A Systematic Review of Artificial Intelligence-Based Computer Adaptive Testing (CAT) and Item Response Theory for Enhancing the Effectiveness of Science Learning Assessment

Muhammad Gibran Alif Prasetya[1], Arif Widiyatmoko[2], Ani Rusilowati[3]
[1,2,3]Universitas Negeri Semarang, Semarang, Indonesia
Email: gibranalif45@students.unnes.ac.id

Abstract

The advancement of technology has accelerated the adoption of Computerized Adaptive Testing (CAT) in educational assessment due to its ability to dynamically adjust item difficulty levels, thereby producing more precise, efficient, and valid measurements compared to conventional tests. While Item Response Theory (IRT) serves as the primary psychometric foundation of CAT, traditional IRT implementation faces computational challenges because ability estimation requires lengthy iterative processes, resulting in reduced system responsiveness. To address this issue, Artificial Intelligence (AI), particularly Fuzzy Logic, offers a promising solution through rapid inference mechanisms and monotonic reasoning that can adaptively map students' cognitive abilities to corresponding item difficulty levels. This study aims to develop a hybrid CAT system that integrates Fuzzy Logic for fast inference with IRT as a robust and valid psychometric framework in the context of science learning. The research employs a systematic literature review using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework, encompassing the stages of Identification, Screening, and Inclusion of relevant studies. The findings indicate that the integration of AI/ML with IRT in CAT consistently enhances assessment accuracy and efficiency. Algorithms such as Maximum Information (MI) and Expected a Posteriori (EAP) effectively reduce test length without compromising reliability, while Fast Adaptive Cognitive Diagnosis (FACD) improves early-stage ability prediction. Furthermore, Fuzzy Logic demonstrates strong effectiveness in selecting adaptive test items aligned with students' ability levels. The study concludes that developing CAT systems based on AI and IRT yields adaptive, personalized, efficient, and diagnostic evaluation mechanisms that support personalized science learning.

Keywords: *Computer Adaptive Testing, Item Response Theory, Artificial Intelligence, Fuzzy Logic, Science Learning.*

A. INTRODUCTION

The rapid advancement of information technology has driven a fundamental transformation in educational evaluation through the adoption of sophisticated adaptive measurement systems. Computerized Adaptive Testing (CAT) is designed to dynamically adjust the difficulty level of each test item based on examinees' responses, allowing every student to receive a set of questions tailored to their individual ability (Cheng et al., 2021; Imawan, Retnawati, & Ismail, 2025; Tian & Dai, 2020). This adaptive adjustment enables more precise measurement, time efficiency, and resource optimization, while simultaneously enhancing the validity of evaluation outcomes compared to conventional fixed-length tests. CAT also supports real-time ability estimation, enabling continuous adaptation of the assessment pathway to

reflect the examinee's actual cognitive ability, minimize mismatches between item difficulty and student capacity, and promote a more personalized and focused learning experience (Ma et al., 2025; Tsaousis et al., 2021).

Item Response Theory (IRT) serves as the primary psychometric foundation in CAT development due to its ability to provide a robust analytical framework for understanding item characteristics. IRT models, including the three-parameter logistic (3PL) model, analyze item difficulty, discrimination, and guessing probability, allowing the system to optimally select items that match the examinee's ability profile (Frick et al., 2024; Iwintolu et al., 2024). This approach facilitates continuous ability estimation as the student progresses through the test, ensuring accurate and responsive item adaptation. Integrating IRT into CAT not only enhances the validity and reliability of assessment but also provides rich diagnostic information on student performance, item properties, and potential pedagogical interventions based on response patterns (C. Huda, 2024). This psychometric foundation is essential for developing CAT systems that are adaptive, personalized, and capable of supporting holistic measurement of student ability.

However, implementing IRT-based CAT in its traditional form presents computational challenges. Ability estimation using Maximum Likelihood Estimation or Bayesian methods requires iterative processes from the initial to the final item, resulting in longer response times for subsequent item selection and reduced system responsiveness. Artificial Intelligence (AI), particularly Fuzzy Logic, offers a solution through rapid inference mechanisms and monotonic reasoning, enabling the adaptive mapping of students' cognitive levels (low, medium, high) to item difficulty (easy, medium, hard) (Göktepe Körpeoğlu et al., 2025; Papadimitriou & Virvou, 2025). Integrating Fuzzy Logic into CAT accelerates item selection and reduces computational load compared to traditional methods. Prior studies have demonstrated the effectiveness of Fuzzy Logic-based CAT in adaptively adjusting item difficulty and optimizing the number of administered items (Maji & Ganguli, 2025; Sathya et al., 2024; Suzuki & Negishi, 2024; Wulansari & Kirana, 2023). This system facilitates monotonic inference pathways, enabling faster and more stable ability estimation. Meanwhile, IRT-based CAT remains essential due to its capacity to provide valid item parameter analyses and accurate ability estimation, making it a strong psychometric foundation for integration with AI.

A research gap persists in the absence of a comprehensive integration between Fuzzy Logic as the core AI component for fast and monotonic inference, and IRT as the psychometric backbone within a single CAT system. Such integration is particularly crucial in science education, where students' conceptual understanding is hierarchical and diverse. Fuzzy Logic offers an efficient adaptive inference mechanism, while IRT provides valid item parameter analysis and ability estimation. The combination of both approaches is expected to enhance the effectiveness of personalized assessment. Therefore, this study aims to develop a hybrid CAT system that integrates Fuzzy Logic and IRT within the context of science learning. The main objectives are: (1) to analyze existing research that designs inference systems

leveraging the speed and monotonic reasoning of Fuzzy Logic while incorporating IRT parameters for optimal item selection, and (2) to evaluate the effectiveness of the hybrid CAT in improving assessment precision and learning personalization. This integrated system is expected to adjust the evaluation pathway to students' abilities in real time, increase measurement efficiency, and support personalized science learning.

B. METHODS

This study employs a systematic literature review approach to map and evaluate global developments related to Computerized Adaptive Testing (CAT) that integrates Artificial Intelligence (AI) and Item Response Theory (IRT) within the context of science education. A systematic review was selected to ensure that the processes of identifying, selecting, and evaluating the literature were conducted in a structured and replicable manner, while also enabling the development of a comprehensive scientific synthesis. The methodological framework of this study was designed in accordance with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), which is widely applied in contemporary educational technology research. In addition, this study incorporates bibliometric analysis using VOSviewer to identify thematic relationships, keyword occurrences, and author networks associated with the development of AI- and IRT-based CAT. Thus, the study integrates systematic and bibliometric approaches to provide a holistic overview of the research landscape spanning more than a decade.

The data for this study were obtained from the Scopus database, which provides peer-reviewed literature across multiple disciplines, particularly in educational technology, artificial intelligence, machine learning, psychometrics, and adaptive assessment. Scopus was chosen due to its high credibility, broad multidisciplinary coverage, and compatibility with bibliometric analysis tools such as VOSviewer. Data collection was conducted directly through the Scopus website, which allows searching, filtering, and downloading article metadata, including author names, affiliations, publication years, journal titles, citations, and keywords. The document search was restricted to publications from 2011 to 2025 to capture key developmental phases of AI/IRT integration in CAT, ranging from the application of fuzzy logic to modern cognitive diagnostic models and adaptive assessment within science education.

The literature selection process followed the PRISMA flow to ensure transparency across the stages of identification, screening, and inclusion. The initial search was conducted in the Scopus database using the following Boolean query: TITLE-ABS-KEY ("*Computerized Adaptive Testing*") AND TITLE-ABS-KEY ("*Item Response Theory*") AND TITLE-ABS-KEY ("*Artificial Intelligence*") AND PUBYEAR > 2020 AND PUBYEAR < 2026

This search yielded a set of documents relevant to research on AI- and IRT-based CAT. In the subsequent step, a language filter was applied to include only English-language publications in order to maintain terminological consistency.

Document-type filtering further narrowed the results to include only research articles and review articles, excluding short notes, commentaries, abstract-only papers, and non-peer-reviewed publications. This screening process ensured that the selected documents meaningfully addressed the integration of adaptive methods, AI, fuzzy logic, recommender systems, cognitive diagnosis models, and IRT implementation in evaluation systems relevant to science learning. The PRISMA diagram illustrating the identification, screening, and inclusion stages is presented in Figure 1.

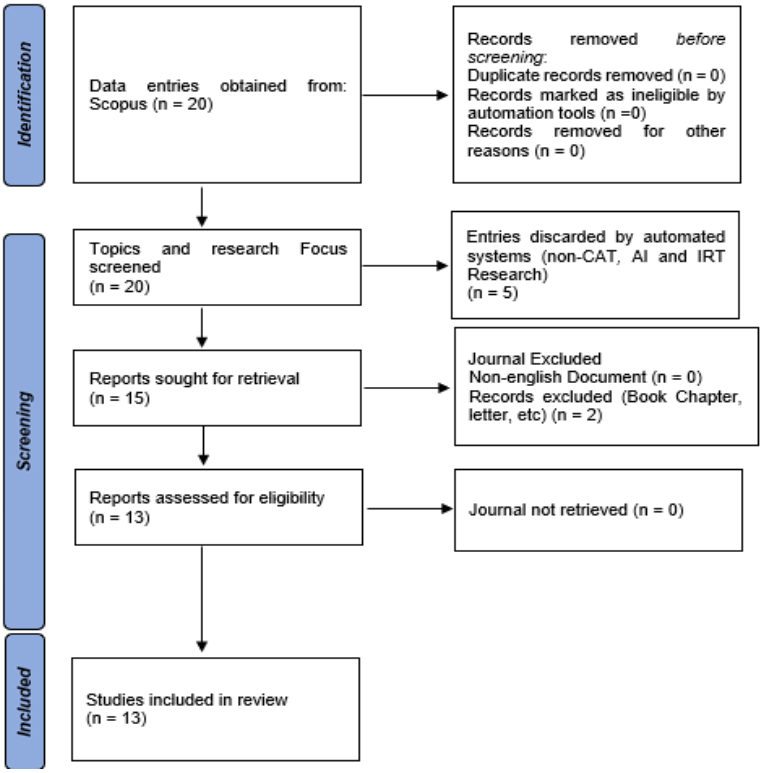

Figure 1. PRISMA Diagram of Article Selection

All search results from Scopus were exported in the Research Information Systems (.ris) format, as this format supports complete metadata processing for bibliometric analysis using VOSviewer. After export, all files were merged and underwent a data-cleaning procedure that included removing duplicates, correcting author name inconsistencies, standardizing keyword terminology, and manually verifying article content to ensure relevance to CAT, AI, and IRT. This stage also involved examining the research focus to determine whether the studies addressed item selection mechanisms such as Maximum Information (MI), ability estimation methods such as Expected a Posteriori (EAP), the use of one- to three-parameter IRT models, and the integration of algorithms such as BOBCAT, FACD, AdaCrowd, or fuzzy inference systems. After the cleaning process was completed, the final dataset was compiled as the basis for content analysis and bibliometric analysis.

Data validity was ensured through cross-verification between Scopus metadata and full article content to confirm that the listed methods, applied IRT models, and implemented AI algorithms aligned with the objectives of this study. The validation process also included checking keyword consistency within the VOSviewer-mapped data, particularly to ensure that thematic clusters emerged from conceptual coherence

rather than metadata errors. Standardization of author names was conducted to prevent unnecessary node fragmentation in co-authorship mapping. Although Scopus provides a highly comprehensive database, certain limitations remain, such as incomplete documentation of specific AI models, variability in methodological quality across studies, and inconsistencies in author-assigned keywords. Nonetheless, these limitations were minimized through selective inclusion of studies that met strict academic standards and demonstrated clear relevance to the research focus.

The analysis in this study employed a combination of bibliometric techniques and qualitative synthesis. On the quantitative side, VOSviewer was used to map keyword co-occurrence, author networks, and institutional affiliations that dominate research on AI- and IRT-based CAT. These visualizations enabled the identification of dominant themes and emerging research clusters. On the qualitative side, each article was analyzed based on the adaptive methods employed, the types of AI algorithms implemented, the IRT models used, the efficiency of item selection, improvements in ability estimation accuracy, and the article's contribution to personalized science learning. The combination of these two analytical approaches provided a comprehensive understanding of methodological developments and scientific trends in the advancement of AI- and IRT-based CAT.

C. RESULTS AND DISCUSSION

1. Publication Outcomes and Subject Area Distribution

Figure 2 presents the distribution of documents by subject area, clearly indicating the dominance of the Computer Science discipline, which accounts for the largest proportion at 36.4% (8 documents), followed by Engineering at 18.2% (4 documents) and Medicine at 13.6% (3 documents). Meanwhile, fields such as Mathematics, Decision Sciences, Multidisciplinary Studies, Neuroscience, Pharmacology–Toxicology–Pharmacy, and Social Sciences contribute smaller proportions, with the latter four areas each representing only 4.5% (1 document) of the total dataset. This distribution underscores the strong concentration of the collected documents within technical and scientific domains, as illustrated in Figure 2. Documents by Subject Area.

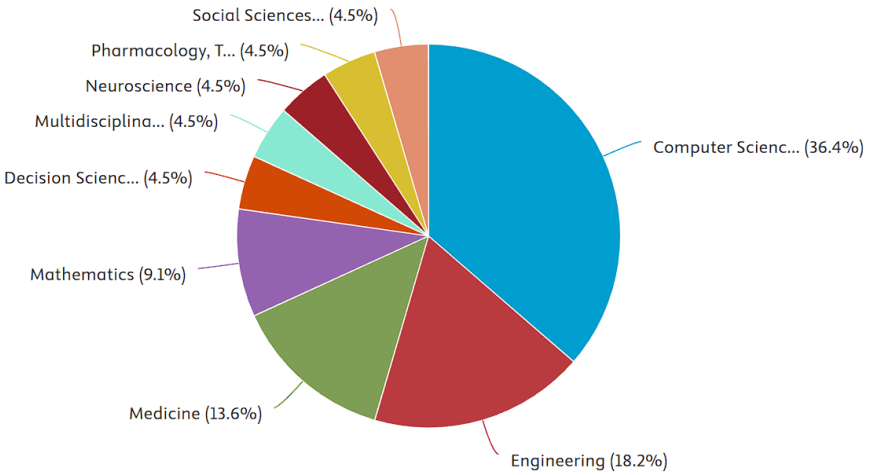

Figure 2. Documents by Subject Area

2. Author Productivity Analysis

The analysis presented in Figure 3 on Documents by the Top 10 Authors reveals a highly dispersed pattern of contributions among the listed authors, where each author in the top ten—such as Abboud, J.A., Albalushi, N., Aoula, E.S., and others up to Ghosh, A.—is recorded as contributing only one document to the dataset. This pattern indicates that the collected research and publications originate from a broad range of sources, with little to no dominance by any single author in terms of publication volume. This distribution is illustrated in Figure 3, Documents by the Top 10 Authors.

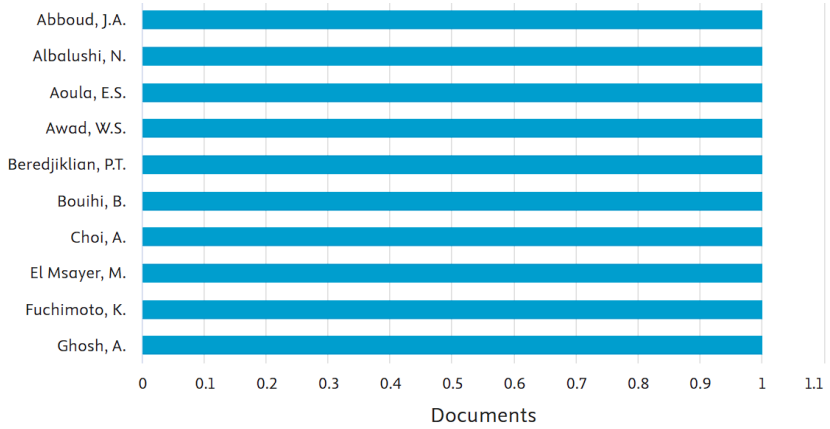

Figure 3. Documents by the Top 10 Authors

3. Institutional and Affiliation Analysis

The dataset also reveals substantial dispersion in terms of institutional affiliations associated with the published or supported documents. As illustrated in Figure 4, which presents the Documents by the Top 10 Publishers/Affiliations, each institution listed—from the Changzhou Vocational Institute of Mechatronic Technology to the Polish Academy of Sciences—contributes only one document. This pattern suggests that the analyzed works originate from a diverse range of academic, research, and commercial institutions across various locations, with no single publisher or affiliation exhibiting a dominant contribution. This distribution is depicted in Figure 4, Documents by the Top 10 Publishers.

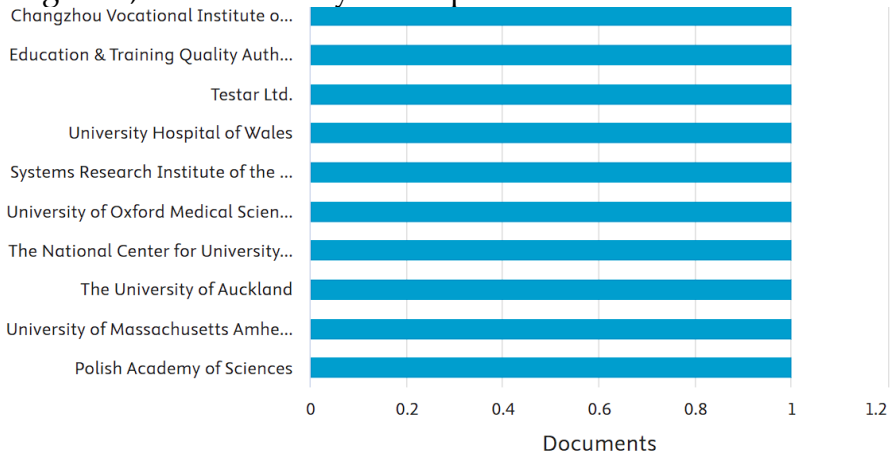

Figure 4. Documents by the Top 10 Publishers

4. Data Visualization Results

The analysis conducted using VOSviewer, as presented in Figure 5, displays a network visualization that maps the conceptual structure of research related to AI-based adaptive testing and Item Response Theory through the co-occurrence relationships of keywords in the literature. This network visualization reveals that key terms such as adaptive testing, item, test, and data form a strong cluster on the left side, while terms associated with analytical approaches—such as review, intervention, and youth—constitute a separate cluster in the central area. On the right side, terminologies such as case and dash appear as a more isolated group, indicating research areas that follow their own developmental trajectory. The connections between nodes, represented by line thickness and spatial proximity, illustrate the strength of relationships among concepts within the publication corpus. This enables the identification of collaborative patterns across research themes and provides a mapping of methodological evolution within the field of adaptive testing and modern psychometric modeling.

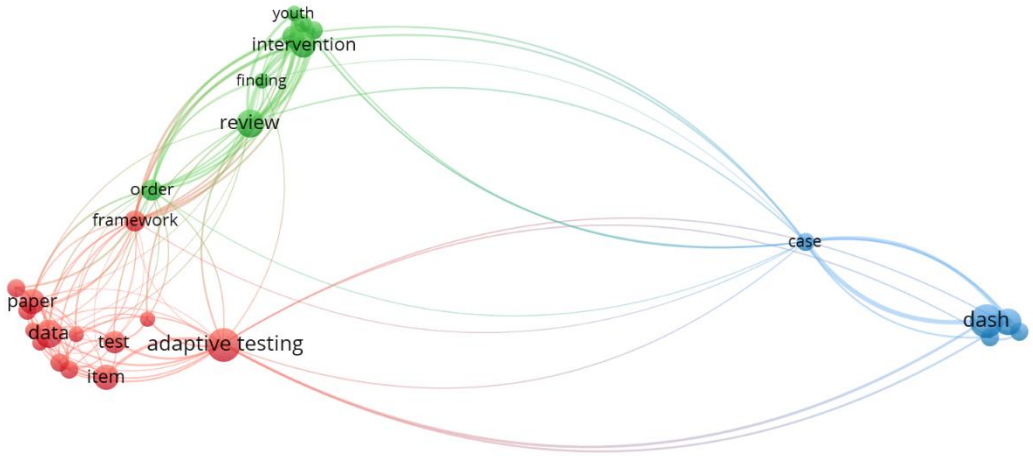

Figure 5. Network Visualization

Figure 6 presents the density visualization, which illustrates the intensity of keyword occurrences within the publication corpus, allowing highly researched areas to be identified through brighter color gradients. Regions highlighted in yellow indicate centers of concentrated scholarly activity, particularly around keywords such as data, paper, test, and adaptive testing, signaling that research efforts are primarily focused on item selection algorithm development and the performance evaluation of Computerized Adaptive Testing. In contrast, areas shaded in green to blue represent lower research density, as seen in terms such as case and dash, positioning these concepts as more specialized or niche research domains. This visualization is crucial for identifying both core and peripheral topics within the research field, as it provides insights into how the literature cumulatively evolves in certain areas, thereby shaping distinct patterns of scientific concentration within the broader bibliometric landscape.

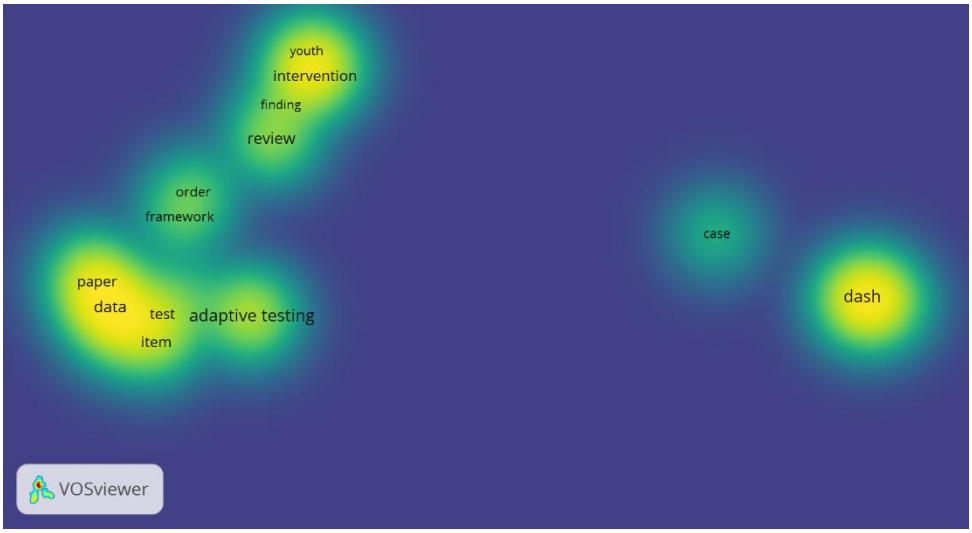

Figure 6. Density Visualization

Figure 7 presents the overlay visualization, which illustrates the temporal dynamics of research topic development by coloring nodes based on their publication year. Purple to blue shades represent earlier-emerging keywords such as paper, data, and test, whereas green to yellow shades indicate more recently developed terms such as intervention, review, and youth. This pattern suggests a shift in research focus from technical psychometric issues toward the application of adaptive testing in educational interventions and modern evaluation frameworks. Meanwhile, keywords such as case and dash remain in darker tones, indicating that these areas have not experienced significant growth over the same period. Accordingly, the overlay visualization enables researchers to observe the conceptual evolution of the literature, identify emerging thematic trends, and recognize potential areas of future inquiry projected to expand within AI-based adaptive testing research.

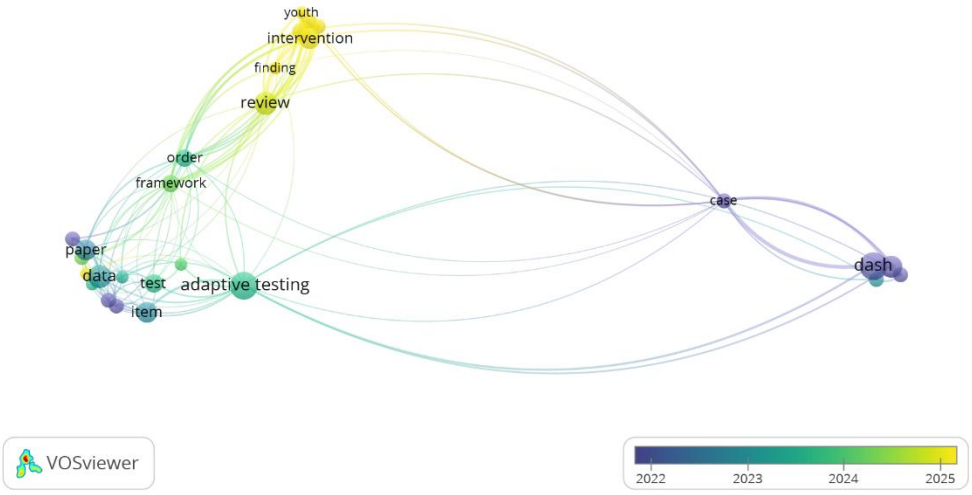

Figure 7. Overlay Visualization

5. Qualitative Analysis of AI and IRT Integration in CAT Development for Enhancing Assessment Effectiveness in Learning

An in-depth analysis of the integration of Artificial Intelligence and Item Response Theory in the development of Computerized Adaptive Testing requires a

systematic mapping of the various algorithmic approaches, psychometric models, and empirical findings reported in international literature over more than a decade. This comprehensive review is essential because the evolution of modern CAT no longer relies solely on conventional IRT parameters; instead, it has expanded through the adoption of artificial intelligence techniques such as machine learning, fuzzy logic, deep adaptive models, and bilevel optimization. These approaches collectively enhance item selection efficiency, ability estimation accuracy, and the quality of personalized assessment, particularly in science education, which demands highly precise competency measurement. The diversity of these approaches has generated a wide spectrum of innovations, including the use of Maximum Information and Expected a Posteriori for optimal item selection, the integration of Cognitive Diagnosis Models for multidimensional analysis, the application of Large Language Models in verbal ability assessment, and the development of probabilistic adaptive systems capable of reducing data annotation requirements without compromising evaluative performance. To systematically present these methodological advancements, this study constructs a qualitative analysis table summarizing the relationships between types of Artificial Intelligence algorithms, the IRT models employed, and the contributions reported in each study, as presented in Table 1.

Table 1. Qualitative Analysis of AI and IRT Integration in CAT Development for Improving Assessment Effectiveness in Learning

No.	Reference (Author & Year)	Type of AI Algorithm	IRT Model Used	Key Findings
1	Harrison & Trickett (2025)	Artificial Intelligence (AI), Computerized Adaptive Testing (CAT), Item Response Theory (IRT)	Classical Test Theory (initial basis) and Item Response Theory (IRT)	Explains the progression from established scientific measurement methods to modern IRT and CAT, highlighting opportunities enabled by AI.
2	Imawan et al. (2025)	Artificial Intelligence (AI), CAT, R programming using the *mirtCAT* package	Not specified explicitly, but *mirtCAT* operationalizes various IRT models	Demonstrates efficient CAT implementation using Maximum Information (MI) for item selection and Expected a Posteriori (EAP) for ability estimation, enabling energy-efficient testing.
3	Kwong & Mohammadi (2025)	Machine Learning (ML), Recommender System (including Truncated Singular Value Decomposition & Collaborative Filtering)	Not explicitly stated	Proposes a computationally efficient ML-based approach for CAT using small datasets, emphasizing item–item relationships rather than learner

				characteristics to improve predictive accuracy.
4	Yang et al. (2025)	Adaptive Learning from Crowds (AdaCrowd), inspired by CAT	Not explicitly stated	Demonstrates effectiveness in reducing annotation requirements without compromising performance by employing probabilistic models to capture the informativeness of instances per worker.
5	Liu et al. (2025)	Fast Adaptive Cognitive Diagnosis (FACD) framework combining dynamic collaborative and personalized diagnostic modules	Cognitive Diagnosis Model (CDM)	FACD achieves superior predictive performance (5–10% improvement in early CAT stages) while maintaining high inference speed compared with static CDMs.
6	Albalushi & Awad (2025)	Artificial Intelligence (AI), Machine Learning (ML)	Item Response Theory (IRT) and Cognitive Diagnosis Models (CDM)	AI/ML-enhanced adaptive assessment (including CAT) can promote equitable and efficient learning environments by improving item generation, difficulty prediction, and proficiency estimation.
7	Wanniarachchi et al. (2025)	Machine Learning (ML), with recommendations for Generative AI integration	Not explicitly stated	Suggests that future interventions could benefit from incorporating Generative AI to personalize user experiences, enhance engagement, and support mental health protocol adherence.
8	Klein & Kovács (2024)	Large Language Models (LLMs) such as ChatGPT and Bing in CAT	Not explicitly stated	Verbal ability CAT is an appropriate tool for critically evaluating LLM performance, though traditional human psychometric instruments have limitations when applied to AI assessment.

9	El Msayer et al. (2024)	Artificial Intelligence (AI), Machine Learning (ML)	Item Response Theory (IRT)	AI/ML enhances traditional CAT approaches by improving item selection, ability estimation, and automated item generation for adaptive testing.
10	Kishida et al. (2023)	Item Difficulty Constrained Uniform Adaptive Testing	Item Response Theory (IRT)	This method mitigates item exposure bias while maintaining low measurement error.
11	Zhu (2022)	Intelligent analysis system using AI, Multi-dimensional dual-objective CD-CAT topic selection strategy	CD-CAT (Cognitive Diagnostic Computerized Adaptive Testing)	The proposed intelligent decision system demonstrates strong performance and effectively improves psychological training outcomes for athletes.
12	Ghosh & Lan (2021)	BOBCAT (Bilevel Optimization-Based Framework for CAT)	Item Response Theory (IRT) as underlying response model (BOBCAT is model-agnostic)	BOBCAT outperforms existing CAT methods—often significantly—by reducing test length through data-driven question selection algorithms.
13	Kane et al. (2021)	Predictive models using Machine Learning applied to CAT	IRT model not specified	Application of CAT to DASH and QuickDASH questionnaires reduces respondent burden (49% and 47% reduction in average item count) with negligible impact on measurement integrity.
14	Huda et al. (2024)	Items Response Theory (IRT) algorithm	Three-Parameter Logistic (3PL) model	CAT/IRT systems are highly practical and effective for achieving measurement objectives, providing more accurate and efficient ability estimation through dynamically adjusted item difficulty.
15	V. Srikanth et al. (2023)	Adaptive Testing, AI-driven learning, Deep Q-Network (DQN)	Three-Parameter Logistic (3PL) Model	AI-driven adaptive testing leveraging IRT personalizes learning in computer science

		Reinforcement Learning		education, achieving 92.8% adaptation accuracy and requiring only 21 items compared to 35 in static tests.

Table 1 provides a comprehensive mapping of the literature on the integration of Artificial Intelligence (AI) and Item Response Theory (IRT) in the development of Computerized Adaptive Testing (CAT), forming the empirical foundation for analyzing the two main foci of this study: (1) the development of inference-system models that leverage the speed and monotonic reasoning of fuzzy logic while incorporating IRT parameters into item selection, and (2) the evaluation of hybrid CAT effectiveness in improving assessment precision and learning personalization. Chronologically, the studies listed—from Haryanto (2011) through Harrison and Trickett (2025)—demonstrate the evolution of intelligent technologies from rule-based inference systems to more complex predictive frameworks, including machine learning (ML), cognitive diagnosis modelling (CDM), large language models (LLMs), and generative AI. Harrison and Trickett (2025), for instance, highlight the shift from Classical Test Theory to IRT as the scientific foundation of modern CAT, while emphasizing how AI expands adaptive capabilities through optimized item selection and more sensitive ability estimation. These findings closely align with the first research focus, as IRT provides formal parameters (discrimination, difficulty, guessing) that can be integrated into fuzzy logic inference mechanisms to produce monotonicity-consistent item selection.

Subsequent studies further reinforce the technical functioning of hybrid adaptive systems. Imawan et al. (2025) show that using mirtCAT with the Maximum Information (MI) criterion and Expected a Posteriori (EAP) estimation can save time and testing energy without reducing accuracy; both MI and EAP align well with fuzzy-based decision structures that prioritize informative items under uncertainty. Kwong and Mohammadi (2025) emphasize that ML and recommender-system approaches operating on small datasets rely more on item–item relationships than student characteristics, thereby accelerating inference and enhancing algorithmic stability. Yang et al. (2025), with the AdaCrowd model, demonstrate how probabilistic modelling reduces manual annotation requirements without degrading adaptive performance, illustrating that hybrid inference systems can function optimally even with limited initial information. Liu et al. (2025), through the Fast Adaptive Cognitive Diagnosis (FACD) framework, show that early-stage CAT prediction accuracy can improve by 5% to 10% while maintaining high inference speed—an advantage that reinforces the value of fuzzy logic, which is designed for rapid and monotonically consistent decision processing under uncertainty.

Findings from Albalushi and Awad (2025) and Zhu (2022) further reveal that integrating AI and IRT broadens the potential for multidimensional assessment through static CDM or Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT). Zhu (2022), in particular, demonstrates that a dual-objective topic-selection

strategy in CD-CAT effectively enhances athletes' psychological training outcomes, a result analogous to science learning contexts where multidimensional competencies such as data interpretation, experimentation, and scientific reasoning must be assessed concurrently. Research on BOBCAT by Ghosh and Lan (2021) shows that direct optimization from data significantly shortens test length without compromising measurement quality. Kane et al. (2021) support this claim by showing that CAT implementation reduces respondent burden by 49% for DASH and 47% for QuickDASH with negligible impact on score integrity. These findings collectively reinforce the argument that hybrid CAT combining AI and IRT is not only more precise but also more efficient, directly addressing the second focus of this study.

Other studies highlight fairness and bias mitigation in adaptive assessment. Kishida et al. (2023), through Item Difficulty Constrained Uniform Adaptive Testing, demonstrate that controlling item exposure can be achieved while maintaining low measurement error, ensuring adaptive systems remain equitable for all examinees. Klein and Kovács (2024) add a critical perspective on using LLMs in CAT, noting that although these models are powerful, traditional psychometric tools may be limited in evaluating AI performance—an insight with serious implications for designing CAT based on language models. Wanniarachchi et al. (2025) recommend integrating Generative AI to personalize user experience, including mental health support, signaling the pedagogical value of learning environments informed by cognitive and affective learner profiles. These findings collectively indicate that personalization in CAT extends beyond item difficulty to encompass the learner's holistic experience.

Haryanto (2011) provides a foundational contribution to the first research focus. His integration of fuzzy logic as an inference system with Two-Parameter IRT yielded a 0.72 correlation between theoretical predictions and system outputs, demonstrating that fuzzy logic can facilitate consistent, adaptive, and participant-aligned item selection. This early integration offers conceptual clarity that fuzzy logic can serve as a rapid, soft, and monotonic decision-making mechanism, while IRT provides a stable and measurable mathematical structure. With subsequent developments, it becomes evident that the integration of AI, IRT, and fuzzy logic in CAT enhances inference speed, accuracy, test efficiency, and generates more personal, objective, and diagnostically rich evaluation systems.

The findings in Table 1 suggest that research is moving toward consolidating inference mechanisms that combine the flexibility of fuzzy logic, the mathematical rigor of IRT, and the predictive strength of AI. Studies such as Haryanto (2011) provide early evidence that fuzzy inference systems can capture gradual variations in examinee ability through membership-degree mapping, while IRT parameters offer precision control via item information functions. Research by Liu et al. (2025), Zhu (2022), and Ghosh and Lan (2021) further expand on this by demonstrating that predictive models based on CDM, CD-CAT, or bilevel optimization can synergize with adaptive inference logic. This indicates that ideal hybrid CAT design should not rely solely on item-selection algorithms, but rather build a multilayered inference mechanism that integrates probabilistic evaluation, fuzzy inference rules, and IRT

information structures. Thus, the studies collectively justify that integrating fuzzy logic with IRT is not merely an experimental approach but a methodological evolution aligned with developments in adaptive technologies.

Beyond technical contributions, the studies in Table 1 offer important pedagogical implications for how hybrid CAT can enhance the quality of science learning. The nature of science education—requiring data-analysis skills, evidence-based reasoning, and interpretation of scientific phenomena—aligns strongly with the strengths of adaptive systems that adjust difficulty levels, provide diagnostic feedback, and map multidimensional learner profiles. Approaches such as CD-CAT and FACD, as demonstrated by Zhu (2022) and Liu et al. (2025), confirm that learners can be mapped according to specific weaknesses, whether in conceptual understanding, quantitative reasoning, or visual–data interpretation. Integrating recommender systems, as shown in Kwong and Mohammadi (2025), further enables the delivery of items aligned with learners' response patterns and learning styles. Thus, hybrid CAT not only enhances measurement accuracy but also fosters a data-informed learning ecosystem where teachers can design evidence-based interventions, adjust practical teaching strategies, and deepen students' scientific understanding through assessments that truly adapt to their cognitive conditions.

D. CONCLUSION

Based on the findings, it can be concluded that the integration of Artificial Intelligence and Item Response Theory within Computerized Adaptive Testing establishes a hybrid evaluation framework that is more precise, efficient, and responsive to learner characteristics. The network, density, and overlay visualizations generated through VOSviewer confirm that topics such as machine learning, fuzzy logic, IRT, cognitive diagnosis, and generative AI constitute dominant and interconnected clusters that continue to evolve. The reviewed literature demonstrates that models such as FACD, BOBCAT, MI–EAP, as well as fuzzy logic–based systems, contribute significantly to improving item-selection accuracy, accelerating ability estimation, and reducing test length without compromising measurement quality. Thus, AI–IRT integration emerges as a robust approach for enabling adaptive and personalized assessment in the context of science education.

Future research is encouraged to design and test prototype hybrid CAT systems that explicitly combine fuzzy logic reasoning with IRT parameters, particularly within multidimensional models that more accurately represent competencies in science learning. Moreover, further development should incorporate large-scale empirical testing to evaluate the stability of item selection, the consistency of test adaptation, and the fairness of measurement across diverse learner profiles. The incorporation of Generative AI and Large Language Models (LLMs) also warrants deeper exploration to enhance diagnostic feedback and support more personalized assessment experiences, thereby ensuring that the resulting CAT systems fully align with data-driven learning environments and the competency demands of the twenty-first century.

REFERENCES

1. Albalushi, N., & Awad, W. S. (2025). *Generating Questions Bank for Adaptive Assessment Using Machine Learning Techniques: Review*.
2. Cheng, S.-C., Cheng, Y.-P., & Huang, Y.-M. (2021). To implement computerized adaptive testing by automatically adjusting item difficulty index on adaptive English learning platform. *Journal of Internet Technology*, *22*(7), 1599–1607.
3. El Msayer, M., Aoula, E.-S., & Bouihi, B. (2024). *Artificial intelligence in computerized adaptive testing to assess the cognitive performance of students: A Systematic Review*.
4. Frick, S., Krivosija, A., & Munteanu, A. (2024). Scalable learning of item response theory models. *International Conference on Artificial Intelligence and Statistics*, 1234–1242.
5. Ghosh, A., & Lan, A. (2021). BOBCAT: Bilevel Optimization-Based Computerized Adaptive Testing. *IJCAI International Joint Conference on Artificial Intelligence*, 2410–2417.
6. Göktepe Körpeoğlu, S., Filiz, A., & Göktepe Yıldız, S. (2025). AI-driven predictions of mathematical problem-solving beliefs: Fuzzy logic, adaptive neuro-fuzzy inference systems, and artificial neural networks. *Applied Sciences*, *15*(2), 494.
7. Harrison, C. J., & Trickett, R. W. (2025). Patient reported outcome measures: from the classics to AI. *Journal of Hand Surgery: European Volume*, *50*(6), 807–813.
8. Haryanto, H. (2011). Pengembangan Computerized Adaptive Testing (CAT) dengan Algoritma Logika Fuzzy. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *15*(1).
9. Huda, A., Irfan, D., Hendriyani, Y., & Sukmawati, M. (2024). Optimizing Educational Assessment : The Practicality of Computer Adaptive Testing (CAT) with an Item Response Theory (IRT) Approach. *International Journal on Informatics Visualization*, *8*(1), 473–480.
10. Huda, C. (2024). *Paradigma Pembelajaran IPA Berbasis Proyek Berdiferensiasi: Menyukseskan Kurikulum Merdeka Belajar Kampus Merdeka*. Penerbit NEM.
11. Imawan, O. R., Retnawati, H., Haryanto, H., & Ismail, R. (2025). Innovations in Assessment Methods: Computerized Adaptive Testing (CAT) for Sustainable Energy Efficiency. *Lecture Notes in Civil Engineering*, *557 LNCE*, 161–168.
12. Imawan, O. R., Retnawati, H., & Ismail, R. (2025). The Challenges of Implementing Computerized Adaptive Testing in Indonesia. *Journal of Education and E-Learning Research*, *12*(2), 124–144.
13. Iwintolu, R. O., Opesemowo, O. A. G., & Adetutu, P. O. (2024). Effect of 2-PL and 3-PL Models on the Ability Estimate in Mathematics Binary Items. *Journal on Efficiency and Responsibility in Education and Science*, *17*(3), 257–272.
14. Kane, L. T., Abboud, J. A., Plummer, O. R., & Beredjiklian, P. T. (2021). Improving Efficiency of Patient-Reported Outcome Collection: Application of Computerized Adaptive Testing to DASH and QuickDASH Outcome Scores. *Journal of Hand Surgery*, *46*(4), 278–286.
15. Kishida, W., Fuchimoto, K., Miyazawa, Y., & Ueno, M. (2023). Item Difficulty Constrained Uniform Adaptive Testing. *Communications in Computer and Information Science*, *1831 CCIS*, 568–573.

16. Klein, B., & Kovács, K. (2024). The performance of ChatGPT and Bing on a computerized adaptive test of verbal intelligence. *PLOS ONE*, *19*(7 July).
17. Kwong, H. Y., & Mohammadi, G. (2025). *Recommender Methods for Computerised Adaptive Testing*. 13–15.
18. Liu, Y., You, Y., Liu, S., Qian, H., Qian, Y., & Zhou, A. (2025). A Fast-Adaptive Cognitive Diagnosis Framework for Computerized Adaptive Testing Systems. *IJCAI International Joint Conference on Artificial Intelligence*, 5824–5832.
19. Ma, W. A., Richie-Halford, A., Burkhardt, A. K., Kanopka, K., Chou, C., Domingue, B. W., & Yeatman, J. D. (2025). ROAR-CAT: Rapid Online Assessment of Reading ability with Computerized Adaptive Testing. *Behavior Research Methods*, *57*(1), 56.
20. Maji, S., & Ganguli, S. (2025). Fuzzy Logic Control for Industrial Applications. *Controller Design for Industrial Applications*, 1–20.
21. Papadimitriou, S., & Virvou, M. (2025). Fuzzy Logic and Applications in Education and Games: Theory, Practical Implementations and a Literature Review. *Artificial Intelligence-Based Games as Novel Holistic Educational Environments to Teach 21st Century Skills*, 95–127.
22. Sathya, D., Saravanan, G., & Thangamani, R. (2024). Fuzzy logic and its applications in mechatronic control systems. *Computational Intelligent Techniques in Mechatronics*, 211–241.
23. Suzuki, A., & Negishi, E. (2024). Fuzzy logic systems for healthcare applications. *Journal of Biomedical and Sustainable Healthcare Applications*, *4*(1).
24. Tian, X., & Dai, B. (2020). Developing a computerized adaptive test to assess stress in Chinese college students. *Frontiers in Psychology*, *11*, 7.
25. Tsaousis, I., Sideridis, G. D., & AlGhamdi, H. M. (2021). Evaluating a computerized adaptive testing version of a cognitive ability test using a simulation study. *Journal of Psychoeducational Assessment*, *39*(8), 954–968.
26. Wanniarachchi, V. U., Greenhalgh, C., Choi, A., & Warren, J. R. (2025). Personalization variables in digital mental health interventions for depression and anxiety in adolescents and youth: a scoping review. *Frontiers in Digital Health*, *7*.
27. Wulansari, A. D., & Kirana, D. P. (2023). *Pengukuran English Vocabulary Size dengan Computerized Adaptive Testing*. Thalibul Ilmi Publishing & Education.
28. Yang, H., Li, Z., & Pedrycz, W. (2025). Adaptive Deep Learning from Crowds. *IJCAI International Joint Conference on Artificial Intelligence*, 4263–4272.
29. Zhu, H. (2022). Research on intelligent analysis strategies to improve athletes' psychological experience in the era of artificial intelligence. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *119*, 110597.