paired student's t-test is suitable since the scores are small [32], and comparing paired observations in the two sets of scores is necessary.

Spearman's correlation coefficient, also known as Spearman's rank correlation coefficient, is a statistical measure used to assess the relationship between two variables by examining the correlation of their respective rankings rather than their raw values [33].

For this research:

- A p-value less than 0.05 indicates a significant difference between the scores.
- A Spearman correlation coefficient close to 1 indicates a perfect positive correlation.
- A Spearman correlation coefficient close to −1 indicates a perfect negative correlation.
- A Spearman correlation coefficient close to 0 indicates an absence of linear correlation between the variables.
- Python, the most widely used and renowned programming language in data science, generates these values [34].

E. Data Collection

For this study, the series of marks that were analyzed and regenerated are derived from tests and FME administered to computer science future teachers in the algorithmics and programming modules for the academic years 2021, 2022, and 2023.

These tests and examinations assessed trainees' understanding and skills regarding algorithmics and the C and Python programming languages.

In total, the results of 72 trainees were considered, including 23 trainees for 2021, 26 for 2022, and 23 for 2023.

IV. Results

A. Student's T-Test: Comparative Analysis of Pre- and Post-Score Regeneration Series

To evaluate the divergence between the series of scores assigned to trainees in the context of MCT before and after score regeneration, the paired Student's t-test was used.

Exploiting the Python programming language, the p-value parameter is calculated. Table 2 illustrates these parameters and the degree of disparity between the scores before and after their regeneration.

Table 2. *P*-value between pre- and post-score regeneration series

Academic Year	*P*-value
2021	0.037
2022	0.012
2023	0.016

Based on the results presented in Table 1, the calculated p-value between the score series after and before score regeneration is less than 0.05 for all three years. This indicates a significant disparity between the two sets of scores.

B. Box Plot: Visualization of Pre- and Post-Score Regeneration Series

The subsequent phase involves evaluating the most reliable and homogeneous series. For this purpose, the use of box plots has been opted. A comparison of the dispersion and the presence of outlier values between the two series was conducted by presenting the MCT score series before and after score regeneration. This visualization allows for a better understanding of each series' consistency and distribution of performance. Figs. 5–7 present successively the pre- and post-score regeneration series box plots for the academic years 2021, 2022, and 2023.

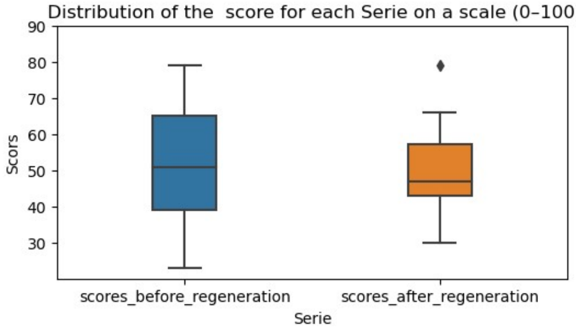

Fig. 5. Pre- and post-score regeneration series for the 2021 academic year.

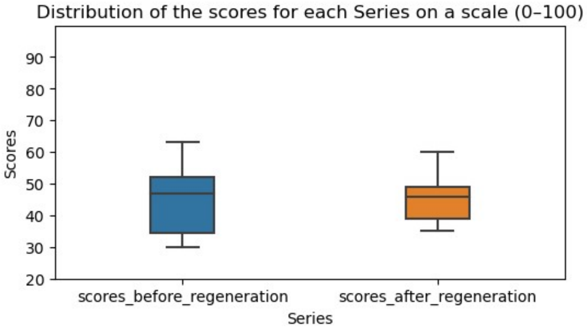

Fig. 6. Pre- and post-score regeneration series for the 2022 academic year.

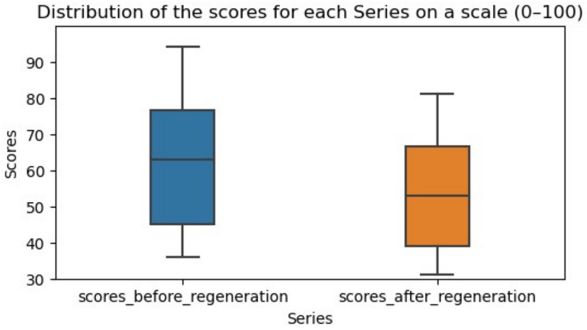

Fig. 7. Pre- and post-score regeneration series for the 2023 academic year.

A significant difference was noted between the series representing the scores before and after regeneration, as supported by the results of the student's *t*-test: the series of scores before regeneration has a higher median and longer whisker, while the series of scores after regeneration show a lower median and shorter whiskers. This comparison suggests that the series representing the scores after regeneration, given its lower dispersion, could offer greater stability in performance. This consistency is often requested in academic work.

C. Spearman's Coefficient Analysis: Correlation between FME Grades and Regenerated MCT Scores

Using the Python programming language, a program was implemented to calculate the Spearman's coefficient between the series of scores obtained by the trainees in the FME and the series resulting from the score regeneration conducted as part of the MCT analysis. This program produced the following results exposed in Table 3:

Table 3. Spearman's rank correlation coefficients

Academic Year	Spearman's rank correlation coefficient
2021	0.51
2022	0.76
2023	0.81

The values 0.51, 0.76, and 0.81 for Spearman's correlation coefficients indicate a moderate to strong correlation between the ranks of trainees in the two data series. This means that there is a tendency for the values in one series to increase or decrease consistently with the corresponding values in the other series. In other words, the trainee ranks in the two series are strongly correlated, but there may still be some dispersion in this correlation.

Based on these results, the similarity in the ranking of trainees across both sets of scores suggests enhanced credibility in evaluating the MCT through the post-item analysis approach and score regeneration.

V. Discussion

The objective of this study is to assess how the post-test analysis impacts the validity of MCTs. Our finding demonstrates that MCTs exhibit enhanced effectiveness when items are carefully selected and when the scores are regenerated, compared to initial scoring results. This underscores the significance of adopting systematic post-test analysis to improve the quality of MCTs-based assessments.

Moreover, our study reveals a notable correlation between the learners' ranking based on MCTs and their ranking on the FME. This correlation suggests that integrating score regeneration into MCT assessment could potentially optimize the candidate selection process for entrance examinations to graduate schools or recruitment competitions. The alignment between MCT outcomes and FME results further supports the efficacy of this approach in identifying candidates with the requisite competencies.

In conclusion, the integration of post-test score regeneration enhances the validity of MCTs-based assessments, offering a promising avenue for refining selection processes in competitive academic and professional settings.

VI. Limitations and Future Work

Although this study revealed significant results on the positive impact of item selection and score regeneration on the validity of MCTs, some limitations were identified. Firstly, concerning item selection, the study focused exclusively on discrimination and difficulty indexes to determine their relevance. Although these criteria are commonly used, other indices could also influence item quality. Another limitation of this study is the restricted sample size. Indeed, the limited number of participants may restrict the generalizability of the results obtained. Finally, it should be noted that this research was conducted within the framework of one discipline. Consequently, the results obtained may be specific to this field of study and may not apply to other fields. Future research should aim to address additional potential sources of bias, such as variations in cultural backgrounds or educational experiences, which could further enhance the generalizability of findings across diverse populations. Moreover, exploring alternative indices for item quality assessment beyond discrimination and difficulty could provide a more comprehensive understanding of effective item selection methodologies. These enhancements would contribute to advancing the robustness and applicability of AI-driven assessment frameworks in various educational and professional settings.

VII. Conclusion

In conclusion, the study highlights the importance of score regeneration in the scoring process of MCTs to ensure the validity and reliability of the results. The results clearly show that incorporating the regeneration phase significantly improves the credibility of MCT-based assessments, providing a sound basis for educational decision-making.

The practical implications of this research are wide-ranging, offering educators and educational decision-makers a valuable opportunity to improve their candidate selection practices. By adjusting their assessment methods to consider the results of this study, they will be able to ensure a more accurate and fair evaluation of candidates' skills.

In addition, the study contributes significantly to the assessment debate by highlighting the benefits of integrating score regeneration into the assessment process of MCTs. These results pave the way for further reflection and discussion on best practices in educational assessment, underlining the importance of continuing to explore and develop innovative approaches to improving the quality of educational assessments.

In summary, although this study has made significant contributions to understanding the impact of score regeneration and item selection on the validity of MCTs, it is important to recognize its limitations, particularly concerning the methodology used, the sample size, and the specific disciplinary area studied. These limitations provide avenues for future research to deepen our understanding of these complex issues.

Overall, the perspective of extending the experience of the proposal to other disciplines and forms of evaluation presents an exciting opportunity to enhance the quality and validity of these assessments. With continued research and development, this approach can potentially transform the field of education and provide more precise and meaningful information about students' learning and performance.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

NH and MA conceptualized, prepared the research background, and conducted experiments using the developed AI model, CE contributed to comparative studies and data analysis; MK verified methods and supervised the findings of this work. All authors approved the final version.

References

[1] T. M. Haladyna, S. M. Downing, and M. C. Rodriguez, "A review of multiple-choice item-writing guidelines for classroom assessment," *Applied Measurement in Education*, vol. 15, no. 3, pp. 309–333, 2002. doi: 10.1207/S15324818AME1503_5

[2] M. Tarrant, A. Knierim, S. Hayes, and J. Ware, "The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments," *Nurse Education in Practice*, vol. 6, pp. 354–363, 2007. doi: 10.1016/j.nepr.2006.07.002

[3] S. M. Downing, "Validity: On the meaningful interpretation of assessment data," *Med Educ*, vol. 37, no. 9, pp. 830–837, Sept. 2003. doi: 10.1046/j.1365-2923.2003.01594.x

[4] N. Hrich, M. Lazaar, and M. Khaldi, "Problematic of the assessment activity within adaptive e-learning systems," *Int. J. Emerg. Technol. Learn.*, vol. 14, no. 17, 133, Sept. 2019. doi: 10.3991/ijet.v14i17.10675

[5] N. Reynolds, A. Diamantopoulos, and B. Schlegelmilch, "Pre-testing in questionnaire design: A review of the literature and suggestions for further research," *Market Research Society Journal.*, vol. 35, no. 2, pp. 1–11, Mar. 1993. doi: 10.1177/147078539303500202

[6] G. Janssen, V. Meier, and J. Trace, "Classical test theory and item response theory: Two understandings of one high-stakes performance exam," *Colomb. Appl. Linguist. J*, vol. 16, no. 2, 167, Sept. 2014. doi: 10.14483/udistrital.jour.calj.2014.2.a03

[7] R. J. Wright, *Educational Assessment: Tests and Measurements in the Age of Accountability*, 1st ed., SAGE Publications, Inc, 2007.

[8] N. Hrich, M. Lazaar, and M. Khaldi, "Assessment process-reflection on pedagogical practices and integration of technologies in education," in *Promoting Positive Learning Experiences in Middle School Education*, C. B. Gaines andt K. M. Hutson, Eds., IGI Global, 2021, pp. 42–65. doi: 10.4018/978-1-7998-7057-9.ch003

[9] R. P. McDonald, *Test Theory: A Unified Treatment*, New York: Psychology Press, 1999. doi: 10.4324/9781410601087

[10] D. R. Eignor, "The standards for educational and psychological testing," *APA Handbook of Testing and Assessment in Psychology, Vol. 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology*, Washington, DC, US: American Psychological Association, 2013, pp. 245–250. doi: 10.1037/14047-013

[11] H. Jiao and R. W. Lissitz, *Application of Artificial Intelligence to Assessment,* IAP, 2020.

[12] N. Hrich, M. Azekri, and M. Khaldi, "Artificial intelligence item analysis tool for educational assessment: Case of large-scale competitive exams," *IJIET*, vol. 14, no. 6, pp. 822–827, 2024. doi: 10.18178/ijiet.2024.14.6.2107

[13] D. W. Zimmerman and B. D. Zumbo, "Rank transformations and the power of the student t-test and welch t' test for non-normal populations with unequal variances," *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie expÉrimentale*, vol. 47, no. 3, pp. 523–539, 1993. doi: 10.1037/h0078850

[14] G. D. Ruxton, "The unequal variance t-test is an underused alternative to student's t-test and the Mann–Whitney U test," *Behavioral Ecology*, vol. 17, issue 4, pp. 688–690, 2006. https://doi.org/10.1093/beheco/ark016

[15] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesthesia & Analgesia*, vol. 126, no. 5, 1763, May 2018. doi: 10.1213/ANE.0000000000002864

[16] J. de Winter, S. Gosling, and J. Potter, "Comparing the Pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data," *Psychological Methods*, vol. 21, pp. 273–290, Sept. 2016. doi: 10.1037/met0000079.supp

[17] T. M. Haladyna and M. C. Rodriguez, *Developing and Validating Test Items,* New York: Routledge, 2013. doi: 10.4324/9780203850381

[18] S. Lane, M. Raymond, and T. M. Haladyna, *Handbook of Test Development*, 2nd ed., New York: Routledge, 2015. doi: 10.4324/9780203102961

[19] T. M. Haladyna and S. M. Downing, "A taxonomy of multiple-choice item-writing rules," *Applied Measurement in Education*, vol. 2, no. 1, pp. 37–50, 1989. doi: 10.1207/s15324818ame0201_3

[20] S. M. Downing, "The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education," *Adv. Health Sci. Educ. Theory Pract.*, vol. 10, no. 2, pp. 133–143, 2005. doi: 10.1007/s10459-004-4019-5

[21] M. C. Rodriguez, "Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research," *Educational Measurement*, vol. 24, no. 2, pp. 3–13, 2005. doi: 10.1111/j.1745-3992.2005.00006.x

[22] M. J. Gierl, O. Bulut, Q. Guo, and X. Zhang, "Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review," *Review of Educational Research*, vol. 87, no. 6, pp. 1082–1116, 2017. doi: 10.3102/0034654317726529

[23] C. Peng, X. Zhou, and S. Liu, "An introduction to artificial intelligence and machine learning for online education," *Mobile Netw. Appl.*, vol. 27, no. 3, pp. 1147–1150, 2022. doi: 10.1007/s11036-022-01953-3

[24] S. Yilmaz and S. Toklu, "A deep learning analysis on question classification task using Word2vec representations," *Neural Comput. & Applic.*, vol. 32, no. 7, pp. 2909–2928, 2020. doi: 10.1007/s00521-020-04725-w

[25] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint, arXiv:1810.04805, 2019. doi: 10.48550/arXiv.1810.04805

[27] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song, "A survey of controllable text generation using transformer-based pre-trained language models," *ACM Comput. Surv.*, vol. 56, no. 3, pp. 64:1–64:37, Oct. 2023. doi: 10.1145/3617680

[28] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM Comput. Surv.*, vol. 52, no. 1, pp. 1–38, 2020. doi: 10.1145/3285029

[29] C.-K. Yeung, "Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory," arXiv preprint, arXiv:1904.11738, 2019. doi: 10.48550/arXiv.1904.11738

[30] S. Cao, X. Sun, L. Bo, Y. Wei, and B. Li, "*BGNN4VD*: Constructing bidirectional graph neural-network for vulnerability detection," *Information and Software Technology*, vol. 136, 106576, 2021. doi: 10.1016/j.infsof.2021.106576

[31] D. Charte, F. Charte, S. García, M. J. del Jesus, and F. Herrera, "A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines," *Information Fusion*, vol. 44, pp. 78–96, Nov. 2018. doi: 10.1016/j.inffus.2017.12.007

[32] Elementary statistical tests. [Online]. Available: https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-inf-tests.pdf (in France)

[33] L. Myers and M. J. Sirois, "Spearman correlation coefficients, differences between," *Encyclopedia of Statistical Sciences*, 2006. https://doi.org/10.1002/0471667196.ess5050.pub2

[34] SciPy API-Statistical functions (scipy.stats)-Spearmanr—SciPy v1.12.0 Manual. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

Copyright © 2024 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).