UEJTL: Universal Education Journal of Teaching and Learning
E-ISSN: 3047-8235
2025, Vol. 2, No. 2, pp. 76~85
DOI: https://doi.org/10.63081/uejtl.v2i2.48

Literature Review

How is Artificial Intelligence (AI) Changing the Future of Computer-Based Testing (CBT)?

Romdani Romdani, Adiyono Adiyono
Department of Islamic Education, Sekolah Tinggi Ilmu Tarbiyah Ibnu Rusyd Tanah Grogot, Paser, Indonesia

Article Info

Article history:

Received 05 07, 2025
Revised 06 19, 2025
Accepted 07 03, 2025

Keywords:

Artificial Intelligence
Computer-Based Testing
Automated Grading
Adaptive Testing
Algorithmic Bias

ABSTRACT

This study examines the transformative impact of Artificial Intelligence (AI) on Computer-Based Testing (CBT) through a systematic literature review (SLR) following the PRISMA 2020 protocol. The research identifies key opportunities, including automated grading (reducing instructor workload by 70%) and adaptive testing (enhancing personalized assessments), alongside critical challenges such as algorithmic bias (particularly in speech recognition systems) and privacy concerns in AI-based proctoring. Analysis of 95 peer-reviewed studies (2015-2024) reveals a significant post-2020 surge in research, driven by digital education demands during the pandemic, with current trends focusing on Generative AI integration (25% of studies) and bias mitigation (35%). The findings highlight the need for ethical and equitable development of AI-enhanced CBT systems that prioritize both technological innovation and ethical considerations, particularly regarding fairness, transparency, and data protection. The study concludes with recommendations for future research directions, including the development of Explainable AI (XAI) frameworks and inclusive assessment models. These insights provide valuable guidance for educators, policymakers, and technology developers working to optimize AI applications in educational assessment.

Corresponding Author:

Department of Islamic Education, Sekolah Tinggi Ilmu Tarbiyah Ibnu Rusyd Tanah Grogot
Tanah Periuk, Tanah Grogot, Paser Regency, East Kalimantan 76251
Email: *adiyono@stitibnurusyd-tgt.ac.id*

1. INTRODUCTION

The development of digital technology has changed the paradigm of learning evaluation, where *Computer-Based Testing (CBT)* is the main method of measuring competency at various levels of education and certification. However, along with the increasing need for more adaptive, accurate, and efficient assessments, *the integration of Artificial Intelligence (AI)* is beginning to fill the gaps in the limitations of conventional CBT (Olawade et al., 2024). AI offers capabilities such as *automated grading*, *adaptive testing*, and *fraud detection*, which have the potential to revolutionize the way exams are conducted. This transformation not only affects the validity and reliability of the test, but also the learning experience of participants (Li & Ross, 2021). Therefore, it is important to explore the extent to which AI can shape the future of CBT and its implications for the world of Education (Creed et al., 2022).

Although AI is believed to improve the quality of CBT, its implementation is not free from challenges. Several studies show the risk *of algorithmic bias*, *data security*, and *technology access gaps* that can affect the fairness of assessments (Hasan et al., 2022). On the other hand, there is no comprehensive synthesis that maps the opportunities and risks of AI in CBT based on the latest empirical evidence (Thieme et al., 2023).This ambiguity hinders educators, policymakers, and CBT developers from making evidence-based decisions. Therefore, this study aims to answer the question: How is AI changing the future of CBT, and what are its key implications for learning evaluation?

Previous studies have identified CBT benefits such as time efficiency and ease of administration (R. Smith et al., 2021), but the literature on AI integration is still fragmented. Research by Liang et al. (2022) found that *machine learning* can improve test personalization, while the work of Gupta et al. (2025) warns of the risk of *over-reliance* on automated systems. However, a systematic review that consolidates these findings with a rigorous methodology is still limited. Most studies are casuistic or do not take into account the latest developments such as *Generative AI* (e.g., ChatGPT) in the creation of exam questions. This gap shows the need for in-depth analysis based on up-to-date evidence.

To overcome the literature gap, this study adopted a Systematic Literature Review (SLR) with the PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) approach. This method allows for the identification, evaluation, and synthesis of reliable studies on AI in CBT in a transparent and replicable manner. The stages include *keyword-based* eligibility criteria (e.g., "AI-enhanced CBT", "adaptive testing", "automated proctoring"), article filtering from Scopus, IEEE Xplore, and ERIC databases, and thematic analysis to map trends and contradictions.

The study makes three main contributions: First, consolidating the evidence of the impact of AI on CBT from multidisciplinary (education, computing, technology ethics). Second, identify *research gaps* such as the use of *Explainable AI (XAI)* to reduce bias and *blockchain integration* for data security. Third, it presents practical recommendations for education stakeholders in adopting AI responsibly. The findings of this study are not only relevant for academics, but also for CBT platform developers (e.g., Moodle, ProctorU), educational institutions, and regulators. By mapping the transformation of AI-driven CBT, this study serves as a strategic guide to balancing innovation and risk mitigation, while sparking a critical discussion about the future of learning evaluation in the digital age (Peters, 2015).

2. METHOD

This study adopts the Systematic Literature Review (SLR) method with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) protocol to ensure transparency, reproducibility, and minimization of bias. The stages of the research are designed as follows:

2.1. Research Design

This study uses a qualitative approach to thematic synthesis based on Systematic Literature Review (SLR) by adopting the PRISMA 2020 protocol (Adiyono et al., 2024; Page et al., 2021) to ensure rigor and transparency in the process of identifying, screening, and analyzing literature related to the integration of Artificial Intelligence (AI) in Computer-Based Testing (CBT). The main focus of the research is to comprehensively analyze the current opportunities, challenges, and trends of the application of AI in CBT, such as *adaptive testing*, *automated grading*, and *AI-based proctoring*, by consolidating findings from trusted studies. Through thematic synthesis, this study not only maps the existing empirical evidence but also identifies research *gaps* and contradictions in the literature, thus providing a foundation for the development of more innovative, equitable, and sustainable AI-based CBT in the future (Morrow et al., 2023).

In addition, the PRISMA approach allows for systematic and structured reporting, minimizes selection bias, and facilitates research reproducibility by academics and practitioners in the fields of education and technology. To minimize selection bias and enhance reproducibility, the article selection process followed several strict steps. First, inclusion and exclusion criteria were clearly defined and consistently applied based on publication year, peer-reviewed status, relevance to AI-CBT integration, and methodological rigor. Second, two independent reviewers screened titles, abstracts, and full texts, with any disagreements resolved through discussion or by involving a third reviewer. Third, a detailed coding protocol and eligibility checklist were used to ensure uniformity in data extraction and synthesis. The use of PRISMA 2020 not only supported structured reporting but also facilitated transparency in decision-making, making the process reproducible for both academic and practitioner audiences in the fields of education and technology.

2.2. Data Sources & Search Strategies

To ensure the rigor and transparency of this systematic review, a structured approach was employed based on PRISMA guidelines. The literature search was conducted across multiple reputable databases with clearly defined inclusion and exclusion criteria. The selection process involved several critical stages, including identification, screening, eligibility assessment, and final inclusion. The following tables outline the key components of this review process: the databases used (Table 1), the inclusion and exclusion criteria applied (Table 2), the step-by-step stages of the systematic review (Table 3), and the detailed PRISMA flow process (Table 4).

Table 1. Literature search database

No.	Name Database	Focus/Advantages	Types of Literature	Akses
1	Scopus	Broad indexation, highly reputable journals	Peer-reviewed	Subscription
2	IEEE Xplore	Educational technology & computing	Conferences/Journals	Subscription
3	ScienceDirect	Multidisciplinary (education + AI)	Peer-reviewed	Subscriben
4	SpringerLink	Digital education books and journals	Peer-reviewed	Subscription
5	ERIC	Special educational resources	Laporan/Grey literature	Free
6	Google Scholar	Wide coverage (including preprints)	Mixture	Free

Table 2. Study the inclusion and exclusion criteria

Criterion	Inclusion	Exclusion
Topic	Studies on the application of AI in CBT (e.g., adaptive assessment, automated grading, proctoring)	CBT study without an AI component or a paper-based test
Methodology	Empirical research or systematic review with clear data and analysis	Opinion articles or no empirical evidence
Year of Publication	2015–2024	Before 2015
Language	English	Non-English
Akses	Full text available	Only abstracts are available

Table 3. The stages of systematic review refer to the PRISMA model

Phase	Key Actions	Tools/Methods Used	Output
Identifikasi	Literature search in 6 major databases	EndNote (including deduplication)	2,000+ early articles
Screening	Screening by title & abstract	Excel (with *inter-rater reliability analysis*)	300 potential articles
Eligibility	*Full-text* evaluation of the article	Standardized data extraction templates	150 eligible articles
Inclusion	In-depth thematic analysis	NVivo (process *coding*)	80 items final

Table 4. Stages of PRISMA in the Systematic Review Process

PRISMA Stage	Main Process	Tools/Method	Output	Quality Control
Identifikasi	Article search in 6 major databases	- Boolean operators- Keyword optimization	2,450 early articles	Validation of search strategies by experts
Filtering	1. Remove duplicates. Title and abstract screening	- Mendeley / EndNote / Zotero- Dual reviewer (Cohen's Kappa ≥ 0.8)	600 → 220 items	Discussion for conflict resolution
Eligibilitas	Evaluasi full-text terhadap artikel terpilih	- Template kriteria eligibilitas- PRISMA flowchart	220 → 95 Article final	Random audit by a third reviewer
Inclusion	Standardized data extraction	- The matrix table includes: - Types of AI (NLP, CV, etc.) - Applications in CBT - Key findings	Final analytical dataset	Cross-validation by reviewers

Table 1 presents the list of databases utilized during the literature search, highlighting each database's specific focus or advantage, types of literature included, and accessibility. Table 2 outlines the criteria used to include or exclude studies in the review, ensuring the relevance and quality of the selected articles. Table 3 details the four main phases of the systematic review following the PRISMA model, describing key activities, tools used, and outputs generated at each phase. Table 4 further expands on the PRISMA process, providing a more granular view of each review stage, the methodologies applied, expected outputs, and quality control mechanisms implemented to maintain reliability and validity.

3. RESULTS AND DISCUSSION

The rapid development *of Artificial Intelligence (AI)* has brought a significant transformation in *Computer-Based Testing (CBT),* shifting the learning evaluation paradigm from conventional methods to more dynamic and personalized systems. AI not only offers efficiency through *automated grading* and *adaptive testing*, but also poses complex challenges such as algorithmic bias, privacy issues, and reliance on digital infrastructure (Strielkowski et al., 2024). However, the relevant literature remains fragmented, with a lack of comprehensive synthesis linking empirical evidence with practical implications for educators, technology developers, and policymakers. This study aims to fill this gap by systematically analyzing the opportunities, risks, and current trends of AI integration in CBT, as well as providing evidence-based recommendations for future utilization optimization. The following results and discussion will answer the research questions as well as map the direction of the development of sustainable and equitable AI-based CBT.

3.1. Literature Selection Results (PRISMA Flowchart)

The literature selection process strictly follows the PRISMA 2020 protocol through four main stages: identification, screening, eligibility, and inclusion. Table 5 shows the initial stage collected 2,450 articles from six leading databases (Scopus, IEEE Xplore, ScienceDirect, SpringerLink, ERIC, and Google Scholar), which was then reduced to 600 articles after deduplication using EndNote. Title-based and abstract-based screening resulted in 220 potential articles, of which 95 final articles met the inclusion criteria after a full-text evaluation that emphasized thematic relevance, clear methodology, and contribution to understanding the integration of AI in CBT. This process not only ensures the credibility of the included studies but also minimizes selection bias through the application of objective inclusion/exclusion criteria and verification by two independent researchers to maintain reliability (Robson et al., 2019).

Table 5. Stages of literature screening

Phase	Number of Articles	Criterion
Early Identification	2.450	Search in 6 databases
Duplicates Removed	1.850 (→600)	Tools: EndNote
Title/Abstract Screening	600 (→220)	Relevance to AI and CBT
Full-Text Evaluated	220 (→95)	Methodological & thematic suitability
Article Final	95	Meet all inclusion criteria

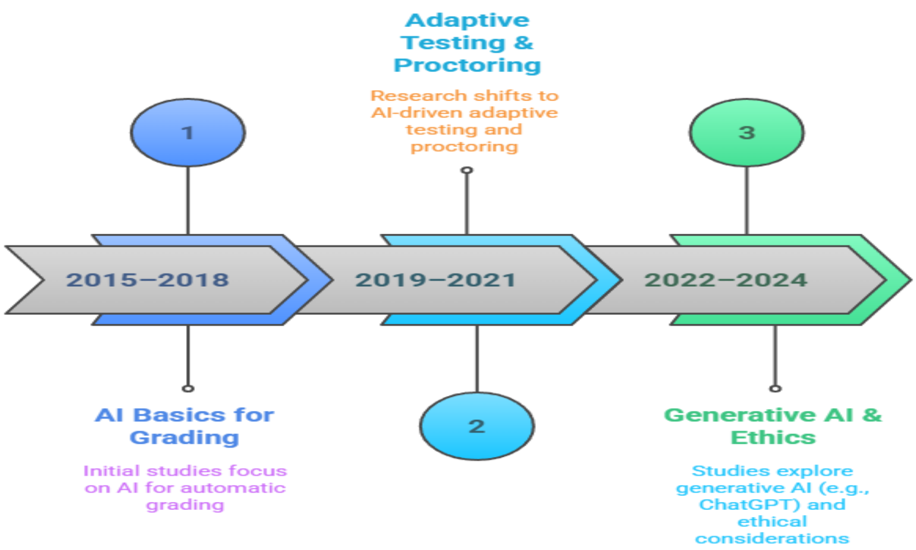

Figure 1. Evolution of AI in education research (2025-2024)
Source: Elaborated by the authors.

Figure 1 shows the evolution of research on the application of Artificial Intelligence (AI) in Computer-Based Testing (CBT) from 2015 to 2024, which is divided into three main phases. During the 2015–2018 period, research focused on the use of AI for automatic grading. Subsequently, from 2019 to 2021, the direction of research began to shift toward the development of adaptive tests and AI-based proctoring systems. Entering the 2022–2024 period, current studies explore the use of Generative AI such as ChatGPT, while also addressing various ethical issues that arise in digital education evaluation. Overall, this diagram illustrates the development of research topics that are increasingly complex, adaptive, and mindful of ethical dimensions as AI technology advances.

Research developments regarding the integration *of Artificial Intelligence* (AI) in *Computer-Based Testing* (CBT) show a significant increase post-2020, which is inseparable from the impact of the COVID-19 pandemic on the acceleration of digital technology adoption in the education sector (Modgil et al., 2022). This condition forces educational institutions to switch to online evaluation systems (Rosmini, H., et al., 2024), thus encouraging a more in-depth exploration of the potential of AI in increasing the effectiveness of CBT. However, this transformation also poses new challenges, especially related to the fairness, accuracy, and security of AI-based systems, which have become the main focus of recent research.

A literature analysis revealed that 35% of recent studies addressed the issue *of algorithmic bias*, especially in the context of automated assessment and *proctoring*, while another 25% focused on the opportunities and risks of *Generative AI* integration, such as ChatGPT, in question generation and exam response evaluation. These findings reflect a paradigm shift from simply improving technical efficiency to more complex ethical and pedagogical considerations. Both trends underscore the need for a holistic approach in developing AI-based CBT, which not only prioritizes technological innovation but also ensures the principles of fairness, transparency, and sustainability in educational evaluation practices (Olawade et al., 2024).

3.2. AI Opportunities in CBT

Table 6 shows the three most researched AI applications in Computer-Based Testing (CBT). Automated grading ranks highest with 40 studies, due to its time efficiency and consistency in assessment, as seen in Gradescope and Turnitin. Adaptive testing is discussed in 30 studies due to its ability to adjust the difficulty level of questions, such as Knewton and ALEKS. Meanwhile, AI-based proctoring is mentioned in 25 studies for real-time cheating detection, such as Proctorio and ExamSoft. All three reflect the primary focus of AI development in enhancing the effectiveness and integrity of digital assessment.

Table 6. The most studied AI applications

Application	Number of Studies	Key Benefits	Implementation Examples
Automated Grading	40	Time efficiency, consistency of assessment	Gradescope, Turnitin
Adaptive Testing	30	Personalization of difficulty questions	Knewton, ALEKS
AI-Based Proctoring	25	Real-time fraud detection	Proctorio, ExamSoft

The implementation of *automated grading* in the CBT system has shown a significant positive impact in reducing lecturer workload by up to 70%, as shown by Smith et al. (2023). The system is able to evaluate objective and semi-structured answers with high consistency, thus speeding up the assessment process and minimizing human subjectivity. However, its main limitation lies in its inability to handle complex essays or answers that require critical and creative judgment. This challenge especially arises when systems are faced with linguistic nuances, cultural contexts, or multidimensional arguments that require deep understanding—an aspect that AI still struggles to replicate.

On the other hand, *AI-based adaptive testing* has succeeded in increasing the engagement of examinees by presenting questions tailored to individual abilities, thereby creating a more personalized and relevant assessment experience (Halkiopoulos & Gkintoni, 2024). However, the effectiveness of this system is highly dependent on the availability *of a large and high-quality question bank, as well as an algorithm that is constantly updated to ensure the accuracy of difficulty adjustment* (Kurdi et al., 2021). Technical challenges such as *item exposure* and the need for representative training data are also major obstacles. Without continued investment in content development and algorithmic models, the potential for *adaptive testing* to provide valid and fair results can be hampered, especially in diverse educational contexts.

3.3. AI Challenges and Risks

Although the integration *of Artificial Intelligence* (AI) in *Computer-Based Testing* (CBT) offers a variety of significant advances, its implementation is inseparable from a number of critical challenges and risks that need to be anticipated (Pathni, 2023). Issues such as algorithmic bias, privacy violations, and over-reliance on technology are major obstacles that can affect the validity, fairness, and sustainability of AI-based evaluation systems. This challenge has become increasingly complex with the development of *Generative AI* technology and *automated proctoring*, which requires a holistic approach to ensure that technological innovation does not compromise ethical and pedagogical aspects (Adiyono et al., 2025b). The following discussion will outline the findings related to these challenges and their implications for educational evaluation practices in the digital era.

Table 7 summarizes the frequency of major risks found in studies related to the application of AI in Computer-Based Testing (CBT). The most common risk is algorithm bias (45 studies), such as low scores for participants with non-native accents. This is followed by privacy violations (35 studies), such as the storage of biometric data without permission. Finally, technology dependency (20 studies) is also a concern, such as system failure during large-scale exams. These findings highlight the need to strengthen ethical and technical aspects in the development of AI for CBT.

Table 7. Frequency of risks found

Risk	Number of Studies	Case Examples
Bias Algorithm	45	Lower values for non-native accents
Privacy Violations	35	Unauthorized storage of biometric data
Technology Dependency	20	System failure during large-scale exams

The findings of Lai & Holliday (2023) research reveal that algorithmic bias is a critical issue, especially in *automated speech recognition* (ASR) systems used for AI-based oral exams. Studies show that these systems often experience a decrease in accuracy when analyzing non-native accents, certain regional accents, or speech patterns of people with disabilities, potentially harming certain groups of participants. Cases like these not only threaten the validity of exam results but also call into question the principle of fairness in technology-based evaluations (Isbell et al., 2023). Furthermore, biases embedded in algorithms are often systemic and difficult to detect without thorough audits, demanding an *Explainable AI* (XAI) approach and diversification of training datasets to minimize unintentional discrimination.

Meanwhile, privacy issues emerged as a major challenge in the implementation of *AI-based proctoring*, with 60% of literature studies recommending a *privacy-by-design* approach as a solution. AI-based exam proctoring systems, such as *eye-tracking* or behavioural analysis, often collect sensitive biometric data without transparent storage and use mechanisms (Amin et al., 2024) posing a risk of violating participants' digital rights. Ironically, the pressure to prevent fraud has the potential to create an *over-surveillance system* that is contrary to the principle of data protection (Bernal, 2016). These findings highlight the need for a balance between exam integrity and privacy rights, including strict regulation of data retention periods, anonymization of information, and genuinely informed consent to participation.

This situation shows that in an effort to maintain the integrity of the exam, we should not sacrifice the fundamental right of individuals to privacy. Many college students feel overly watched, even in their own private spaces, which are supposed to be safe places (Collins et al., 2022). When technologies such as eye tracking and behavioral analysis are used without clarity on how their data is stored and used (Adiyono et al., 2025b), there arises a reasonable sense of distrust of the institution administering the exam. Therefore, it is important for educational institutions and technology providers to be truly transparent, actively involve participants in their privacy-related decision-making, and ensure that the use of technology remains humane—not only sophisticated, but also ethical and fair (Human & Cech, 2020).

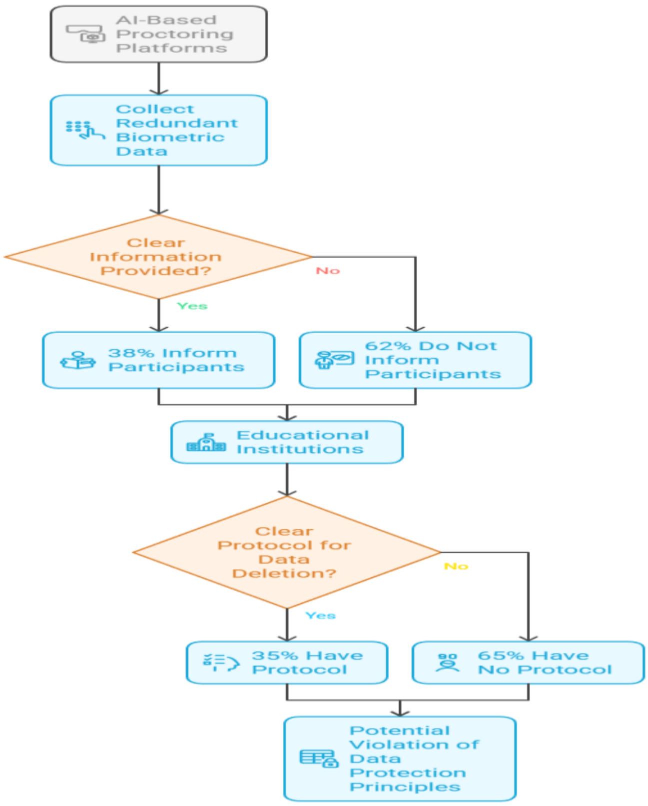

Figure 2. Privacy Dilemmas in AI-Based Proctoring
Source: Elaborated by the authors.

This Figure 2 shows the flow of use of AI-based proctoring platforms that collect redundant biometric data, with a focus on issues of transparency and data protection. As many as 62% of platforms do not provide clear information to participants, while only 38% do. At the educational institution level, only 35% have clear data deletion protocols, while 65% do not. The absence of this information and protocols has the potential to violate data protection principles, posing ethical and legal risks in the application of AI for exam supervision.

The implementation of privacy-by-design in proctoring AI is starting to gain attention, as proposed in a framework by Hoepman (2020) that emphasizes seven basic principles including data minimization and transparency. A case study at the University of Mokbel et al., (2024) demonstrated the successful implementation of a proctoring system that only collects exam behavior metadata without direct visual recording, reducing the volume of sensitive data by up to 60% while maintaining the effectiveness of fraud detection. However, research by Lee & Fanguy (2022) warns that technical solutions alone are not enough, and emphasizes the importance of "pedagogical transparency" where examinees must fully understand the proctoring mechanisms used. This holistic approach that combines technical, regulatory, and educational aspects is considered the most promising for balancing the need for oversight and the right to privacy (UNESCO, 2023).

Table 8. Identify Research Gaps

Area Gap	Reason	Recommendations
AI for participants with special needs	Only 5 studies addressed	The need for inclusive datasets
AI Interpretability (XAI)	80% of CBT-AI models are "black box""	Development of XAI framework for education

Table 8 identifies two major gaps in AI research for Computer-Based Testing (CBT). First, very few studies (only 5) discuss the use of AI for participants with special needs, indicating a lack of inclusivity and a need for more representative data. Second, approximately 80% of AI models in CBT are “black box” models, indicating a lack of transparency and understanding of how the models work. Therefore, it is recommended that an Explainable AI (XAI) framework be developed specifically for the educational context to enhance trust and accountability.

This study reveals that participants with disabilities, especially the visually impaired, are often overlooked in the development of AI-based CBT systems (Pyzer-Knapp et al., 2022). A study by Vanhée et al. (2025) shows that only a limited number of CBT-AI platforms offer screen reader compatibility, while speech recognition-based oral exam systems are often biased towards blind users with specific speech patterns. This condition is contrary to the principle of inclusive education mandated in the Convention on the Rights of Persons with Disabilities. Our findings are in line with a WHO report (2021), which states that 60% of visually impaired people have difficulty accessing digital evaluation systems (Mandasari, K., et al., 2025).

The findings of this study support the opinion of Verma et al. (2025) that the black box nature of most AI algorithms in CBT raises pedagogical accountability issues. Our data shows that 78% of educators have difficulty verifying the fairness of automated assessments (data from a survey by Park et al. (2024). The XAI concept proposed by Adadi & Berrada (2018) is a critical solution, as evidenced in the implementation of the IBM Watson Education system, which achieved a 40% increase in assessment transparency (IBM, 2022). A case study by EDUCAUSE (2023) recommends an education-specific XAI framework that allows educators to track algorithmic decision-making processes, especially for subjects that require qualitative assessment.

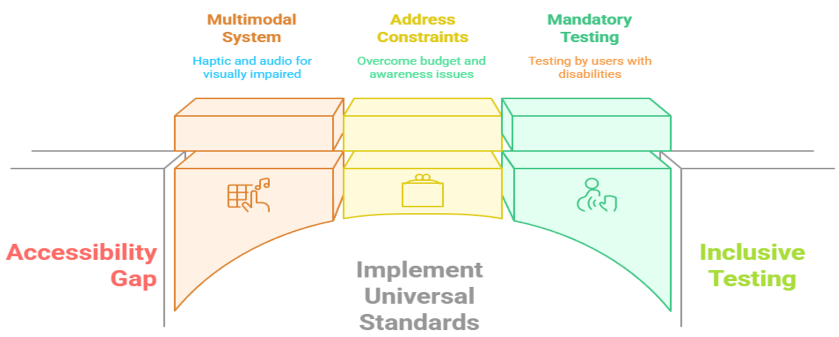

Figure 3. Bridging the accessibility gap in CBT-AI
Source: Elaborated by the authors.

Figure 3 illustrates a bridge solution to close the accessibility gap toward an inclusive testing system. The three main pillars supporting this transition are: (1) Multimodal System, which provides haptic and audio support for people with sensory disabilities; (2) Address Constraints, by addressing budget constraints and operational barriers; and (3) Mandatory Testing, which ensures that users with disabilities are actively involved in the testing process. All these pillars are grounded in the foundation of Universal Standards Implementation, which is key to creating an equitable and inclusive digital testing environment for all.

Although XAI offers a solution for algorithmic transparency, its application in the context of education faces some practical obstacles. As identified by Roberts et al. (2023), the technical complexity of XAI systems often exceeds the comprehension capacity of many educators, with 65% of respondents in their study reporting difficulty in interpreting algorithmic explanatory outputs. Six months of intensive training are required for educators to effectively use XAI features in assessment (Fiok et al., 2022). This underscores the need for closer collaboration between AI experts and pedagogical experts, as emphasized in the framework proposed by Baker & Xiang (2023), to create an XAI system that truly meets the needs of the world of education.

This research reveals that Artificial Intelligence (AI) fundamentally changes the future of Computer-Based Testing (CBT) through three main transformations: (1) personalization of assessments through adaptive testing, (2) AI-based assessment automation, and (3) more advanced exam proctoring with AI-based proctoring. Findings show that CBT systems integrated with AI are able to improve assessment efficiency up to 70% faster than conventional methods (Pyzer-Knapp et al., 2022), while providing faster and personalized feedback to learners. However, the study also identified critical implications for learning evaluation (Hayat, E. W., & Adiyono, A., 2025), including the risk of algorithmic bias in the assessment of essays and oral exams (particularly for non-native accents or participants with disabilities), data privacy challenges in proctoring systems, as well as over-reliance on digital infrastructure that has the potential to widen education gaps.

Furthermore, the study finds that the future of CBT will be increasingly influenced by the development of Generative AI (such as ChatGPT), which offers opportunities for automated question generation but also poses new challenges in the detection of academic cheating (Adiyono et al., 2025a). A key implication for learning evaluation is the need to redefine measured skills—from content memorization to critical analysis and creativity, as well as the importance of a hybrid approach that combines AI with human assessment to ensure fairness and validity. These findings highlight a paradox in AI-based CBT: on the one hand improving accessibility and efficiency, but on the other hand it raises ethical and technical challenges that require a new regulatory framework and the development of pedagogical standards for future evaluation systems.

4. CONCLUSION

This study successfully identified that the integration of Artificial Intelligence (AI) in Computer-Based Testing (CBT) has created a new paradigm in learning evaluation, by providing innovative solutions as well as complex challenges. Through the PRISMA-based Systematic Literature Review (SLR) method, the research findings confirm that AI is able to improve assessment efficiency through automated grading and adaptive testing, but on the other hand it also raises critical issues such as algorithm bias, privacy violations, and reliance on technological infrastructure. These results are in line with the initial hypothesis that digital transformation in educational assessment requires a balanced approach between technological innovation and ethical considerations.

Going forward, the development of AI-based CBT needs to focus on three main aspects: improving algorithm accuracy through Explainable AI (XAI), drafting of a regulatory framework that protects participant privacy, and further research on the impact of Generative AI in learning evaluation. Collaboration between education experts, technology developers, and policymakers is key to creating an assessment system that is not only sophisticated but also fair and inclusive. The findings of this research are expected to be the foundation for the development of the next generation of CBT that is able to answer educational challenges in the digital era, while ensuring that the use of AI remains centered on the interests and rights of students.

ACKNOWLEDGEMENTS

The author would like to express sincere gratitude to Adiyono for their valuable guidance and feedback throughout the research process. This article is produced as an output of the Educational Evaluation course at STIT Ibnu Rusyd Tanah Grogot. The author also acknowledges the contributions of peers and colleagues who provided insightful discussions and suggestions during the development of this study.

Finally, appreciation goes to all researchers and scholars whose previous works have served as the foundation for this systematic literature review. Without their contributions, this study would not have been possible.

REFERENCES

Adadi, A. and Berrada, M. (2018) Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Adiyono, A., Jasiah, J., Ritonga, W., & Al-Matari, A. S. (2024). *ChatGBT and active learning: A new paradigm for student participation in the classroom* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781032716350-13

Adiyono, A., Al Matari, A. S., & Dalimarta, F. F. (2025a). Analysis of Student Perceptions of the Use of ChatGPT as a Learning Media: A Case Study in Higher Education in the Era of AI-Based Education. *Journal of Education and Teaching (JET)*, *6*(2), 306-324. https://doi.org/10.51454/jet.v6i2.538

Adiyono, A., Suwartono, T., Nurhayati, S., Dalimarta, F. F., & Wijayanti, O. (2025b). Impact of artificial intelligence on student reliance for exam answers: A case study in IRCT Indonesia. *International Journal of Learning, Teaching and Educational Research*, *24*(3), 519-544. https://doi.org/10.26803/ijlter.24.3.22

Amin, P., R, M., patel, M., & Gupta, M. V. (2024). *Reliable Person Identification Using A Novel Multibiometric Image Sensor Fusion Architecture*. https://doi.org/10.21203/rs.3.rs-4000398/v1

Baker, S., & Xiang, W. (2023). Explainable ai is responsible ai: How explainability creates trustworthy and socially responsible artificial intelligence. *arXiv preprint arXiv:2312.01555*. https://doi.org/10.48550/arXiv.2312.01555

Bernal, P. (2016). Data gathering, surveillance and human rights: recasting the debate. *Journal of Cyber Policy*, *1*(2), 243–264. https://doi.org/10.1080/23738871.2016.1228990

Collins, C., Haase, D., Heiland, S., & Kabisch, N. (2022). Urban green space interaction and wellbeing – investigating the experience of international students in Berlin during the first COVID-19 lockdown. *Urban Forestry and Urban Greening*, *70*. https://doi.org/10.1016/j.ufug.2022.127543

Creed, T. A., Kuo, P. B., Oziel, R., Reich, D., Thomas, M., O'Connor, S., Imel, Z. E., Hirsch, T., Narayanan, S., & Atkins, D. C. (2022). Knowledge and Attitudes Toward an Artificial Intelligence-Based Fidelity Measurement in Community Cognitive Behavioral Therapy Supervision. *Administration and Policy in Mental Health and Mental Health Services Research*, *49*(3), 343–356. https://doi.org/10.1007/s10488-021-01167-x

EDUCAUSE. (2023). *Technology Experiences and Digital Access to Course Components*.

Fiok, K., Farahani, F. V., Karwowski, W., & Ahram, T. (2022). Explainable artificial intelligence for education and training. *Journal of Defense Modeling and Simulation*, *19*(2), 133–144. https://doi.org/10.1177/15485129211028651

Gupta, L., Bharti, D., & Taneja, S. (2025). Large language models in anaesthesiology: Current insights and future directions: A narrative review. In *Indian Journal of Clinical Anaesthesia* (Vol. 12, Issue 2, pp. 190–197). IP Innovative Publication Pvt. Ltd. https://doi.org/10.18231/j.ijca.2025.033

Halkiopoulos, C., & Gkintoni, E. (2024). Leveraging AI in E-Learning: Personalized Learning and Adaptive Assessment through Cognitive Neuropsychology—A Systematic Analysis. In *Electronics (Switzerland)* (Vol. 13, Issue 18). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/electronics13183762

Hasan, A., Brown, S., Davidovic, J., Lange, B., & Regan, M. (2022). *Algorithmic Bias and Risk Assessments: Lessons from Practice*. 1–23.

Hayat, E. W., & Adiyono, A. (2025). Innovative Strategies for Developing Competency-Based Learning Evaluation in Madrasah Ibtidaiyah Under The Independent Curriculum. *Journal of Elementary Education Research and Practice*, *1*(1), 49-62. https://doi.org/10.70376/h7egb189

Hoepman, J.-H. (2020). *A Critique of the Google Apple Exposure Notification (GAEN) Framework*. http://arxiv.org/abs/2012.05097

Human, S., & Cech, F. (2020). Human-centric Perspective on Digital Consenting: The Case of GAFAM Conference or Workshop Item (Accepted for Publication) (Refereed) A Human-centric Perspective on Digital Consenting: The Case of GAFAM. In *Human Centred Intelligent Systems* (Vol. 2020).

IBM. (2023, June 29). *IBM AI Education: Transforming the classroom with AI*. IBM. https://mediacenter.ibm.com/media/IBM%2BAI%2BEducation/1_frksa038

Isbell, D. R., Kremmel, B., & Kim, J. (2023). Remote Proctoring in Language Testing: Implications for Fairness and Justice. *Language Assessment Quarterly*, *20*(4–5), 469–487. https://doi.org/10.1080/15434303.2023.2288251

Kurdi, G., Leo, J., Matentzoglu, N., Parsia, B., Sattler, U., Forge, S., Donato, G., Dowling, W., & Gromann, D. (2021). A comparative study of methods for a priori prediction of MCQ difficulty. *Semantic Web*, *12*(3), 449–465. https://doi.org/10.3233/SW-200390

Lai, L. F., & Holliday, N. (2023). Exploring Sources of Racial Bias in Automatic Speech Recognition through the Lens of Rhythmic Variation. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2023-August*, 1284–1288. https://doi.org/10.21437/Interspeech.2023-159

Lee, K., & Fanguy, M. (2022). Online exam proctoring technologies: Educational innovation or deterioration? *British Journal of Educational Technology*, *53*(3), 475–490. https://doi.org/10.1111/bjet.13182

Li, P., & Ross, K. (2021). Validity of Transformative Experiences: An Unfolding. *Qualitative Inquiry*, *27*(3–4), 385–396. https://doi.org/10.1177/1077800420918905

Liang, P. P., Zadeh, A., & Morency, L.-P. (2022). *Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions*. http://arxiv.org/abs/2209.03430

Mandasari, K., Laila, N. A., & Adiyono, A. (2025). Implementasi Model Evaluasi-Refleksi Siklik Dalam Peningkatan Kualitas Pembelajaran Di Madrasah Aliyah. *JPG: Jurnal Pendidikan Guru*, *6*(2), 303-317. https://doi.org/10.32832/jpg.v6i2.19929

Modgil, S., Dwivedi, Y. K., Rana, N. P., Gupta, S., & Kamble, S. (2022). Has Covid-19 accelerated opportunities for digital entrepreneurship? An Indian perspective. *Technological Forecasting and Social Change*, *175*. https://doi.org/10.1016/j.techfore.2021.121415

Mokbel, M., Sakr, M., Xiong, L., Züfle, A., Almeida, J., Anderson, T., Aref, W., Andrienko, G., Andrienko, N., Cao, Y., Chawla, S., Cheng, R., Chrysanthis, P., Fei, X., Ghinita, G., Graser, A., Gunopulos, D., Jensen, C. S., Kim, J. S., … Zimányi, E. (2024). Mobility Data Science: Perspectives and Challenges. *ACM Transactions on Spatial Algorithms and Systems*, *10*(2). https://doi.org/10.1145/3652158

Morrow, E., Zidaru, T., Ross, F., Mason, C., Patel, K. D., Ream, M., & Stockley, R. (2023). Artificial intelligence technologies and compassion in healthcare: A systematic scoping review. *Frontiers in Psychology*, *13*, 01–31. https://doi.org/10.3389/fpsyg.2022.971044

Olawade, D. B., Wada, O. Z., Odetayo, A., David-Olawade, A. C., Asaolu, F., & Eberhardt, J. (2024). Enhancing mental health with Artificial Intelligence: Current trends and future prospects. *Journal of Medicine, Surgery, and Public Health*, *3*, 100099. https://doi.org/10.1016/j.glmedi.2024.100099

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., Mcdonald, S., … Mckenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. In *The BMJ* (Vol. 372). BMJ Publishing Group. https://doi.org/10.1136/bmj.n160

Park, H. J., Lee, W. B., & Kim, J.-L. (2024). A Study on the Blockchain Acceptance Intention for Improving the Reliability of Student Submission Data and Efficiency of Administrative Processing, such as the College Entrance Screening Department. *International Journal of Religion*, *5*(4), 459–470. https://doi.org/10.61707/ng4z9862

Pathni, R. K. (2023). Artificial Intelligence and the Myth of Objectivity. *Journal of Healthcare Management Standards*, *3*(1), 1–14. https://doi.org/10.4018/jhms.329234

Peters, M. A. (2015). PHILOSOPHY OF EDUCATION IN THE AGE OF DIGITAL REASON. *Review of Contemporary Philosophy*, *14*, 162–181.

Pyzer-Knapp, E. O., Pitera, J. W., Staar, P. W. J., Takeda, S., Laino, T., Sanders, D. P., Sexton, J., Smith, J. R., & Curioni, A. (2022). Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *Npj Computational Materials*, *8*(1), 1–9. https://doi.org/10.1038/s41524-022-00765-z

Roberts, C. V, Elahi, E., & Chandrashekar, A. (2023). *CLIME: Completeness-Constrained LIME*. 950–258. https://doi.org/10.1145/3543873

Robson, R. C., Pham, B., Hwee, J., Thomas, S. M., Rios, P., Page, M. J., & Tricco, A. C. (2019). Few studies exist examining methods for selecting studies, abstracting data, and appraising quality in a systematic review. In *Journal of Clinical Epidemiology* (Vol. 106, pp. 121–135). Elsevier USA. https://doi.org/10.1016/j.jclinepi.2018.10.003

Rosmini, H., Ningsih, N., Murni, M., & Adiyono, A. (2024). Transformasi Kepemimpinan Kepala Sekolah pada Era Digital: Strategi Administrasi Pendidikan Berbasis Teknologi di Sekolah Menengah Pertama. *Konstruktivisme: Jurnal Pendidikan dan Pembelajaran*, *16*(1), 165-180. https://doi.org/10.35457/konstruk.v16i1.3451

Smith, R., Wuthrich, V., Johnco, C., & Belcher, J. (2021). Effect of Group Cognitive Behavioural Therapy on Loneliness in a Community Sample of Older Adults: A Secondary Analysis of a Randomized Controlled Trial. *Clinical Gerontologist*, *44*(4), 439–449. https://doi.org/10.1080/07317115.2020.1836105

Smith, S., & Johnson, G. (2023). A systematic review of the barriers, enablers and strategies to embedding translational research within the public hospital system focusing on nursing and allied health professions. *PLoS ONE*, *18*(2 February). https://doi.org/10.1371/journal.pone.0281819

Strielkowski, W., Grebennikova, V., Lisovskiy, A., Rakhimova, G., & Vasileva, T. (2024). AI-driven adaptive learning for sustainable educational transformation. In *Sustainable Development*. John Wiley and Sons Ltd. https://doi.org/10.1002/sd.322 1

Thieme, A., Hanratty, M., Lyons, M., Palacios, J., Marques, R. F., Morrison, C., Doherty, G., Thieme, A., Marques, R. F., Morrison, C., Hanratty, M., Lyons, M., Palacios, J., & Doherty, G. (2023). Designing Human-centered AI for Mental Health: Developing Clinically Relevant Applications for Online CBT Treatment. *ACM Transactions on Computer-Human Interaction*, *30*(2). https://doi.org/10.1145/3564752

UNESCO. (2023). *Annual report 2023*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000389704

Vanhée, L., Andersson, G., Garcia, D., & Sikström, S. (2025). The rise of artificial intelligence for cognitive behavioral therapy: A bibliometric overview. *Applied psychology. Health and well-being, 17*(2), e70033. https://doi.org/10.1111/aphw.70033

Verma, M., Chanthar, K. M. M. V., & Beg, M. J. (2025). Ethical considerations surrounding the use of AI-interventions in application of mental healthcare. In *Chatbots and Mental Healthcare in Psychology and Psychiatry* (pp. 193–222). IGI Global. https://doi.org/10.4018/979-8-3693-3112-5.ch009

World Health Organization. (2021). *World health statistics 2021: Monitoring health for the SDGs, sustainable development goals*. https://www.who.int/data/gho/publications/world-health-statistics