Proceedings of the 5th International Conference on Education Innovation and Philosophical Inquiries
DOI: 10.54254/2753-7048/58/20241716

Evaluating Human and Machine Assessment: Introducing a Hybrid Approach for Enhanced Educational Evaluation

Nan Zhou[1,a,*]

[1]*Institute of Education, University College London, London, WC1E 6BT, The United Kingdom*
a. zn2725757066@163.com
**corresponding author*

Abstract: This essay examines the evolving landscape of educational assessment, focusing on the analysis and comparison of human expertise and machine capabilities. Traditional educational assessment has predominantly relied on human evaluators who utilize their judgment and experience to assess student performance. However, advancements in technology have introduced machine assessments that leverage artificial intelligence to analyze data and provide feedback efficiently. This paper explores various assessment methods, including formative, summative, and peer assessments, and the role technology plays in enhancing these processes. A critical analysis of the strengths and weaknesses of both human and machine assessments is presented, highlighting scenarios where machines excel in efficiency and objectivity, particularly in handling large datasets and standardizing evaluations. Conversely, the essay emphasizes the irreplaceable depth of human insight necessary for assessing complex cognitive skills like creativity and critical thinking. The research advocates for a hybrid approach that combines the rapid analytical capabilities of machines with the nuanced judgment of human assessors. This integrative strategy aims to enhance reliability, personalize learning feedback, and optimize strategic educational planning. The proposed hybrid model not only addresses the shortcomings of both systems but also underscores the potential for improved assessment practices in various professional fields beyond education. This study calls for further exploration into the implementation of such hybrid assessment methods across diverse educational settings to maximize the benefits of both human and machine contributions.

Keywords: Educational Assessment, Human Assessment, Machine Assessment, Hybrid Approach Assessment

1. Introduction

In both educational contexts and everyday life, human development is inextricably linked with learning, accompanied by corresponding evaluations and challenges. Scholars have long debated the diverse learning methodologies and patterns exhibited by different age groups, emphasizing not only the importance of the learning process itself but also the critical role of assessing its outcomes. Reflecting this importance, Staake [1] highlights that assessment entails the systematic collection of data to gauge progress and efficacy, underscoring its essential role in educational settings where a variety of methods, such as formative, summative, and peer assessments, are crucial for evaluating student learning.

DOI: 10.54254/2753-7048/58/20241716

The integration of technology has significantly expanded assessment methodologies, with machines employing artificial intelligence to swiftly evaluate tasks and provide feedback, as demonstrated by platforms like ChatGPT that can review an essay in less than a minute [2]. Unlike human assessment, which relies on the judgment and experience of educators or examiners to measure students' knowledge, skills, and abilities, machine assessment predominantly utilizes AI and machine learning algorithms for this purpose [3]. This technological advancement not only enhances the speed of feedback delivery but also ensures consistent and objective evaluation across diverse educational contexts.

Previous research, including findings by [4], consistently suggests that machines are not capable of fully replacing humans in the assessment and evaluation of students' work. However, detailed investigations into the specific aspects where humans excel over machines, and vice versa, are often missing. This article aims to explore various assessment types and critically analyze the ongoing debate between machine versus human evaluation across different task contexts. Additionally, it seeks to investigate whether a hybrid approach that combines both human insight and machine efficiency could be effectively implemented.

2. Assessment Methods and the Introduction of the Debate

In educational settings, prevalent assessment methods such as formative, summative, and peer evaluations play pivotal roles. Formative assessments, integral during the instructional process, are designed not for grading but to facilitate ongoing adjustments in teaching and learning activities [1]. They are strategically incorporated into lessons to provide immediate feedback, allowing teachers to promptly gauge both individual and collective student progress, thus optimizing learning during a specific educational period. Examples of formative assessments include classroom quizzes and weekly presentations, which furnish timely feedback on students' strengths and weaknesses. In contrast, summative assessments are essential at the end of an instructional unit or academic term, providing a comprehensive evaluation of a student's knowledge and skills against established standards or benchmarks. Typical examples are final or annual examinations that measure educational progress and outcomes over an extended period and offer summary feedback to delineate students' long-term learning trajectories [5]. Other assessment types, such as peer evaluations, involve students in assessing each other’s learning effectiveness, adding a communal dimension to the assessment process. Practical assessments test the applicability of learned skills through activities like conducting research and presenting findings, further diversifying the methods and dimensions of evaluation tailored to specific educational needs.

The critical role of these assessment methods in gauging learning progress and effectiveness raises important considerations about the future of educational assessment. As digital technologies advance, machines are increasingly able to perform assessments with greater speed and efficiency, prompting a reevaluation of the balance between human-driven and technology-assisted assessment strategies. This technological shift, highlighted by González-Calatayud et al. [6], propels the debate on whether machines could, or should, replace human assessors in educational contexts, setting the stage for a detailed analysis of the interplay between machine and human assessment capabilities.

3. Human Assessment and Machine Assessment

3.1. Human Assessment

3.1.1. Advantages of Human Assessment

It is important to recognize that many assessment criteria are set by humans, which allows for a more nuanced application of these criteria when tasks are evaluated by humans rather than machines. This

DOI: 10.54254/2753-7048/58/20241716

flexibility is particularly critical in contexts where understanding the underlying principles of the criteria is essential. Human assessors can interpret and apply criteria in ways that consider the full range of performance dimensions, leading to more accurate and fair evaluations. For instance, when assessing students' language abilities, as Davis and Papageorgiou [7] note, well-trained raters, equipped with suitable scoring tools, are believed to be capable of appraising relevant aspects of performance. This ensures that the scores awarded are consistent and truly reflective of the language ability construct measured by the assessment. Human raters can evaluate not only the grammatical correctness and vocabulary use but also the subtleties of language such as tone, style, and context appropriateness. These aspects are critical in language assessments as they reflect the practical use of language in real-life situations.

Moreover, human evaluators bring a depth of understanding and contextual knowledge that machines typically lack. They can consider the socio-cultural context of the test-taker, which is particularly important in language assessments [8]. For example, idiomatic expressions, cultural references, and regional variations in language use are areas where human assessors can apply their nuanced understanding to provide a more accurate assessment of a student's language ability. This cultural sensitivity ensures that assessments are fair and that the results genuinely reflect the individual's proficiency. Human raters also have the ability to adjust their evaluations based on real-time interactions and observations. During oral language assessments, for example, human assessors can consider non-verbal cues such as body language, eye contact, and facial expressions, which are significant components of communication but often overlooked by automated systems. This holistic approach ensures that all relevant aspects of language use are considered, leading to a more comprehensive evaluation.

3.1.2. Challenges of Human Assessment

One significant drawback of human assessment is its time-consuming and labor-intensive nature. Human assessing does not scale well, particularly in larger educational settings or where detailed feedback is necessary. This limitation is pronounced in scenarios that demand quick turnarounds and personalized attention to a large number of students. It requires substantial time and effort, which can be a constraint in resource-limited environments [9]. For instance, if a teacher needs to grade 40-50 essays and aims to provide personalized, detailed feedback, students may have to wait approximately 20 days to receive their feedback. This extended timeline is not only frustrating for students who need timely feedback to improve but also burdensome for teachers who have to manage large workloads. This situation is cumbersome for both teachers and students, appearing as an inefficient method for both parties. The delay in receiving feedback can hinder students' learning processes, as they may not be able to address their mistakes or misconceptions promptly. Human assessment requires significant resources, including time, energy, and often financial investment in the training and development of assessors. In resource-limited environments, these requirements can be particularly challenging to meet. Schools and educational institutions may struggle to allocate sufficient resources to maintain a high standard of human assessment, leading to potential compromises in the quality of education and feedback provided to students.

Moreover, teachers generally exhibit limited capacity to appropriately collect, analyze, and plan in response to both formal and informal assessment data [10]. The process of manually handling large volumes of data is not only time-consuming but also susceptible to human error. When teachers need to analyze extensive statistical data, such as means, outliers, and variances of scores, manual calculations are more prone to errors, whereas machine computations can be more precise and convenient. This limitation affects the accuracy and reliability of the assessments, potentially leading to biased or incorrect conclusions about student performance. Manual data analysis and grading processes are inherently prone to errors. Human assessors might make mistakes due to fatigue,

DOI: 10.54254/2753-7048/58/20241716

oversight, or bias, which can affect the fairness and accuracy of the assessment. In contrast, automated systems can handle large datasets and perform complex calculations with high precision and consistency, reducing the likelihood of errors. Combining these two disadvantages, in terms of correction duration and data calculation efficiency, human assessment appears to be less powerful. The inefficiencies in human assessment can lead to broader implications for educational quality. Delays in feedback and potential errors in grading can undermine students' trust in the assessment process and affect their motivation and engagement. Additionally, the inability to provide timely and accurate feedback can hinder students' academic progress and learning outcomes.

3.2. Machine Assessment

3.2.1. Advantages of Machine Assessment

When it comes to privacy protection, avoiding examiner bias, and handling data, machine assessing demonstrates significant advantages. This technology represents a notable shift from traditional security solutions such as user authentication, access control, and personal information protection systems. According to Ahsan et al. [11], machine evaluations minimize subjective biases and enhance the protection of personal data, adhering more closely to ethical standards in data handling. Machine assessments are programmed to evaluate data based on predefined criteria, which helps to reduce these subjective biases. By ensuring that each student's work is assessed against the same objective standards, machines can contribute to a fairer evaluation process. This is particularly important in high-stakes testing environments where unbiased results are critical. Additionally, in terms of data security, machine assessment systems are designed to protect personal information through advanced encryption and secure data storage protocols. These systems can limit access to sensitive data, ensuring that only authorized personnel can view or manipulate student records. This enhanced level of security is crucial in maintaining the confidentiality and integrity of student information, thus fostering a trustworthy environment for data handling.

Moreover, in the academic sector, AI applications process large datasets to represent diverse student characteristics. These methods are utilized not only to extract patterns and predict behaviors but also to identify trends, thereby enabling educators to apply the most effective teaching strategies and monitor student progress comprehensively [12]. AI-driven assessment tools can handle vast amounts of data far more efficiently than humans. This capability is particularly beneficial for large-scale standardized testing, where quick turnaround times are essential. By automating the grading process, machines can provide immediate feedback to students, allowing them to understand their performance and areas for improvement without delay. This prompt feedback loop is instrumental in supporting continuous learning and development. Machine learning algorithms excel at recognizing patterns within complex datasets. For example, AI can identify correlations between student performance and various demographic or behavioral factors. By analyzing these patterns, educators can gain insights into factors that influence learning outcomes and adjust their teaching strategies accordingly. Predictive analytics can also be used to identify students at risk of falling behind, enabling early interventions to support their academic success.

3.2.2. Limitations of Machine Assessment

In discussing the complexities of assessment and the crucial role of human interaction, the limitations of machine assessing become evident. Although AI-powered systems have made significant strides in evaluating subjective assessments such as essays and open-ended questions, they struggle with fully grasping the depth of creativity, critical thinking, and nuanced arguments that are often essential in academic assessments. Machines typically rely on algorithms that are excellent at recognizing patterns and processing structured data, but they fall short when it comes to interpreting the subtle

DOI: 10.54254/2753-7048/58/20241716

and multifaceted nature of human thought and creativity. As González-Calatayud et al. [6] point out, while AI can typically assess basic writing skills such as grammar, syntax, and spelling, it lacks the sophistication to evaluate whether tasks possess the critical ability, creativity, and organizational skills that align with assessment requirements and criteria. This limitation underscores the necessity for human judgment and expertise in accurately evaluating these complex aspects of student work. Humans can recognize and appreciate the nuance in arguments, the originality in thought, and the coherence in complex projects, making them indispensable in higher-level assessments.

Furthermore, the lack of human interaction in machine evaluations poses another significant challenge. Research by Hooda et al. [4] indicates that the majority of students place greater trust in teachers who can interact with them directly, compared to machines. This preference highlights an inherent trust deficit in machine evaluations, attributed to their inability to engage in meaningful human interactions. The value of personal interaction in the learning process cannot be overstated; it helps build rapport, provides emotional support, and allows for immediate clarification and personalized feedback. Consequently, the credibility and reliability of machine evaluators remain questioned, necessitating enhancements in their design to incorporate elements of human interaction and intuition in the assessment process. To bridge this gap, integrating AI tools with human oversight might be a viable solution, where teachers use AI to handle routine tasks but remain actively involved in interpreting and responding to more complex and subjective aspects of student performance.

3.3. Short Reflection of All Advantages and Challenges

Based on the detailed analysis presented, it becomes evident that human assessing offers a more comprehensive and interactive evaluation of student abilities. Most assessment criteria and requirements are defined by humans, highlighting a significant advantage of human assessors in understanding and interpreting the full spectrum of student capabilities. However, the drawbacks of human assessing include time consumption and the susceptibility to errors in data analysis. On the other hand, although machines and AI technology can address these shortcomings and offer enhanced data privacy protections, they lack the capability to fully assess the broad scope of student learning abilities, raising concerns about their reliability. Consequently, both human assessors and machine evaluations have their respective strengths and weaknesses. This duality raises an important question for further research in the field of educational assessment: Is there a hybrid method that could combine the strengths of both human and machine assessments to maximize benefits and minimize drawbacks? Exploring such a possibility could lead to more effective and efficient assessment strategies in education, warranting deeper investigation into the potential integration of these approaches.

4. Hybrid Approaches of Both Assessment Methods

Machine assessments are highly advantageous for evaluating standardized responses such as those found in multiple-choice questions and the listening and reading sections of language proficiency tests like IELTS and TOEFL. Due to their capacity for rapid processing and advanced data analysis, machines can significantly reduce both the time and human resources needed to administer and score these types of assessments. This efficiency allows for quick turnaround times in producing test results, which is crucial in high-stakes testing environments. For more complex assessment tasks such as essay or passage writing, a hybrid method known as "double marking" has proven effective [7]. The dual scoring system, which incorporates both human and machine assessments, is highly beneficial. It enhances the reliability of evaluations by adding redundancy and ensures the consistent and objective measurement of specific performance features across various testing scenarios. The integration of machine scoring offers an added layer of reliability through redundancy, enhancing the

DOI: 10.54254/2753-7048/58/20241716

overall validity of the assessment. It is particularly valuable for ensuring that performance features are consistently and objectively measured, thereby increasing the assessment's reliability [7].

Beyond double marking, the division of work between humans and machines depending on their respective strengths is a strategic approach. For instance, Enright and Quinlan [13] suggest that while humans excel at evaluating the creativity and organization of written content, machines are more efficient at assessing specific linguistic phenomena. Similarly, in assessments of speaking and communication, machines can effectively measure technical aspects like pronunciation and fluency, whereas human raters are better suited to judge communicative effectiveness and task accomplishment [14]. This collaborative approach between human intuition and machine precision creates a robust framework for educational assessments, leveraging the unique capabilities of each to enhance the fairness and accuracy of test outcomes.

The synergistic application of human and machine assessment transcends educational environments, providing substantial benefits in various sectors including healthcare and human resources. In the healthcare domain, machine learning algorithms are employed to parse extensive datasets, facilitating the diagnosis of diseases, evaluation of patient outcomes, and customization of treatment plans [15]. However, the interpretation of these machine-generated predictions and the final decision-making process remains heavily reliant on human expertise. This human involvement is essential to ensure that the provided care is precisely tailored to meet individual patient needs, thus enhancing the effectiveness of treatments. Similarly, in the realm of human resources, AI-driven systems are utilized for resume screening, candidate evaluation, and initial job interviews. These systems streamline the recruitment process and improve the efficiency of candidate selection. Despite these advancements, the final decisions regarding hiring and nuanced aspects of workplace management, such as conflict resolution and employee relations, require the discernment and interpersonal skills of human HR professionals. This human input is critical for maintaining the integrity and personal touch of the recruitment process. These instances underscore the potential of integrating human and machine assessment to enhance operational efficiency across diverse professional fields. This collaborative approach not only leverages the analytical capabilities of machines but also harnesses the interpretative and decision-making prowess of humans.

5. Conclusion

In summary, this essay delves into the integration of human expertise and machine capabilities in the assessment of educational outcomes, highlighting a transformative approach in pedagogical evaluation methods. It critically analyzes the dynamic interplay between human evaluators and advanced algorithmic processes, proposing a paradigm where the precision of machine assessment complements the irreplaceable depth of human insight. In educational contexts, the deployment of machines and related AI technology to process and analyze extensive data sets represents a significant innovation, facilitating rapid feedback mechanisms and providing objective measures of student performance. These technologies are adept at quantifying explicit knowledge and can efficiently manage routine tasks such as grading and basic skills assessment. However, this essay underscores the indispensable role of human judgment, particularly in evaluating complex cognitive skills such as critical thinking and creativity. These human faculties are crucial for assessing abstract and higher-order thinking skills that machines might overlook or misinterpret.

The research also advocates for two hybrid assessment strategies where the efficiency of machine assessment is balanced with the nuanced judgment of human educators. The strategies aim to leverage the strengths of both approaches to create a more effective, efficient, and holistic evaluation system. The potential benefits of such strategies include increased reliability in assessments, the ability to provide personalized learning feedback, and the enhancement of strategic educational planning. Further investigation into this integrative approach is suggested, with a focus on its implementation

DOI: 10.54254/2753-7048/58/20241716

across various educational settings. The research also highlights the importance of a hybrid approach that combines machine efficiency with human insight to optimize the assessment process across different professional fields beyond education. This integrated approach not only leverages the rapid analytical capabilities of machines but also benefits from the nuanced understanding and judgment of human assessors, enhancing both operational efficiency and the quality of outcomes.

References

[1] *Staake, J. (2023, March 2). Types of Assessments for Education (Plus How and When To Use Them). We Are Teachers. https://www.weareteachers.com/types-of-assessments/*

[2] *Ibrahim Adeshola, & Adeola Praise Adepoju. (2023). The opportunities and challenges of ChatGPT in education. Interactive Learning Environments, 1–14. https://doi.org/10.1080/10494820.2023.2253858*

[3] *Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. Research and Practice in Technology Enhanced Learning, 12(1). https://doi.org/10.1186/s41039-017-0062-8*

[4] *Hooda, M., Rana, C., Dahiya, O., Rizwan, A., & Hossain, M. S. (2022). Artificial Intelligence for Assessment and Feedback to Enhance Student Success in Higher Education. Mathematical Problems in Engineering, 2022(5215722), 1–19. https://doi.org/10.1155/2022/5215722*

[5] *Barron, J. (2023, September 20). The 7 Different Types Of Assessment In Education. Start Teaching. https://start-teaching.com/the-7-different-types-of-assessment-in-education/*

[6] *González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial Intelligence for Student Assessment: A Systematic Review. Applied Sciences, 11(12), 5467. https://doi.org/10.3390/app11125467*

[7] *Papageorgiou, S., Davis, L., Norris, J., Gomez, P., Manna, V., & Monfils, L. (2021). Design Framework for the TOEFL® EssentialsTM Test 2021. https://www.ets.org/Media/Research/pdf/RM-21-03.pdf*

[8] *Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2012). A comparison of two scoring methods for an automated speech scoring system. Language Testing, 29(3), 371–394. https://doi.org/10.1177/0265532211425673*

[9] *Galloway, R., Reynolds, B., & Williamson, J. (2020). Strengths-based teaching and learning approaches for children: Perceptions and practices. Journal of Pedagogical Research, 4(1), 1–15. https://doi.org/10.33902/jpr.2020058178*

[10] *Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. Teaching and Teacher Education, 58, 149–162. https://doi.org/10.1016/j.tate.2016.05.010*

[11] *Ahsan, M., Nygard, K. E., Gomes, R., Chowdhury, M. M., Rifat, N., & Connolly, J. F. (2022). Cybersecurity Threats and Their Mitigation Approaches Using Machine Learning—A Review. Journal of Cybersecurity and Privacy, 2(3), 527–555. https://doi.org/10.3390/jcp2030027*

[12] *Gray, C. C., & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. Computers & Education, 131, 22–32. https://doi.org/10.1016/j.compedu.2018.12.006*

[13] *Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. Language Testing, 27(3), 317–334. https://doi.org/10.1177/0265532210363144*

[14] *Isaacs, T. (2018). Shifting Sands in Second Language Pronunciation Teaching and Assessment Research and Practice. Language Assessment Quarterly, 15(3), 273–293. https://doi.org/10.1080/15434303.2018.1472264*

[15] *Moreira, R., Teles, A., Fialho, R., Baluz, R., Santos, T. C., Goulart-Filho, R., Rocha, L., Silva, F. J., Gupta, N., Bastos, V. H., & Teixeira, S. (2020). Mobile Applications for Assessing Human Posture: A Systematic Literature Review. Electronics, 9(8), 1196. https://doi.org/10.3390/electronics9081196*